

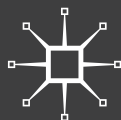
INVESTIGATING SPOKEN ENGLISH

A PRACTICAL GUIDE TO PHONETICS
AND PHONOLOGY USING PRAAT



ŠTEFAN BEŇUŠ

MOREMEDIA



Investigating Spoken English

“The book brings a practical approach to understanding spoken English in a compelling and lucid way. It is filled with simple, hands-on suggestions for engaging with your own speech and testing a wide range of ideas in phonetics and phonology. Its presentation of intonation and prosody is exciting and masterful. A must for any students wishing to gain a rapid, broad-ranging and very accessible understanding of how sounds are put to use in speech communication.”

—Jonathan Harrington, *Ludwig-Maximilians University of Munich*

“A wealth of practical hands-on activities and follow-up commentaries guide the reader toward their own discoveries and understanding the foundations that provide for our ability to communicate in English in its true form as a spoken language. The phonetic analyses make the book valuable for anyone interested in speech and its relation to language in general. This includes all speakers of English! Moreover, the book appears to be excellent also for self-guided and distance-based learning.”

—Matti Vainio, *University of Helsinki*

“Praat is today the most widely used program for the automatic analysis of speech. The software is extensively documented but does not provide the basic knowledge of phonetics and phonology needed to use it efficiently. The first book in English to cover both Praat and the fields of phonetics and phonology, this will be an extremely valuable resource for the many thousands of Praat-users who do not have a solid background in speech sciences.”

—Daniel Hirst, *French National Centre for Scientific Research (CNRS) & Aix-Marseille University*

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC)

© 2015 IPA

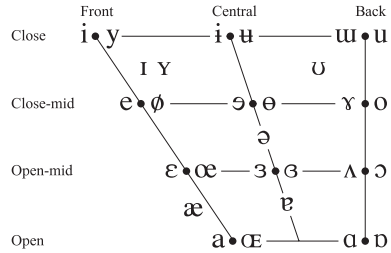
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			ʀ					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
◌ Bilabial	ɓ Bilabial	ʼ Examples:
◌ Dental	ɗ Dental/alveolar	ɸ' Bilabial
◌ (Post)alveolar	ɟ Palatal	t' Dental/alveolar
◌ Palatoalveolar	ɡ Velar	k' Velar
◌ Alveolar lateral	ɠ Uvular	s' Alveolar fricative

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

- ɱ Voiceless labial-velar fricative
- ɥ Voiced labial-velar approximant
- ɰ Voiced labial-palatal approximant
- ʜ Voiceless epiglottal fricative
- ʕ Voiced epiglottal fricative
- ʡ Epiglottal plosive
- ɠ Alveolo-palatal fricatives
- ɩ Voiced alveolar lateral flap
- ɺ Simultaneous ʃ and ʒ

ts k̟p

SUPRASEGMENTALS

- ˈ Primary stress
- ˌ Secondary stress
- ː Long
- ˑ Half-long
- ◌ Extra-short
- ◌ Minor (foot) group
- ◌ Major (intonation) group
- ˑ Syllable break
- ◌ Linking (absence of a break)

TONES AND WORD ACCENTS

- | | | | |
|-------|------|------------|-----------------------|
| LEVEL | | CONTOUR | |
| ẽ | or ̃ | Extra high | ẽ or ̃ Rising |
| é | ̂ | High | ẽ or ̂ Falling |
| ē | ̄ | Mid | ẽ or ̄ High rising |
| è | ̌ | Low | ẽ or ̌ Low rising |
| ẽ | ̎ | Extra low | ẽ or ̎ Rising-falling |
| ↓ | | Downstep | ↗ Global rise |
| ↑ | | Upstep | ↘ Global fall |

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ɲ̰

◌ Voiceless	◌ Breathy voiced	◌ Dental	◌ Apical
◌ Voiced	◌ Creaky voiced	◌ Laminar	
◌ Aspirated	◌ Linguolabial	◌ Labialized	◌ Nasalized
◌ More rounded	◌ Labialized	◌ Palatalized	◌ Nasal release
◌ Less rounded	◌ Velarized	◌ Pharyngealized	◌ No audible release
◌ Advanced	◌ Retracted	◌ Centralized	◌ Mid-centralized
◌ Syllabic	◌ Non-syllabic	◌ Rhoticity	

Typefaces: Doulos SIL (metatext); Doulos SIL, IPA Kiel, IPA LS Uni (symbols)

Štefan Beňuš

Investigating Spoken English

A Practical Guide to Phonetics
and Phonology Using Praat

palgrave
macmillan

Štefan Beňuš
Department of English and American Studies,
Faculty of Arts
Constantine the Philosopher University in Nitra
Nitra, Slovakia

Institute of Informatics
Slovak Academy of Sciences
Bratislava, Slovakia

ISBN 978-3-030-54348-8 ISBN 978-3-030-54349-5 (eBook)
<https://doi.org/10.1007/978-3-030-54349-5>

© The Editor(s) (if applicable) and The Author(s) 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

This book draws heavily on my understanding of human speech and the cognitive system of phonetics-phonology that underlies it when applied to English. There have been numerous fellow academics working on speech who sparked my curiosity and provided inspiration, guidance, expertise, or a friendly ear over the course of its development. Additionally, many students at Constantine the Philosopher University in Nitra, Slovakia, mostly unknowingly, helped identify and fix dead ends, sharpen the focus or the pedagogical approach. I am deeply grateful to all of them.

In addition to all the unnamed above, let me also express my gratitude to the following people. Adamantios Gafos was my teacher and advisor at New York University and he has been pivotal in my interest and basic views regarding phonetic, phonology, and their relationship. I am indebted to the entire faculty and fellow students in the NYU department of Linguistics for great formatting five years of graduate school. I was also incredibly lucky to have Julia Hirschberg as a remarkable and caring post-doc supervisor and eventually a colleague in prosody research. I benefited enormously also from interactions with the members of Julia's Spoken Language Processing group at Columbia University, especially Agustín Gravano, Rivka Levitan, Frank Enos, and Andrew Rosenberg. Juraj Šimko, Katalin Mády and Uwe Reichel selflessly endured and commented on drafts of my text and also kindly shared time while hiking, doing manual labour, or over beer/wine for many exciting discussions regarding prosody or speech in general.

I am also thankful to my colleagues in the department of English and American studies at Constantine the Philosopher University in Nitra and the department of speech analysis and synthesis in the Institute of Informatics, Slovak Academy of Sciences, who have kept very open and supportive atmosphere over the years.

Lisa Davidson's lecture slides and Adrian Underhill's workshop provided various inspirations for my teaching. Péter Siptár, Katri Hiovain, and Heini Kallio commented on drafts of some chapters of the books. Marcos Perez generously donated his time and voice for some of the recordings. The manuscript benefited

greatly also from feedback and comments of three anonymous reviewers and invaluable help provided by the editorial team.

My son Matej patiently read and edited the entire first draft of the book and my entire family provided much needed mental life support in the process of writing and beyond.

Thank you!

Contents

1	Introduction	1
1.1	What to Expect from This Book?	1
1.2	Who Is This Book for?	3
1.3	How Is This Book Structured?	4
2	Fundamental Concepts	7
2.1	Speaking as Acquired Habits	7
2.2	Continuity and Contrastiveness of Speech	11
2.3	Context and Variability	18
2.4	Transcription of Speech	23
	References	26
3	Articulatory Mechanisms in Speech Production	27
3.1	Introduction	27
3.2	Breathing and Airstream Mechanisms	28
3.3	Voicing	30
3.4	Other Activity Inside the Larynx	34
3.5	Articulation and the Activity Above the Larynx	36
	References	43
4	Acoustics and Introduction to Praat	45
4.1	Introduction	45
4.2	Praat Basics	46
	Record Your First Sound in Praat	47
	First Steps in Navigating Praat User Interface	47
4.3	Sound Waves and Their Visual Representation	49
4.4	Periodicity, Frequency and Amplitude	52
4.5	Sources, Filters and Speech	54
4.6	Visualizing Speech with Spectrograms	59
5	English Vowels	63
5.1	Introduction	63
5.2	Vowel Space	64

5.3	English Front Vowels	73
5.4	English Back Vowels	79
5.5	English Central Vowels	86
5.6	English Diphthongs	87
	References.	92
6	English Consonants.	93
6.1	Introduction	93
6.2	Stricture, Voicing, Place	94
6.3	English Consonants by Their Manner	98
	Stops (Plosives)	99
	Fricatives	100
	Affricates	104
	Nasals	105
	Approximants.	106
6.4	Sound–Spelling Correspondences for English Consonants.	111
	References.	113
7	Allophonic Variation in English.	115
7.1	Introduction	115
7.2	English Stops, Phonemes and Allophones.	116
7.3	Other Common Allophonic Patterns	131
	References.	136
8	Syllables	137
8.1	Introduction	137
8.2	Mental Knowledge of Syllables.	138
8.3	Sonority	140
8.4	English Syllables	144
8.5	Syllable Boundaries	147
8.6	Syllable Weight	151
	References.	153
9	Word Stress	155
9.1	Introduction	155
9.2	Mental Knowledge of Word Stress	156
9.3	Phonetic Aspects of Word Stress	159
9.4	Limitations of the Phonetic Aspects of Word Stress	166
9.5	Basic Word Stress Patterns in English	168
9.6	IPA Transcription of English Words	172
	References.	172
10	Connected Speech Aspects	175
10.1	Introduction	175
10.2	Preview of the Speaking Habits Under Investigation	176
10.3	Assimilations	180
10.4	Elisions.	186

10.5	Linking	190
10.6	Processes Within Words	192
	References	193
11	Prosody I: Basics	195
11.1	Introduction	195
11.2	Primer to Prosodic Analysis: Pitch Accents and Disjunctures	196
11.3	Weak Forms	201
11.4	Global Trends Within Intonational Phrases	205
	References	209
12	Prosody II: Intonation	211
12.1	Introduction	211
12.2	Fundamental Frequency and Pitch in Praat	212
12.3	Approaches to Marking Intonation	217
12.4	Simplified ToBI	221
12.5	Basic Functional Meanings of Intonation	230
	References	234
13	Prosody III: Beyond Intonational Phrase	235
13.1	Introduction	235
13.2	Discourse Markers	236
13.3	Conversational Fillers	238
13.4	Turn-Taking	239
13.5	Discourse Organization	243
13.6	Speech Entrainment	245
	References	248
14	Exemplifying the Book Material in Real Interviews	251
14.1	Introduction	251
14.2	The Grinch	253
14.3	Bakeoff_Dialogue	257
14.4	Jeeves and Wooster	261
	Further Reading	265
	Index	267

List of Figures

Fig. 2.1	Acoustic and articulatory data from the production of the word ‘iba’ in Slovak. See text for explanation.	13
Fig. 2.2	The first frame of b-closure in ‘biography’. See Activity 2-4 for explanation	15
Fig. 2.3	Five successive x-ray frames during ‘ba’. See text for explanations	15
Fig. 3.1	Illustration of the inverse relationship between volume and pressure, 10 molecules of air in a smaller (left) and larger (right) container resulting in greater (left) and smaller (right) pressure, respectively	28
Fig. 3.2	Schematic illustration of the structures in the larynx (top) and the activity of the muscles attached to the arytenoid cartilages, shown as arrows, for opening and closing the vocal folds (bottom).	32
Fig. 3.3	Stylized pictures for settings allowing modal voicing, breathy voice, whisper and creaky voice	34
Fig. 3.4	Schematic illustration of the major parts of the tongue for speech description	36
Fig. 3.5	Schematic illustration of the articulators above the larynx	37
Fig. 3.6	Active and passive articulators.	39
Fig. 3.7	English adjectives describing the major places of articulation	40
Fig. 3.8	Schematic representations of the vocal tract for humans (left) and chimpanzees (right); the lips and nose are on the right. The relevant articulators include: NC (nasal cavity), RN (roof of the nasopharynx), P (pharynx), VF (vocal folds), O (opening of larynx into pharynx), E (epiglottis), SP (soft palate), HP (hard palate), OC (oral cavity), T (tongue). Reproduced with permission, Lieberman (1975, Fig. 9-4, p. 108)	41

Fig. 3.9	Schematic representations of a vocal tract. See text of the activity. Reproduced with permission, Lieberman (1975, Fig. 9-4, p. 108)	42
Fig. 4.1	Screenshot of the sound wave corresponding to ‘Discovering spoken English’ in Praat.	48
Fig. 4.2	Visualization of sound waves. A tuning fork on the left with air particles as vertical lines in silence (a), when the fork is hit and its tines vibrate by cyclically moving outwards (b) and inwards (c). The result of this vibration shown in upper right and the travelling of sound waves is similar to the travelling of waves in water	49
Fig. 4.3	The movement of a single air particle. On the left, the pendulum-like horizontal movement from the rest position to the right, back, then left, and back again. The same stages of the particle position in time shown with the sine wave on the right	50
Fig. 4.4	Zoomed-in section corresponding to ‘eng’ of ‘English’ in the recording ‘Discovering spoken English’ in Fig. 4.1	51
Fig. 4.5	Two zoomed-in sections of the original ‘Discovering spoken English’ from Fig. 4.1. On the left, the interval during the hissing sound of the initial sound in ‘spoken’, on the right, the vowel portion of the oh-like sound in ‘spoken’	52
Fig. 4.6	Frequency and amplitude of sound waves are independent. Same frequency but different amplitude (top left and right), and same amplitude but different frequency (left top and bottom)	54
Fig. 4.7	Illustrations of fundamental frequency and its harmonics. On the left, vibrations of a string, on the right, sine waves of the fundamental and first four harmonic frequencies and their summation into the composite complex waveform	56
Fig. 4.8	Composite waveform created by summing five sine waves (200, 400, 600, 800, 1000 Hz). See Activity 4-4 for instructions and description	57
Fig. 4.9	Schematic illustration of the Source-filter theory of speech production in a nutshell. See text for explanations	58
Fig. 4.10	Waveform (top) and spectrogram (bottom) of ‘Discovering spoken English’ in Praat. The (blue) bolder contour corresponds to pitch (f_0), the (yellow) thinner line to intensity and (red) speckles to formant frequencies	60
Fig. 4.11	Waveform (top) and narrowband spectrogram (bottom) of ‘Discovering spoken English’ in Praat.	61
Fig. 5.1	Schematic illustration of the tongue positions for vowels [i], [u], [æ] and [a]/[ɒ] in the left panel in black solid, and neutral schwa [ə] in grey dotted lines. The centre	

	is the ellipsis covering the space defined by the highest points of the tongue in the four vowels. The right, reproduced with permission Culpeper et al. (2018), shows the abstraction from the ellipsis to the traditional trapezoid shape used for describing vowels.	66
Fig. 5.2	Illustration of the spectrogram and formants of ‘he’ in Praat	67
Fig. 5.3	Bar graphs created from formant values in Table 5.1	69
Fig. 5.4	Scatterplot showing the vowel space created from formant values in Table 5.1. See Activity 5-6 for guidelines.	70
Fig. 5.5	Primary cardinal vowels in a quadrilateral. Note that ‘high’ and ‘low’ are commonly also described as ‘close’ and ‘open’	72
Fig. 5.6	Front English vowels. Schematic articulatory position of the tongue on the left, vowel quadrilateral with dots indicating horizontal and vertical position within the vowel space, and the drawing of approximate formant frequencies F1 and F2	74
Fig. 5.7	Illustration of measuring duration in Praat	76
Fig. 5.8	Example of a spreadsheet file with formant values extracted (left) and the corresponding plot done in Praat (right) following the steps outlined in Activity 5-12	79
Fig. 5.9	Back English vowels. Schematic articulatory position of the tongue on the left, vowel quadrilateral with dots indicating horizontal and vertical position within the vowel space, and drawing of approximate formant frequencies F1 and F2	81
Fig. 5.10	Interval of the first word ‘morning’ in file ‘good_morning.wav’ with the [ɔ:] vowel highlighted and formant tracking problems marked with white ellipses.	82
Fig. 5.11	Visualizing NORTH (n) and FORCE (f) vowels in my pronunciation; see Activity 5-16	85
Fig. 5.12	Visualizing the movement of the tongue in ‘bide’ via the transition of formants in Praat. Spectrogram with dotted formants on the left, and the corresponding tracing of the formant movement in the 2D vowel space.	88
Fig. 5.13	Empty vowel trapezoid	89
Fig. 5.14	Schematic visualization of English diphthongs. The closing diphthongs on the left, centring on the right (see the text above). The grey tokens refer to the vowel in goat lexical set [əʊ] and [oʊ] in British and American pronunciation, respectively	90
Fig. 6.1	Acoustic characteristics in ‘fiber-viper.wav’ of the book companion. The combined periodic (shown with vertical pulses) and aperiodic components of [v] in the left panel. The shaded closure duration for [b] and the release burst of [p] with the ellipsis in the right panel	95

Fig. 6.2	Illustration of the boundary between [s] and the following vowel with the vertical solid line in ‘seed’. Praat bars to click for hearing [s] as #1 and hearing the rest of the word as #2	97
Fig. 6.3	Formant transitions in the production of ‘sod’. The left panel shows the transition with white boxes in the regular Praat visualization with time on the x-axis, the right shows the movement of the tongue in the schematized vowel space familiar from the previous chapter.	98
Fig. 6.4	Position of active articulators for three places of articulation for English stop consonants	100
Fig. 6.5	Diagrams of 4 places of articulation for English fricatives	101
Fig. 6.6	Illustration of the two phases of English affricates: complete palate-alveolar closure on the left, and narrow release resulting in the fricative on the right	104
Fig. 6.7	Illustration of bilabial nasal [m] on the left with the velum lowered and the resulting velic opening, and an oral bilabial stop on the right with the velum preventing the passage of air to the nasal cavity	105
Fig. 6.8	MRI images of the mid-sagittal plane from the sustained production of [ɹ] by 22 speakers of North American English, Tiede et al. (2004)	109
Fig. 7.1	‘a pie’ and ‘a bye’. See text for explanation of the panels and numbered vertical lines	117
Fig. 7.2	Waveforms and spectrogram of ‘spy’ with the (pink) shaded interval indicating the target interval to listen to in Activity 7-2	118
Fig. 7.3	Illustration of phonemes (abstract cognitive units) and allophones (actual realizations of the sounds produced and perceived)	120
Fig. 7.4	Phonemes and allophones of bilabial stops in Hindi or Thai.	123
Fig. 7.5	Screenshots of annotating words in Praat. See activity 7-7 for explanation. The black circle marks the dot to be clicked in order to create a boundary	128
Fig. 7.6	Illustration of the articulatory activity resulting in t-glottalization. On the left, fully released alveolar closure, in the middle (pre-)glottalized [t] and on the right full glottal stop without alveolar closure. See text for detailed description	129
Fig. 7.7	Glottalization of /t/ in ‘start’ marked with the ellipsis	130
Fig. 8.1	Sonority hierarchy with major classes of sounds. The numbers correspond to sonority levels used in the remainder of the chapter	141
Fig. 8.2	Visualization of sonority in four English words	142
Fig. 8.3	Syllable structure in ‘cramp’	144

Fig. 8.4	Phonological representations for ‘dinner’, ‘deaner’, ‘diner’ and ‘den’	152
Fig. 8.5	Phonological representations of ambisyllabicity in ‘dinner’ [din.nə]	153
Fig. 9.1	Three examples of non-sense word ‘fofofo’ with varying primary stress annotated in Praat. The solid black curve shows intensity and the greyish line shows pitch.	161
Fig. 9.2	Manipulation of pitch and duration in Praat. The original soundwave with pulses in the top panel, its pitch curve in the middle panel (larger dots represent the new and the smaller the original pitch points), and the duration panel in the bottom (lengthening of the first syllable and shortening of the third one)	163
Fig. 10.1	Visualization of the sound wave, spectrogram, f_0 tracking and text-to-speech alignment for the utterance ‘in the top right hand corner of the page’ discussed in the text.	177
Fig. 10.2	Characterization of nasal place assimilation in ‘ten million people’	181
Fig. 10.3	Characterization of manner assimilation in ‘in the’.	183
Fig. 10.4	Visualization of three target junctures in Praat. See text for detailed description.	186
Fig. 11.1	A single sentence with varying placement of pitch accents marked a dots	198
Fig. 11.2	Smoothed and interpolated pitch of an utterance consisting of five intonational units shown with a black curve. Pitch declinations over the units separately as well as over the entire utterance are shown with grey thin lines	207
Fig. 11.3	Stylized production of a nursery rhyme with dots marking strong stressed syllables.	208
Fig. 12.1	Praat’s pitch track of ‘warning_siren.wav’ showing pitch halving with 90–250 Hz range and the adjusted 90–300 Hz pitch range	213
Fig. 12.2	File ‘DSE.wav’ with the original f_0 (white larger dots) and the interpolated one (black smaller dots)	214
Fig. 12.3	Examples of unreliable f_0 contours in the vicinity of consonants. The grey box marks the interval and the dashed line through this box traces the idealized contour	216
Fig. 12.4	Alignment of pitch contours with text for ‘welcome_to_the_program_triplet.wav’ analysed in Activity 12-5. The top panel traces the raw f_0 in Hz and the bottom shows interpolated, smoothed pitch contours in semitone (based 100 Hz) scale that corresponds more to the perception of melody	218

Fig. 12.5	Examples of interlinear systems traditional in the British school of intonation analysis with dots/dashes corresponding to syllables, their thickness or size to the word stress and prominence, horizontal lines to pitch range, and the major contour movement described with stylized tails or curves	219
Fig. 12.6	Screenshot from manipulating pitch in ‘fan_of_yours.wav’. See Activity 11-8 for a detailed description.	228
Fig. 12.7	Manipulation of stylized pitch in Praat for generating various ToBI annotations	229
Fig. 13.1	Illustration of gap labelling for Activity 13-4 with the problematic overlaps discussed in the text.	242
Fig. 13.2	The input and output of the script ‘extract_gaps.praat’; see Activity 13-5 for details.	243

List of Tables

Table 3.1	The articulators, associated adjectives and the descriptions of their activity.	39
Table 5.1	Example of formant frequencies extracted with Praat in Activity 5-4 from the recording done in Activity 5-3	68
Table 5.2	Summary of the front English monophthongs.	80
Table 5.3	Summary of the back English monophthongs.	86
Table 5.4	Summary of English diphthongs	89
Table 6.1	Summary of English stop consonants	99
Table 6.2	Summary of English fricative consonants.	103
Table 6.3	Summary of English affricate consonants.	104
Table 6.4	Summary of English nasal consonants	106
Table 6.5	Summary of English approximants	111
Table 7.1	Summary of alveolar stop allophones	132
Table 8.1	Summary of the main phonotactic patterns characterizing English syllables	146
Table 9.1	Examples of derived forms with suffixes that shift the stress to the syllables preceding the suffix. IPA transcription uses Southern British English	170
Table 10.1	Environments to consider for t/d realizations. See text of Activity 10-6	187
Table 11.1	Most common weak forms in English.	203
Table 12.1	Traditional contour-based labels, ToBI annotation, and schematized f_0 for basic nuclear tones	223
Table 12.2	Summary of the schematized f_0 contours predicted in the adapted ToBI framework described in this section. As with Table 12.1, the box represents a two-syllable word with the stress on the first syllable (the shaded part) and the speaker's pitch range as the dotted midline.	230
Table 12.3	Sample tunes and meanings for 'It's nine o'clock'	231
Table 13.1	Prosodic features characterizing breath-delimited units extracted with Praat. See text and Activity 13-6 for details.	245



Introduction

1

This book assumes that you are fairly proficient, native or non-native, speakers of English. But that despite this proficiency you are not consciously aware of how speaking and pronunciation are done. It is similar to you walking automatically but not consciously grasping all the neural commands, muscle activity, the kinematics and dynamics that underlie it, or how the physical surroundings or the walker's intentions affect walking. Speaking is similarly unconscious and automatic, but also an intriguingly complex activity and behaviour guided by various types of acquired competence. The primary goal of this book is to ***bring this unconscious knowledge into your conscious awareness***.

Many traditional textbooks provide some content in a structured and pedagogically reasonable plan in an appealing form and include exercises or problems to help students apply their knowledge to novel situations or problems. It is very reasonable to 'learn about' things in this way. For example, in a textbook on anatomy, you cannot dissect your own body to explore its insides and thus you first learn about them and then apply your knowledge to practice. Or, a textbook on history normally cannot ask you to go to the libraries, archives and archeological digging sites to find out the facts and relationships among events and thus the book presents these facts and relationships. But spoken English is different. All of you can already do it and you experience it first-hand! This book thus can differ from traditional approaches and it guides you to engage with your own body and mind and to explore both physical and mental aspects of your ability to communicate in English.

1.1 What to Expect from This Book?

Instead of just learning about things in a traditional way, we will ***learn by doing*** things. This book contains many activities, asks you to engage, observe and evaluate your own spoken English or record and analyse the speech of others.

It presents tools that help you visualize and analyse speech and guides you to explore and discover on your own. The activities thus appeal to your intellectual curiosity, invite you to explore topics further, and compare your observations with the text. Hence, the book expects you to be active, and guides you towards *hands-on exploration* of speaking behaviour. Through this exploration you become consciously aware of the amazing system of subconscious speaking patterns.

In this approach, our ultimate goal is your *skill* and *understanding* and the theoretical knowledge provides a background to the skills and understanding. We are after your awareness and ability to describe speech to non-experts. I sometimes tell students that you do not learn math or programming by reading about it, you have to be doing it, you have to solve problems and exercises, produce code, etc. In this book, you should engage with speech in a similar manner through the multiple interspersed activities in the main text rather than digesting some content first and practice with exercises at the end of chapters, of which there are only a few.

A strong emphasis is therefore placed on *discovery activities* in the main text. It is essential to actually do the activities before reading the text that follows, and only after doing so, compare the observations with the text. You could also redo the activity after reading the commentary. Equally important is to attempt to *explain* these concepts in your own words to non-expert listeners (members of your family or friends; imagine you have to tell them what you are learning when you read the book). A satisfactory explanation to such people is a sure way to your understanding and it is a great mental exercise to try to imagine such a listener and describe the concepts in detail. Many activities and suggestions for projects in the text are easily adaptable for homework, labs or term projects suiting the instructor's or course's needs.

The goal of many activities is to facilitate your awareness and understanding of speech through employing all available senses. Primarily, we will use *Praat*, which is the widely available freeware tool that will be our working horse throughout the book. It has become a standard and user-friendly speech analysis and visualization software of the speech community. The functionality of Praat will be presented step-by-step starting in Chapter 4, first visualizing acoustic concepts and speech, then inspecting and measuring values of interest, annotating relevant characteristics, to basic scripting for extraction features of interest from the speech signal. The book offers a window to speaking behaviour that links the auditory information provided by the ear, the visual information provided by Praat and kinesthetic information from introspecting your own speaking. I believe that exploring a concept from these various angles brings an understanding that is deeper, more tangible, lasting and stimulates curiosity more than the understanding offered just by reading about the concept.

All of the above suggests that when reading the descriptions and deepening your understanding of speaking behaviour, the aim is not to categorize 'correct' and 'incorrect' speaking patterns. In other words, the approach of the book is *not prescriptive*. Rather, the focus is on building your awareness of speech patterns and understanding their realizations and functions in a particular context. This, in

turn, will enable you to analyse novel excerpts of speech or dialogue in different contexts on your own and describe in your own words what, and why, most likely is happening in the speaker's mouth and mind. In other words, our approach is *descriptive* in that we are trying to understand how the skilled behaviour of speaking is effortlessly adapted to the situation in a similar way to other forms of skilled behaviour (sport, playing an instrument, handicraft, etc.). Hence, rather than saying what a speaker should do, we are trying to understand the choices she has at her disposal and the reasons leading her to pick some of them.

Consequently, the patterns arising from exploring your own recordings in the activities with Praat might not always correspond perfectly (or even generally) with the discussion in the text that is based on other phonetic studies. It is natural that speech varies; a single observation from you might be different from mine due to various reasons (specificity of the tokens, recording context, the speaker, the quality of the recording, etc.). We assume that if you recorded more people, e.g., from your family and friends, most renditions would be like those discussed in the text. The patterns discussed in this book are those that are most likely to stand out, or those that majority of recordings typically have in common. The book also gives you the tools for designing and carrying out small projects when a particular speech pattern might be explored with more speakers and tokens for gaining more objective information.

1.2 Who Is This Book for?

Both *native (L1)* and *non-native (L2)* speakers of English are the target audience of this book. For native speakers, the activities and discussions strengthen the awareness and self-exploration of their everyday speaking by focusing consciously on the production and auditory characteristics of speech. The non-native speakers will discover how the speaking patterns of English differ from the patterns they have in their native language, and become more aware of the interference the native language patterns might have on speaking English. For non-natives I might occasionally give some guidelines to working on these new patterns to approximate their non-native English to native English if they wish to do so. However, the book is not going to provide specific instructions or discuss the interference of various L1 mother tongues on speaking L2 English. For this awareness, the instructor of the course is an expert. That said, the skills in experimentally investigating speaking patterns can be put to use in multiple mini-projects comparing L1 and L2 English.

Many students approaching a course or a module using this book naturally ask: Why do I need this, or how will this be useful in my future life? Whether you are an L1 or L2 student of English, linguistics, other languages, speech pathology and audiology, interested in developing applications in human-machine spoken interaction, attending a general education course or just simply interested in speech, engaging with the material in this book will help you learn more about yourself by better understanding the underlying principles of spoken communication,

and also introduce you to the scientific approach to studying speech. While the book content covers the basics of English phonetics and phonology, the skills and understanding are applicable very broadly and generally; they include, but are not limited to, observing in detail, finding patterns in data, applying understanding to novel situations, forming hypotheses and potential alternatives and checking them with data (evidence), using computer software to build support for conclusions, discovering the relationships between seemingly separate phenomena or actions, and hopefully many others. These, I believe, are valuable in essentially any job or position and expected from university-educated people in the twenty-first century.

1.3 How Is This Book Structured?

Chapters 2–13 cover most of the material found in modules on introductory English phonetics and phonology but use the novel approach described above. I tried to present material in a way that is as theory-neutral as possible, and provide good descriptions of relevant observations rather than elements of any particular theory. Chapter 14 brings all the concepts and understanding together with commentaries and descriptive analysis of three short excerpts of dialogues whose parts were previously discussed in chapters as individual files. Both British and American English are covered as the main varieties. My intention is that readers will read the commentaries while listening to, visually examining, and in the case of L2 speakers possibly imitating, the audio signal of the speech under discussion using the Praat interface.

The *companion* website for the book is hosted at chapter level on the Palgrave and SpringerLink websites. It contains all sound material, associated Praat annotation files or Praat scripts if appropriate, relevant for the discussion in each chapter. Please do not treat it as an extra or additional supplement but as an integral part of your engagement with the book. Each chapter opening provides the link for that chapter material, and individual activities will direct you to the appropriate companion files. Additionally, all internet links can be accessed from there as well.

The sound material in the companion comes from several sources. First, there are sections from three interviews aired in National Public Radio (NPR).¹ Several

¹I am thankful for the permission to use these excerpts from NPR identified as follows:

©2016 National Public Radio, Inc. Excerpts from news report titled “‘British Bake Off’ Winner Takes On The Toughest Judge Of All: The Queen” were originally broadcast on NPR’s Morning Edition on April 21, 2016, and are used with the permission of NPR. Any unauthorized duplication is strictly prohibited.

©2018 National Public Radio, Inc. Excerpt from news report titled “‘It’s Lovely Being Mean’: Benedict Cumberbatch Gets Into Character As The Grinch” was originally broadcast on NPR’s Weekend Edition Sunday on November 11, 2018, and is used with the permission of NPR. Any unauthorized duplication is strictly prohibited.

©2018 National Public Radio, Inc. Excerpt from news report titled “Jeeves And Wooster, But Make It A Modern Spy Novel” was originally broadcast on NPR’s Weekend Edition Sunday on December 2, 2018, and is used with the permission of NPR. Any unauthorized duplication is strictly prohibited.

sound clips come from Columbia Games Corpus and are used with permission from the creators of the corpus, Julia Hirschberg and Agustin Gravano. Then there are individual short excerpts from other native speakers and myself as an L2 speaker. The quality of the recordings varies, which is intentional since your own recordings might not reach the professional sound quality.

In addition to the core activities-centred discussion covering the basic material in English phonetics and phonology, there are also several sections marked as ‘advanced’ or ‘find out more’ offering more in-depth discussion or examples on selected points and not critical for the discussion in the rest of the book. The references in the main text are kept at a minimum. Instead of suggested readings for each chapter, I provide a very short general list of potentially useful further reading suggestions at the end of the book.

Finally, the book uses the following notational scheme. Square brackets for [phonetic realization], slashes for /intended sounds/, single quotes for English ‘utterances, phrases, or words under the discussion’, pointy brackets for <spelling>, bolded italics for ***important concepts***, plain italics for *Praat commands* or *buttons*.

I hope you enjoy our exploration and find it intellectually stimulating.



Fundamental Concepts

2

In this chapter we will frame the discussion in the book by:

- describing speaking as a skilled habitual activity similar to playing a musical instrument or performing any sporting activity,
- presenting two hallmark features of speaking—its continuous overlapping nature and necessity to be contrastive,
- discussing the concepts of variability and context that are essential for exploring speech,
- introducing and motivating the use of International Phonetic Alphabet (IPA) for transcribing speech.

2.1 Speaking as Acquired Habits

Activity 2-1: Catching a ball

Ask somebody to throw a ball, a pen, or any small object in your direction so that you can catch it with two hands. Repeat two or three times. Try to **describe verbally** how you achieve this simple action. Literally, STOP here and describe before you continue reading. Remember from Chapter 1, doing what activities ask you to do is extremely important for making most of this book. Can the desired goal of catching the object be broken to subroutines? Note how each of the throws was slightly different and how you made the adjustments to catch it each time. Or, if you failed to catch, what went wrong?

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_2) contains supplementary material, which is available to authorized users.

Here is a subset of things you might have mentioned in Activity 2-1. You have to visually observe the person throwing the object and deduce the speed of his/her arm movement and the angle at which the object is released to estimate the direction and the velocity of the object at the release. Then you have to estimate and constantly monitor the trajectory and the approximate time at which the object will appear within the reach of your hands. If your estimation does not fall within this reach, you have to move your body. With this assessment, and before the object reaches the location in front of your hands, you have to send signals through your nervous system to the right muscles so that your hands approximate at the precise time point, precise location, and with precise force to successfully catch the object. If anything is off, you are either too early, or too late, or too much to the right or left, or apply too little or too much force, you fail to catch it. Hence, you might be unsuccessful in any number of possible ways and some people might laugh at you. In short, catching an object thrown at you requires *perception*, *action*, and a very tight *coordination* at multiple levels.

You probably never analysed simple ball catching in this way. You simply have the skill for catching it, you know how to do it, but you do it intuitively, automatically, without grasping rationally what is actually done in order to be successful. Hence, in this sense, ball catching is *subconscious knowledge*.

Speaking your native language is a similarly awesome and exciting feat that we do unconsciously, automatically, but we hardly ever pause to consciously ponder how we actually do it. Speaking is thus also a form of subconscious activity and saying even simplest words or phrases like ‘bag’ or ‘on the top’ is similar to catching a ball. It involves articulatory actions of various organs and requires precise coordination among these actions and monitoring. When speaking is part of interpersonal communication, it also involves the perception of the interlocutor’s actions and the adjustments of your speech production in view of these actions.

Activity 2-2: Introspecting speaking

Say a simple phrase such as ‘I’m thinking of you’, or something similar in your native language, and introspect your own speaking. Try to consciously be aware of the movement of your tongue, the lips, the jaw and how they fluidly coordinate with one another. In what way is this introspection different from introspecting the ball catching in Activity 2-1? Is it more difficult?

Find Out More 2-1

You can find several online resources such as YouTube videos that provide a nice illustration that might help the introspection of your own speaking. For example, <https://www.youtube.com/watch?v=Z8yysjQeYT4> shows human speech with the help of MRI technology. Even without knowing about the names of the organs involved in speaking and physiological processes, you can appreciate how the various actions have to be precisely coordinated, akin to a gymnast or a dancer performing her routines, for the desired speech to come out.

Now, think about how you developed this skill and knowledge. First, you probably do not remember any conscious effort spent on learning how to catch objects thrown at you or how to speak. You acquired object-catching through gradual steps in learning how to grasp, developing hand–eye coordination and other necessary subroutines while being a baby. You developed this *habit* through countless trials and errors, making adjustments, gradually increasing difficulty.

In speaking, habit formation also starts early. If you ever were around small infants you may notice their experimentation with vocalizing called *babbling*. In this stage of language development, typically around 4–12 months of age, babies learn the correspondence between various combinations and timings of articulatory movements and the resulting acoustic and perceptual output they achieve. In most cases, this is a very repetitive process in which the baby introduces small variations to its production, after which it might go back, and then again try something new. Of course, infants also grow rapidly and thus the habits they are acquiring in producing their first simple syllables like ‘ba-ma-da-ga’ must be robust and adjustable to a constantly growing vocal tract. Babbling is typically a solitary yet extremely pleasing activity when the baby is alone but eventually it tries to produce sounds similar to the ones she hears around and responds to feedback provided by the caregivers (imagine the response of parents to baby’s first recognizable words).

Find Out More 2-2

Babbling can be seen in several videos like the following. Note how the first one nicely captures the solitary stage and the second already involves the communication with a caregiver.

- https://upload.wikimedia.org/wikipedia/commons/8/84/Infant_babbling_in_crib.ogv
- <https://www.youtube.com/watch?v=sMaxy8uaJjY>

Consider now how many times in your life you have said and practiced your ‘ma’ or ‘ba’ and under how many various circumstances. Think of all words that begin with these syllables, or end with them, whether you said them loudly, when you were excited, or emphasized them to get your point across to your interlocutor. And consider the variability provided by the ever-changing context of the words that were preceding and following these syllables. Speaking is **habitual subconscious knowledge** being practiced daily for years and fluidly adapting to changes in environments.

However, characterizing speaking habits only from the more physical aspects of articulation and perception would be incomplete since these habits also involve multiple levels of *mental knowledge*.

Activity 2-3: English plurals

As speakers of English you know that the plural of ‘dog’ is ‘dogs’ and the plural of ‘cat’ is ‘cats’. Do you say the final <s> in these two words identically? If you

are a native speaker introspect your own speaking and the non-native speakers are advised to elicit these words from native English speakers if available. Check with other simple words like ‘tables’, ‘tablets’, ‘kids’, ‘guys’, ‘taps’, etc.

You probably found that words like ‘dogs’, ‘tables’, ‘guys’ or ‘kids’ actually end with a [z]-like sound (the initial sound of ‘zebra’) while words like ‘cats’, ‘tablets’ or ‘taps’ end with a [s]-like sound (the initial sound of ‘snake’). Hence, despite identical spelling of the plural marker <s>, the pronunciation differs. Some non-native speakers might have been told by their English teachers, and some native speakers might know for other reasons, but there might be many English speakers who were not aware of this before reading this paragraph. Crucially, native speakers **know** subconsciously whether to make a plural of any English word with a [s] or [z], or possibly [ɪz] for plurals of words like ‘rose’ or ‘church’. This is an example of systematic mental knowledge involved in speaking habits.

The explanation of what it is the native speakers actually know regarding the pronunciation of plurals will come later. Let’s consider now how they acquired this habit. A plausible hypothesis is that they heard from people around them during speech development and remembered. This hypothesis was tested in a famous wug test experiment in 1950s when Berko-Gleason created cute hand-drawn pictures of animal-like creatures and gave them nonsense names such as ‘wug’ or ‘blick’. She showed a picture with one creature to 3–4-year-old kids and said ‘Look, this is a wug’. She then showed a picture with two such creatures and analysed whether kids said ‘wug[s]’ or ‘wug[z]’ and ‘blick[s]’ or ‘blick[z]’; a small percentage also said ‘wug[ɪz]’ and blick[sɪz] like Gollum in the Tolkien’s Lord of the Rings. Berko-Gleason found out that kids systematically said ‘wug[z]’ and ‘blick[s]’ despite the fact that these kids never heard these words, much less their plural forms, before. Hence, habits like these are not acquired by repetition or memorization but by unconsciously extracting patterns from the speech around the kids during their language acquisition.

Habits involved in speaking have developed unconsciously for native (L1) speakers. This applies to both the physical and mental aspects of speaking individual words as well as social communication skills in understanding when to say things or in what tone or intonation. For non-native (L2) speakers things are a bit different. Languages differ in the speaking patterns their L1 speakers use; some of the habits in your L1 speech are necessarily different from the habits of the L2 language. This is essentially the nature of the L2 accent that native speakers are so good at recognizing almost immediately after several utterances despite flawless grammar and style. The habits that work perfectly for native speech, for example, ending every sentence with a rising intonation or ignoring all /h/’s in the pronunciation, might be considered ‘bad’ or interfering when speaking a non-native language.

Therefore, mastering habits in L2 typically requires conscious effort, awareness of the differences between L1 and L2 habits, and practice in performing the habits naturally and smoothly. Consider playing a musical instrument or being good at any type of sport or craft. Those of you who possess skills like these can bear

witness that in most cases, you become good at something by training and practicing certain moves and their sequences over and over and over again. Think of how much practice must have gone into mastering the smooth and elegant back-hand stroke by Roger Federer, stylish and graceful turns on skis by Mikaela Shiffrin, or almost indefensible free kicks by Cristiano Ronaldo. Or think of your favourite musician, painter, or craftsman and the routines they must have been practicing for hours daily to make the results look effortless and beautiful.

These long-term repeated actions slowly develop into unconscious automatic robust habits, both physical and mental, primarily by countless *adjustments* to various conditions and environments in performing these actions.

Find Out More 2-3

A Washington Post wonkblog piece reviews research suggesting that learning a skill (e.g. swinging a baseball bat, hitting tennis backhands or playing difficult guitar chord successions) is more efficient if you introduce subtle variations into the practice routines (different bat weights, sizes of the tennis racket or playing cord sets in different tunes) https://www.washingtonpost.com/news/wonk/wp/2016/02/12/how-to-learn-new-skills-twice-as-fast/?utm_term=.b9822520b364.

Speaking English, or any other language, as your L2 involves similar procedures. To make it even worse, L2 speakers do not need to just develop new habits, they have to change the already deeply ingrained ones, which is very hard. Have you ever tried to change an already well-developed habit? Biting your nails, smoking, eliminating late-night snacking or any routine in sports/music/crafts that was hampering your progress. If you have gone through this, you know how difficult it is. If you have not done so, you probably know of the experience of somebody close to you who has. Just like habits mentioned above, changing already mastered subroutines in speaking requires the level of commitment that is the same, if not greater, than that involved in building a skill anew. Hence, by building conscious awareness of the habits involved in speaking English in this book, non-native speakers will, indirectly, become more aware of the unconscious habits involved in their own native language, and native English speakers will find out more about their unconscious habits.

2.2 Continuity and Contrastiveness of Speech

In tennis, the players' movements are smooth and fluid and they do not stop their run towards the ball to hit a forehand and break again to run back to the centre of the court. Rather, they run and start the motions involved in executing the forehand while they are still running and commonly start running back while still involved in the follow-through motions after the stroke. We do recognize that there is 'running' and 'striking the ball' involved since both of these could be performed

separately: a tennis player can just run, or she can hit the forehand strokes while standing. Crucially, despite the clear intuitive identification of these two macroscopic routines, the clear boundary separating the two actions during the game cannot be easily identified. This is because there will always be a temporal interval during which the tennis player is both running and in the motion of hitting at the same time.

In speaking, exactly the same system is at work as well. Find out for yourself in Activity 2-4 before reading further.

Activity 2-4: Lips and tongue in ‘beet’, ‘boot’ and ‘bat’

Say these three words: ‘beet’, ‘boot’ and ‘bat’. Now say the words again slowly and stop just at the point you were to open the lips. Feel the position of the lips and the tongue. Do you notice the differences? Now repeat the same but this time observe your lips in a mirror when pronouncing these words. How does your proprioceptive ‘feel’ compare to the visual information? Describe the differences in producing these three words informally and then compare with the text below.

At the temporal point of opening your lips, in ‘beet’, the lips are likely slightly spread while your tongue is up and front with its sides firmly touching your upper molars. In ‘boot’, you notice your lips slightly protruded and rounded while you do not feel any contact of your tongue with the teeth and might feel a somewhat retracted position of the tongue. And in ‘bat’ you probably feel how your tongue is quite low. Crucially, I asked you to observe these differences while the lips are still closed, hence before the release of /b/. And thus, in each of the three cases, there was an interval when you were still saying a /b/, since the lips were closed, but at the same time you have started the motions for saying the following vowel, since the positions of the lips and the tongue were different.

The crucial take-home point is that speaking is *continuous* in the sense that the production of individual sounds, or subroutines, is always influenced by the context in which they occur. And this is fundamentally different from what we might think based on letters on a page. If you take the word ‘bat’, it seems that just like there are three letters <b, a, t>, there are also three sounds corresponding to the letters and that one could easily differentiate a /b/ from the following /a/. However, this is not how it works in real speech. It is extremely difficult, sometimes impossible, to say at which point the /b/ ends and the /a/ starts.

To corroborate your own feelings and observations, below we discuss two different ways of examining the articulatory production of ‘ba’; one is in the main text, the second in Activity 2-5 below that. This is the first encounter with this book’s approach to understanding speech that is based on complementing your own intuitions and self-observations with the visualization and examination of speech data. Although this might be challenging at times, the goal is to grasp the gist of the discussed pattern seen in the data and associate it with your own speaking. There is no need to completely understand every single aspect of speech visualization.

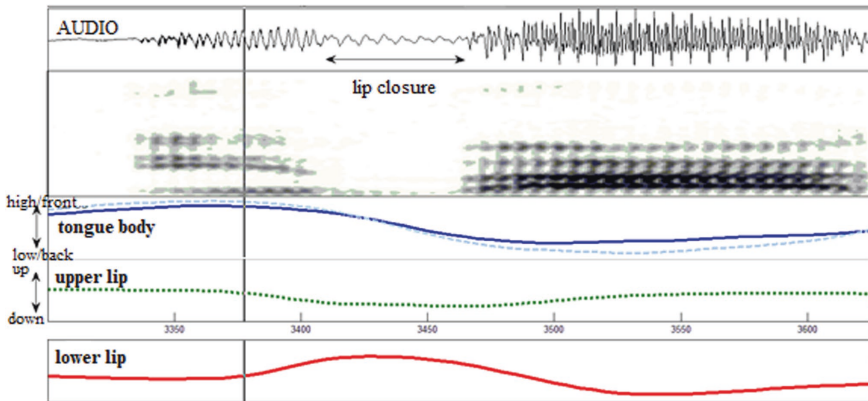


Fig. 2.1 Acoustic and articulatory data from the production of the word ‘iba’ in Slovak. See text for explanation

Figure 2.1 shows five vertically arranged panels of acoustic and articulatory data when I say ‘iba’, which is a Slovak word meaning ‘only’. The x-axis shows time in milliseconds. The first and second panels show the acoustic record captured with a microphone (oscillogram and spectrogram) and are not crucial at this point (we will cover acoustics in Chapter 4 and this display will make more sense then).

Let’s focus on the bottom three panels that show data captured with electromagnetometry. This is a technique in which small sensors (about 2 mm) are glued to articulators like the lips, the jaw, and the tongue, and due to electromagnetic fields we are able to visualize the movements of these small sensors. Hence, the third panel with one solid and one dashed line shows the vertical and horizontal movements of the sensor attached to the tongue body, respectively, with the tongue being in the high and front position when the values are high and in the low and back position when the values are low. The bottom two panels show the vertical movement of the sensors glued to the upper and lower lip, respectively. The horizontal bidirectional arrow in the first panel shows the interval during which my lips are closed for /b/ sound of ‘iba’. You may spend a minute trying to follow the lines from left to right as you say /i/, like the first sound of ‘eat’ with the tongue high and front, then gradually closing the lips for /b/ and opening them saying /a/, like ‘bar’ in British English with the tongue low and back.

Now that you have a good idea of the speech data represented in the picture, let’s look at the vertical solid line. It shows the approximate time point at which the lips started their closing movement: we see a very robust raising of the lower lip and somewhat less pronounced lowering of the upper lip. The crucial point of interest is the movement of the tongue. While the initial vowel /i/ requires a front and high position of the tongue (seen as the high values of the tongue body sensor before the lip closure), /a/ is the vowel requiring a somewhat low and back position of the tongue in the mouth, which is shown with the low values of the tongue body after the lip closure. You might try alternating [ii-aa-ii-aa], which is

roughly the sound of donkeys in English ('hee-haw', 'eeyore'), to feel this movement of the tongue better. Note in the figure that the tongue starts moving downwards and backwards well before the lips become open and we 'hear' the sound /a/. It is even before the lips close together, approximately around the same time as the lips themselves start closing. Hence, I start producing the third sound /a/ of 'iba' during the production of the first vowel /i/. This shows that the *continuity* of speech does not include only immediately adjacent sounds, as we saw in Activity 2-4 but may involve also sounds that are not adjacent in the traditional spelling. This discussion thus illustrates the physical evidence that the production of /b/ and /a/ specifically, and other adjacent (and possibly non-adjacent) sounds in speaking, *overlap in time* just like hitting and running overlap in tennis.

Importantly, notice also the smoothness of the lines representing the vertical and horizontal movement of the sensors attached to the articulators. This is the evidence of fluid motion without any of the jerkiness that would be expected if we produced individual sounds separately. Smoothness and fluidity are hallmarks of virtually any human physical activity including speaking.

Activity 2-5 looks at continuity with slightly different data to facilitate building awareness of these crucial characteristics of speech.

Activity 2-5: Examining 'ba' with an x-ray movie

Locate the 'biography.mp4' file in the book companion and open it.¹ It is important to be able to move frame-by-frame, which is easily done with arrows in Windows Media, but any other software with this capability that you are familiar with is fine. Play the file several times and then navigate by forwarding frame-by-frame to the initial frame of lip closing for /b/ in 'biography' using the bottom x-axis of time roughly when the vertical line reaches 0.9 as shown in Fig. 2.2. Now move frame-by-frame 3–4 times and observe that lips are still closed but the tongue moves slightly. Describe how the tongue moves and speculate why it does. Then continue reading the main text.

In Activity 2-5 you inspected an x-ray movie of producing the first syllable of the word 'biography' from an utterance 'Is that biography?' Hence, you saw an example of a 'ba' production with different methodology from that in Fig. 2.1.

Figure 2.3 shows 5 successive frames from this 'ba'. The outer edges of the lips in the grey shade on the right side of the pictures and the edge of tongue back in the left side of the pictures are highlighted in the first frame for better illustration. The white areas in the middle are fillings in the top and bottom molars. The

¹I am grateful to Phil Hoole for allowing me to use this movie clip. A detailed information about the origin and history of the movie can be found at his website <https://www.phonetik.uni-muenchen.de/~hoole/kurse/movies/xray/xrayreadme.pdf>. Also consult the following link and the references there for the description of the original material: <https://www.queensu.ca/psychology/speech-perception-and-production-lab/x-ray-database>.

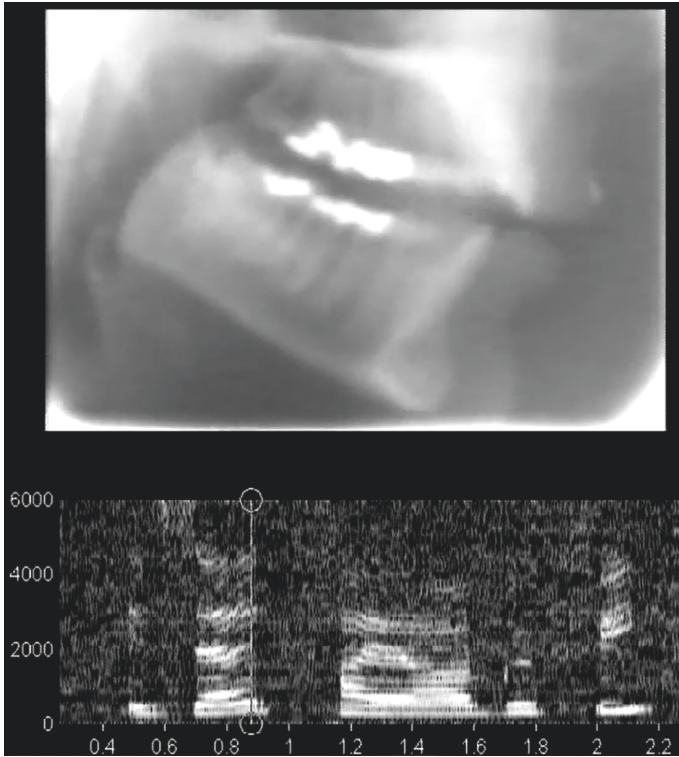


Fig. 2.2 The first frame of b-closure in ‘biography’. See Activity 2-4 for explanation



Fig. 2.3 Five successive x-ray frames during ‘ba’. See text for explanations

black dot is a fixed reference point. During the first frame the lips close and they remained closed until the fifth frame in which they are already slightly open.

The movie from Activity 2-5 and still frames in Fig. 2.3 both show that in each frame the tongue moves progressively more backwards as can be seen by the decreasing distance between the tongue edge and the fixed black dot. This is because [a] is a vowel requiring the back position of the tongue and the preceding vowel [æ] of ‘that’ requiring the front position and thus the tongue moves backwards during the lip closure of [b]. So when the lips finally open in the fifth frame, the tongue is almost ready in the position needed for [a].

We have experienced continuity both through introspection with the three /b/s in ‘beet’, ‘boot’ and ‘bat’ that had all quite different makeups of the lips and

tongue configuration, as well as ‘ba’ with the x-ray or electromagnetometry data above. They all show that we do not first say /b/ and then /a/ but that in the production of ‘ba’ the articulators move in a highly overlapped, continuous and coordinated way.

Some of you might find solace in the fact that at least in /b/ the lips are always closed. Hence, it makes sense that /b/ is ‘separate’ from /a/ because in /b/ the lips are closed and in /a/ they are not and thus irrespective of the protrusion, rounding, spreading of the lips or the movement of the tongue, /b/ is stable in this way. Well... yes and no. Although normally, the complete lip closure is the *goal* that we are trying to achieve, we sometime either might not succeed or simply do not have this goal. Consider, for example, situations when /b/ is produced by somebody who smokes and speaks while having a cigarette in his/her mouth, or is panting heavily, or speaks Spanish and says ‘trabajo’ for ‘work’, etc. In all these cases, the complete lip closure normally does not take place. Additionally, there are sounds other than /b/ that do have the closed lips. Hence, we are really talking about our understanding of how a /b/ is expected to be produced.

If it is difficult to identify the boundaries between individual sounds, then the very reality of these sounds as separate and independent units might be questionable. However, language and our spoken communication depend on there being something we can call /b/. This is because language is a collection of arbitrary mappings between sounds and meanings, and to convey different meanings, the sounds must be *contrastive*. What I mean is that if we want to convey, for example, two different meanings like ‘bet’ and ‘vet’, the interlocutors must perceive a clear contrast between a /b/ and a /v/ to convey the difference between the two words. This is an example of one of the hallmarks of speech and language; we use a small finite set of discrete units (in this case sounds like /b/, /v/, /e/ or /t/) and following certain principles, both general and particular for each language, these units may be combined into larger units like syllables or words.

But wait! You should probably be somewhat puzzled by now. Just two paragraphs above I was questioning the very reality of something like a /b/ since we cannot identify its beginnings and ends and it is always differently produced depending on the context in which it appears. But now I am also saying that units like /b/ are indispensable for our speaking and communication since without perceiving them as contrastive units we would not be able to convey differences in meanings and recombine them to form a virtually infinite number of new possible words. These two claims seem to be in conflict that is an instance of a rather old conflict in the philosophy of mind and cognitive science relating to the differences in, and possible interactions between, our *bodies* and our *minds*. We will not venture into these debates in detail here but the awareness of the mind–body dichotomy in speech is crucial for our project of investigating spoken English.

On the one hand, there is speech observable by analysing our articulatory movements. These aspects of speaking we can physically measure by applying the methods of natural sciences. For example, we can measure the tongue movements from x-ray or ultrasound images or very precisely track the movement of the sensor attached to the tongue with electromagnetometry. The same applies to

measuring the acoustic records of speech that we will introduce later, or even neurological activity in our brains associated with our speaking. Importantly, all such measurements show that speech is continuous and varies a lot. This is, of course, not surprising given the similarity of speech with other skilled physical movements like sports or music.

On the other hand, there are also *mental shortcuts* of these continuous events in our minds like ‘this is a /b/’, or ‘say [z] when creating the plural of “wug”’. We cannot really directly measure our mental processes like these in the physical sense. Therefore, this type of our subconscious knowledge is fruitfully approached by leaving the continuity of speech aside and assuming stability and contrastiveness of the units corresponding to individual sounds.

Hence, we resolve the basic conflict between the continuous nature of speech and the necessarily discrete character of sounds, syllables or words as follows. When talking about individual sounds, we will be aware that their discreteness is an *abstraction* from our bodies (our muscles, movements, ears, neurons firing) to our minds. In other words, we will assume that any sound, syllable or word can be identified as a stable unit and that the continuous flow of speech can be analysed as a sequence of these units similar to the impression given by separate letters or words appearing as a text on a page. But we will know that in actual speaking, the realization of these abstract units is always continuous and variable. For example, we will know that although the goal for the lips is to be closed during /b/, they are always slightly more/less protruded, more/less rounded, more/less spread, the tongue is more/less forward, upward, etc. Returning again to tennis, we know that all the movements during the match are fluid and continuous in real life and that there is no single boundary between running towards a ball and hitting it. On the other hand, we also know that the tennis routines can be broken to abstract discrete mental representations like RUNNING, HITTING or JUMPING and this is in fact how our minds tend to perceive the actions when we watch a tennis match.

This discussion corroborates the end of the previous section in that speaking involves both (1) acquired habitual skills of continuous physical activity influenced by context, like articulating ‘iba’, and (2) mental knowledge like knowing if the plural should end with a [z] or a [s]. *These two domains of mind–body dichotomy form two interacting sides of a single complex cognitive system of speech.* This is the essence of the terms phonetics and phonology from the subtitle of this book. The continuous, directly observable and measurable aspects of speaking represent the object of study of *phonetics*. The discrete, contrastive aspects that are not directly observable, and also the modelling of these aspects, is the domain of *phonology*. Both separately, and even better when considered together, provide a fruitful approach to understanding the cognitive system underlying our speaking behaviour. And this approach will be used to describe various aspects of speaking English in this book.

To appreciate the beauty and challenges of this approach and also relate it to almost any other scientific endeavour, I think that A. Einstein and L. Infeld put it extremely aptly:

In our endeavour to understand the reality we are somewhat like a man trying to understand the mechanism of a closed watch. He sees the face and the moving hands, even hears its ticking, but he has no way of opening the case. If he is ingenious, he may form some picture of a mechanism which could be responsible for all the things he observes, but he may never be quite sure his picture is the only one which could explain his observations. [...] But he certainly believes that, as his knowledge increases, his picture of reality will become simpler and simpler and will explain wider and wider range of his sensuous impressions. (Einstein and Infeld 1938, p. 33)

For me, the ‘closed watch’ represents the phonological knowledge, our mind. In the quest for trying to understand these mental aspects of our speaking competence we have to rely only on analysing the observable physical aspects of our everyday speaking behaviour, that is the ticking of the watch and moving of the hands. Hence, just like the ticking and moving represent the window into the closed clock, phonetic observable features of speaking provide a window into our minds. This book is your guide to peeking through that window.

2.3 Context and Variability

The previous section introduced continuity as the fundamental feature of speech. Continuity also serves as the first dimension of how *context* affects speech. The observation that the lip positions during the production of [b] is affected by the quality of the following vowel in words like ‘beet’ or ‘boot’ can be formulated as follows: in the context of the following [u], the closed lips during [b] are slightly rounded and protruded. Hence, lip closure for [b] and lip rounding/protrusion for [u] overlap in time just like running and hitting in tennis and this temporal overlap in speech is called *coarticulation*.

Another fundamental feature of speech is *variability*. Let me illustrate using the context again. The characteristics of [b] *vary* depending on what vowel follows it in ‘beet’ or ‘boot’. Hence, context and variability are intimately intertwined in speech: different contexts produce systematic variation in our speaking behaviour. The relationship between context and variability is the essence of the cognitive approach to speech. We are interested to find out which contextual factors are cognitively meaningful and result in the systematic and predictable speaking habits.

So far I have mostly implied that context can be construed in a narrow sense; for example, when talking about the way [b] and the immediately adjacent vowel influence each other. However, I will argue below that context is best understood as a *multidimensional continuum* of effects influencing every aspect of speaking. The rest of the section is devoted to illustrating this idea and by doing that, to framing much of the discussion in the book.

We start with inspecting the first vowel of the words ‘object’. Activity 2-6 describes the tasks.

Activity 2-6: Contextual variability in the word ‘object’

First, say a short phrase ‘a long object’. Try to verbalize how the immediate preceding and following context affects the production of the initial vowel in ‘object’. Hint: try to vary the context and introspect how your movements change. For example, try alternating ‘a long object’ with ‘the object’ and ‘long odds’, respectively, and placing your finger on the jaw.

Second, compare the production of the same initial vowel in words ‘object’ and ‘option’. Do you observe any differences? What about the jaw movement?

Third, you probably know that ‘object’ is an ambiguous word and might refer to a thing when used as a noun but can also mean to oppose something if used as a verb. Describe informally in your own words how the production of the first vowel differs in these two meanings. It is useful to look at yourself in the mirror saying ‘object’ in these two meanings.

We take the three questions from Activity 2-6 in turn. You might feel that your jaw makes a more pronounced downward movement in ‘the object’ than in ‘a long object’ since it is relatively high for the vowel of ‘the’ and starts moving down much later than in ‘a long object’ for which your jaw is already pretty low during the production of ‘long’. On the other hand, comparing ‘object’ and ‘odds’ you might notice the upward jaw movement during the vowel in ‘object’ (to assist the creation of the complete lip closure for /b/) but slightly smaller jaw movement in ‘odds’.

While the first question about ‘object’ focused on the contextual variability in the spatial characteristics of speaking (what moves to what position), the second question targeted the temporal variability. In both ‘object’ and ‘option’ the initial vowel is followed by closing the lips and thus spatially the initial vowels in both words should be similar. But when native speakers say these words the vowel in the first word is typically slightly longer than the same vowel of the second word. I downloaded the sound files of these two words from the online Cambridge Dictionary and then measured the duration from the beginning of the word to the time point when the lips are closed.² This interval was approximately 20% longer in ‘object’ than in ‘option’. Hence, the contextual effects of immediately adjacent sounds can result in either *spatial or temporal variability* of the individual sounds. These so-called allophonic patterns will be discussed in more detail in Chapter 7.

The third task of Activity 2-6 probably elicited your descriptions in which both spatial and temporal variability was mentioned since the production of the first sounds in these two meanings is radically different. In the noun you say a full vowel that is similar to the vowel in the word ‘dog’ with an open mouth. If, however, you said it as a verb, to object to something, the first sound is a so-called

²We will explain how measurements like these can be done in Chapter 4 and subsequent chapters of this book hence you will be able to do this yourself. You can listen to ‘object’ in <https://dictionary.cambridge.org/dictionary/english/object> and ‘option’ in <https://dictionary.cambridge.org/dictionary/english/option>.

schwa [ə] with a very neutral position of the jaw and the lips, and its duration is significantly shorter than the first vowel of the noun.

This noun–verb difference is the contrast arising from *grammatical categories*. In English, the mental knowledge of how to signal this difference involves knowing whether to highlight the first syllable, as in the noun, or the second one, as in the verb. This will be discussed in Chapter 9 showing how discrete-like knowledge about word stress interacts with continuous-like patterns of articulatory activity.

Word stress is an example of a prosodic contrast that is an extremely potent source of contextual variation. Understanding how our speaking habits are modified under various *prosodic contexts* will be the main focus of the chapters in the second half of this book. Here we illustrate this dimension of context with two more examples in Activity 2-7. As before, you should try to describe informally your intuition and assessment of your pronunciation of these words and then compare with the discussion below.

Activity 2-7: Realization of ‘object’ in additional prosodic contexts

First, consider excerpt #1 below, extracted from the International Corpus of English (Nelson et al. 2002; unit S1B-002 #68, speaker Prof. Neil Smith, UCL, 1991). In your own production of this utterance, which of the two mentions of ‘object’ are more prominent and/or more highlighted? How does this difference transfer to the articulation and quality of the word in general and the first vowel of this word in particular?

Second, consider excerpt #2 below, extracted from the Columbia Games Corpus (Gravano and Hirschberg 2011). What differences do you expect based on the transcribed text? What contextual aspects might affect the way the two tokens of ‘object’ are phonetically realized?

1. So now we can have people object to it, if anybody wishes to object to it.
2. The blinking object is the M&M man. A smaller object, while this is like longer

Only after considering these tasks and verbalizing your responses, listen to the sound files in the book companion. You can use any software for playing sounds you are comfortable with. Excerpt #1 refers to ‘object_1_ICE.wav’ and I apologize for poor sound quality. Excerpt #2 refers to ‘object_2_CGC.wav’. Do you feel you need to amend or adjust your original evaluation?

In excerpt #1 of Activity 2-7, the word ‘object’ is itself highlighted as important when mentioned the first time. It is common that when some word or concept is mentioned for the first time in some conversation, it would be highlighted and salient. When mentioned for the second or subsequent times, it might be considered less important in a given situation since it has already been mentioned recently and is active in the interlocutor’s mind. This difference is reflected in many phonetic aspects of the word production. For example, in excerpt #1 the first mention of ‘object’ is longer, on average with higher pitch and intensity than the second mention.

Another prosodic contextual dimension involves the way people chunk their utterances in smaller units and where the words are with respect to the boundaries of these units. For example, words preceding a chunk boundary are typically longer than those not preceding one. In excerpt #2 of Activity 2-7, the first mention of ‘object’ is about 75% of the length of the second one. Of course, there are many other differences between the two mentions including pitch or vocal quality since the second mention has a distinctly creaky voice. Note also how the position within the chunks and the first or second mentioning might mutually interact. In excerpt #2, the first mention is shorter than the second one, which is in contrast to how we described excerpt #1. This is due to the position of ‘object’ within the chunks that affects the duration of ‘object’ more than the fact that is mentioned for the first or second time.

The investigations of how the measurable differences in various continuous features like duration, or pitch relate to different intentions of the speakers feature prominently in the book. In this book, you will find multiple hands-on activities and step-by-step instructions and opportunities for engagement with various types of sound files, measuring various aspects of speaking habits, thus exploring the link between how things are said and what they mean.

Switching now to yet another type of contextual dimension, consider the variability provided by analysing the *communicative situation*. Our target word ‘object’ might be said in a lecture or a public presentation; essentially a monologue directed to some audience. Think of your expectations for speakers in these situations regarding the importance of clarity in pronunciation for getting the message across. This again might include several habitual strategies including, for example, slower, louder speech implemented through a tighter and longer contact of your lips during [b] or greater jaw lowering during the initial vowel. Contrast this with a communicative situation involving a group of friends who discuss rules for a board game. This production would probably not be enunciated, and might even be muttered or mumbled, using thus less articulatory effort and the opposite tendencies as those described for enunciation. It is quite a feat to observe on your selves, and others, how flexibly and effortlessly we adjust our speaking habits to conform to the expectations for different communicative situations.

The final contextual dimension linked to the participants in communication that will be mentioned in this section includes the *sociolinguistic* aspects of the speakers’ *identities*. Still using ‘object’, British and American speakers tend to differ in the lip activity during the first vowel: The Brits round the lips more than the Americans. Apart from regional differences, other factors like socio-economic status, race, age, ethnicity and many others greatly influence our speaking habits.

► **Advanced Section: Sociolinguistic aspects of speaking**

First, consider the sound /r/. In the varieties of English spoken in Great Britain, the production of this sound is a good indicator of the **regional dialect** of the speaker. Most of you are familiar with a distinctive sound produced with the tip of the tongue by speakers of Scottish English, so-called rolled or tapped [r]. Not so much the quality of /r/ but rather its presence or absence in the ends of syllables is a major factor separating the southern

and northern varieties in England. So words like ‘car’, ‘north’ or ‘arm’ are typically pronounced with a /r/ sound in the north of England in addition to Scotland and Ireland and without it in the south.

However, this general picture has changed a lot in the last 60 years. In the 1950s the majority of the south-west, together with specific dialects of Newcastle or Merseyside area (Scouse dialect typical for Liverpool and north of it) were predominantly r-full, i.e. the /r/ in ‘car’, ‘north’ or ‘arm’ was produced. However, by 2016, the percentage of r-full speakers has dropped significantly in favour of the r-less variety.³ Of course, the regions are still distinctive due to other specific pronunciation features, but the presence or absence of /r/ has become less distinctive regarding the regional varieties in England.

So in this example, we see not only differences of regional accents but the association with another social factor, **age**, where it is typically the younger generation that is the initiator and carrier of the change while the older generation is typically more conservative and maintains older patterns.

The situation with r-full and r-less patterns in the US is a nice example of another social factor affecting pronunciation: **socio-economic status**. Of course, the same factor affects varieties in Britain as well. In the classic study of /r/ in three New York departmental stores, Labov (1966) showed that the r-full/less variation in the speech of shop assistants corresponded to the socio-economic status of the customers in the stores. Labov also hypothesized that white females might be the driving force behind the *change* of the prestige variety from r-less before World War 2 to r-full in nowadays New York area, which exemplifies race and gender as other social factors affecting our speaking habits.

Our speech is an essential part of our identity, of who we are. Studying speech thus presents an opportunity to explore how our cognitive system works, how we get things done through spoken interactions, or how we function in our immediate communities of family and peers as well as larger societies or less tight social networks. Each act of speaking is couched in multiple dimensions and layers of context, which makes context indispensable for exploring speech. Investigating systematic patterns in how context in the broad sense and our speaking behaviour interact is an exciting way to find out more about ourselves.

³This situation is very nicely illustrated in a map by Cambridge Department of Theoretical and Applied Linguistics; <https://www.businessinsider.com/maps-english-dialect-pronunciation-regional-variations-2016-6>.

2.4 Transcription of Speech

Whereas the first three sections of this framing chapter introduced and discussed general conceptual notions, this last section is devoted to a more technical aspect of capturing speech ‘on paper’ and the importance of it.

The continuity of speech tells us that it essentially never happens that even two identical sounds would be pronounced completely identically; recall the way we say /b/ in ‘beet’, ‘boot’ or ‘bat’. Now that we are aware of these differences in speaking, we will be assuming that we might disregard the differences that do not lead to any misunderstanding in terms of what sound, and, ultimately, what word is intended. By the same token, we will assume that there is a stable ‘mental’ representation of the /b/ sound. Hence, instead of drawing x-ray or MRI pictures and extracting the numbers corresponding to the precise position of the lips or the tongue at precise times, we will simplify by writing [b]. This symbol corresponds to the sound at the beginning of ‘beet’, ‘boot’ or ‘bat’ but we know that these [b] sounds, despite transcribed with the same symbol, may be (sometimes sharply) different.

However, while the symbol [b] is, coincidentally, identical to the letter of our alphabet, it will not be sufficient to use only the letters of the Roman alphabet in describing the sounds when speaking English. English is notorious for its weak correspondence between spelling and pronunciation and non-native speakers commonly struggle with this discrepancy, especially if their native languages have a stronger spelling–pronunciation link than English. But native English speakers have also struggled when they learned how to read and write, they might just not remember it so well since they were younger. The reasons for this discrepancy are complex but include the general lagging of the spelling changes in writing behind the natural development of spoken language, or the tendency to reflect the origin of words in their spelling despite the lack of this evidence in speaking. Note, however, that also in languages with more phonetic spelling like Italian, German, Finnish or Slovak, the correspondence is never one letter to one symbol.

In exemplifying the complex relationship between pronunciation and spelling in English we don’t have to go far from our, now familiar, focus on /b/. Most of you are already aware that /b/ is sometimes pronounced, as in ‘beet’, ‘obey’ or ‘tab’, and sometimes it is silent as in ‘debt’ or ‘plumber’. An even better example might be the sound [f]. There are at least five different ways this single sound could be spelled in English. Consider the words ‘food’, ‘affair’, ‘tough’, ‘physics’ and ‘giraffe’ in which the [f] sound corresponds to the letter(s) <f>, <ff>, <(u)gh>, <ph> and <ffe>, respectively.

Looking at the opposite direction, there are multiple ways in which a single letter (or the same sequence of letters) can be pronounced. For example, consider the pronunciation of <ough> in English. In past tense of many verbs like ‘bought’ or ‘thought’ it is pronounced with the same vowel as in ‘saw’. In words like ‘rough’ it

is pronounced as the vowel of ‘cut’. In words like ‘through’ it has the same vowels as in ‘cue’. In yet another group of words like ‘though’ or ‘dough’ it is pronounced the same as in ‘coat’. Finally, and this is admittedly somewhat archaic, <ough> may also be pronounced with the same vowel as ‘cow’ in words like ‘bough’ meaning a large branch of a tree.

Find Out More 2-4

A funny story refers to the times when a spelling reform was discussed in England and somebody suggested mockingly that English ‘fish’ could very well be spelled as ‘ghoti’ since ‘gh’ commonly refers to /f/ in words like ‘rough’ or ‘tough’, ‘o’ might be pronounced as [ɪ] in ‘women’, and ‘ti’ is commonly pronounced as [ʃ] (a sound starting ‘shelf’ or ‘chauffer’) in words like ‘nation’. This suggestion is commonly attributed to G. B. Shaw but as B. Zimmer found out, it probably originated earlier and may be attributed to Charles Ollier in 1855; <http://languagelog.ldc.upenn.edu/nll/?p=81>.

The topic of ‘absurdities’ of English spelling is quite alive on the internet and there exist many poems and examples showing the discrepancies between spelling and pronunciation. One of the most well-known is the poem Chaos attributed to Gerard Nolst Trenité and I think it is both instructive and fun to read, or better listen, to the poem at least once. I include here 2 links to the poems performed by a British and an American speaker with added subtitles, but typing ‘youtube chaos poem’ into a search engine gives multiple hits, abridged versions and performances.

- <https://www.youtube.com/watch?v=1edPxKqiptw> (American)
- https://www.youtube.com/watch?v=r_FBhIcg95M&t=19s (British)

Hence, even within one language and one orthographic system, there might be radical differences in spelling a single sound and thus a separate set of symbols referring to specific sounds is needed. Moreover, languages do not have the same orthographic systems, so we need a standard for transmitting pronunciation that would be accessible to speakers of any language. For example, languages might differ widely in how they pronounce certain (groups of) letters and <ch> might correspond to the different sounds of ‘church’ or ‘chemistry’ in English but also to yet another sounds in German ‘Bach’ ‘stream’ or French ‘chien’ ‘dog’.

It is also clear that even if we use all letter-like symbols of the traditional ASCII keyboard, the number would be neither sufficient nor intuitive. Consider, for example, the ‘click’ languages of Africa like Ju’hoansi described by A. Miller. In this language there are, in addition to over 20 ‘typical’ consonants also, at least 12 contrastive clicks for which we do not have intuitive letter-like symbols, but these clicks function just like English /b/ or /v/ in producing contrasts in words.

Find Out More 2-5

You may listen to an excerpt from this language in ‘juhoansi_example.wav’ in the book companion and appreciate the difficulty it would take to transcribe this language using only the ASCII letters. The internet also provides good examples including two links below.

- <https://www.britannica.com/topic/Khoisan-languages/Classification-of-the-Khoisan-languages> (30 s sound clips)
- <https://www.youtube.com/watch?v=W6WO5XabD-s> (1-minute video)

The discussion above shows that it is not a good idea to use English spelling or alphabet for describing the pronunciation of English or other languages. There is already a well-established system called International Phonetic Alphabet (IPA) with symbols and diacritics for transcribing every meaningful contrast in the languages of the world. The traditional chart that can be found in almost all books on phonetics/phonology is shown in the prelims of this book.⁴ A more interactive chart with clickable symbols that you can listen to can be found here: <http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/>.

Although some dictionaries used to use alternative ways of showing pronunciation, currently IPA is the standard way for transcribing words in most common dictionaries and relevant online resources such as phone applications or pages helping non-native speakers who want to master a new language.

In this book, we will use only a subset of the symbols and diacritics that are relevant for English. In the dedicated chapters on English consonants and vowels we will review the taxonomy of these sounds and their IPA symbols. While IPA takes us a long way when discussing sounds and words separately, there is no yet generally agreed-upon system for transcribing sentence prosody. When the need arises in Chapter 13, we will add Tones and Break Indices (ToBI) as one possible framework for capturing linguistically meaningful discrete aspects of naturally continuous prosodic variation in speech.

Transcribing speech is a necessity in any discussion of the phonetic and cognitive aspects of our speaking habits. Additionally, fluently producing transcriptions of speech as well as fluently reading them goes a long way towards building awareness regarding speaking habits. This is of course due to the fact that transcriptions correspond to actual speaking more than writing. Before reading this, you might have not thought that native English speakers use roughly the same vowel in the beginnings of words as different-looking as ‘thumb’, ‘rough’ or

⁴IPA Chart, <http://www.internationalphoneticassociation.org/content/ipa-chart>, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2015 International Phonetic Association.

‘someone’. But if you are forced to transcribe these words or check their pronunciation with a dictionary, it will help build your awareness of this regularity.

In sum, transcription is a tool that helps us describe speaking behaviour and build awareness regarding the habits involved in speaking. Its discrete nature as a set of finite symbols allows us to abstract away from continuous physical events to discrete representations characterizing meaningful contrasts and variability.

Exercises

- 2-1 Compare the number of characters with the number of sounds in some selected lines of the poem Chaos linked in Find Out More box 2-4. Additionally, consider if the mismatches between sound and spelling are truly ‘chaotic’ or if some generalizations can be made. We will return to them in Chapters 5 and 6.
- 2-2 The text mentioned that several social variables, e.g. regional affiliation, age or socio-economic status, might be reflected in speech. What other similar variables might participate in forming our identity and how they affect our speaking behaviour?

References

- Einstein, Albert, and Leopold Infeld. 1938. *The evolution of physics*. Cambridge: Cambridge University Press.
- Gravano, Agustín, and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25: 601–634.
- Labov, William. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.
- Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.



Articulatory Mechanisms in Speech Production

3

In this chapter we will

- build better understanding of our bodies and become more familiar with the organs of speech and some of their functions
- unpack the three main processes related to speaking: breathing, voicing and articulation

3.1 Introduction

There are two fundamental ways of exploring the continuous phonetic aspects of speech: we investigate the physical movements and actions of our bodies involved in producing speech, and analyse the acoustic characteristics of speech arising from these movements and actions. The first area thus involves understanding how our bodies create speech and this chapter provides a foundation for that understanding. The second area will be the topic of the subsequent chapter. These two chapters thus build a basis that we will rely on when discussing many speaking habits in subsequent chapters. We begin with the three most relevant subroutines included in the act of speaking: *breathing*, *voicing* and *articulation*.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_3) contains supplementary material, which is available to authorized users.

3.2 Breathing and Airstream Mechanisms

Breathing is a fundamental biological process. Although we naturally breathe without speaking, we cannot speak without the air. Since speech requires air and breathing, let's start with reviewing how the air gets into the lungs.

Activity 3-1: Breathing in

Take a deep breath either through your nose or the mouth. Observe the associated movements in your body. Now imagine the air molecules in front of your face and how they rush inside your body. What makes them rush in? If you know or remember from your previous schooling, try to explain to a person who doesn't know. If you do not know or don't remember, try to brainstorm the plausible account of how these molecules get in. Remember, they cannot do it on their own!

In order to make the air molecules in front of our nose and mouth rush inside our lungs, we have to create *low pressure* in the lungs and the rest is pretty much taken care of. Think of a simple air pump for inflating bicycle tubes or balls. To get the air into the pump we elevate the valve creating a vacuum-like low pressure inside the pump since we have increased the volume inside the pump. Hence, there is a reverse relationship between the pressure inside a container and the container's volume: the larger the volume the lower the pressure. Figure 3.1 illustrates this with two closed containers each containing 10 air molecules. If the container is small, as in the left of the figure, the molecules are quite squished, pushing on each other, which results in higher pressure. However, if the container is larger, the molecules have more room and thus the pressure is lower.

Returning to the air pump, the fundamental characteristic of pressure is that it wants to balance, and thus available air molecules from the area of higher pressure (around the pump) will rush inside to balance the pressure inside the pump when we elevate the valve.

The very same mechanism applies when we breathe. We use the muscles affecting the size of our chest, namely the *diaphragm* below the ribcage and *intercostal*

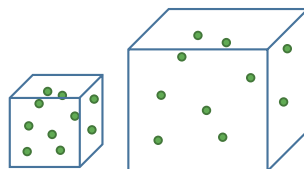


Fig. 3.1 Illustration of the inverse relationship between volume and pressure, 10 molecules of air in a smaller (left) and larger (right) container resulting in greater (left) and smaller (right) pressure, respectively

muscles in between the ribs, to increase the volume of our chest. As a result of the increased volume, the pressure inside the lungs becomes lower than the pressure of air in front of your face. Hence, the air molecules will rush inside through the airways to balance the pressure and increase it inside the lungs. This process of breathing in is also called *inspiration*. Hence, the activity of the diaphragm and intercostal muscles is essential for breathing in. There are other actions, for example, the expansion of the nostrils, which many people do automatically when taking a deep breath, but these are not essential.

Activity 3-2: Feel your muscles involved in breathing

Take a breath and then blow all the air out. Then, without taking in any more air, try to read some text loudly. Continue reading without taking any more air in. After a couple of seconds you should start feeling the muscles, even pain inside your belly and around the chest. What you are feeling are the diaphragm and intercostals. While reading, these muscles push the air out of the lungs but the same muscles are used to enlarge the chest when you breathe in.

Actually, in normal situations, we do not actively push the air out of the lungs when we speak (as we did in Activity 3-2). The lungs contain elastic tissue that is stretched during inspiration. For speaking with an *egressive* mechanism (= air goes out), we use this elasticity to release the air in a controlled way. Of course, if you are out of breath or want to finish saying something important in one chunk without an interruption for breathing, you can use the muscles to push air out. To illustrate this, imagine opening an inflated air balloon. The elasticity of the balloon walls is responsible for the air rushing out although you can add an external squeeze with your hands if you want. However, note that this elasticity in the balloon is much greater than in our bodies and thus the air rushes out of the balloon much faster than it happens in our bodies.

Activity 3-3: airstream mechanism in tut-tuting

Produce the click sound typically expressing (ironic, mocking) disapproval commonly transcribed as *tut-tut* or *tsk-tsk* and accompanied with head shaking. In which situations do you produce these sounds? I, for example, typically produce a single click when I make some error, for example, in typing on a keyboard, or mis-hitting a stroke in tennis. But reduplicating the sound two-three times has a clear sarcastic or mocking function for me. How is it for you?

After considering the function of this sound, think of its phonetic form. Say it again, this time very slowly, and introspect how the airstream mechanism for this sound works. Describe in as much detail as possible, what happens in your mouth in order to produce this sound and then compare with the description below.

The airstream mechanism for this sound relies on creating the body of trapped air between the tongue and the palate and by lowering the tongue, this trapped area increases, causing the low pressure, just as in the discussion of Fig. 3.1. When we release this constriction, the air rushes into balance the pressure and this is realized with an audible burst called the click. Hence, we did not need the air escaping from the lungs, we used the air we already had in our mouth and trapped it with the tongue. The clicks are thus not pulmonic (i.e. using the lungs) but *velaric* sounds (because the back of the tongue seals off the constriction against the velum; Sect. 3.5) and the IPA symbol for the click from Activity 3-3 above is [ɰ]. In addition to velaric sounds, *glottalic* airstream mechanism used in many non-Indo-European languages involves the movement of the larynx (next section), but it will not be discussed further.

3.3 Voicing

Once we have air in the lungs and release it in a controlled way, it travels through the *bronchi*, to the *trachea* (windpipe) and reaches the *larynx*. This is the hard bone-like structure also called Adam's apple or the voice box that you can feel by pressing the fingers against the middle of the neck below the jaw and swallow. The organ that moves up and down when swallowing is your larynx. It does not contain bones but *cartilages*, a pair of tissue stripes attached to cartilages called *vocal folds* (or vocal cords), and muscles that move the cartilages. The fine interplay of these organs results in *voicing*, sometimes also called *phonation*.

Let's start with feeling the voicing. Put your fingers on the larynx and make a prolonged [zzzz] sound. You can imagine imitating a large swarm of bees or flies and the loud sound they make. You should clearly feel the *vibrations* inside the larynx. To make the sensation more robust, contrast the [zzzz] sound with the [ssss] sound imagining a big barrel of snakes, or a single gigantic snake making the loud noise. Alternating [zzzz] with [ssss] you should feel the vibrations with the former and no vibration with the latter despite the fact that the two sounds are roughly equally loud.

Note how you switch the vibrations on and off for [zzzz] and [ssss] habitually and without conscious awareness of the processes involved in doing so. Our goal in the rest of the section is to become aware and understand how we produce this effortless switch.

Find Out More 3-1

The following links show movies of vocal fold vibration; they are great for building a tangible awareness of this phenomenon. The first one shows a real image in slow motion without sound, the second is a very instructive animation of the entire phonation process and the third one is from an online e-learning course.

- https://www.youtube.com/watch?v=Drns_eV9wWg
- <https://www.youtube.com/watch?v=kfkFTw3sBXQ>
- <https://www.youtube.com/watch?v=f62dq-L36o>

Vibration is essentially a regular repetition of a certain activity. With string musical instruments, the movement of a string is initiated by plucking (guitar), bowing (violin) or hitting (piano). With the human voice, the vocal folds that repeatedly open and close produce the vibration you feel during [zzzz]. But how do we open and close the vocal folds as seen in the videos linked in the Find Out More box above?

Activity 3-4: How fast can you contract your muscles?

Let's select a repeated movement that you are comfortable with and can do very fast. For example, open and close the mouth (e.g. silently mouthing 'babababa' as fast as you can), or simply tap your finger on the table as if you were nervous. How many times a second you think you might reach if you do this as fast as you can? A good way to check is to pair with a companion and ask him/her to time 10 seconds and count the repetitions while you move as fast as you can. You then switch turns. You can also do it on your own but it is a bit more challenging to keep the time and the count at the same time. Finally, divide the resulting total of repetitions managed in 10 seconds by 10 to get your number of repetitions in one second.

You should realize two things. First, you are probably tired from just 10 seconds of repeated movement. Try taking a deep breath and saying [zzzz] for 10 seconds. Are you tired? Second, the number of cycles per second that you have managed is probably less than 10.

After Activity 3-4 above, the first key feature of voicing should be clear: vibrations of the vocal cords heard and felt during [zzzz] cannot be solely due to voluntary muscle activity. First, your count of cycles likely resulted in some single-digit or low number. However, the vocal cords open and close commonly 100–400 times a second in normal speech. Second, repeated muscle contraction brings fatigue and although you were probably tired after finger-tapping or mouthing 'babababa' for 10 seconds, you were not tired after a [zzzz] of the same duration.

Although we do not repeatedly and voluntarily contract our muscles to vibrate the vocal cords, we certainly can voluntarily start and end sustained voicing (think of switching between [ssss] and [zzzz]). Hence, we do need muscles. Figure 3.2 illustrates the placement of the core structures needed for understanding voicing.

The large *thyroid cartilage* is the one you feel as the V-notch of Adam's apple. Both tissues of the vocal cords attach to this cartilage together in a more or less fixed manner. On the other side, the ends of the vocal cords are attached to the triangular-shaped *arytenoid cartilages* that in turn form a joint with the *cricoid cartilage* and can thus pivot around. The voluntary contraction of the muscles

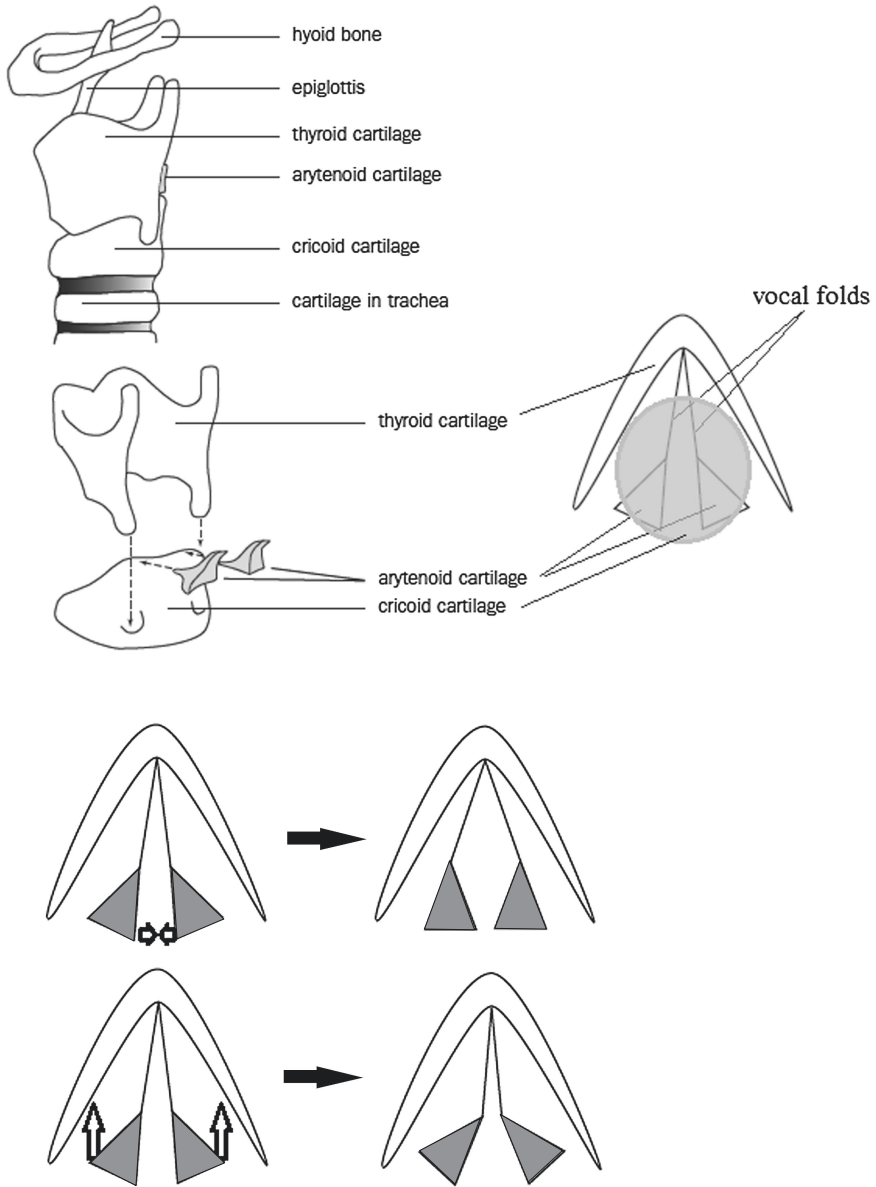


Fig. 3.2 Schematic illustration of the structures in the larynx (top) and the activity of the muscles attached to the arytenoid cartilages, shown as arrows, for opening and closing the vocal folds (bottom)

attached to these cartilages make the vocal cords opened (**abduction**) as in breathing, or closed and ready for voicing (**adduction**).

The closed vocal cords create a complete obstruction to the airflow being pushed from the lungs. The air accumulates below the vocal cords increasing thus the air pressure. Since the vocal folds are not rigid but elastic, at some point the pressure below them is greater than the forces that hold them together and the elasticity of the vocal folds enables their opening. At this moment, air molecules below the vocal cords start rushing through the opening to balance the pressure. Since the muscle force and the elasticity of the vocal cords still apply, the vocal cords close again.

Voicing is thus another automatic speaking habit involving intricate changes in the tension of the vocal cords to initiate or stop their vibrations and a tight coordination of all this with the respiratory system. We characterize voicing as a **dynamic** process of vocal cord opening and closing that requires **air** (pressure), the activity of the **muscles** for vocal folds adduction, and their **elasticity** for returning to the closed state after the forced opening. Hence, voicing is an **aerodynamic myo-elastic process**.

Activity 3-5: Describing voicing

The last two sentences preceding this activity are rather heavy in terminology. Recall that our main goal is your understanding of the concepts underlying our speaking habits. I strongly recommend to review the text above, and then watch the entire process in a very nice visualization in the YouTube demonstration from the second link of the Find Out More box above and linked again here: <https://www.youtube.com/watch?v=kfkFTw3sBXQ>.

Once you feel you are comfortable with the material and you feel you understand, I urge you to try to explain how your vocal folds vibrate to a non-expert friend or a family member (of course, you cannot just tell them that voicing is an aerodynamic myo-elastic process).

After fully grasping the process of voicing, there is one more feature of the emitted sound that plays an important role in speech: **intensity** or **loudness**. Try saying a prolonged [zzzz] or a vowel [aaaa] and vary its loudness, as if the sound source moved (like a swarm of bees) alternating its movement towards and away from you. Speech loudness is fully under your control. The most available sensation and the mechanism mostly responsible for variation in loudness is the amount of pressure below the vocal folds. Greater pressure means that more air at greater speed is emitted during each vocal folds opening, which translates into greater loudness of the resulting sound. In addition to the primary function of the diaphragm and intercostal muscles in increasing the loudness of emitted sounds, fine adjustments to vocal cord tension with the laryngeal muscles also influence the loudness of the sound.

3.4 Other Activity Inside the Larynx

The previous section described the processes involved in the production of *modal voicing* when the vocal folds open and close regularly along their length, which is common in the production of vowels like ‘ah’. The contrast between voiced and voiceless sounds, in English, for example, between [s] and [z], is one of the major contrastive features of sounds that we will cover in more detail in Chapter 6. Before discussing the movements of other articulators above the larynx in Sect. 3.5, let’s briefly consider variability in voice quality afforded by the differences in the settings of vocal cords inside the larynx and their functions in speech.

The *glottis* is the space between the vocal cords. When we breathe, the glottis is open. We close the glottis, for example, when you try to cough on purpose. Or try imitating the sound that people sometimes make when they do something bad or by accident: ‘uh-oh’; the first syllable with higher pitch than the second syllable. Both syllables start with a closed glottis; a sound called a *glottal stop*, [ʔ] in IPA. People familiar with the Cockney accent may recognize the glottal stops instead of the [t] sounds in words like ‘bottle’ or ‘cotton’ (Fig. 3.3a).

In between these two ends of the continuum (wide open and tightly closed), the glottis may assume several settings depending on the activity of the muscles controlling the vocal cords. *Whispering* is a common mode of speaking if we don’t want to wake a sleeping baby or to signal para-linguistically that what we are saying should be treated confidentially. This mode is achieved with an incomplete closure of the vocal cords at the arytenoid end of the glottis. The passing of air through this narrow opening produces friction-like noise characteristic of whispering. You may try to blow some air as if you were trying to blow away a fly or a speck. You can clearly hear that this creates some noise. The source of that noise is the narrow opening of the mouth between the lips. This is similar in the larynx when the vocal cords are very close and the air passing the narrowing creates whisper. This setting is illustrated in Fig. 3.3c.

Breathy voice is achieved when the vocal cords are not adducted (closed) with sufficient muscle tension. As a result, less pressure is needed to open the vocal cords and the opening phase of the cycle is long enough, or not complete enough, to allow for air leaking through the glottis to add a whisper-like quality. Alternatively, a portion of the vocal cords might be closed and vibrating while another

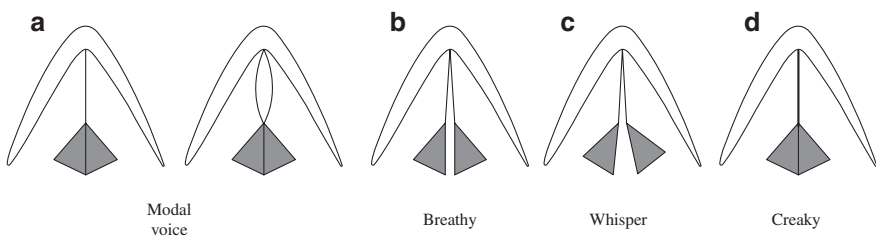


Fig. 3.3 Stylized pictures for settings allowing modal voicing, breathy voice, whisper and creaky voice

portion has a narrow opening, or vocal folds might also be flapping during incomplete closure. All these settings result in breathy voice that sounds like a combination of voicing and whispering. In English, a /h/ sound in between vowels, as in ‘behind’ or ‘ahead’ has this quality. Sustained breathy voice can be found in acting and singing of Marilyn Monroe. Through her sex symbol image and the cultural iconicity, breathy voice became linked to sexual attractiveness in females, but it is also linked to youthfulness in general since breathy voice is associated with younger and thinner vocal cords. However, in sustained breathy speaking much more air is required than for speaking in modal voice. This setting is illustrated in Fig. 3.3b.

Another common setting of the glottis is referred to as *creaky voice*. This is the situation when the vocal folds are made very thick by the muscles and open only partially. If you tried imitating the opening of the old squeaky door in a horror movie, you would feel slight pressure inside the larynx and almost hear individual cycles because the frequency of vibrations is very low (below roughly 50 open–close cycles per second for males and 100 for females). Due to its low frequency of vocal cord vibration, creaky voice is typical in the final syllables of statements ending with and intonation fall expressing finality.

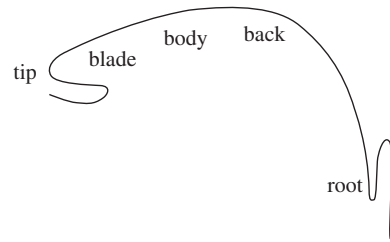
Just like breathy voice, creaky voice is also linked to several sociocultural and communicative aspects. For example, another cultural icon Kim Kardashian is famous for her *vocal fry*, which is another term used for creaky voice. The perception of creakiness differ, some females might do it to come across as more confident and authoritative since creak and low pitch are typically connected to masculinity and high social status (e.g. Yuasa 2010). On the other hand, several studies suggest that people perceive creak negatively especially with young females who come across as less competent, less educated, less trustworthy, less attractive and less hireable (Anderson et al. 2014). Creaky voice is also one of the cues to signal that a speaker finished her turn and another interlocutor may assume speaking in English (Laver 1980; Wolk et al. 2011) and many other languages. The setting for creaky voice is illustrated in Fig. 3.3d.

In short, the setting of the larynx and the glottis make a powerful device for displaying our identity, emotions, communicative intentions and many more. And we are not even mentioning here the melody of speech and intonation, also primarily linked to the setting of the vocal cords, and discussed in detail in Chapter 12.

Activity 3-6: Compare your modal, whisper, breathy and creaky voice

Start saying a regular long [aaaaa]. Now try to alternate between this modal voice and whisper. Next, try to lower your pitch continuously until you hear a ‘rattle’ or ‘throttle’ coming out of your larynx similar to squeaking of an old door opening. You should feel some tension in your larynx that is greater than with the modal or whisper voices. Finally, try to combine the modal and whisper voices together into the breathy quality of your [a]. What emotions, or situations, do you associate with these voice qualities in your experience in addition, or different from, the ones mentioned in the text?

Fig. 3.4 Schematic illustration of the major parts of the tongue for speech description



3.5 Articulation and the Activity Above the Larynx

After breathing and voicing, the third essential component of speaking is *articulation*. It involves the activity of the organs above the larynx that we can see easily in the mirror (the lips, the tongue) and lend themselves more naturally for introspection.

We focus our discussion around the *active articulators*, that is, the organs that we voluntarily move in order to produce speech sounds. Along the way, however, we will mention other *passive articulators* participating in producing speech sounds, that is, serving as the spatial targets for the movements of the active articulators.

The most versatile and agile organ of speech is the *tongue*. The tongue can form various shapes or touch various other parts inside the mouth. It is a collection of various muscles of the tongue itself (*intrinsic muscles*) that are attached to other parts of the body (such as the chin or the hyoid bone) with another set of *extrinsic muscles*. All these muscles are involved in intricate coordinations for achieving the desired shapes or actions required of the tongue when speaking. For describing speech, it is not crucial to understand the activity of the individual muscles.¹

Although the tongue is a single organ, it is useful to divide it into regions that participate in producing various speech sounds. The five major regions are: the tip, the blade, the body², the back and the root, which is shown in Fig. 3.4.

When describing the tongue's regions, it is useful to imagine the rest position of the tongue and the structures above it. The *tip of the tongue* is the very small area of the very tip. In the rest position it is typically below the edge of the upper (incisor) *teeth*, and possibly touching lightly the lower incisors. The *blade of the tongue* is a small area behind the tip and in the rest position it is normally positioned below the upper teeth and the *alveolar ridge*. Alveolar ridge is the region of the gum extending from the upper teeth to the somewhat sharp edge as seen in Fig. 3.5. The middle portion called the *body of the tongue* lies opposite to the *hard palate*, which is the hard arch of the roof of the mouth. The *back of the tongue* is the section of the tongue's surface resting opposite to the *soft palate*. Finally, the *root of the tongue* is positioned against the *pharyngeal wall*.

¹But interested readers can find many great illustrations of the anatomy of the tongue in the internet, for example here <https://www.yorku.ca/earmstro/journey/tongue.html>.

²Many authors refer to this part of the tongue as the 'front'. However, this is not very intuitive and thus the term 'body' will be used in this book.

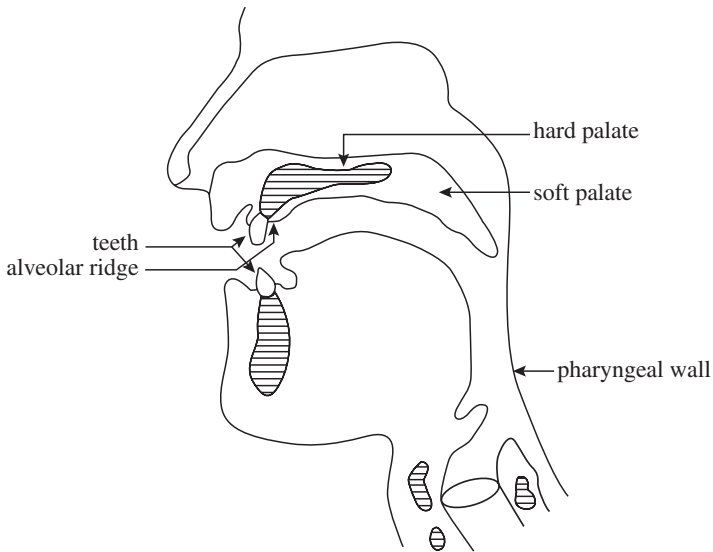


Fig. 3.5 Schematic illustration of the articulators above the larynx

The second major moving articulator are the *lips*. The most common task for lips during speaking is rounding, commonly accompanied with protrusion (i.e. forward movements of the lips), and vertical movements participating in closing the lips or pressing the lower lip against the upper incisor teeth. The lips are also formed by muscles, mainly the orbicularis oris muscle.

The *jaw*, or the *mandible*, is the massive bone that participates primarily in the opening and closing of the mouth. The jaw plays a crucial supportive role in the movements of both the lips and the tongue. With the lips, this is obvious since the jaw is physically connected with the bottom lip and thus opening and closing of the mouth involves primarily the jaw movement. The physical connection between the jaw and the tongue is less rigid but clearly present since some extrinsic tongue muscles attach to the jaw. Despite these physical linkages, the lips and the tongue can be moved independently of the jaw, as explored in Activity 3-7 below.

Activity 3-7: Separating the lips and the tongue from the jaw

Try to put a wooden pencil in between your upper and lower incisor teeth. Look at the mirror and open and close the lips. Since the teeth clench the pencil, the movements of the lips are independent of the jaw. If you like a challenge and want to practice face gymnastics, you might try separating the movements of the two lips. Try pressing the bottom lip against the pencil and move the top lip up and down. Now switch and press the upper lip against the pencil and move the bottom lip up and down. You notice that opening/closing the mouth and

simultaneously clenching the pencil between the teeth is rather natural if both lips coordinate. However, separating the lower lip both from the jaw and the upper lip requires more effort. This suggests that the independence of the lips from each other and the lower lip from the jaw are not habits naturally present in speaking.

Now without a pencil, try to alternate saying ‘bean’ and ‘barn’ and notice the movements of the jaw. You can either observe yourself in the mirror, or touch the bottom part of your chin with your fingers while alternating ‘bean’ and ‘barn’. You see or feel the vertical movement of the jaw that supports the vertical movement of the tongue required for producing the vowels in these two words; [i] and [ɑ], respectively, in IPA. While you see the jaw movement, you can only feel the tongue movement. Now bite the pencil as before and alternate saying ‘bean’ and ‘barn’ again. You should feel the vertical movement of the tongue more clearly. Also notice that now you are producing this vertical tongue movement without the supporting movement of the jaw.

The last active articulator participating in producing English speech sounds is the *soft palate* or *velum* already mentioned above when discussing Fig. 3.4 and the parts of the tongue. This is a somewhat elusive articulator for two reasons. First, its most obvious function is to be a passive articulator in sounds like [k] or [g] in which the tongue back presses against the soft palate. This is also a very good way of feeling the soft palate. However, the soft palate can also move, somewhat vertically lowering and raising, and this ability is crucial for the production of nasal sounds like [m] or [n]. Activity 3-8 helps increase your awareness of this function of the velum.

Activity 3-8: Feel your soft palate (velum)

First start with your tip of the tongue and explore the upper area of the mouth. Feel the edge of the upper incisors, continue to the alveolar ridge through the arch of the hard palate and then further back until you feel the palate is no longer hard and softens up.

The next activity gives the sensation of how the soft palate moves. Prepare as if you were going to say [p] (taking a deep breath, close the lips and feel the pressure behind the lips) but instead of opening the lips, say [m] keeping the lips closed. Hence, the air leaves through the nose. Once you get used to this strange sensation, try alternating [pm pm pm]. What you feel in the back of your mouth is the soft palate functioning like a valve directing the air to the oral cavity when raised for [p] or to the nasal cavity when lowered for [m].

When you have a good grasp of the soft palate and its movement, it is good to face a mirror and open your mouth wide with your tongue low. You will see the soft palate and the *uvula* ‘hanging from’ the soft palate. To summarize the articulators above the larynx, we have the active articulators (the tongue, the lips, the jaw, the soft palate), the passive articulators (the teeth, the alveolar ridge, the hard and soft palate, the uvula, the pharynx). There are also the cavities participating in the acoustic

filtering of sound discussed more in the subsequent chapters (the oral, nasal and pharyngeal cavities). These articulators are schematically depicted in Fig. 3.6.

In describing sounds, we will also need adjectives for various places at which the sounds are produced. The adjectives usually come from the Latin names of the organs. Now that we are familiar with the vocal tract and its major articulators, Table 3.1 and Fig. 3.7 review those adjectives since we will need them in subsequent chapters.

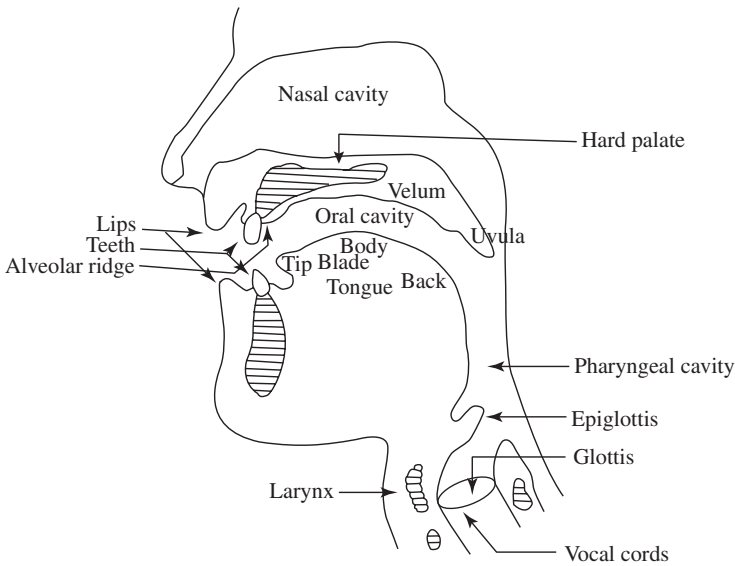


Fig. 3.6 Active and passive articulators

Table 3.1 The articulators, associated adjectives and the descriptions of their activity

Articulator		Adjective	Description
English	Latin		
Lips	Labia	Labial	Involving one or both lips, rounding
Teeth	Dentes	Dental	Involving the teeth, e.g. touching w/the tongue or the lips
Tongue	Lingua	Lingual	Involving the tongue for constriction inside the mouth
Tip	Apex	Apical	Activity of the tip of the tongue; touching, curling
Blade	Lamina	Laminal	Constriction between the blade and the dental-alveolar-palatal area
Back	Dorsum	Dorsal	Constriction between the back of the tongue and the soft palate
Root	Radix	Radical	Constriction between the tongue root and the pharyngeal wall
Vocal cords	Glottis	Glottal	Constriction between the two vocal cords

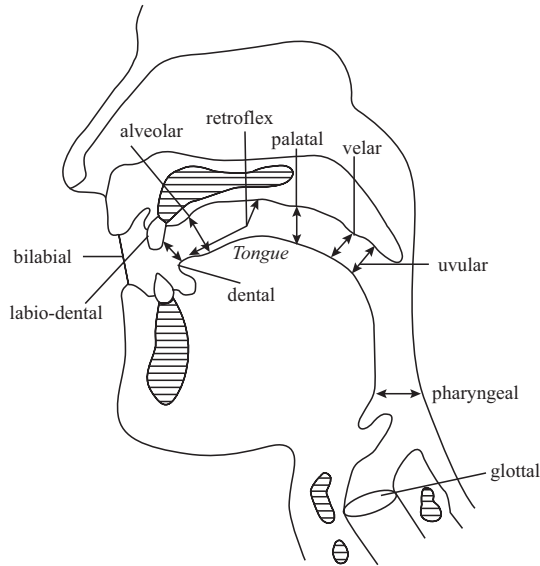


Fig. 3.7 English adjectives describing the major places of articulation

► Advanced Section: Evolution of speech organs

Consider the functionality of the articulators and organs participating in speech production. The first observation relevant for us is that none of the organs we have discussed above is used uniquely for speech. For example, the tongue functions in food digestion or tasting, the jaw serves primarily for chewing food, or **epiglottis**, as seen in Fig. 3.6, prevents food from entering the airstream pathways and the larynx. It is reasonable to assume that these organs evolved primarily for these functions while speaking was a relatively late invention of homo sapiens.

This conclusion is corroborated by comparing the vocal tracts of humans and higher primates. As shown in Fig. 3.8, all essential components required for speaking are present in the vocal tracts of both humans and chimpanzees. Nevertheless, the production of speech sounds is common and effortless for humans whereas non-existent, or highly laborious, for higher primates. It is true that many primates are able to communicate (with humans or among each other) through a semi-complex system of signs that resembles in some aspects human language and speech characteristics. However, all these communication systems are primarily based on gestures and signs rather than on speech sounds.

The second relevant observation from comparing humans and higher primates regarding speech is that despite similarities in the inventory of speech organs between the two, there are essential differences. Consider Fig. 3.8 and engage with the linked Activity 3-9 before continuing to read further.

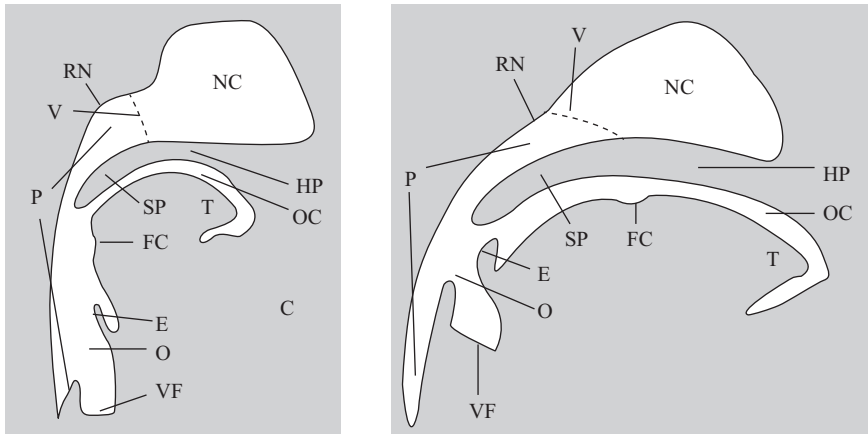


Fig. 3.8 Schematic representations of the vocal tract for humans (left) and chimpanzees (right); the lips and nose are on the right. The relevant articulators include: NC (nasal cavity), RN (roof of the nasopharynx), P (pharynx), VF (vocal folds), O (opening of larynx into pharynx), E (epiglottis), SP (soft palate), HP (hard palate), OC (oral cavity), T (tongue). Reproduced with permission, Lieberman (1975, Fig. 9-4, p. 108)

Activity 3-9: Spot the difference!

Consider the two representations of vocal tracts in Fig. 3.8. Try to identify the major differences between the two vocal tracts. How many were you able to list?

Now that you have the differences, try to identify the biological changes that most likely lead to the evolution of the human vocal tract from that of the chimpanzee.

Continue reading the text to compare your observations with mine.

Probably the most apparent difference between the two pictures is the size and the shape of the tongue. The chimp's tongue appears larger, spreads horizontally to extend beyond the nasal cavity on the right and as a result it is flatter than the human's tongue. You might also notice that the human larynx with the vocal folds is significantly lowered; for example, by comparing the distance between vocal folds and the soft palate in the two pictures. Consequently, a human's vocal tract has an upside-down L-shape while the chimp's vocal tract is much flatter. Another visible consequence of the lowered larynx in humans is the relationship between the epiglottis and the soft palate. While in humans they are far apart and cannot touch, in the chimpanzee's vocal tract they are close to each other and can easily touch.

The lowered larynx of humans is a surprising evolutionary change for at least two reasons. First, the airstream pathways of humans are less direct, resulting in lower intake of air compared to chimpanzees. This may create a disadvantage in terms of lower amount of oxygen in the bloodstream and consequently lower power output of the muscle activity for humans compared to chimpanzees. Informally, the vocal tract of the chimps facilitates faster running, either to catch the prey or outrun danger or a predator. Second, the human vocal tract has an extended pathway used for the transport of both air and food in the pharynx. The human

epiglottis prevents food from entering the larynx and the air passages but its function is not as good as in the chimpanzees in which the epiglottis and the soft palate may separate food from air completely. As a result, humans might choke when food enters the airstream pathway, which in extreme cases may lead to death. Hence, evolutionary changes increasing the risk of death, either by choking or by being eaten by a predator, are quite surprising.

To summarize our observations so far, the human larynx evolved to be located in a much lower position than that of the chimpanzee's larynx, which creates evolutionary disadvantages for survival. Additionally, all organs needed for speaking are found in both species, but only humans developed the complex communicative system of speaking. A plausible hypothesis offered by Ph. Lieberman that explains this apparent paradox is that our **speech apparatus evolved adaptively favouring the communicative function** over the more basic ones linked to survival (Lieberman et al. 1972, Lieberman 1975). It is important to note, however, that the changes in the physiology of the vocal tract were most likely not sufficient for speech evolution in humans and changes in brain were also needed. See the articles by Fitch et al. (2016) and Lieberman (2016) in *Science Advances* for more insights.

The last observation relevant to our discussion concerns the task in Activity 3-10 that you should do before continuing reading the text.

Activity 3-10: Match the picture

Figure 3.9 shows a vocal tract of an unknown origin. Your task is to compare it with the vocal tracts of human and chimpanzee in Fig. 3.8 and decide which of these two pictures is more similar to the current picture. Once you have the answer, try to verbally support your choice by describing the similarities and differences between the current picture and the two pictures in Fig. 3.8.

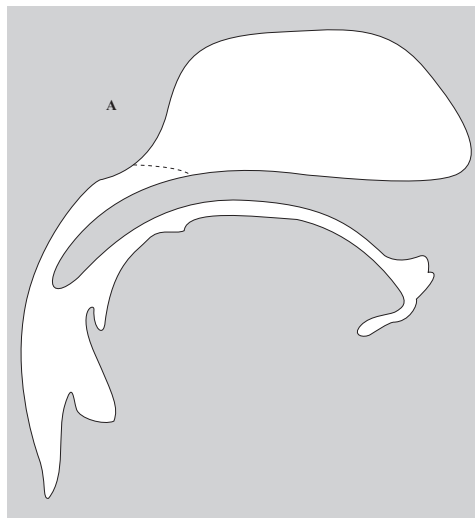


Fig. 3.9 Schematic representations of a vocal tract. See text of the activity. Reproduced with permission, Lieberman (1975, Fig. 9-4, p. 108)

So, did the picture above resemble more the human or the chimpanzee's vocal tract? The majority of you probably opted for the chimpanzee. The position of the epiglottis with respect to the soft palate and the general flat nature of the entire vocal tract has more in common with the chimpanzee than with the human. Nevertheless, the picture in Activity 3-10 above depicts the vocal tract of a new-born human baby!

Hence, the larynx descent can be observed both in the long-term evolution of tens of thousands of years from higher primates to homo sapiens as well as in the short-term development from the birth to adolescence in humans. These observations all support the hypothesis that the anatomy of the human vocal tract has evolved, at least partly, to facilitate the contrastiveness of speech sounds increasing thus the effectiveness of spoken communication. Of course, this proposal by Lieberman is not uncontroversial. Probably the most relevant issue is the timescale of the developments. Some suggest that reconfiguring the anatomy of the vocal tract must have been a long process starting earlier than the development of speech, which was a more recent evolutionary change. In any case, the link between the anatomical evolution of speech organs and the social advantage posed by speech hypothesized by Ph. Lieberman remains an intriguing one.

Exercises

- 3-1 Think of the active and passive articulators and how they make constrictions for major sounds, like both lips together for [p] or [b], or the tongue back against the velum for [k] or [g]. Explore various constrictions and group them into (1) those you are sure, or reasonably sure, occur in human languages, (2) those that are possible to make but are quite unlikely people are using for making meaning contrasts, e.g. touching the blade of the tongue against the upper lip and (3) those that are physically impossible, for example, a consonant produced by touching the uvula with the tongue tip.

References

- Anderson, Rindy C., Casey A. Klofstad, William J. Mayew, and Mohan Venkatachalam. 2014. Vocal fry may undermine the success of young women in the labor market. *PLoS ONE* 9 (5): e97506. <https://doi.org/10.1371/journal.pone.0097506>.
- Fitch, Tecumseh W., Bart de Boer, Neil Mathur, and Asif A. Ghazanfar. 2016. Monkey vocal tracts are speech-ready. *Science Advances* 2 (12): e1600723. <https://doi.org/10.1126/sciadv.1600723>.
- Laver, John. 1980. *The phonetic description of voice quality*. Cambridge: Cambridge University Press.
- Lieberman, Philip. 2016. Comment on "Monkey vocal tracts are speech-ready". *Science Advances* 3 (7): e1700442. <https://doi.org/10.1126/sciadv.1700442>.
- Lieberman, Philip. 1975. *On the origins of language: An introduction to the evolution of human speech*. New York: Macmillan.

- Lieberman, Philip, Edmund S. Crelin, and Dennis H. Klatt. 1972. Phonetic ability and related anatomy of the newborn and adult human, neanderthal man, and the chimpanzee. *American Anthropologist* 74: 287–307.
- Wolk, Lesley, Nassima B. Abdelli-Beruh, and Dianne Slavin. 2011. Habitual use of vocal fry in young adult female speakers. *Journal of Voice* 26(3): 111–116.
- Yuasa, Ikuko P. 2010. Creaky Voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech* 85(3): 315–337.



In this chapter we will

- introduce speech visualization in Praat as our workhorse for the remainder of the book
- cover the basic physical processes of sound creation and propagation related to human speech
- exemplify sound wave characteristics; e.g. their frequency or amplitude and visually and auditorily explore and combine the waves in Praat
- link the understanding of the articulatory aspect of voicing and articulation from Chapter 3 to the acoustic aspects of sources and filters

4.1 Introduction

In the previous chapters, I asked you to observe the movements of your articulators when saying some words (e.g. ‘beet’ vs. ‘boot’), watch a video, trace the movements of the tongue and lips in a graph, or introspect your own body (e.g. feel the muscles participating in breathing). Starting this chapter, in addition to the tactile and auditory information regarding speech movements, we will add a tool for utilizing the visual modality in understanding and appreciating speech: looking at speech through its *acoustics* using the freeware for analysing speech called *Praat*.

The chapter assumes no prior knowledge of either acoustics or Praat. It is important to know, however, that Praat and acoustics are intertwined in the chapter. Praat is a software analysing the acoustic characteristics of speech and relevant acoustics features might be exemplified through Praat. Hence, the chapter will

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_4) contains supplementary material, which is available to authorized users.

proceed as if we started swinging on a children swing: some familiarization with Praat (forward), a little acoustics (backwards), more Praat (forward), more acoustics (backward), etc. In the rest of the book we will increase the extent of swinging, and hopefully bringing more fun and enjoyment with every move. The goal is to gradually build your proficiency and make you comfortable with using Praat in exploring and discovering patterns of spoken English.

4.2 Praat Basics


Praat is computer software and thus requires access to a computer. It is also recommended to have headphones with a microphone to ensure a decent quality of sound when recording speech as well as during sound playback. While computers are expensive, basic headsets with a microphone are cheap (costing only the equivalent of several beers or cappuccinos). They can, however, dramatically improve your experience with Praat. So I strongly advise using headsets when working with Praat.

Praat was developed by P. Boersma and D. Weenink and can be downloaded from <http://www.fon.hum.uva.nl/praat/> for any computer by selecting the operating system in the top left corner (Macintosh, Windows, Linux). My description is based on the Windows version; minor differences regarding the colour visualization or shortcuts might apply for other versions. The program is extremely compact. In addition to Praat's own manual and tutorials (*Help* → *Praat Intro*), there are many tutorials for Praat by others if you get stuck at any point or wish to explore on your own. I list three of them in the Find Out More box 4-1. Moreover, there is a Praat-users list which has over 2700 members and an archive of messages dating back to 2000; the link is included in the box as well.

Find Out More 4-1

Useful manuals for using Praat available online

- Will Styler's tutorial that is being maintained and updated
 - <http://savethevowels.org/praat>
- Jennifer Smith's hand-outs:
 - <https://users.castle.unc.edu/~jlsmith/ling520/praat.html>
- Maria Gouskova's tutorial:
 - <https://www.gouskova.com/2016/09/03/praat-tutorial/>
- Praat-users list:
 - <https://groups.io/g/Praat-Users-List>

Once you have Praat on your computer, open by clicking the 'mouth & ear' icon . Two windows will pop up: *Praat Objects* and *Praat Picture*. You can close *Praat Picture* for now.

Record Your First Sound in Praat

Let's start with recording your own voice. Connect a microphone to your computer. Position the mic slightly below the corner of your mouth rather than directly in front of the mouth in order to avoid puffs. If you don't have a headset with a microphone, using the internal built-in microphone of your laptop is an option but the quality of the recording might be significantly downgraded.

Press *New* in *Praat Objects* and the first option of the pull-down menu is *Record mono sound*. This opens a new window *SoundRecorder*. Hit *Record* and say, for example, 'Discovering spoken English'. You should see a moving green bar during voice recording, and then hit *Stop*. If the green bar was very low and didn't move much, you should increase the gain on your computer sound mixer. If, on the other hand, the bar was high and often yellow or red, you should lower the gain or move the mic away from your mouth. Adjusting the gains as well as selecting the input device if you have a Windows machine is described in *Help* → *Intro* → *Intro 1 How to get a sound* → *Recording a sound*; point #2. Once you are happy with the recording, click *Save to list*. This transfers the recorded sound into the list of objects in *Praat Objects*. You can play the recorded sound (select the sound object so that it is highlighted (in blue) and hit *Play*; 3rd button from the top). If you are satisfied with the recording and its quality, save it by *Save* → *Save as WAV file...*

► **Praat never automatically saves anything and to prevent losing your work or files, always manually and regularly save your work.**

You can also use other devices for recording your voice if you are already familiar or comfortable with them. Then transfer the recorded sound to the computer and open the file in Praat: *Open* → *Read from file...* and navigate where you have stored your recording. If you currently cannot record your own sound, you can use sound 'DSE.wav' recorded by me in this chapter of the book companion.

First Steps in Navigating Praat User Interface

Now that you have a sound, let's familiarize yourselves with its visualization. When the sound is highlighted in *Praat Objects*, click *View & Edit*. This opens a new Editor window with new menu options. Make sure that none of the five boxes are checked in *View* → *Show analyses...* You should then be looking at an image similar to Fig. 4.1.

You can play the sound by clicking the bar *Visible part* or *Total duration* in the bottom of the image, or place a cursor somewhere in the middle of the sound and play the part up to that point by clicking the bar with the interval duration left of the horizontal red line (0.807113 in the figure above) and the part starting here until the end by clicking the bar with the interval duration right of the vertical

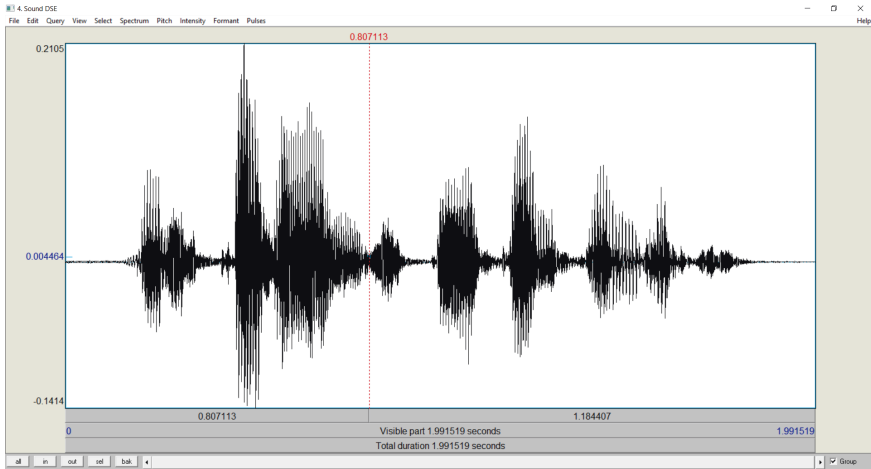


Fig. 4.1 Screenshot of the sound wave corresponding to ‘Discovering spoken English’ in Praat

red line (1.184407). I placed the cursor (and thus the vertical line) approximately between the words ‘discovering’ and ‘spoken’ in the figure. You should also try zooming in and out (either *View* → *Zoom in/Zoom out*, using the shortcuts *Ctrl-I* and *Ctrl-O*, or pressing the buttons *in* and *out* in the bottom left of the window, respectively). You can also select an interval with a mouse by dragging it, play the selected interval by clicking the bar below or above it and zoom into this selection by *Ctrl-N* or the *sel* button.

Activity 4-1: Cut & Paste sound sections

There are lots of fun things that could be done with sounds in Praat. For example, you can cut the word ‘spoken’ and paste it to the end of the sentence producing ‘Discovering English spoken’ from the original ‘Discovering spoken English’.

First, experiment with placing the cursor and identifying a rough start and tend of the word ‘spoken’. Then drag the mouse to the target interval and check auditorily that you have only selected the word ‘spoken’. Now go to *Edit* → *Copy selection to Sound clipboard*, or use the universal shortcut *Ctrl-C*. Then cut the selection (*Edit* → *Cut*, *Ctrl-X*), place the cursor at the end of the final ‘sh’ sound of ‘English’ and paste (*Edit* → *Paste after selection*, *Ctrl-V*). Listen to your creation. What aspects, if any, sound unnatural?

Finally, while ‘Discovering English spoken’ is not a well-formed English phrase, ‘Discovering English’ is. Could this utterance, created by cutting, pass for an original recording?

4.3 Sound Waves and Their Visual Representation

Now is a good time to switch gears and explore acoustics in order to understand what the waves and lines in Praat's interface (and Fig. 4.1) represent. For example, it is quite obvious after doing Activity 4-1 that the x-axis represents time. But what about the y-axis? Here we start with discussing basic physical characteristics of sounds and gradually build the appreciation of speech visualization as a useful facilitator in our explorations of the phonetic characteristics of speaking.

A sound is caused by the **vibration** of some object. The vibration is initiated by some force; e.g. by plucking a string on a guitar, somebody slamming their palm on the table, or a bee flapping its wings. This initiation then causes the vibration of the surrounding medium, most commonly air, but possibly also water or other substances, and sound becomes a **travelling pressure variation** in the medium. Let's illustrate these concepts with a tuning fork.

This is a device producing a fixed tone that was commonly used for tuning instruments, checking hearing loss, or cuing a choir leader to give proper notes to the singers. Before initiation, the two tines of the fork are not moving and the air molecules are stationary (shown in Fig. 4.2 with straight lines in (a) upper left). When the tuning fork is hit against a hard object, it starts vibrating by repeatedly increasing and decreasing the distance between the tines at the top of the fork; shown in Fig. 4.2 (b) moving outward and in (c) inward. In (b), the movement of

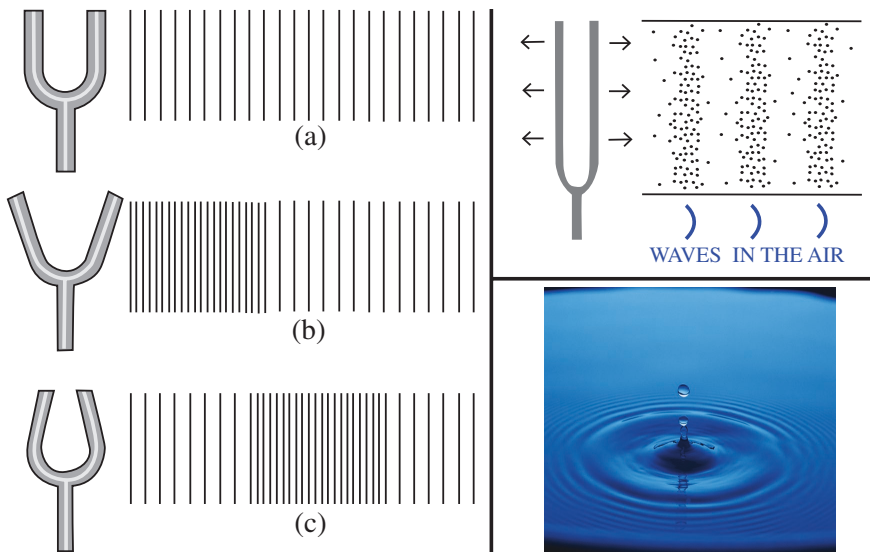


Fig. 4.2 Visualization of sound waves. A tuning fork on the left with air particles as vertical lines in silence (a), when the fork is hit and its tines vibrate by cyclically moving outwards (b) and inwards (c). The result of this vibration shown in upper right and the travelling of sound waves is similar to the travelling of waves in water

the right tine pushes the air molecules rightward, causing a small area of increased pressure; i.e. more molecules in roughly the same area. When the tines go inwards as in (c), this creates an area of low pressure next to the right tine since the molecules that were there were pushed rightward before. Meanwhile, the molecules that were originally pushed rightward in (b) bumped into those in the middle (roughly above the letters (a), (b), (c)), which created another area of higher pressure in the middle of (c) further away from the original area of high pressure in (b). The next stage would be again (b) pushing those original molecules, which returned to their original positions, again rightward and the *cycle* would continue. The sound thus travels as the areas of increasing and decreasing air pressure progressively moving away from the *source* of vibration.

The pictures in the right panel of Fig. 4.2 illustrate the travelling sound wave (top) or familiar water wave (bottom) that are based on the same principle. It is important to realize that air molecules do not travel anywhere; they just move away from the fork and back due to the movement of the tines. What travels is pressure fluctuation, not the air particles themselves.

Now, let's take a single air particle and imagine its movement. The particle and table on the left in Fig. 4.3 illustrate the particle being pushed from its rest position (0) to the right (1) and then it goes back to the rest position (2), continues overshooting to the left (3) and then again back to the rest position (4).

If we graph the horizontal position of the air particle in time so that its rest position corresponds to zero, movement to the right corresponds to positive numbers and movement to the left to negative ones, we get a *sine wave* (Fig. 4.3 right) with the five stages in the particle movement represented as (blue) circles.

The final mental leap we need is to treat the y-axis on the sine wave as the air pressure in the vicinity of the particle rather than the particle position. Zero corresponds to the atmospheric pressure and values above zero correspond to *compression*, i.e. pressure higher than the atmospheric pressure, and values below zero to *rarefaction*; i.e. pressure that is lower than the atmospheric pressure. We can now see sound as local changes in air pressure caused by a vibrating source (a tuning fork or a guitar string).

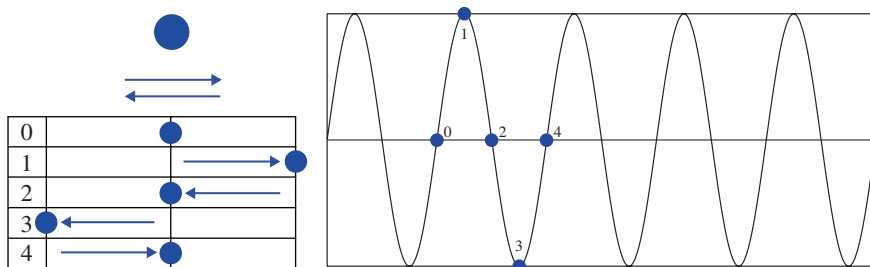


Fig. 4.3 The movement of a single air particle. On the left, the pendulum-like horizontal movement from the rest position to the right, back, then left, and back again. The same stages of the particle position in time shown with the sine wave on the right

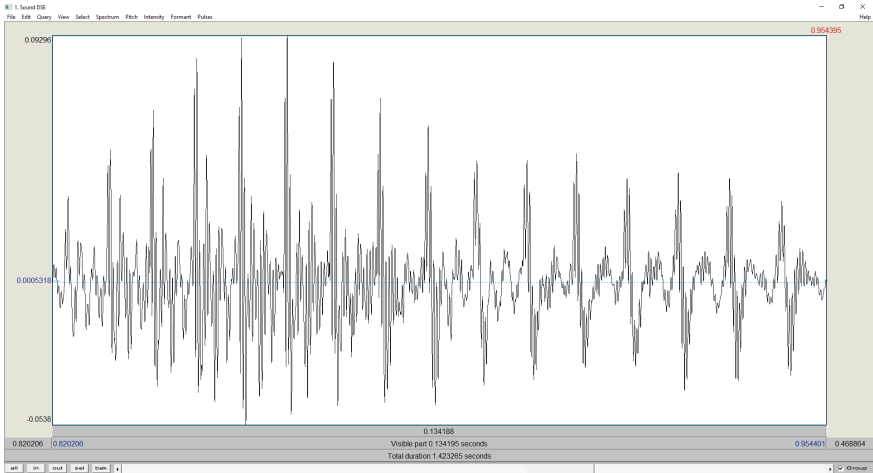


Fig. 4.4 Zoomed-in section corresponding to ‘eng’ of ‘English’ in the recording ‘Discovering spoken English’ in Fig. 4.1

Going back now to Fig. 4.1 representing the recording of ‘Discovering spoken English’, the x-axis is time, y-axis is air pressure and the squiggly lines represent the pressure fluctuation picked up by the microphone membrane during speaking. Now in your recording, try placing the cursor at the beginning of the word ‘English’ and drag the mouse to select the interval roughly corresponding to the first syllable of the word (‘eng’). Check with your ears by clicking the bar before the interval, the interval itself and the bar following it, and adjust your selection until you are happy with the result. Then zoom into this selection by Ctrl-N. You should see something similar to the image in Fig. 4.4.

We clearly see the sound wave corresponding to pressure fluctuations. We can also see the **periodicity** of the wave since we can identify a repeating pattern (more on this later). However, it is also clear that the repeating pattern is not a nice simple sine wave from Fig. 4.3 but something much more complex. This is because only certain specific sounds correspond to simple sine waves and we call these sounds pure tones. You can create a pure tone in Praat to check for yourself.

Activity 4-2: Create pure tones of various frequencies in Praat

Go to *Praat Objects* window and select *New* → *Sound* → *Create Sound as pure tone...* A table with several parameters pops up. Let’s call the sound you are creating ‘tone100’ and keep the defaults for the following two values as they are (creating a mono sound starting from 0). Set the *End time (s)* to 1 so that your sound is 1 second long. Leave the sampling frequency value and specify *Tone frequency (Hz)* to 100 so that the resulting tone will have the frequency of

100 Hz. Leave the remaining values as they are and hit *OK*. The *Praat Objects* window now has your tone100 and you can listen to it. Also click View & Edit and then zoom in several times (Ctrl-I) to verify that the sound wave of pure tone is indeed a sine wave.

You can play around with creating multiple sounds of various frequencies.

4.4 Periodicity, Frequency and Amplitude

Now that you can create sounds of various frequencies, let's look at these waves in more detail. Speech sounds correspond to two major types of waves: *periodic* and *aperiodic*. Periodic waves can be characterized by a repeating pattern, or a cycle. A typical example is a sine wave of a pure tone. In human speech, periodicity comes from the vibration of the vocal cords, which is a cyclic opening and closing based on aerodynamic principles that we discussed in Chapter 3. The continuous stream of air is chopped by the glottis closures into periodic puffs of higher pressure. Hence, all speech sounds in which the vocal folds vibrate regularly should be represented by periodic waves in Praat.

Aperiodic waves are irregular non-repetitive waveforms that can be seen when we zoom to sounds like /s, sh/or the burst when you open your lips during words like 'pan'. An example in the non-speech domain is radio static or white noise.

The left panel of Fig. 4.5 shows a section during the first sound of 'spoken' (i.e. a snake-like [s] sound). We see that it is not possible to identify a repeated pattern and the wave is thus irregular. The right panel shows a section of similar duration from the first vowel in 'spoken' (i.e. a sound like 'oh') and the vertical dashed lines illustrate approximately individual repeated cycles corresponding to the vibration of vocal cords. I suggest you inspect these portions in your own recording of 'Discovering spoken English'.

For every periodic wave we can identify its *fundamental frequency* that corresponds to the number of cycles per one second. In speech it is thus the rate of vibration (open–close cycles) of the vocal folds. The duration of one cycle corresponds to *period* (T). Frequency (F) is measured in *Hertz (Hz)* and the relation between frequency and period is expressed mathematically as $F=1/T$. Hence, in

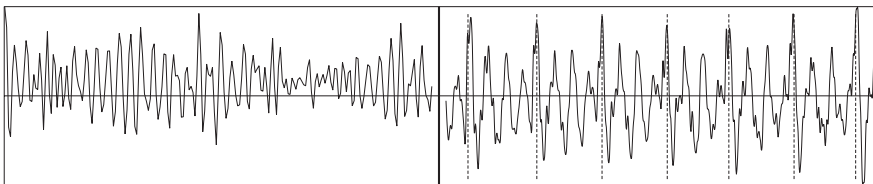


Fig. 4.5 Two zoomed-in sections of the original 'Discovering spoken English' from Fig. 4.1. On the left, the interval during the hissing sound of the initial sound in 'spoken', on the right, the vowel portion of the oh-like sound in 'spoken'

the pure tone with frequency 100 Hz that you have created in Activity 4-2, one cycle repeats 100 times in one second and the duration of one cycle (= period) is thus 0.01 s or one-hundredth of a second.

Humans can hear sounds whose frequency ranges roughly from 20 to 20,000 Hz, though speech sounds essentially utilize the lower half of this range. Old telephone lines could transmit frequencies up to 4000 Hz and the comprehensibility of such a signal was sufficient for phone calls.

Activity 4-3: Can you match the frequency of a pure tone?

From the frequencies of pure tones with which you played around in the previous Activity 4-2, pick one you are comfortable with (100 Hz, 150 Hz, 200 Hz or 440 Hz) and try to sing ‘aahhh’ in this frequency. If you are happy with the result, record yourself singing this ‘aahhh’ and do not forget to save the resulting sound as a wav file to your computer. Now verify the frequency of the pure tone by zooming in until you can identify individual cycles and measuring the period T (duration of one cycle). This is best done by dragging the mouse from the beginning to the end of the cycle at zero crossing and Praat shows the duration of this interval in the bar above and below the interval. You can calculate frequency using the formula $F=1/T$. It is likely that the resulting number is slightly different from the whole number used for creating the file. This is the ‘measuring error’ of not placing the start and/or the end of the interval at precisely zero crossings.

Now repeat the same with the file you have recorded, zoom in somewhere in the middle of ‘aahhh’ until you can recognize cycles. Although the sound wave is complex and not a simple sine wave, you should be able to identify the repeating pattern and measure the duration of one cycle by mouse dragging. Use $F=1/T$ again. How close is the fundamental frequency of your singing to the frequency of the pure tone?

In addition to frequency, the other major parameter characterizing sound waves is their **amplitude**. This corresponds to the force of the excitation for the air particles: greater force corresponds to greater displacement of the particles and thus greater fluctuation of air pressure. Hence, amplitude corresponds roughly to the loudness of sounds. You can create a pair of pure tones following the steps in Activity 4-2, keep all parameters for the two tones identical, except for a change in the 7th parameter *Amplitude (Pa)*. For example, keep the default 0.2 for one sound and set it to 0.6 for the other sound. Now play the two sounds in succession and notice how the sound with the amplitude of 0.6 is louder than the one with the amplitude of 0.2.

It is important to understand that frequency and amplitude are **independent** of each other. In the top left and right panels of Fig. 4.6, we see two waves that have the same frequency (440 Hz) but different amplitude: the left one has greater amplitude and the resulting sound is thus louder than the right one. The

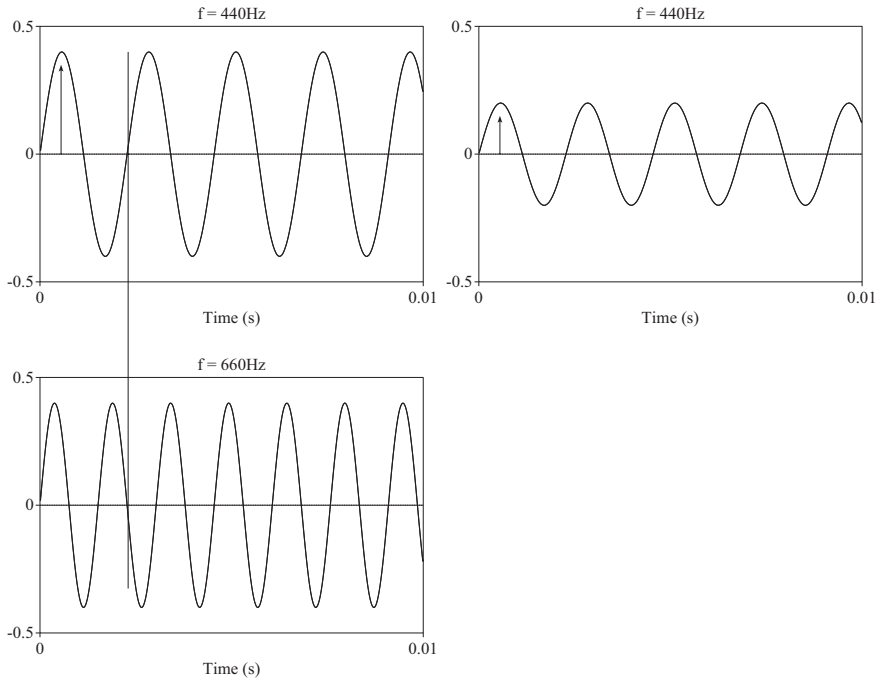


Fig. 4.6 Frequency and amplitude of sound waves are independent. Same frequency but different amplitude (top left and right), and same amplitude but different frequency (left top and bottom)

comparison of the left top and bottom panels illustrates the opposite situation in which the two sound waves have identical amplitudes but different frequencies; 440 and 660 Hz, respectively. In the same amount of time it takes the top wave to complete one cycle, the bottom wave completes one and a half cycles.

You can experiment with your own voice and try singing two notes, one lower and one higher, and singing them with roughly the same volume. This corresponds to the top left and bottom left panels of Fig. 4.6. Singing the same note but varying its volume is depicted in the two images on the top. Recall from the previous chapter that increased loudness results primarily from increased sub-glottal pressure.

4.5 Sources, Filters and Speech

We have already seen that human speech sounds do not correspond to nice simple sine waves of pure tones but to *complex waves*. In this section we discuss the origin of this complexity by reviewing the functions of the sound *source* (vocal cords) and the *filter* (cavities above the larynx).

Let's start with characterizing the *source* of vibration in human vocal folds and how it differs from a tuning fork. It is instructive here to see again a slow-motion view of vocal folds vibration in the first link of Find Out More 3-1 (Chapter 3) in the companion. There are three major differences between the tuning fork and the vocal folds. The first is the rigidity of the tuning fork that is made of metal and the elasticity of the vocal folds made of soft tissue. The second is the complexity of the open–close movement itself. The video above showed us that the vocal folds do not open abruptly along the entire length of the glottis (the space between the vocal folds). Rather, they start opening at one end, let's call it the bottom, and gradually open towards the top while closing has already begun at the bottom. The third difference is that in tuning forks it is the oscillating body itself causing the pressure fluctuations. For vocal folds it is not them pushing the air but their closing, which causes a negative pressure that contrasts with the positive pressure of the air released after the vocal folds opening. Due to these differences, the complexity of the resulting pressure fluctuation is much greater in human voice than in the tuning fork. Similar to strings, vibrating human vocal folds do not produce a single frequency but multiple frequencies. We call them *harmonics*, and in music they might also be called *overtones*.

All musical instruments likewise produce multiple frequencies when their vibration is initiated. If you pluck a guitar string or bow a violin string, the fundamental frequency that we hear as pitch (or the height) of the sound corresponds to the length, tightness and thickness of the string. However, a vibrating string also produces more complex movements and it is not only the entire string that vibrates, also both of its halves, and all of its thirds, quarters, etc. This is illustrated in the left panel of Fig. 4.7.

These harmonic frequencies, both in humans and musical instruments, thus correspond to whole-number multiples of the fundamental frequency. And if we combine a fundamental frequency with several of its harmonics, all of them being simple sine waves of pure tones, the resulting wave is a complex periodic wave that begins to resemble the actual sound waves of the human voice. The right panel of Fig. 4.7 shows how summing five sine waves (the fundamental of 200 Hz and its harmonics of 400, 600, 800 and 1000 Hz) produces a complex periodic wave with the fundamental frequency of 200 Hz. It is important to realize that despite the fact that all five sine waves appear to have the same amplitude in the figure, the amplitudes of harmonic frequencies are in fact decreasing so that the fundamental frequency is the loudest, 1st harmonics less loud, and so on.

Activity 4-4 guides you to combine sine waves of multiple frequencies on your own in Praat.

Activity 4-4: Combine simple sine waves into a complex periodic wave

Another fun and instructive thing to do in Praat is to combine sounds. Let's create the sound depicted at the bottom of the right panel of Fig. 4.7 under

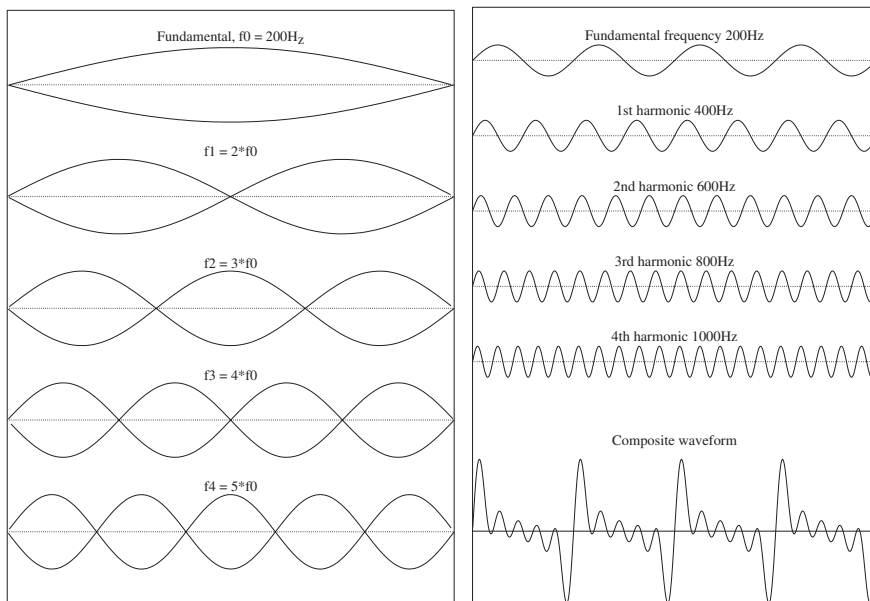


Fig. 4.7 Illustrations of fundamental frequency and its harmonics. On the left, vibrations of a string, on the right, sine waves of the fundamental and first four harmonic frequencies and their summation into the composite complex waveform

‘Composite waveform’ and also listen to it. The only difference we will introduce to model real life more closely is the decreasing amplitude of the sine waves with the fundamental being the loudest and the 4th harmonic the quietest. As we did in Activity 4-2, create the fundamental frequency (*New* → *Sound* → *Create Sound as pure tone...*) of a one-second long tone with a frequency 200 Hz and an amplitude 0.8 and name it *tone200*, leaving other parameters in their default values. Then create the 1st harmonic of 400 Hz and 0.6 amplitude, and name *tone400*. In the same way continue with *tone600* with 0.5 amplitude, *tone800* with 0.4 and *tone1000* of 1000 Hz and 0.2 amplitude. Listen to them and notice how artificial they sound.

Now select all five sound objects in the Object window with the mouse, so that they are all highlighted (in blue), and click *Combine* → *Combine to stereo*. This creates a new Sound in the Praat Objects window called *Sound combined_5*. You can click *View & Edit* and see all five channels representing all five waveforms, respectively. Now back in the Praat Object window, select *Sound combined_5* and then click *Convert* → *Convert to mono*, which sums all five waveforms together and creates *Sound combined_5_mono*. If you click *View & Edit* and then zoom in, you should see the complex periodic wave with a period of 0.005 s and a fundamental frequency of 200 Hz that is shown in Fig. 4.8 and resembles the composite waveform from the bottom of the right panel of Fig. 4.7.

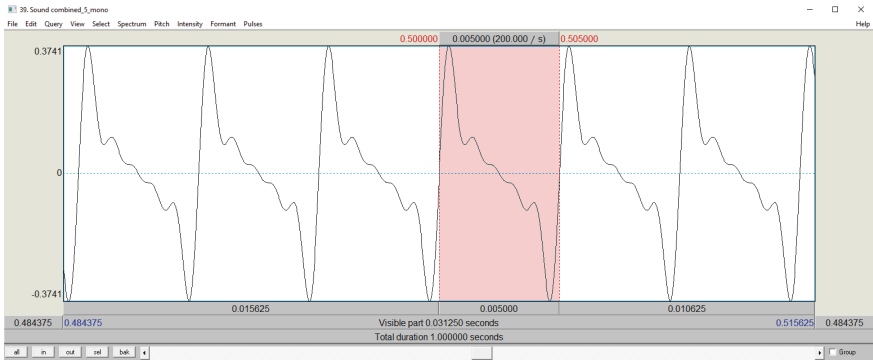


Fig. 4.8 Composite waveform created by summing five sine waves (200, 400, 600, 800, 1000 Hz). See Activity 4-4 for instructions and description

More importantly, you can listen to your own creation. How does this sound compare to how the individual sine waves sound? Also, note how the composite wave below more closely resembles the visual representation of real speech on the right side of Fig. 4.5.

Just as we can combine sine waves of known frequencies and amplitudes into a complex periodic wave, there is a mathematical process called **Fourier Transform** that can take a complex periodic wave and determine the individual waves such that summing these individual waves creates this complex wave.

Now that we understand that the source of vibration is created by the vibrating vocal folds in the larynx, the only remaining concept we have to introduce to understand the visualization of waveforms is the **filter** afforded by the cavities above the larynx.

Let's call on musical instruments again for help in understanding this concept. Imagine the difference between a classical and electric guitar. The classical guitar has a hollow body with a hole in the middle. The electric guitar only has a solid piece of wood or plastic for its body. In the classical guitar, the body serves as the resonator and the amplifier. The frequencies (fundamental and harmonics) created by the source (vibrating string) enter the hollow body and due to the size and shape of this body, some frequencies are selectively strengthened, or amplified, and other frequencies are dampened. Now imagine that a string on an unplugged electric guitar is plucked. The sound would be extremely quiet because there is only a solid hard body that can resonate and no hollow area filled with air in which the sound could be amplified. This is why we need to plug electric guitars into electric amplifiers to truly hear the sound they produce.

Now let's imagine a violinist playing the basic note a1 (440 Hz, the empty second thinnest string) and a cello player playing the very same note. Experienced musicians would be able to tell these two sounds apart and even most novices, after a couple of practice trials, should be able to hear the difference. The difference, of course, comes in part from the different sizes of the bodies of these

instruments since the players play exactly the same note. In the larger body of cello, lower (harmonic) frequencies are better amplified while in the smaller violin higher frequencies are happier bouncing around and resonating. Hence the instrument's body acts as a selective filter that participates in creating the timbre of the sound by amplifying certain harmonic frequencies and damping others.

The very same principle is at work in human speech. Essentially, the difference between saying [i] and [a] in the same pitch and loudness is of the same nature as the difference between the violin and the cello when they play the same note with equal loudness. This is amazing since we have no trouble recognizing the difference between the two vowels but the difference in the timbre of string instruments might be more challenging. Here comes the long-practiced habit again when contrasts with vowels have been produced and perceived countless times, and linked to the articulatory actions making this difference, whereas the contrast between cello and violin only few experts practiced sufficiently.

In short, the movements of the active articulators – the tongue, the lips and the jaw – affect the shape of the resonating cavities, and these different shapes select different frequency ranges for amplification. The amplified frequencies of human speech are called *formants* and are typically referred to as F1, F2, F3, etc. with f_0 being reserved for the fundamental frequency. And what our ears decode as /i/ or /a/ is essentially the relationships between the formant frequencies.

Now we have all the pieces together for summarizing the *source-filter theory of speech production*. Figure 4.9 illustrates this summary. The vibration of the

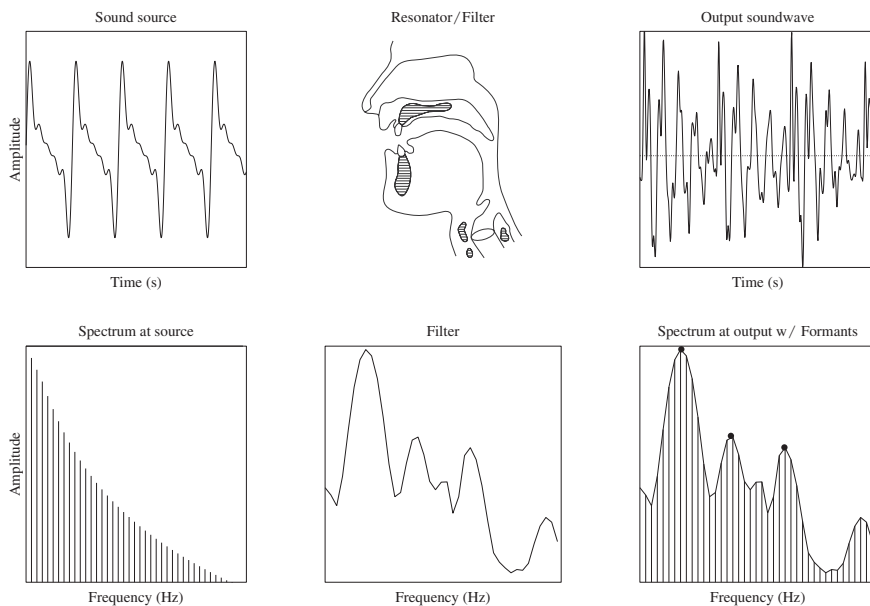


Fig. 4.9 Schematic illustration of the Source-filter theory of speech production in a nutshell. See text for explanations

vocal folds creates multiple simultaneous frequencies with the lowest fundamental (f_0) and whole-number multiples harmonic frequencies (H1, H2, ...). The top left panel illustrates this with the composite waveform we created in Activity 4-4 from $f_0=200$ Hz and its first four harmonics with time and amplitude as x- and y-axis, respectively. The bottom left panel graphs the vocal source as a *spectrum*, which shows frequency on the x-axis and amplitude on the y-axis. Note that this is a representation of a single point in time. The vertical lines are equally spaced and the first fundamental frequency has the greatest amplitude and the subsequent harmonics decrease in amplitude.

This acoustic source created at the vocal cords then enters the resonator, or an acoustic filter, illustrated in the middle panels of Fig. 4.9. This is formed primarily by the oral and pharyngeal cavities, with the option of also including the nasal cavity. The shape of the first two cavities is determined by the movements of the active articulators and is frequency-selective, which means that it selects only certain frequencies to be amplified in the resonator of this shape. The articulatory shape of the resonator in the top corresponds to the acoustic spectrum of the filter at the bottom.

In the rightmost panels we have the output when the filter acts on the source frequencies during the real a-like vowel from the second syllable of ‘discovering’ in the file ‘DSE.wav’ analysed throughout this chapter. The top image is again a time-by-amplitude graph that we will be using a lot in Praat and shows a complex periodic wave for the resulting a-like sound. The bottom right is again the spectrum representation in which the peaks, marked with dots, represent the most amplified frequencies called formants. The first three formants are marked and in the analysed vowel they correspond to 613 Hz, 1374 Hz and 2144 Hz. These formant frequencies then primarily guide the decoding process in our ears resulting in identifying this sound as an a-like sound.

The concepts behind the source-filter theory of speech production might feel heavy and complex and this is a good place to stop and try to informally explain to a non-expert listener to strengthen your own understanding of these concepts.

4.6 Visualizing Speech with Spectrograms

Let us now return to the beginning of the chapter and Fig. 4.1 that represented the wave form of the recording ‘Discovering spoken English’. This is essentially a time by amplitude representation of the waveform. There is a lot of useful information in this visualization of speech. We can zoom in and see if certain sections are periodic, and thus correspond to vocal fold vibration, or aperiodic. We can also see that loudness generally decreases over the phrase but also that locally there is great variation of increasing and decreasing loudness (corresponding more or less to vowels and consonants; more on this in the next chapters). However, we do not really see the frequencies that were discussed above. We could zoom in and calculate the fundamental frequency of vocal folds vibration, but the plain waveform

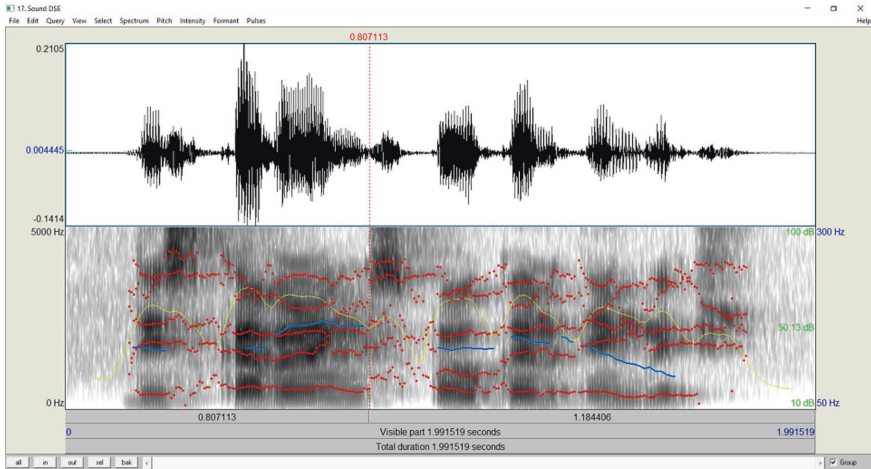


Fig. 4.10 Waveform (top) and spectrogram (bottom) of ‘Discovering spoken English’ in Praat. The (blue) bolder contour corresponds to pitch (f_0), the (yellow) thinner line to intensity and (red) speckles to formant frequencies

is not very useful for observing how formant frequencies change when we speak. Hence, we introduce here the last concept of this chapter: a *spectrogram*.

This is a three-dimensional representation of time on the x -axis, frequency on the y -axis, and amplitude represented on the greyscale such that dark areas correspond to louder frequencies and lighter areas to quieter ones. In Praat, a spectrogram is visible when you select *Spectrum* \rightarrow *Show spectrogram* while in the Editing window. Figure 4.10 shows the waveform, the spectrogram, pitch, amplitude and formants. You can also select what representation to see by *View* \rightarrow *Show analyses...* and checking the appropriate boxes.

Visualizing speech with both the waveform and the spectrogram, possibly also with f_0 and amplitude contours, provides a much more complete picture than just the waveform alone. For example, we can see that /s/ of spoken, starting roughly at the cursor shown with the dashed red vertical line in Fig. 4.10, has distinctly loud, i.e. dark, frequencies in the high-frequency range. The sound /p/ immediately following is very quiet, which is shown by the light-grey of the spectrogram and minimal amplitude in the waveform. This is because the lips are closed, vocal folds do not vibrate and thus essentially no sound is emitted. After ‘p’ we see the oh-like vowel, characterized with very salient formant frequencies shown as dark bands in the lower part of the spectrogram and emphasized in Praat with (red) dots.

Activity 4-5: Strengthening the eye–ear–mouth links with Praat

In Praat, open the target file of the chapter, the recording of ‘Discovering spoken English’ – the best is your recordings or use the one provided in the companion – and produce the representation identical to Fig. 4.10. The goal is to

become comfortable with this visualization so it does not feel daunting, and reinforce the link between what you see on the screen with your eyes, what you hear with your ears, and what you feel in your vocal tract. Select various meaningful intervals (words, sounds) and play them, say them yourself and trace the visualization. Discuss why some sections are louder, some have clear formants, and also look at formants without the red dots (unchecking *Formants* → *Show formants* or in *View* → *Show analyses...*).

Spectrograms come in two basic flavours. The one shown in Fig. 4.10 is called a **broadband** spectrogram and will be used most in the book and in your work with Praat. Here we mention also the **narrowband** spectrogram whose settings of the window size allow for finer illustration of frequencies, both the fundamental and its harmonics. The spectrogram in Fig. 4.11 is from the same utterance but the window length was not the default 0.005 s as in Fig. 4.10 but 0.05 s. You can experiment with window lengths in *Spectrum* → *Spectrogram settings...*

Figure 4.11 illustrates better the fundamental frequency, the bottom-most dark stripe around 140 Hz (the y-axis in the bottom panel represents frequencies), and its harmonics that are whole-number multiples of the fundamental, and are thus equally spaced dark stripes. The figure also nicely shows the movements of the fundamental frequency due to the variation in the frequency of the vocal cord vibration. On the other hand, the formant frequencies as amplified harmonics are less clearly seen compared to the broadband spectrogram.

In the following chapters we will discuss consonants and vowels in more detail, and present them first through articulatory descriptions, i.e. what our articulators are doing building on Chapter 3, and also through acoustic descriptions building on this chapter. Importantly, you will be also guided to record your own sounds,

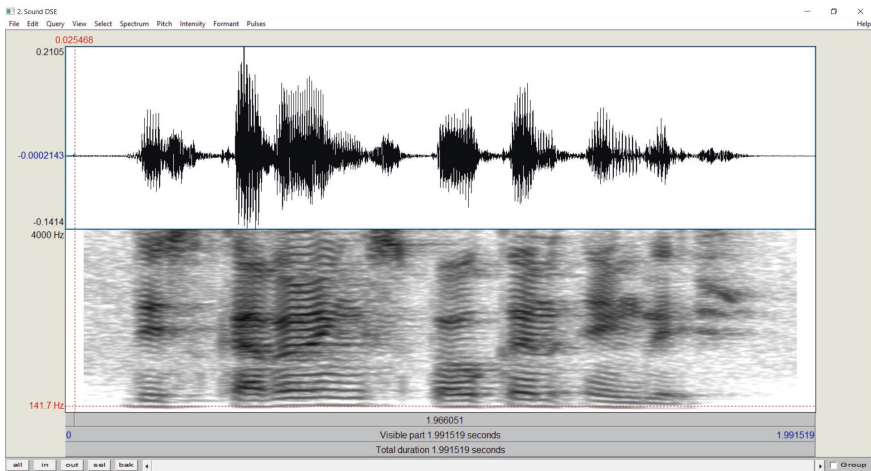


Fig. 4.11 Waveform (top) and narrowband spectrogram (bottom) of ‘Discovering spoken English’ in Praat

and inspect their characteristics in Praat both visually and auditorily. And then we do the same with syllables, words and entire sentences. At every step we will discuss systematic patterns in the speaking behaviour, visualize and engage with them using Praat. This will gradually build not only proficiency with using Praat, but also conscious awareness of the habits we acquired unconsciously for our native speech, and somewhat more consciously for non-native speech.

Exercises

- 4-1 Strengthen your proficiency with Praat and understanding of the concepts from the chapter by recording a new short utterance, saving it, zooming into inspect periodic and aperiodic sections of the waveform, understanding the nature of the waveform as showing moving pressure fluctuation in time, calculating the fundamental frequency somewhere during a vowel, looking at both broadband and narrowband spectrograms, identifying the fundamental, the harmonic and the formant frequencies and being able to describe how the visual information in the spectrogram correspond to the articulatory activity in the vocal track.
- 4-2 Return to Activity 4-1 and the ‘Discovering English spoken’ sound that was a result of your cut-and-paste manipulation of the original ‘Discovering spoken English’. The newly created utterance sounds presumably quite different compared to how you would actually say ‘Discovering English spoken’. What do you hear as the main differences and what might be the reasons for these differences? You may help yourself by using Praat to record ‘Discovering English spoken’, and comparing this with the file you have created by cutting & pasting.



In this chapter we will

- explore both articulatory and acoustic characteristics of English vowels building on the awareness gained in Chapters 3 and 4
- expand skills in Praat mainly in visualizing and measuring the acoustic characteristics in vowel quality or duration
- begin using IPA for transcribing English vowels, and strengthen the awareness of sound–spelling correspondences and differences between the British and American varieties of English.

5.1 Introduction

In presenting individual sounds, many books dealing with English phonetics and phonology opt to start with consonants and follow with vowels. It makes a lot of sense since consonants are much friendlier for introspection and present less variability in terms of dialectal differences. You ‘feel’ when your active articulators touch and how they create obstructions to the airflow. In this book we will reverse that order for two main reasons. First, vowels can be produced in a steady state without consonants and thus out of context, which enables concentrating solely on the contrasts among vowels. Several consonants, on the other hand, are difficult, or unnatural to produce without vowels and this order enables the discussion of consonants in the next chapter to be richer. Nevertheless, keep in mind that even vowels are almost never produced without consonants in natural speech.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_5) contains supplementary material, which is available to authorized users.

Understanding gained from studying vowels separately in this chapter builds the foundations for the subsequent chapters gradually expanding the focus to syllables, words or phrases, and the rich variability of vowel production there.

Second, the acoustic nature of vowels provides a natural continuation of the previous chapter on acoustics. This allows for further strengthening and explanation of the concepts such as formant frequencies introduced there and of Praat as a useful tool for their examination. We thus aim at a holistic picture of vowels through both articulatory, acoustic and kinaesthetic examination.

The issue of the rich dialectal variability mostly exemplified by differences in vowels, is unavoidable. Have you noticed how saying ‘good’, ‘dog’ or ‘hurry’ might be very different for native English speakers from different regions and countries? This book concentrates on the two varieties of English most studied phonetically (British and American) and limits the discussion to their ‘standard’ varieties traditionally referred to as Received Pronunciation (RP) and General American (GA). It is hoped that the skills and awareness gained in this limited scope enable you to explore other varieties, L2 speech, or other factors reflecting the identity of speakers in speech in general, and vowels in particular.

5.2 Vowel Space

Vowels are usually defined as sounds made without a significant obstruction in the vocal tract. This, as mentioned above, makes them less yielding for articulatory introspection. Nevertheless, we will start with activities that help us to become aware of the range of articulatory movements by the active articulators when producing vowels. There will be pictures depicting these movements later on but it is important to actually feel and experience these movements on your own not just intellectually learn about them from the text and the pictures.

Activity 5-1: he vs. who

Experiment with producing the words ‘he’ and ‘who’ containing [i:] and [u:] vowels, respectively. Say them aloud and then whisper them, and use a different emphasis on the word, for example, in ‘WHO did it? HE did’. It would be useful to have a look at the mirror as well. Now, take the really emphasized tokens of ‘he’ and ‘who’ and alternate the two words several times. The activity of the lips is the most easily observable either in the mirror or by putting your index finger on the lips. Do you feel any other changes apart from lips rounding and spreading when alternating ‘he’ and ‘who’? Concentrate on the position of the tongue and its possible contact with the teeth.

With ‘he’ you should feel a firm touch of the sides of tongue against the upper molars. With ‘who’ there is no such firm contact and you might feel the sides of the tongue are more *retracted* and lightly touching (or not even) only the very

back molars. ‘Mouthing’ these two words silently strengthens the awareness of this horizontal movement of the tongue.

Activity 5-1 helped in identifying two of the three major articulatory dimensions for describing vowels: lip position and the horizontal tongue position. Separating these two movements increases your awareness and understanding of vowel production in other languages or English dialects. First, try to fix your tongue by pressing it against the alveolar ridge and the hard palate and then mouth [i:] and [u:] silently, just with your lips. Since the lips activity is clearly obvious, it might feel a bit awkward suggesting that the tongue movements are an integral part of producing these vowels. Now take a deep breath and try to say [i:] with the normal position of the tongue and the lips and slowly and continuously un-spread to neutral and then round your lips while keeping your tongue in the same fixed position. You now approached a front rounded vowel, [y] in IPA, and present in many languages like German, Finnish or Turkish. Now try to alternate between [i:] and [y:] to fully grasp the separation of the tongue and the lips. After mastering this, the next level of difficulty would be to move your tongue forwards and backwards while keeping the position of the lips fixed (first rounded, which is a bit easier, and then spread).

Do not worry if the sensation is weird or if you struggle with separating the tongue and the lip movements. If your native language does not have front rounded vowels and back unrounded ones, you are trying a combination of muscle activity you probably have not tried before and thus are producing a totally novel combination of actions. This is extremely difficult due to the presence of old habits practiced millions of times, in which your tongue and lips moved in coordination for [i] and [u]; that is the tongue moving forwards accompanies lip spreading and the backward movement accompanies lip rounding.

Activity 5-2: has & hop with he & who

Similarly to the previous activity, try producing the words ‘has’ and ‘hop’ separately in an emphasized way, both aloud and silently. Now say ‘HE HAS’, stressing each word but linking them together and repeat 3–4 times. Concentrate on the movements of your lips, jaw and tongue (with a mirror if possible). Verbalize the difference between the two vowels before continuing to read further.

A useful way of feeling the tongue movement not seen in the mirror is to say [i:] of ‘he’ and then raise your tongue to touch the alveolar ridge while noticing how much distance needs to be covered until the tongue touches. Now say [æ] of ‘has’ and do the same: keep raising your tongue until it touches the alveolar ridge. You notice that the distance covered in the latter case is much greater than in the former, which means the tongue is much lower for [æ] than for [i].

In Activity 5-2 you noticed that in the transition from ‘he’ to ‘has’, your lips become slightly less spread and more open, which is accompanied by a quite

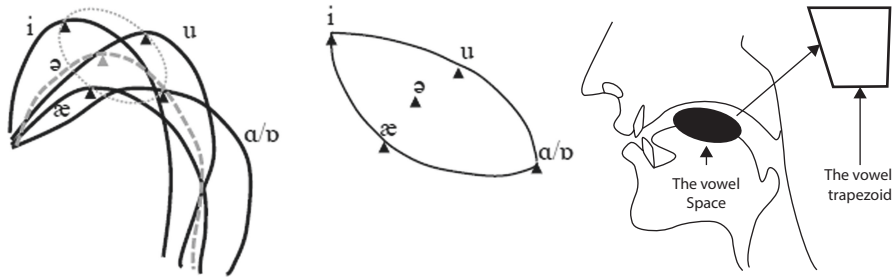


Fig. 5.1 Schematic illustration of the tongue positions for vowels [i], [u], [æ] and [ɑ]/[ɒ] in the left panel in black solid, and neutral schwa [ə] in grey dotted lines. The centre is the ellipsis covering the space defined by the highest points of the tongue in the four vowels. The right, reproduced with permission Culpeper et al. (2018), shows the abstraction from the ellipsis to the traditional trapezoid shape used for describing vowels

pronounced downward movement of the jaw and the tongue. This activity helps us to feel the third major dimension in the production of vowels: the vertical position of the tongue. Combining all four target words and the associated activity of the active articulators (lips, jaw, tongue) gives a good idea of the available possibilities for vowel production in English. Try enunciating phrases like ‘he, who hopped’, ‘who has hopped’ or ‘who hopped, he has?’ and note that for speakers of British English, the vowel in ‘hop’ is rounded, [ɒ] in IPA, while for speakers of General American it is unrounded, [ɑ] in IPA.

After acquiring a solid practical ‘feel’ for the ranges of activity of the articulators when producing vowels, we may proceed to the theoretical description. As in any scientific endeavour, **abstraction, generalization or model** are essential concepts for progress in understanding the world. We start with the schematic pictures of the tongue positions for the four vowels we introduced above: [i], [u], [æ] and [ɑ]/[ɒ]. When looking at the vertically highest sections of the tongue, shown with triangles in the left panel of Fig. 5.1, in [i] it is high and front, for [u] it is high and back, for [æ] it is low and front and for [ɑ]/[ɒ] it is low and back.

These drawings schematically show the target positions for the tongue when the goal is to produce one of these four vowels. We will call the space defined by these positions **vowel space**, within which the tongue manoeuvres. Figure 5.1 provides an abstraction from abundant degrees of freedom for the tongue shapes afforded by the complexities of the intrinsic and extrinsic tongue muscles to an essentially two-dimensional space of the horizontal and vertical position of the top point on the tongue surface. While this is a gross simplification, the concept of the vowel space with these two principal dimensions offers a very useful tool for describing the production of vowels.

We started with the introspection of the articulatory movements during vowel production and the abstraction from that to the vowel space. Now we are in position to switch to the acoustic description of vowels using Praat and test the notion of vowel space from the acoustic viewpoint.

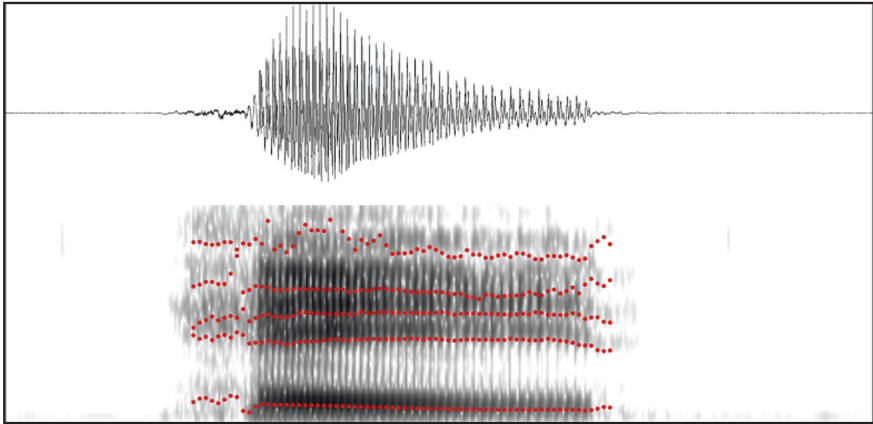


Fig. 5.2 Illustration of the spectrogram and formants of 'he' in Praat

Activity 5-3: Recording vowels in 'he', 'who', 'has', 'hop' and 'uh/er'

Following the steps in Sect. 4.3 of the previous chapter, record the four words we inspected above (he, who, has, hop), plus the plain hesitation usually written in text as 'uh' or 'err' commonly heard when people stall for time when planning what to say or searching for a particular word. Say the words in prominent highlighted enunciated way and when recording you might say each word twice so that later you can select the best example of each word. Pay attention to the gain metre staying within the green/yellow colour and do not forget to save as a wav file into your computer.

Inspect your recording in Praat by selecting the interval for individual words and listening to them separately. Also select each pair of the identical words, listen to this bigger interval and compare the two renditions. Do you perceive differences when saying the identical words? Try to describe them informally.

Recall from Sect. 4.7 that the vibration of vocal folds produces multiple frequencies and those that are amplified based on the shape of the vocal tract are called formants. When you zoom in to one of the words you recorded and select to view the spectrogram and indicate formants (*View* → *Show analyses...*), you should get an image similar to Fig. 5.2, which depicts an example of 'he'.

In the figure on the left there is a silence shown as a flat line in the waveform and a mostly white region in the spectrogram. If your recording is noisier, the line would be less flat and spectrogram more greyish. Then a brief interval of an aperiodic wave of low amplitude corresponding to [h] and then an interval of periodic high amplitude wave shown in the spectrogram with clear bands of amplified loud (=dark) formant frequencies marked with (red) dots.

We said in Chapter 4 that these formant frequencies correspond to the shapes of the resonator in the oral cavity, which in turn corresponds to the position of the tongue and other articulators. So let's determine those frequencies and explore what they show.

Table 5.1 Example of formant frequencies extracted with Praat in Activity 5-4 from the recording done in Activity 5-3

Word	F1	F2
he_1	422	2068
who_1	438	1131
has_1	705	1575
hop_1	677	998
uh_1	530	1307

Activity 5-4: Extract formant frequencies with Praat

Select your recording from Activity 5-3, view it in Praat and enable the spectrogram and formant analyses. Then zoom to a single word and put the cursor within the vowel around the first 1/3 of its duration in the vicinity of the highest amplitude in the region where the red formant frequencies are quite clear and stable. Then click *Formant* → *Formant listing*. Praat will open a new Info window with the information of the time point of the cursor and the values for the formant frequencies (F1_Hz, F2_Hz,...) determined by Praat. Open an Excel (or other spreadsheet) file and record the F1 and F2 frequencies rounded to whole numbers. Repeat this process for all 10 words that you have recorded producing data similar to Table 5.1.

Of course, the numbers extracted from your recordings will be different from those in the table partly due to plain anatomical differences in the vocal tracts and partly due to the difference in how you and I produce these vowels.

For the following steps in the analysis I will use Excel. If not available on your platform use another software of your choice. Copy the table into an Excel sheet, select the entire area of the table and ask Excel to produce a simple grouped bar graph that should be similar to the left panel of Fig. 5.3.

There are several observations we can make. First, the formant frequencies differentiate the vowels relatively well. If two vowels are similar in one formant, they are differentiated in the other formant. For example, comparing [i] of ‘he’ with [u] of ‘who’ we see that the first formant (F1) values are quite close but the second formant (F2) values are clearly different. Similarly, if F2 values are relatively close, as in [u] of ‘who’ and [a] of ‘hop’, the F1 values are different. Hence, calling *formants the acoustic signatures of vowels* is justified.

Second, the potential of formants to differentiate vowels is fine, but we still do not know if the *acoustic* formant values have a *systematic* relationship to the *articulatory* characteristics in terms of the horizontal and vertical tongue positions examined in Activities 5-1 and 5-2. To explore this relationship, the smaller graphs in the right-hand panel of Fig. 5.3 illustrate the same data but now ordered according to the first formant values (top) and the second formant values (bottom). Activity 5-5 guides your exploration before reading the main text.

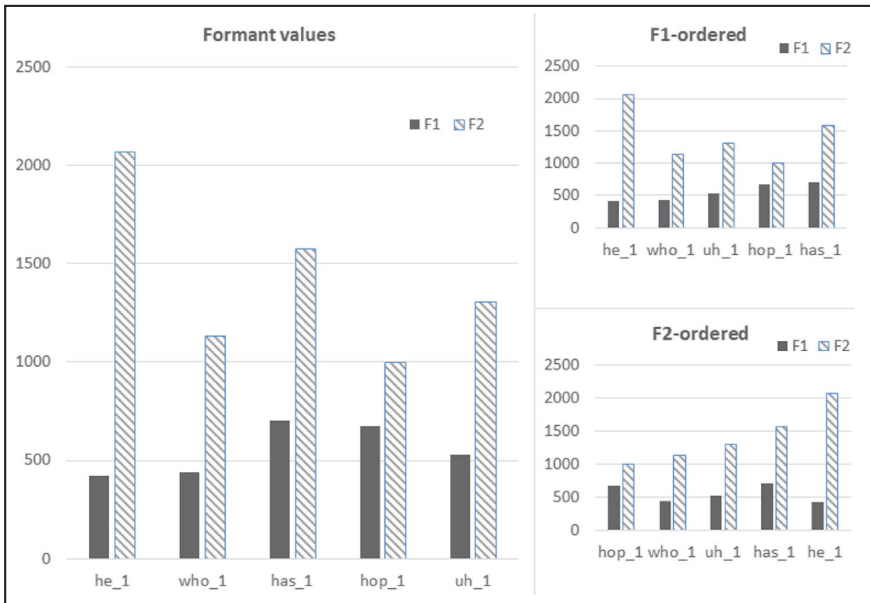


Fig. 5.3 Bar graphs created from formant values in Table 5.1

Activity 5-5: How do F1 and F2 relate to the tongue position?

Take a minute and compare the smaller graphs in Fig. 5.3 with the schematic illustrations of the tongue positions in Fig. 5.1. First take the top graph and F1 values. Note how they raise from left to right. Does this ordering of vowels produce any pattern in one of the two dimensions for the tongue position? Now take the bottom graph of Fig. 5.3 and look at the systematically rising values of F2 from left to right. Is there a pattern in the horizontal or vertical position of the tongue in vowels corresponding to such ordering?

After looking for the patterns (the entire study of language is very often a puzzle-solving activity that involves looking for patterns), compare with the commentary in the text below.

Comparing Figs. 5.1 and 5.3 we can see that the tongue is highest in [i] and [u], medium in a mid-central schwa [ə] and lowest in [æ] and [ɑ]. F1 values ordered in this way show a clear rising tendency. Hence, F1 is inversely related to the vertical position of the tongue: the higher the tongue is, the lower the F1 value. Additionally, F2 is related to the horizontal tongue position: the more forward the tongue is, the higher the F2 value. The tongue is most forward for [i] a bit less for [æ], central for schwa and most backward for [u] and [ɑ]. The F2 values in the bottom right graph follow this ordering. Hence, both articulatory and acoustic observations support the two-dimensional model of the vowel space and the relevance of generalizing the description of vowels.

With this understanding, we use Excel to graph the extracted formant values in a way that visualizes the correspondence of these values to the vowel space in Activity 5-6.

Activity 5-6: Find where your vowels are, Part I

With the values extracted from your vowels like in Table 5.1 in an excel sheet, click on any empty cell and insert a scatterplot graph. Excel creates an empty graph that we will fill in the following way. Right-click on the empty graph and select *Data...* and then *Add* a row in the left box. Excel opens a new window asking for the label, x-value and y-value of the data point. You enter the vowel (e.g. [i] for ‘he_1’), F2 and F1 value, respectively. **Important: make sure F2 is selected for the x-value and F1 for the y-value.** After clicking ok, Excel creates the corresponding data point in the graph.

Repeat this process for all 5 vowels in Table 5.1. Finally, we need to move the origin of the axes from the traditional bottom left to the top-right corner. We do this by right-clicking on the x-axis and ticking the box *Values in reverse order*. Repeat the same for the y-axis. You might want to add the axis names, adjust the minimum values, or add legend labels with the data points. Figure 5.4 shows one way how this vowel chart could be done. There are also other links to suggestions how the extracted formant values can be graphed in Find Out More 5-1.

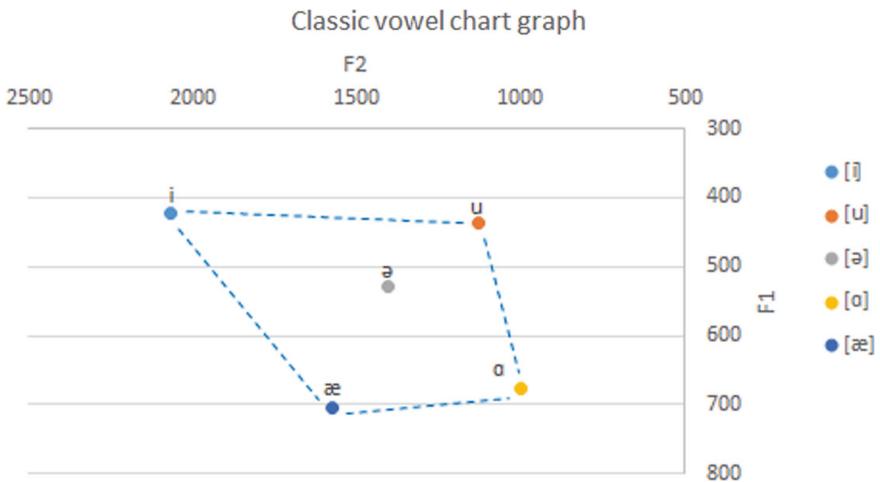


Fig. 5.4 Scatterplot showing the vowel space created from formant values in Table 5.1. See Activity 5-6 for guidelines

Find Out More 5-1

There are several online resources helping with creating a vowel chart in Excel:

- <https://colangpraat.wordpress.com/part-5-3-using-formants-to-plot-vowels/>

If you wish to plot formants from several speakers, this site from the University of Oregon (Thomas and Kendall 2007) is useful since it provides templates for data entry, plotting and, importantly, various types of normalizations, so that the physiological difference among speakers might be filtered out.

- http://lingtools.uoregon.edu/norm/about_norm1.php

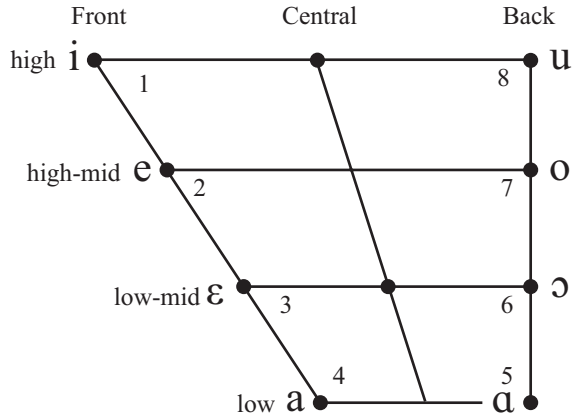
Finally, the widely used and free statistical software R has also multiple tools and packages for drawing vowel charts. These sites provide nice examples:

- <https://guilhermegarcia.github.io/vowels.html>
- <http://drammock.github.io/phonR/>

In Fig. 5.4 I have linked the four vowels around the mid-central schwa with the dotted lines. The resulting shape is similar to the vowel space we have obtained from the articulatory descriptions of the vowels in ‘he’, ‘who’, ‘has’ and ‘hop’. The extracted formant data show that ‘hard’ acoustic data might *correspond well* with the introspected articulatory tongue positions and the simplified 2D model of vowel space proposed in Fig. 5.1. Importantly, due to this correspondence, the visualization of the acoustic data, which is made possible with Praat, can tell us a lot about the articulatory characteristics. Investigating your own speech with Praat thus complements the auditory and kinaesthetic information you have from your speech.

There are two important things to keep in mind. First, the quality of vowels is a complex notion in which horizontal and vertical tongue positions, as well as other important factors, play a role; for example, lip position, vowel duration, pitch and others. Additionally, the third and other formants also participate in the acoustic characteristics of vowels and some consonants. Hence, the 2D model of vowel production is truly a gross but useful simplification. Second, in addition to formant values, ‘hard’ articulatory data regarding the positions of active articulators can also be collected and extracted in a very rigorous way, for example, by using electromagnetometry, mentioned in Chapter 2, or ultrasound and MRI. Unfortunately, these techniques are not readily available to typical students of speech and thus the acoustic characterization and Praat will be the working horse of the scientific exploration of speech in this book.

Fig. 5.5 Primary cardinal vowels in a quadrilateral. Note that ‘high’ and ‘low’ are commonly also described as ‘close’ and ‘open’



We will close the section with *cardinal vowels* as a useful notion related to the vowel space; read Find-Out-More 5-2 below for more information. The concept of primary Cardinal vowels helps with the introspective definition of the vowel space and attempts to delimit the boundaries of the vowel space. Additionally, it facilitates the description of vowels occurring in natural languages. The concept is rather intuitive. Produce an [i] vowel in such a way that your tongue is as front and as high as possible, and your lips are as spread as possible. You need to find the position at which the resulting sound still passes as a vowel but beyond which, i.e. if you move your tongue even a bit higher or more front, or spread your lips a bit more, you would hear a hissing noise and the sound would remind you more of a consonant than a vowel. Experiment with feeling this boundary between a vowel and turbulent noise. The result is the Cardinal vowel #1. You try to do the same but this time with the tongue as low and as front as possible and the lips spread and the result is Cardinal vowel #4. Now explore the vertical dimension between these two extremes and identify Cardinal vowels #2 and #3 in such a way that your tongue is still as front as possible and the four vowels are roughly the same distance from each other in the vertical dimension. Repeat the same with back vowels: Find the most elevated and most retracted tongue position with rounded lips that still sounds as a vowel and you have [u] as Cardinal vowel #8. Then identify the lowest retracted vowel with the neutral lip configuration, [ɑ] as Cardinal vowel #5 and the remaining two rounded vowels # 6,7 in equidistant steps between #5 and #8. Figure 5.5 provides a visualization.

In addition to delimiting the space of possible vowels by identifying the most extreme vowel production, cardinal vowels facilitate the description of naturally occurring vowels. For example, we might say that the English [æ] of ‘has’ explored in this section is close to cardinal vowel #4 but it is slightly higher, with the lips less tense as in the aforementioned vowel. In a similar way we might describe the other three vowels of ‘who’, ‘he’ and ‘hop’. That brings us to the discussion of specifically English vowels in the next sections. We start with *monophthongs* as vowels with more or less stable quality and finish with *diphthongs* that contain salient change of the quality between the start and the end.

Find Out More 5-2: Cardinal vowels

The system of reference vowels was invented by the English phonetician Daniel Jones to delimit the vowel space inside the mouth. Catford (1988) also stresses the importance of muscle strain and even tiredness when producing these vowels in describing their peripheral locations and tense character. The notion of equidistance used in the text should also be taken to mean primarily the articulatory dimension but several authors also include the auditory dimension.

Although the tongue positions are crucial in defining cardinal vowels, the lips also play an important role. CVs # 1–5 have a spread or unrounded lip configuration while the remaining three vowels (CVs # 6–8) have rounded lips. In addition to the eight primary CVs with this lip configuration, there are 10 secondary CVs; 8 have the opposite lip rounding with respect to primary vowels and two more vowels are high and central. For example, CV #9 [y] has identical front and high tongue position but the lips are rounded instead of spread as in CV #1.

There is one notational confusion that should be mentioned. The symbol [a] refers to the front low vowel CV #4 in the context of cardinal vowels. Hence, its quality of low front vowel is somewhat similar to the English [æ] of ‘has’. However, in the context of English vowels, many authors and resources use the symbol [a] for a low central vowel; for example, the first part of the diphthong [aɪ] in pronouncing the personal pronoun ‘I’.

Interested students are encouraged to imitate the production of these vowels using the model pronunciations available at the link below. It is important to concentrate not only on the quality of the vowels but also on their steady monophthongal nature.

<http://www.phonetics.ucla.edu/course/chapter9/cardinal/cardinal.html>

5.3 English Front Vowels

The *front high* vowel with spread lips can be found in a stressed enunciated production of the pronoun ‘he’ or the word ‘heed’ exemplified in Fig. 5.6. In describing English vowels we will use the *Wells’ lexical sets* (see the Find Out More box below) and include them also in the summary tables for front, back, central vowels and the diphthongs. For the front high [i:] it is the FLEECE set: {‘creep’, ‘speak’, ‘leave’, ‘feel’, ‘key’, ‘people’}.

Find Out More 5-3: Wells’ lexical sets

Given great dialectal differences and commonly confusing transcriptions of vowel quality when describing these differences, John C. Wells introduced 24 lexical sets such that each set provides multiple examples of a single vowel quality in a stressed position. Similarly to the current book, the basic distinction was between RP and GenAm. Each set was given a name after a representative

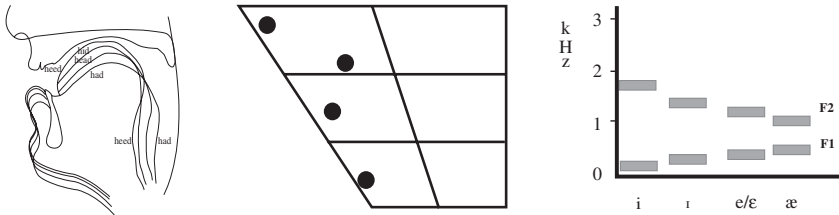


Fig. 5.6 Front English vowels. Schematic articulatory position of the tongue on the left, vowel quadrilateral with dots indicating horizontal and vertical position within the vowel space, and the drawing of approximate formant frequencies F1 and F2

keyword that denotes the set. For example, the DRESS set includes {step, neck, edge, shelf, friend, ready} and denotes the vowel [e] and the LOT set identifies the difference between British [lɒt] and American [lɑt] in words like {‘stop’, ‘sock’, ‘dodge’, ‘romp’, ‘possible’, ‘quality’}. The sets are commonly referred to with the keywords in the capital letters.

Especially useful is the selection of the keywords to prevent mixing them with other words in dialectal realizations, and its consonantal context providing a clear auditory experience despite sometimes low frequency of a keyword.

The articulatory description of vowels is most natural when we do not talk about *absolute* positions of the articulators but *relative* differences between (pairs of) vowels. This approach was already mentioned when discussing Cardinal vowels and how the vowels of natural languages can be described in relation to these cardinal vowels. Following this approach, the English [i:] is very close in quality and articulatory positions to Cardinal vowel #1 but the tongue is slightly less advanced and lower with lips slightly less tensely spread.

Continuing with this approach of identifying the contrast and relative differences between sounds let’s consider the contrast between [i:] and [ɪ], vowels in ‘heed’ and ‘hid’ or the FLEECE and KIT lexical sets, respectively, in Activity 5-8. In the speech of many L2 speakers the contrast between these two vowels is either non-existing or very weak. If you do not feel the contrast, perform Activity 5-8 with a speaker who produces the contrast. See also the note for non-native speakers below.

Activity 5-8: Introspect the contrast between ‘heed’ and ‘hid’

Alternate ‘heed’ and ‘hid’ in front of the mirror, or with a finger on the lips. What is happening to your lips? Consider your tongue using the experience gained with Activities 5-1 and 5-2 and concentrate on the contact between the sides of the tongue and the molars, and the position within the vowel space. How does your tongue participate in producing this contrast? Finally, put your

index finger on the soft area below your jaw when alternating [ɪ] and [i:]. Do you observe any systematic variation?

Activity 5-8 above likely pointed to differences in multiple dimensions. The tongue is slightly more retracted and slightly lower in [ɪ] than in [i:]. Additionally, the lips are much less spread and may be even quite neutral for many native speakers for [ɪ]. The articulatory difference that is not clearly obvious is also the position of the tongue root that is more advanced for [i:] than for [ɪ] due to the musculature of the tongue in elevating and fronting the tongue.

Figure 5.6 provides a three-way image of basic articulatory differences among front vowels: the schematic illustration of the tongue position within the vocal tract (left), the schematized vowel space (centre) and the differences among the first and second formants corresponding to the vertical and horizontal position of the tongue (right). For now, inspect the differences between [i] and [ɪ] and how these schematized illustrations correspond to your auditory and kinaesthetic impressions when you say these vowels. We will cover the acoustics and articulation of all front vowels in Activities 5-10 to 5-12 below.

Activity 5-8 focused on the articulatory differences between [i:] and [ɪ] in terms of the position of the articulators. However, the duration is an important feature of vowels that we have not discussed yet. Set a hypothesis – whether the vowel in ‘heed’ or ‘hid’ is longer – and test it with Praat in Activity 5-9.

Activity 5-9: Measuring the duration of vowels in ‘heed’ and ‘hid’

Record how you say ‘heed’ twice and then ‘hid’ twice. Aim at producing a list containing these four elements in roughly the same way. Employ the now-familiar steps and precautions to assure a high-quality recording. Open in Praat and identify the duration of the four vowels as accurately as possible. Zoom in to clearly identify the boundary between the friction noise of [h] and the high amplitude and clear formant structure of the vowel as well as the end of the vowel and the closure for [d] appearing in the same view. Now drag the mouse from the beginning to the end of the vowel and log the duration of the interval showing in the top and bottom bars of this interval as shown with the ellipses in Fig. 5.7 rounded to three decimal places. Apply the same steps in measuring the duration of all four vowels *consistently*.

Did you observe a robust difference between [i:] and [ɪ] vowels? Does it support or go against your hypothesis? What about the difference between the vowels in the two tokens of the same vowel? It is quite likely that you measured differences even for the same vowels. Try to brainstorm the nature of these differences.

You likely found out that the [i:] in ‘heed’ is robustly longer than [ɪ] in ‘hid’. The colon is the IPA symbol for long sounds. The differences in the duration of the two tokens of identical vowels were almost certainly observed as well and might have stemmed, for example, from the position in the list or a measuring error.

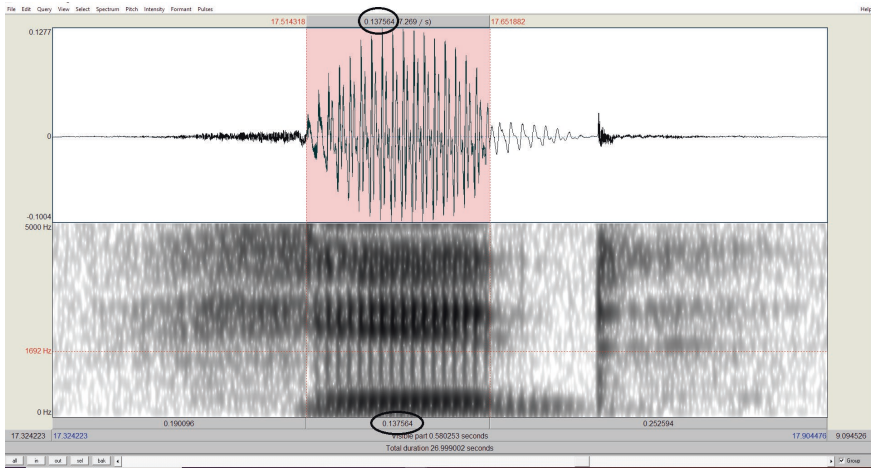


Fig. 5.7 Illustration of measuring duration in Praat

When alternating ‘heed’ and ‘hid’ you should also feel much more muscle tension in the former than the latter since the more peripheral position and longer duration requires greater exertion of muscles. This is one of the reasons why English vowels are commonly described as either *tense* or *lax*, with [i:] being a prototypical example of the tense vowels and [ɪ] of the lax vowels.

Additionally, the more peripheral [i:] is also longer than [ɪ], which underlines the tenseness of [i:] and laxness of [ɪ]. Native speakers should try to produce a long [ɪ] of ‘hid’ and they might find it quite odd. This awkwardness is precisely the evidence of how the habits of native speech are thoroughly complex and ingrained in our behaviour such that consciously trying to change only a single parameter in the articulatory orchestration involved in producing [ɪ] renders a clearly odd result, both in an articulatory and auditory sense.

► **Advanced Section: Note for non-native speakers**

Non-native speakers, depending on their native language, might also find it revealing that these two vowels are so different in terms of quality rather than just duration. Many languages treat vowel duration as a meaningful contrast that can differentiate different words (minimal pairs) like in Slovak, Finnish or others. Speakers of these languages have a tendency to produce [i:] and [ɪ] with a similar quality (i.e. similar position of the tongue and the lips) cuing the contrast primarily through duration. For these speakers the exercise with making the long [ɪ] vowel is also very useful since it makes them aware of the need to separate the vowel quality of the two vowels.

As we progress in our exploration of speech, we will see that duration of sounds is affected by many different contextual influences. It is good to remember that some sounds are, or are expected to be, longer or shorter than other sounds even in very similar or identical contexts.

Activity 5-10: Front vowels in ‘head’ and ‘had’

Look at Fig. 5.6 again and check how the visual representation of the two vowels in ‘head’ and ‘had’ in the figure corresponds to your proprioceptive and auditory sensations with these two vowels.

The remaining two English vowels from the front vowel series in Fig. 5.6 are [e] and [æ]. Starting with [e], this vowel is considered mid in terms of tongue height, lax and produced with a neutral lip position. In addition to ‘head’, it is found in the DRESS lexical set in words like {‘step’, ‘neck’, ‘edge’, ‘shelf’, ‘friend’, ‘ready’} and is considered a short vowel.¹

The vowel [æ] is the lowest of the front vowels, but not as low as Cardinal vowel #4. The lips are commonly characterized as unrounded, but most native speakers feel them slightly spread, which, accompanied by jaw lowering, gives the feel of tenseness in the lips. In addition to ‘has’ or ‘had’ in hyper-articulated stressed pronunciation, this vowel appears in the TRAP lexical set of both British and American pronunciation, and the BATH lexical set of just American variety. The duration is typically longer than [e] but not as long as [i:]. As with all longer English vowels, many native speakers have a distinctive diphthongal character; i.e. the articulators move during the production of this vowel. This is different from speakers of many other languages, e.g. Italian, Spanish or Hungarian where vowels are rather steady and monophthongal even in cases of their prolonged production. Despite this tendency we will consider English vowels [i:] or [æ] as monophthongs.

Whereas [e] occurs in most languages, in some with distinct tongue height variation like in French or Hungarian, the vowel [æ] is commonly missing in many languages and thus non-native speakers of English should pay careful attention to the production of this vowel. The most common interference comes through replacing [æ] either with [e] or with an [ɑ]/[a] vowel, which might also be found in some English dialects, like Dublin or California, for example. Activity 5-11 reinforces the awareness of the continuity and discreteness in vowel realization.

Activity 5-11: Perception-production of the English front vowel series

It is important to realize the essential continuum in vowel production. Try to imagine [i:] and [æ] as endpoints of a line segment. Take a deep breath, start with [i:] and continually and gradually lower your tongue, keeping it in the front position all the way down to [æ]. Do the same in the opposite direction. Now include the medium steps of [ɪ] and [e], gliding smoothly along the [i:] – [ɪ] – [e] – [æ] series.

¹Some sources differentiate slightly higher (close-mid) position in British English and slightly lower (open-mid) in American English accompanied with the IPA symbol [ɛ]. This may also be related to the use of [e] for diphthong [eɪ] in American English (FACE lexical set), the first part of which is clearly higher than the sound of the DRESS set. But for the purposes of this book, we will use the symbol [e] for the DRESS set in both varieties.

Despite this continuum, the individual points represented by the four vowels are different and contrastive and there should be a clear auditory change when moving between any two adjacent pairs of vowels. This activity is also relevant for the non-native speakers who need to establish a habit for four different, and auditorily distinguishable, configurations of the tongue and lips in this front vowel series.

Activity 5-11 above shows that producing vowels is somewhat similar to playing musical instruments like the violin or trombone. The musicians of these instruments need to learn where to put their fingers or how far to extend the main slide to achieve the auditorily desired pitch of the resulting sound without any indication of these positions on the instrument itself. Similarly, as speakers of English, we have no indication where to put our tongues or how to set our lips to produce vowels and have to rely on our ears for minute adjustments. Compare this with guitars or trumpets that are made in such a way that the continuum of resulting pitch is already discretized through frets or valves but pitch variability is still possible through other means.

Activity 5-12 explores further the opportunities Praat offers in visualizing vowel production.

Activity 5-12: Find where your vowels are Part II (precondition: Activities 5-4 and 5-6)

Building on your familiarity with Praat, record three series of front vowels; for example, ‘heed’, ‘hid’, ‘head’, ‘had’; ‘deed’, ‘did’, ‘dead’, ‘dad’ and ‘seed’, ‘Sid’, ‘said’, ‘sad’. View in Praat as a waveform and spectrogram, and extract F1 and F2 values following the steps described in Activity 5-4 above and log them into a spreadsheet file with column labels producing the left-hand panel of Fig. 5.8, which are data from my own recording of the three series. Note that I used ‘\ic’ as a label for [i]. Save the file as a tab-delimited text file.

Move back to Praat Objects window and select *Open* → *Read Table from tab-separated file...* and navigate to your saved file. The file opens as a Table object. When highlighted select from the right-hand list *Draw* → *Scatter plot ...* Enter “F2” for *Horizontal column* and 2500–1000 for *Horizontal Range*, “F1” for *Vertical column* and 800–200 for *Vertical Range*, and ‘IPA’ for *Column with marks*. This effectively reverses the axes in a similar way as we did in Excel in Activity 5-6. If some of your values exceed these ranges, adjust accordingly, F2 values might go above 2500, especially for females. You should see a plot in which the three tokens of each vowel are more or less clustered and the position in the chart resembles the middle panel of Fig. 5.6 with [i:] vowels upper left and the other vowels in the series gradually going lower and to the right.

Finally, highlight the same Table object in the Praat object window and now select *Draw* → *Draw ellipses...* and make sure all values are entered

	A	B	C	D
1	IPA	series	F1	F2
2	i:	h_d	306.3	2408
3	ɪ	h_d	385.1	2031.3
4	e	h_d	558.9	1682.7
5	æ	h_d	650.8	1702.1
6	i:	d_d	292.5	2328.8
7	ɪ	d_d	370.3	1795.7
8	e	d_d	522.2	1687.3
9	æ	d_d	629.3	1718.8
10	i:	s_d	279.5	2394.6
11	ɪ	s_d	374.3	1730
12	e	s_d	535.2	1615.8
13	æ	s_d	659.9	1634.4

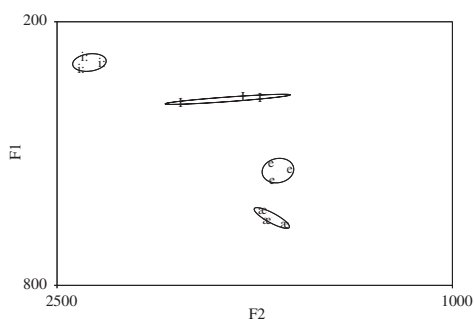


Fig. 5.8 Example of a spreadsheet file with formant values extracted (left) and the corresponding plot done in Praat (right) following the steps outlined in Activity 5-12

identically as in the previous drawing, and select number of sigmas as 1.5 and font size 1. This will draw ellipses, as in the right-hand panel of Fig. 5.8, projecting the region within the vowel space that each vowel occupies with 1.5 standard deviation from the centre of the ellipsis as the small dot.

How do your vowels correspond to the ideal RP-style? If you are a native speaker, do you see specific difference(s) between your dialect and RP? If you are a non-native speaker, can you identify potential traces of the non-native accent and relate them to the front vowels of your native language? For example, for me as a non-native speaker I have a rather wide horizontal spread of [ɪ] and maybe my [æ] could be lower with respect to [e]. Neither [ɪ] nor [æ] are present in my native (Slovak) vowel inventory and it is thus plausible that my articulatory habits in producing them are less consistent than [i:] and [e] that Slovak contains.

Table 5.2 summarizes the characteristics of front monophthongs in the two major varieties of English.

In this book, we are not primarily concerned with the correspondence between English spelling and pronunciation (see Find Out More box 2-3 in Chapter 2). But it is useful to be consciously aware of the major spelling variations for any sound that we cover in Chapters 5 and 6. For front vowels, many indications can be observed in the Wells' lexical sets and some less frequent cases include words like 'any', 'plaid' or 'bury'.

5.4 English Back Vowels

We follow the similar route as with the front vowels of the previous section. Activity 5-13 asks you to introspect your vowels before engaging with the discussion below.

Table 5.2 Summary of the front English monophthongs

IPA	Tongue	Lips	Duration	Lexical set	
				BE	AE
[i:]	high (close), front, tense	spread	long	FLEECE: creep, speak, leave, feel, key, people	
[ɪ]	high (close), front, lax	neutral	short	KIT: ship, sick, bridge, milk, myth, busy	
[e]	mid, front	neutral	short	DRESS: step, neck, edge, shelf, friend, ready	
[æ]	low (open), front	slightly spread	medium	TRAP: tap, back, badge, scalp, hand, cancel	
				BATH: staff, brass, ask, dance, sample, calf	

Activity 5-13: Back vowel series introspection [u:]-[ʊ]-[ɔ:]-[ɒ]/[ɑ] in ‘who’d’, ‘hood’, ‘hawed’ and ‘hod’

Employ the now-familiar introspective activities for four vowels in ‘who’d’, ‘hood’, ‘hawed’ and ‘hod’. Observe/introspect the activity of the lips, the tongue and the jaw, trying to separate these actions from the rest of the articulatory plan. Try then to informally describe them on the three dimensions (horizontal and vertical position of the tongue and the lip position), considering also their duration and tenseness.

We start with [u:] which is the high back long vowel with rounded lips occurring in the words belonging to the GOOSE lexical set {‘loop’, ‘shoot’, ‘tomb’, ‘mute’, ‘huge’, ‘view’}. It is important to mention that in many present-day dialects of English this vowel is relatively fronted and only weakly rounded. As a result, English [u:] maybe considerably different from Cardinal vowel #8 since the latter has a fully retracted tongue and fully rounded lips.

When discussing the vowel space we have already explored the contrast between [i:] and [u:] in terms of both the lip configuration and the horizontal tongue position. The [i:] – [ɪ] pair of the tense and lax front vowels has a corresponding counterpart with the tense [u:] and lax [ʊ]. The English [ʊ] is completely unrounded with a more centralized tongue position (i.e. lower and less retracted) and less rounding when compared to the traditionally more peripheral and rounded [u:]. Moreover, [ʊ] is considerably shorter than [u:] as indicated also with the IPA lengthening ‘colon’ mark of the latter and its absence in the former. The lax [ʊ] can be found in the FOOT lexical set of words like {‘put’, ‘bush’, ‘full’, ‘good’, ‘look’ or ‘wolf’} (Fig. 5.9).

The contrast between [u:] and [ʊ] poses similar challenges as that between [i:] and [ɪ]. It is important to dissociate the quality of the vowels from their duration.

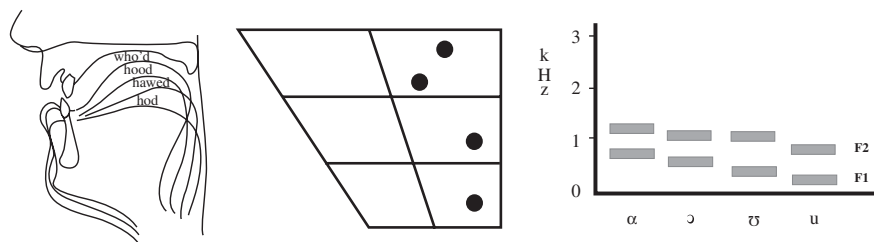


Fig. 5.9 Back English vowels. Schematic articulatory position of the tongue on the left, vowel quadrilateral with dots indicating horizontal and vertical position within the vowel space, and drawing of approximate formant frequencies F1 and F2

A good approach is to exaggerate and lengthen [ɔ], for example, in ‘good!’ as in praising someone for an extremely good job done. Then compare the quality of this lengthened [ɔ:] with the quality of [u:] in ‘goose’.

The mid and low back vowels display several major differences between British and American dialects (BE and AE). First, the long rounded vowel [ɔ:] is mid in terms of the tongue height, quite retracted and rounded in the British variety. It appears in the THOUGHT lexical set: {‘taught’, ‘sauce’, ‘hawk’, ‘jaw’, ‘broad’}. In the American dialects, however, the vowel of this set is a bit lower and with decreased rounding of the lips [ɑ].

The most open back vowels present one of the most salient differences between the two major varieties in the vowel of the LOT lexical set: {‘stop’, ‘sock’, ‘dodge’, ‘romp’, ‘possible’, ‘quality’}. The difference is mostly attributed to the lip configuration: rounded [ɒ] in BE and unrounded [ɑ] in AE. Moreover, the vowel qualities in THOUGHT and LOT sets have merged for many American speakers (the cot-caught merger).

Finally, both BE and AE use the low back unrounded [ɑ:] in the PALM lexical set: {‘psalm’, ‘father’, ‘bra’, ‘spa’, ‘lager’}, and BE uses it also in the BATH set: {‘staff’, ‘brass’, ‘ask’, ‘dance’, ‘sample’, ‘calf’}. Recall from the front vowels discussion that AE uses the front [æ] vowel here.

As a result, the low back rounded [ɒ] is typically not present in the speech of most Americans whereas the [ɑ] quality appears in several types of words: THOUGHT, LOT and PALM lexical sets.

We will not repeat the same activities but you are encouraged to reinforce your command with Praat and your skills and awareness by following the activities with Praat from Sect. 5.3 and replacing the front vowel series with the back vowel series introduced so far. The activities in this section present some novel issues and build on your rapport with Praat. We start with the first caveat relevant to working with Praat: the analyses that Praat show might include errors or artefacts, especially if recordings are in poor quality, come from rapid speech or include non-modal voicing such as creaky voice. Activity 5-14 looks at one example of that with formant values. It is important to try as much as possible to take all the

precautions so that the values we extract and visualize are affected minimally by these errors and artefacts and be always aware of their potential presence.

Activity 5-14: Formant tracking errors in Praat

In the book companion, locate the file ‘good_morning.wav’ containing only initial greetings from a larger interview. Open the file in Praat and click *View & Edit*. Zoom in to the first ‘morning’ token and view the spectrogram and formants. Describe the formant characteristics of the vowel [ɔ:] in the first syllable.

You will notice that formants are not steady. This can be seen in the first 1/3 of the vowel for F1, and around the middle portion of the vowels F1 and F2 seem to ‘cross’. Both issues are indicated in Fig. 5.10 with white ellipses. You notice that placing the cursor at various points during the vowel would result in extracting radically different values for the formants. There are several ways how to deal with such issues.

First, we might try adjusting the settings for the formant analysis. Click *Formant* → *Formant settings...* and adjust the maximum formant value from the default 5500 for female voices to 5000 for male voices. Note how the artefacts in formant tracking changed. It is not perfect, but better than before. Second, to eliminate the effect of spurious values, we might ask Praat to give us the mean or median value for some interval. If you select some portion of the vowel with a mouse and ask for *Formant* → *Formant listing*, Praat returns the values of all formants at all timepoints, and if you just select *Get first formant*, Praat returns a mean value of F1 in the given interval. Getting median, which is even more robust, requires to create a separate Formant object from the

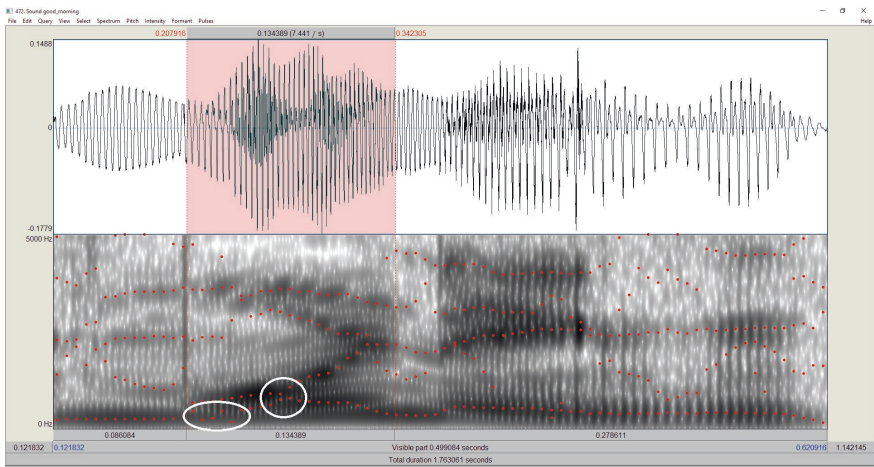


Fig. 5.10 Interval of the first word ‘morning’ in file ‘good_morning.wav’ with the [ɔ:] vowel highlighted and formant tracking problems marked with white ellipses

sound with *Analyse Spectrum* → *To Formant (burg)*... and adjusting maximum formant for the sex of the speaker (5500 Hz predefined for females, 5000 for males). Clicking then the newly created Formant object, we can then query for median by *Query* → *Get quantile* and indicating the formant number, the temporal interval from which we want to extract the median, and specifying the quantile at 0.5, which equals the median. Hence, we have to know the start and end times of the vowel interval we want to investigate.

The last issue that we need to introduce when talking about back vowels, and that will play an important role also for the central vowels and diphthongs in the following two sections, is **rhoticity** or **r-colouring**. This is a feature of those English varieties in which a spelled <r> that is NOT followed by a vowel is retained in pronunciation, for example, in the words ‘start’, ‘car’ or ‘north’. These dialects contrast with **non-rhotic** varieties in which these instances of <r> are not pronounced. In both types of dialects, however, <r> followed by a vowel in pronunciation is always realized, as is the case in words like ‘rude’, ‘raw’ or ‘cross’. Note that <r> is not followed by a vowel in pronunciation, only in spelling in words like ‘core’ or ‘care’.

Note that this is the first time we need to talk specifically about the context surrounding a vowel. As discussed in Chapter 2, the role of context in talking about speech is inevitable and will be an integral part of the discussion for the rest of the book.

Rhoticity is another very salient feature that differentiates many BE and AE dialects. While the southern non-rhotic BE dialects are considered ‘standard’ and some northern BE rhotic dialects ‘non-standard’, the situation in the US is the opposite with major ‘standard’ dialects being rhotic and non-rhotic dialects like that of Boston and New England, are considered ‘non-standard’.²

Rhoticity is very difficult to describe articulatorily since there are multiple ways how r-quality might be achieved even in prevocalic positions (see next chapter on consonants). Some are familiar with a trill-character of ‘r’ in Scottish English, and its ‘softer’ realization, i.e. when the tongue only approaches the alveolar or palatal area rather than when it actually touches it. In addition to many possibilities with the tongue, ‘r’ might be accompanied by some degree of lip rounding or does not have to. Activity 5-15 explores the acoustic characteristics of rhoticity.

Activity 5-15: Acoustic signature of rhotic vowels, r-colouring

Open the same file as in the previous activity ‘good_morning.wav’ and inspect the realization of the word ‘morning’ and rhoticity there. Listen to both tokens from the two speakers, do they have different or the same dialects based on this word?

²Here the words ‘standard’ and ‘non-standard’ are in quotes to highlight the *descriptive* approach of this book (Chapter 2) that should be in no way taken to prescribe some dialects are better or ‘more correct’, etc.

Now, the first speaker is definitely a speaker with a rhotic dialect, she's American; the other speaker is British and has a non-rhotic dialect. Look at the formants during this vowel and describe informally if there is a systematic difference.

The acoustic signature of rhoticity is most commonly observable as F3 lowering. Although not perfect, you should see a salient lowering of F3 at the end of the [ɔ:] of 'morning' in the first speaker, also visible in Fig. 5.10, and a much flatter trajectory of F3 with the same vowel of the second speaker. Hence, measuring F3 for relevant vowels might provide information and test perceptual feeling regarding rhoticity in the speech of English speakers.

There are several ways how rhoticity is transcribed in IPA and none of them are without issues. We will discuss the two most common ways. First, we might indicate the [r], or better [ɹ] (see next chapter on consonants), as a separate symbol; for example, [stɑ:ɹt] for 'start', or even more generally [stɑ:(ɹ)t], indicating that some dialects are rhotic and tend to 'pronounce' <r> while others are non-rhotic in which these instances of <r> tend to be absent. The advantage of this approach is that it could be applied uniformly to all rhotic vowels. The disadvantage, however, is that it gives the impression that the speakers of rhotic dialects produce a non-rhotic vowel first, which is followed by a consonant [ɹ]. But this does not reflect how speakers actually speak. The term r-colouring already indicates that the presence of [ɹ] affects the quality of the vowel very early and there is no 'boundary' between the vowel and [ɹ]. The second way of transcribing rhotic vowels is thus to use a diacritic accompanying the regular vowel symbol: [stɑ:ɹ̥t]; this is especially common for mid-central vowels [ə] and [ɜ] (see the next section). If the transcription is used as a way of showing awareness of the processes involved in producing speech, the second way might be preferable. The first one is acceptable if we do not want to overload students with many symbols and diacritics.

To conclude the discussion of English back vowels we mention the back rhotic vowels. The original Wells' approach includes three lexical sets: NORTH {'for', 'war', 'short', 'scorch', 'born', 'warm'}, FORCE {'four', 'wore', 'sport', 'porch', 'borne', 'story'} and START {'far', 'sharp', 'bark', 'carve', 'farm', 'heart'}. However, for most native speakers, the pronunciation of the first two sets merged and thus the vowel in both sets is the mid-long [ɔ:]. Whether you are native or non-native speaker of English (and speaking a rhotic or non-rhotic variety), you might find it interesting to check, quantitatively in addition to your auditory impression, if this merger applies to your own pronunciation as well. Activity 5-16 guides you to this goal using also the notion of *minimal pair*, that is, a pair of words that are different only in one sound and otherwise are identical.

Activity 5-16: Are NORTH and FORCE vowels merged in your pronunciation?

First select several similar words from the NORTH and FORCE lexical sets. Ideally they should form minimal pairs. For example, you may pair the stressed

versions of ‘for’ and ‘or’ with ‘four’ and ‘oar’, ‘horse’ with ‘hoarse’, ‘born’ with ‘borne’, ‘course’ with ‘coarse’.

Following the skills acquired in the previous activities, record 3–4 of these pairs twice giving you in total 6–8 vowels from each set. Then extract first two formants from each vowel making sure you do the measurements consistently for both members of the pair. Particularly, you should put your cursor in roughly the same distance from the onset of the vowel. Pay attention to possible formant tracking problems since the two formants will probably be quite close to each other; adjust Formant settings following Activity 5-14 if needed. Also notice if you see F3 lowering if your dialect is rhotic or its absence if it is not. Create a table similar to Fig. 5.8 and plot in Praat following the steps in Activity 5-12.

Analyse your results. Do you observe an overlap of the ellipses for the two lexical sets? Does it correspond to your auditory perception of the two vowels? Is there a difference in the variability (the spread) of the two vowels and if so, can it be due to outliers stemming from formant tracking errors? Figure 5.11 shows my pronunciation (adjusted axes to 1500-500 and 700-300 and sigma 2.0) clearly suggesting the presence of the merger in my speech.

Similarly to front vowels, the lexical sets cover most of the major spelling–pronunciation generalizations. Non-native speakers especially should be aware of the overlap between the GOOSE and FOOT sets, particularly for <oo>.

Table 5.3 summarizes the situation with back monophthongs in the two major varieties of English.

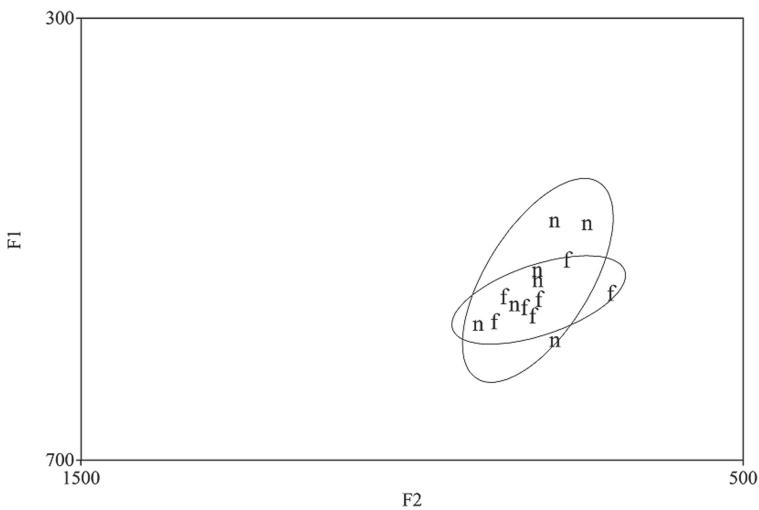


Fig. 5.11 Visualizing NORTH (n) and FORCE (f) vowels in my pronunciation; see Activity 5-16

Table 5.3 Summary of the back English monophthongs

IPA	Tongue	Lips	Duration	Lexical set	
				BE	AE
[u:]	high (close), back (central- ized), tense	(slightly) rounded	long	GOOSE: loop, shoot, tomb, mute, huge, view	
[ʊ]	high (close), centralized, lax	neutral	short	FOOT: put, bush, full, good, look, wolf	
[ɔ:] [ɔ:ɪ]/[ɔ:]	mid, back	rounded	long	NORTH: for, war, short, scorch, born, warm THOUGHT: taught, sauce, hawk, jaw, broad	
[ɒ]	low (open), back	rounded	short	LOT: stop, sock, dodge, romp, possible, quality	
[ɑ] [ɑ:ɪ]/[ɑ:]	low (open) back	unrounded	(variably) long	PALM: psalm, father, bra, spa, lager START: far, sharp, bark, carve, farm, heart BATH: staff, brass, ask, dance, sample, calf THOUGHT: taught, sauce, hawk, jaw, broad LOT: stop, sock, dodge, romp, possible, quality	

5.5 English Central Vowels

There are three remaining English monophthongs to introduce: [ə], [ɜ:] and [ʌ]. All of them are broadly considered central in terms of the horizontal tongue position but each one requires a slightly different discussion. Schwa [ə] is the only English vowel that does not occur in stressed syllables. All words exemplifying vowels so far were monosyllables, and when pronounced in isolation in a dictionary style, each of these words is stressed in English. Hence, no English monosyllabic word pronounced in isolation contains [ə].³ The non-rhotic variant

³There are several monosyllabic function words that can be pronounced with schwa in context, for example ‘for’ mentioned as an example of [ɔ:] is commonly pronounced with [ə] if the word ‘for’ is not stressed or highlighted. The subsequent chapters of the book, particularly Chapter 10, will provide ample examples.

is represented by the vowels in the second syllable of the *COMMA* lexical set like ‘quota’ or ‘vodka’ and the more common variant with plausibly rhotic realization at the end of words in the *LETTER* lexical set: {‘paper’, ‘metre’, ‘calendar’, ‘stupor’, ‘martyr’, ‘figure’}. Schwa is said to correspond to a neutral relaxed position of the tongue: central in the horizontal dimension and mid in the vertical one. It is also the vowel commonly present in short hesitation sounds, for example, when speakers search for a correct word, and is transcribed as ‘er’ or ‘uh’.

The long counterpart of schwa [ɜ:] does occur in monosyllabic words but only in the rhotic context as in the *NURSE* lexical set {‘hurt’, ‘lurk’, ‘urge’, ‘burst’, ‘jerk’, ‘term’}. Despite the different symbol from [ə], the quality is very similar for most native speakers. More importantly, the short [ə] is so affected by context that specifying its quality makes sense only in well-defined research questions.

Finally, [ʌ] is lower, i.e. more open, than schwa and can be found in the words of the *STRUT* lexical set {‘cup’, ‘suck’, ‘budge’, ‘pulse’, ‘trunk’, ‘blood’}. The production of this vowel presents another difference between dialects. While [ʌ] is predominantly central in British English, most American English speakers pronounce it more retracted. Moreover, in the British north and Midland dialects this vowel is essentially replaced with [ʊ] (*STRUT-FOOT* merger).

5.6 English Diphthongs

Consider vowel sounds in words like ‘high’, ‘hey’ or ‘how’. Start with the introspection again observing the activity of the active articulators. How do you make these vowel sounds and how would you characterize the differences between the three words? Consider also the duration and loudness in their realization.

All three vowels belong among *diphthongs*, which are vocalic sounds that contain a salient transition in vowel quality between two articulatory targets within one syllable. In other words, a diphthong starts as one vowel quality and finishes as another one. The IPA transcription of diphthongs thus combines two monophthongs as in ‘buy’ [aɪ]. Activity 5-17 guides you to visualizing this transitory nature of diphthongs in Praat.

Activity 5-17: Visualizing diphthongs in Praat

Record the word ‘bide’ (or ‘eye’, ‘high’) in a slow, enunciated way. Save to your computer, open in Praat, and click *View* and *Show formants* to obtain an image similar to the left panel of Fig. 5.12.

Note the most salient difference between the monophthongs of the previous sections and this diphthong: The first two formants show a clear transition from the first steady section, in which F1 and F2 are close to each other in roughly the first third of the vowel duration, to the end where F1 and F2 are quite apart.

The movements of the tongue within the vowel space are clearly indicated by the changes in the formants in time. To get a more tangible impression,

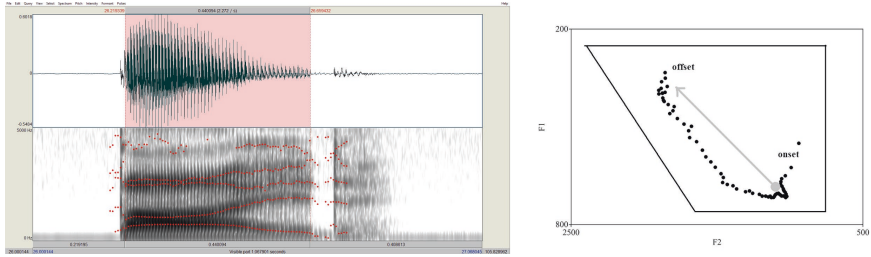


Fig. 5.12 Visualizing the movement of the tongue in ‘bide’ via the transition of formants in Praat. Spectrogram with dotted formants on the left, and the corresponding tracing of the formant movement in the 2D vowel space

the right panel of Fig. 5.12 shows the formant transitions within the familiar $F1 \times F2$ vowel trapezoid. To obtain this picture, drag the selection between the start and the end of the diphthongs with your mouse as in the left panel of Fig. 5.12. Ask Praat for *Formant listing* in the *Formant* menu and the Info window shows a list of values corresponding to the time steps corresponding to the sampling frequency of your recording within the selected window. From these values you can create a tab-separated text file with at least the F1 and F2 columns (either by copying to a spreadsheet program, or saving the Praat Info window into a txt file and then adjusting to make it a tab-separated file with headers, for example, in Excel).

Now open this file in Praat (*Open* → *Read Table from tab-separate file...*) and draw: *Draw* → *Scatter plot (mark)...* entering now familiar required fields for the horizontal (2500-500) and vertical (800-200) ranges and the columns for their values (F2 and F1, respectively) and selecting the marker and its size (dot and 1.5 for the Fig. 5.12).

Before continuing to read the text, try to list other characteristics beyond the transitory nature that you can detect from the visualizations you created.

Both panels support the characterization of the diphthongs as moving between two salient targets. The first third of the vowel’s duration is rather steady indicated with a tightly clustered points in the low back area of the quadrilateral. Then the tongue raises and moves forward, which is indicated also through the spatial distance between individual points (recall that the temporal step between successive points is always the same). When the tongue reaches [ɪ], the formants are again more stationary and stable. This real movement of the articulators is traditionally simplified and depicted with arrows within the schematized vowel space. The grey arrow in Fig. 5.12 shows this schematization and we will use this approach when describing the English diphthongs with the vowel chart. Keep in mind, however, that this production is rather slow, enunciated and out of context while real speech is much more fluid and faster.

Table 5.4 Summary of English diphthongs

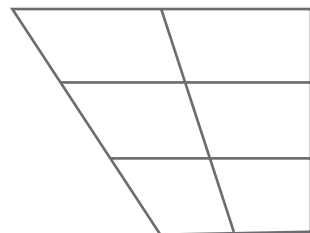
IPA	Tongue	Lips	Direction	Lexical set	
				BE	AE
[eɪ]	Front mid, raising	neutral	closing	FACE: tape, cake, raid, veil, steak, day	
[aɪ]	Low central, raising & fronting	neutral	closing	PRICE: ripe, write, arrive, high, try, buy	
[ɔɪ]	Mid back, raising & fronting	unrounding	closing	CHOICE: adroit, noise, join, toy, royal	
[aʊ]	Low central, raising	slightly rounding	closing	MOUTH: out, house, loud, count, crowd, cow	
[əʊ] [oʊ]	Mid (back) central, raising	slightly rounding	closing	GOAT: soap, joke, home, know, so, roll	
[ɪə] [ɪə̯]/[ɪəɪ]	Front high, centering	neutral	centring	NEAR: beer, sincere, fear, beard, serum	
[eə] [eə̯]/[eəɪ]	Front mid, centering	neutral	centring	SQUARE: care, fair, pear, where, scarce, vary	
[ʊə] [ʊə̯]/[ʊəɪ]	Back high, centering	slightly unrounding	centring	CURE: poor, tourist, pure, plural, jury	

English is typically described as containing 8 diphthongs in these lexical sets: FACE, PRICE, CHOICE, GOAT, MOUTH, NEAR, SQUARE and CURE (see, examples, in Table 5.4).

Activity 5-18: Diphthongs in the vowel space

Following the dot-and-arrow illustration in the right panel of Fig. 5.12, draw arrows representing all eight English diphthongs into the provided vowel trapezoid. Can you determine a systematic way of dividing these eight diphthongs into two categories? In other words, you want two groups and need to identify a unifying feature that characterizes all the members of one group and excludes all the members of another group. Identifying patterns from data is a very useful general skill and pattern recognition is a fundamental approach to studying speech and language (Fig. 5.13).

Fig. 5.13 Empty vowel trapezoid



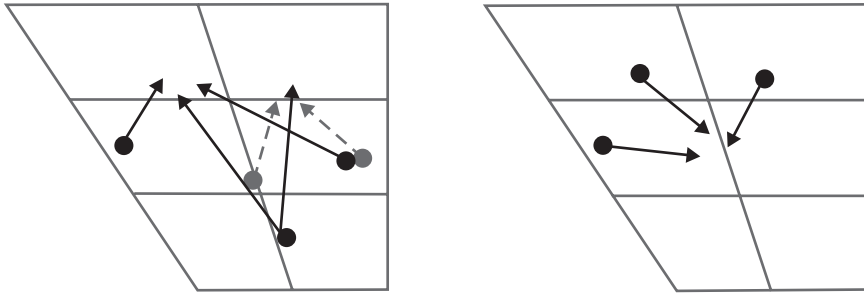


Fig. 5.14 Schematic visualization of English diphthongs. The closing diphthongs on the left, centring on the right (see the text above). The grey tokens refer to the vowel in GOAT lexical set [əʊ] and [oʊ] in British and American pronunciation, respectively

You might notice that the first five diphthongs are non-rhotic while the last three all contain a pre-final <r> in spelling, which yields the difference in pronunciation between the rhotic and non-rhotic varieties. In the non-rhotic ones the tongue raises (*closing diphthongs*), and in the rhotic ones the tongue moves from a more peripheral to a more central position (*centring diphthongs*). The first group can be further divided into those ending in [ɪ] and those ending in [ʊ]. This division is illustrated in Fig. 5.14.

The figure also shows another salient difference between BE and AE in the GOAT lexical set: British speakers tend to start with a central schwa whereas American speakers with a more retracted [o]. In addition to the changes of vowel quality in monophthongs and diphthongs, another salient feature distinguishing various dialects is the degree of monophthongization in diphthongs. For example, in Southern American English [aɪ] of ‘my’ or ‘eye’ is commonly pronounced as prolonged central long [a:]. Similarly, [əʊ] of ‘go’ or ‘know’ in Scottish English is realized as a long [o:].

Returning to the visualization in the left panel of Fig. 5.12 you may have noticed the change in the amplitude. Recall, this corresponds to the y-values in the waveform (top of the left panel) and signals the loudness of the sound. Do you see how the diphthong starts with a high amplitude and gradually becomes less loud? This is a typical feature of English diphthongs that their first part is louder and more prominent than their second part. English diphthongs are thus described as *falling*, which refers to the loudness and prominence fading between the first and the second element. Many languages contain exclusively or partially also *rising* diphthongs such as ‘bueno’ in Spanish.

Finally, we discuss symbols used for standard IPA transcription of English diphthongs. First, the initial vowel quality for PRICE and MOUTH lexical sets is commonly [a] in IPA rather than [ɑ]. This is to indicate that this initial target is not as back as [ɑ] but much more central and close to the low central IPA vowel [ɐ]. Replacing the ‘upside-down’ with regular [a] is simpler but unfortunately creates slight confusion with cardinal vowel #4 [a] that is a front vowel. Second, the first

target of the GOAT lexical set in American English is transcribed as [o] that replaces the [ɔ] as the quality of the English monophthong. Third, in the rhotic dialects the transcription of the r-colouring of the centring diphthongs in the NEAR, SQUARE and CURE lexical sets is also not uniform. Some use the first part of the diphthongs plus [r] or [ɹ], some use the schwa either with the IPA rhotic diacritic [ə̃] or the full [r] or [ɹ] symbol, either replacing the schwa or following it; e.g. [ɪɹ], [ɪə̃] or [ɪɹɪ] for ‘ear’. With the back monophthongs, I suggested using either [ɪ] or the diacritic, and it’s good to be consistent with whatever choice you have adopted there: [ɪə̃], [eə̃], [ʊə̃] or [ɪəɹ], [eəɹ], [ʊəɹ].

Exercises

- 5-1 Measure the duration of [æ] in the recordings from Activity 5-12 and compare with measured durations of [i:] on the one hand and [ɪ] and [e] on the other hand. Check whether the characterization of ‘medium’ duration is supported in your data.
- 5-2 We discussed the variability in low back vowels [ɑ] and [ɒ] such as cot-caught merger in American English. Using the concepts from NORTH-FORCE merger in Activity 5-16 check if the vowel in the THOUGHT lexical set is the same as the NORTH set for the non-rhotic British dialects and/or with the LOT set in the American variety.
- 5-3 The contrast between [ʌ] and [ɑ] in terms of vowel quality, not duration, is problematic for many non-native speakers. Imagine you are a teacher in a class of non-native English speakers and want to use Praat for 1) determining quantitatively which students have a problem differentiating these two vowels, and 2) showing the production of this contrast in students’ speech visually. Think of steps in which these tasks might be achieved using the concepts and skills build in this chapter.
- 5-4 Explore the vowel quality in monophthongs versus the first/second part of the diphthongs. For example, compare ‘blah’ vs. ‘plow’ in Praat, or ‘aha’, a sound when you discover something, vs. ‘au’, when you signal pain. Discuss also how you select a fixed temporal point from which to extract formants and ways of limiting the arbitrariness of this choice.
- 5-5 Create C_C lists of English words with all 20 vowels between identical consonants (like ‘h_d’ words in this chapter). There are few consonantal environments that allow for all, and with others you should search for as close environments as possible. With that list, record a single speaker production of these words, extract formant values and create a vowel chart for that speaker.
- 5-6 Compare English vowel production for two speakers, either a native and a non-native speaker, or two native speakers of different dialects of English. Use the vowel charts and compare so that relative positions rather than absolute numbers can be examined.

References

- Carley, Paul, Inger M. Mees, and Beverly Colins. 2018. *English phonetics and pronunciation practice*. London: Routledge.
- Catford, John. 1988. *A practical introduction to phonetics*. Oxford: Clarendon Press.
- Culpeper, Jonathan, Paul Kerswill, Ruth Wodak, Anthony McEnery, and Francis Katamba. 2018. *English language: Description, variation and context*, 2nd ed. London: Palgrave Macmillan.
- Ladefoged, Peter. 2001. *Vowels and consonants*. Malden, MA and Oxford: Blackwell.
- Thomas, Erik R., and Tyler Kendall. 2007. NORM: The vowel normalization and plotting suite. <http://lingtools.uoregon.edu/norm/norm1.php>. Accessed 13 April 2020.
- Wells, John C. 1982. *Accents of English I: An introduction*. Cambridge and New York: Cambridge University Press.



In this chapter we will

- Explore English consonants building on the awareness of the articulatory and acoustic characteristics in Chapters 3 and 4
- Balance the introspective activities and articulatory descriptions with visualizations afforded by Praat
- continue expanding the command of Praat introducing new functionalities

6.1 Introduction

Similar to the discussion of vowels in the previous chapter, English consonants are presented here in an idealized way, highlighting the articulatory mechanisms in their productions and subsequent acoustic signatures observable in Praat. Here we focus on the contrastive characteristics of consonants and the contextual factors influencing everyday speaking behaviour will be investigated in more detail in the subsequent chapter.

One of the goals of this book is to become consciously aware and able to describe in your own words the feat of saying a simple word. The descriptions of separate segments and the articulatory gestures required for their production are an important prerequisite for discussing how these actions are coordinated and form habitual patterns in speaking of complete words and longer utterances in the subsequent chapters of the book.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_6) contains supplementary material, which is available to authorized users.

Finally, contrary to the chapter on vowels, the production of consonants does not represent major variability in terms of English dialects. Hence, barring occasional mentions of differences, for example with /r/, the material covered in this chapter is similar in both British and American varieties.

6.2 Stricture, Voicing, Place

When delimiting vowel space and discussing cardinal vowels in Sect. 5.2, we explored the effects of slight adjustments in the vertical position of the tongue on the quality of the resulting output sounds. You experienced the thin boundary between a fully resonating vowel sound [i:] and a sound containing clearly turbulent airflow heard as hissing. To refresh this sensation, you might lick your index finger (the last two phalanges) and then place it on your lower lip as if you were gesturing for silence or hushing. Now repeat the exploration from the previous chapter by starting to say the [i:] of ‘he’ and gradually, yet slowly, raise and front the tongue until you hear a change in the quality of sound. This change should be accompanied by a clear sensation of the faster air stream, felt as coldness on your wet finger. You may alternate between the vowel and this turbulent sound to solidify the sensation and awareness.

This articulatory adjustment (primarily in the vertical tongue position) corresponds to the degree of *stricture* in the vocal tract. If you keep raising your tongue even more, at some point the central part of the tongue blade touches the alveolar ridge area and the resulting sound naturally stops and the stricture is thus greater than for the previous turbulent sound. Note that even for [i], the sides of the tongue were already touching your molar teeth and thus the stricture in this case refers to the central, or mid-sagittal, axis of the tongue.

Stricture is an important notion since it allows the most common definition of consonants as sounds in which the degree of stricture (obstruction) is greater than that for the vowels. Additionally, the degree of stricture is one of the features that differentiates consonants among themselves.

In addition to stricture, the second dimension in which vowels contrast from consonants is the activity of the vocal cords. While all vowels are normally *voiced*, which requires the adduction of the vocal cords and the aero-dynamic process described in Sect. 3.3, consonants exhibit salient and contrastive differences in how the activity of the vocal cords relates to the articulatory gestures in the mouth. Recall our discussion of the sounds of snakes and bees, [ssss] and [zzzz], in Sect. 3.3 and the sensation of the vibrating vocal cords felt on your larynx. Informally, we can differentiate consonants to *voiced*, like [z, b, v], and *voiceless*, like [s, p, f]. The visualization of voicing in Praat is demonstrated in Activity 6-1 below.

Activity 6-1: fibre vs. viper

Start with saying the first sounds of ‘fibre’ and ‘viper’ – [ffff] and [vvvv] – very forcefully and clearly to get a similar sensation as in [s]/[z] in snakes/bees

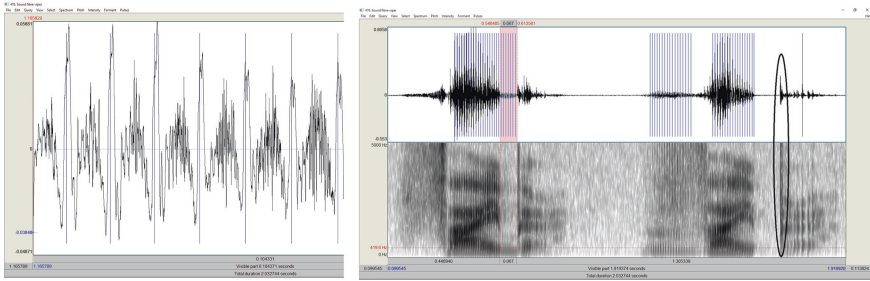


Fig. 6.1 Acoustic characteristics in ‘fiber-viper.wav’ of the book companion. The combined periodic (shown with vertical pulses) and aperiodic components of [v] in the left panel. The shaded closure duration for [b] and the release burst of [p] with the ellipsis in the right panel

above. Then record your production of these two words with prolonged initial consonants and inspect in Praat. If you cannot record yourselves, inspect the recording ‘fibre-viper.wav’ in this chapter of the companion.

First, zoom in to [f] of ‘fibre’ and describe verbally the characteristics of the waveform. Now compare that with [v] of ‘viper’ characterizing the difference.

Now look at another consonant pair in these words: [b] in ‘fibre’ and [p] in ‘viper’. What salient acoustic features, common for both ‘b’ and ‘p’, allow you to identify the beginning and end of these sounds? How do these features link to articulatory movements? Now concentrate on their difference. What features cue the contrast between the two sounds?

Regarding the characteristics of [f], you should have noted the irregularity of the sound wave and thus the aperiodic character typical of these turbulent sounds. The comparison with [v] shows how [v] combines the aperiodicity of the noise component with an observable periodic component in at least a portion of the sound. In Praat, select *Pulses* → *Show Pulses* to help you with the identification of periodic cycles. This sine wave-like component corresponds to the vibration of the vocal cords. The left panel of Fig. 6.1 shows the pulses as vertical lines and illustrates both the periodic and aperiodic quality of the soundwave.

The identification of the [b] and [p] sounds is easy with the rapid decrease of amplitude (energy) corresponding to the rapid increase in the degree of stricture since in both [b] and [p] the lips are completely closed. This closure is illustrated with the shaded (pink) region in the right panel of Fig. 6.1. The opening of the lips is also very clearly observable with a rapid burst of energy indicated in the spectrogram as a salient vertical bar following the low-energy region of the closure. The right panel of Fig. 6.1 shows this burst for [p] with the vertical ellipsis.

But note also the clear differentiating features of [b] and [p]. First, it is the presence of periodicity, and thus vocal cord vibration, during [b] of ‘fibre’ but its absence as more-less straight horizontal line during [p] of ‘viper’ in the waveform. Additionally, some of you might have noted in Activity 6-1 also the difference in

the duration of the closure. In the companion file ‘fibre-viper.wav’, selecting the interval of the closure by dragging the mouse shows the durations of the first closure at about 60 ms while the second one is about 100 ms.

Finally, you may have noticed in Activity 6-1 that [f] requires more articulatory effort, more energy, than [v], which can be felt as the pressure of the lower lip against the upper teeth. Similarly, [p] feels ‘stronger’ than [b]. The consonants that are voiceless and require greater articulatory energy are commonly called *fortis*, from Latin strong, and those that are commonly voiced and require less effort *lenis*, from Latin weak. This is similar to the tense/lax contrast we discussed with vowels. Although the voiced/voiceless labels nicely describe the situation in many languages, in English (and Dutch, German, etc.) the overall energy, or force, in producing these contrasts provides a better general description. The phonetic realization of the fortis-lenis contrast differs in various contexts and across languages and the patterns of English speakers will be discussed more in the next chapter.

Observe also the movements of formants in [ar] familiar with the description of diphthongs in the previous chapter. Interestingly, Praat does not identify pulses, and thus vocal fold vibrations, in the final syllables of both words in this production of ‘fibre’ and ‘viper’. Since all vowels, as we said, are voiced in English, how do you analyse this aspect of the final schwa in this file? Was your recording similar or different in this respect?

After stricture and voicing, the next dimension in describing consonants is the *place of articulation* which is the location of the primary stricture. This notion yields straightforwardly to introspection and self-awareness. Recall the sensations of the lower lips touching the upper teeth in [f] and [v] or the complete lip closure for [b] and [p] in ‘fibre’ and ‘viper’ above. With the complete closure and the clear sensation of the contact between two articulators, we can easily feel the difference between [p] of ‘pea’, [t] of ‘tea’ and [k] of ‘key’.

Activity 6-2 investigates the role that formant frequencies might play in visualizing the location of the stricture in a similar way to how they describe the horizontal and vertical tongue position with vowels.

Activity 6-2: Ghost sounds: seed, side and soap

Record the words ‘seed’, ‘side’ and ‘soap’. Open in Praat and listen to each word separately. Take a moment to strengthen the link between the articulatory sensations, particularly those connected with the formation and release of [s], and the perceptual sensation associated with listening to it.

Now select the interval of the first word and zoom to selection (*Ctrl-N*). Identify the approximate point at which the aperiodic noise of [s] changes to the periodic sound of the vowel and place the cursor at this point. Click the bar to the left of the cursor to hear and verify the voiceless [s]. This boundary is shown with the vertical solid line in Fig. 6.2 and the bar to click with the ellipsis and #1.

The core question of this activity is: What do you think you should hear after clicking on the bar to the right of the cursor shown as #2 in Fig. 6.2? That

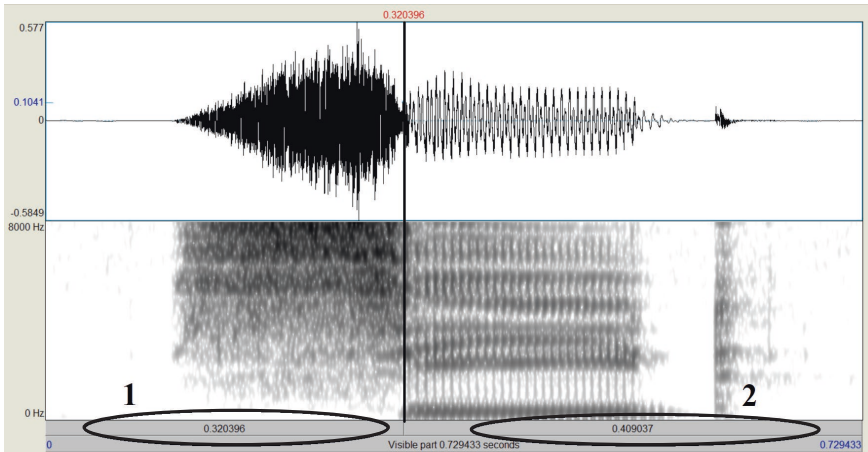


Fig. 6.2 Illustration of the boundary between [s] and the following vowel with the vertical solid line in ‘seed’. Praat bars to click for hearing [s] as #1 and hearing the rest of the word as #2

is, what should be left of the sound without the initial [s]? For ‘seed’ it should be [i:d], right? Now click the right bar. What do you hear? Repeat the same process consistently with ‘side’ and ‘soap’.

You probably hear something extra in addition to the expected [i:d], [aid] and [oʊp]/[əʊp]. If you do, brainstorm about the reasons for this ghost sound before reading the text below.

When I do Activity 6-2, I hear a ghost sound [d] replacing [s] and thus with the cursor at the onset of the vowel I hear ‘deed’, ‘died’ and ‘dope’ ([di:d], [daid] and [doop]/[dəʊp]). So there are two questions you should address:

1. How come there is something there ([d]) that you did not say when pronouncing the words?
2. What is the acoustic signature of this [d]; in other words, can we see it visually in Praat?

Say a prolonged [s] as in the beginning of ‘seed’ and during this [s] raise your tongue to touch it against the roof of the mouth. Note and remember the point of contact. Now say an enunciated [d] as in the beginning of ‘deed’. Compare the point of contact with that from ‘seed’ before. You probably notice they are extremely similar. Hence, the place of articulation of [d] and [s] is similar despite the difference in both stricture and voicing of the two sounds.

Now, consider what movements the tongue has to make to say ‘seed’. It has to create the narrow stricture resulting in turbulent airflow at this place of articulation and then move away (i.e. primarily lower and maybe retract slightly the tongue) for [i:]. In a word like ‘sod’ the tongue would have to lower and retract significantly. In the previous chapter we found out that formant frequencies can reveal a lot in terms of the tongue position (for monophthongs) and tongue movement

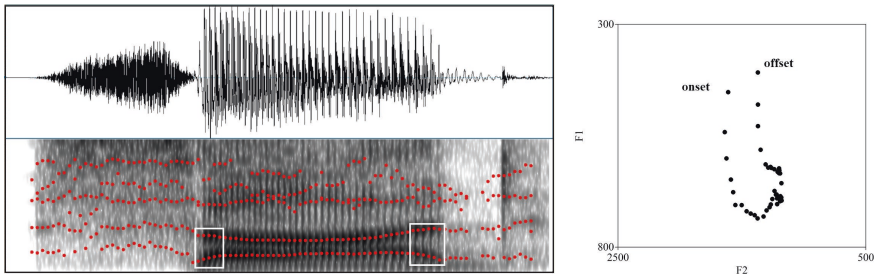


Fig. 6.3 Formant transitions in the production of ‘sod’. The left panel shows the transition with white boxes in the regular Praat visualization with time on the x-axis, the right shows the movement of the tongue in the schematized vowel space familiar from the previous chapter

(for diphthongs) and that tongue lowering can be observed with F1 raising. This rapid movement of the tongue away from the consonantal stricture and towards the steadier vocalic target is characterized by very dynamic *formant transitions* observable 30–50 ms after the release of the consonant or before the forming of the constriction. Figure 6.3 shows the formant transitions in the word ‘sod’. Note that these transitions are faster and shorter than those seen with the diphthongs. Also, the right panel of Fig. 6.3 shows that the transitions are fast at the onset and offset of the vowel whereas in diphthongs the two vowel targets were steady and the transition in the middle of the vowel was fast.

Hence, the information about the place of the stricture is present in the dynamic formant transitions during the short interval at the onset of the vowel. When we placed the cursor at the onset of the vowel in ‘seed’ or ‘soap’ in Activity 6-2, we thus included those transitions. Moreover, by clicking the right bar (#2 in Fig. 6.2), this transition followed after a silence and we know that consonants with a complete stricture have minimal energy during this closure. These two acoustic signatures of the articulatory habits (low energy during the closure and formant transitions at the onset of the vowel) thus combined into giving the perception of the ‘ghost’ [d] (maybe for some non-native speakers of English also a [t]) preceding the vowel in Activity 6-2.

To close this discussion, recall the notion of continuity in speech and overlap in the articulatory activity introduced in Chapter 2. Looking at the formant transitions both at the onset and the offset of the vowels in words examined in Activity 6-2 we see yet another example of this continuity. The formants in Fig. 6.3 show that the consonant ‘leaks’ into the vowel from both sides corroborating the evidence of continuity from the articulatory data in Chapter 2 now in the acoustic data. For example, we see that the consonant begins well before the vowel ends in the transition from the vowel to the consonant.

6.3 English Consonants by Their Manner

Since we now have a good grasp of the stricture, place and voicing of the consonants, we are in good position to provide a systematic description of English consonants based on their manner of articulation.

Stops (Plosives)

Recall the discussion of the words ‘fibre’ and ‘viper’ and particularly the consonantal [b] and [p] in the middle of these words. These are representatives of *stop consonants*, or *plosives*. Stops are produced in two major phases. First, a complete obstruction stops the egressive (outward) airflow. Since the air continues to be expired from the lungs and through the larynx, the air molecules accumulate in the oral cavity behind this stoppage. If the soft palate is raised to prevent the air from leaking through the nose, an area of higher pressure behind the obstruction develops. The second phase is a fast release of the obstruction that rapidly lowers the degree of the stricture. Because of the high pressure before the release, the air molecules rush outside the vocal tract.

Acoustically, the first phase corresponds to the interval of low energy (either complete silence for the fortis stops or possibly low-intensity vibration for the lenis stops) followed by a very distinct burst of acoustic energy (explosion) during the release, also seen in Fig. 6.1. Hence, stops are perceptually very salient sounds due to this contrast between the closure of very low energy and the subsequent burst of very high energy.

In English, there are three major places at which stops are produced: *Bilabial* [p] and [b], *alveolar* [t] and [d], and *velar* [k] and [g]. The two members of each pair differ in how the vocal cords are set – open glottis vs. closed glottis and vibration – and the timing of this action with respect to the stricture formation in the mouth. For now we can refer to this difference as voicing and we discuss the realization of this fortis/lenis contrast in greater detail in the next chapter.

There are no lexical sets commonly used to describe the usage of the consonants, but Table 6.1 summarizes these articulatory descriptions and provides examples of the consonants in words. Figure 6.4 illustrates the three places of articulation of English plosives.

Table 6.1 Summary of English stop consonants

IPA	Manner	Place	Fortis/Lenis	Examples
[p]	Stop (plosive)	Bilabial	Fortis	Pay, appease, pray, plot, pew, speak, spray, splash, police copy, cap, carp, tramp
[b]	Stop (plosive)	Bilabial	Lenis	Bay, above, rabbit, brow, blue, believe, rob, carb, ruby
[t]	Stop (plosive)	Alveolar	Fortis	Ten, attempt, tray, tune, today, steak, street, cat, cart, tent
[d]	Stop (plosive)	Alveolar	Lenis	Dot, idle, draw, duke, delete, wedding, add, card, reading
[k]	Stop (plosive)	Velar	Fortis	Key, car, chaos, accuse, cry, clay, cue, ski, school, screen, collide, pocket, sock, stark, desk, box
[g]	Stop (plosive)	Velar	Lenis	Guy, again, grey, ghost, glue, together, trigger, figure, exaggerate, mug

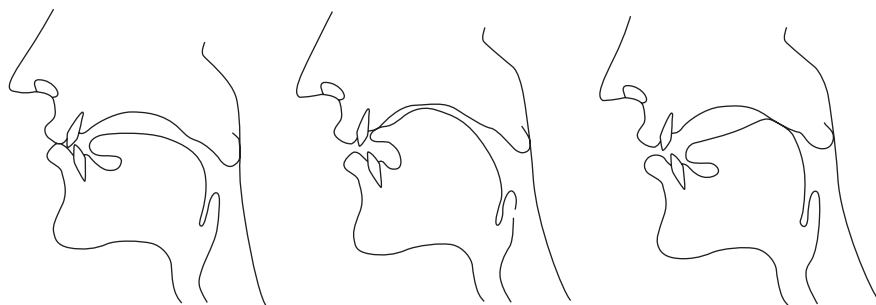


Fig. 6.4 Position of active articulators for three places of articulation for English stop consonants

There is one more stop sound in English but it is not a contrastive sound and depends on many contextual factors. The *glottal stop* [ʔ], already introduced in Chapter 3, is a sound in which the complete obstruction is formed by pressing the vocal cords together and thus tightly closing the glottis. Due to the tight closure of the vocal cords, glottal stops can only be voiceless.

Fricatives

Recall again the discussion of the words ‘fibre’ and ‘viper’, particularly their initial consonants [f] and [v], or the initial consonants in words ‘seed’ or ‘sod’. These are representatives of *fricative* consonants. Fricatives are produced with an incomplete closure but the stricture is very narrow. This narrowing causes the increase in the speed of the airstream. The faster air becomes turbulent, which causes the friction-like noise. It’s similar to water in a creak: as soon as the speed of water increases at some constriction, the water becomes turbulent and noisy.

As illustrated in the discussion of cold air sensation on a licked finger at the start of Sect. 6.2, fricatives are produced in a single phase and thus produce a sustained noise, which contrasts with stops that cannot be sustained temporarily the way fricatives can.

Activity 6-3: Characteristics of English fricatives

There are nine fricative sounds in English. Before peeking at the following text, try to list all of them and describe their voicing and the place of articulation using the adjectives from Table 3.1 and Fig. 3.6 in Chapter 3.

Then, imagine you need to instruct your pet robot, that has all human-like articulatory organs and functionality, to produce these consonants; what detailed instructions do you give?

In English, there are five major places at which fricatives are produced and four of them have pairs of lenis and fortis sounds: *labio-dental* [f] and [v], *(inter-)dental* [θ] and [ð] starting words like ‘thick’ and ‘then’, respectively, *alveolar* [s] and [z], and *post-alveolar* (or palato-alveolar) [ʃ] and [ʒ] starting words like ‘shoe’ and

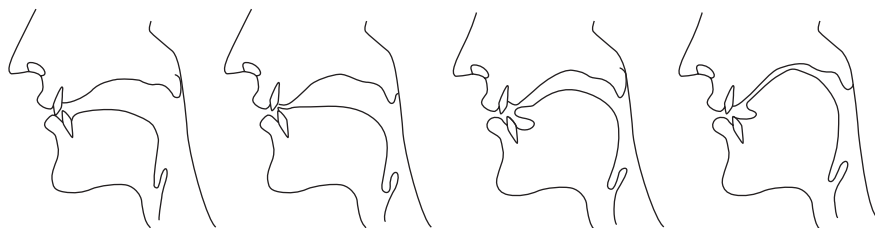


Fig. 6.5 Diagrams of 4 places of articulation for English fricatives

‘genre’, respectively. The reason for classifying [θ] and [ð] as (inter-)dental is that many speakers, mostly of the western-US dialects, place the tip of the tongue in between the upper and lower incisors; hence inter-dental, while other speakers just target the contact between the tongue tip and the edge of the upper incisors; hence simply dental. Where did you instruct your robot to place the tongue for [θ] and [ð] in Activity 6-3?

The fifth place is **glottal** and the narrow stricture is formed by approximating the vocal cords such that the passing air creates friction noise as in loud breathing. In English this sound is normally fortis and voiceless [h], as in ‘he’, ‘who’ or ‘head’ explored in Chapter 5 but in between vowels, e.g. in ‘behind’ or ‘ahead’, the glottal fricative is lenis and voiced for which IPA uses the symbol [ɦ].

Acoustically, the major component of fricatives is aperiodic noise possibly combining with periodicity from vowel cord vibration (Fig. 6.5).

We already know that formant transitions might provide indication of the place of articulation, and periodicity during the stricture of the voicing of consonant. The distribution of acoustic energy across a frequency range in the fricative noise provides an additional indication of the place of articulation. First, there is much more energy in general in so called **sibilants** ([s], [z], [ʃ], [ʒ]) than in the remaining fricatives. This is because the narrow stricture is formed behind the teeth that direct the fast stream of air to hit the front incisors, resulting in a miniature whirlwind that strengthens the energy in the noise. In other English fricatives the noise source is created either directly with the incisors or in case of [h], too far from them, causing the dispersion of the energy.

Second, within the group of sibilants and non-sibilants distinguishable by general loudness (i.e. amplitude height in the waveform and darkness in the spectrogram), the location of energy concentrations in the spectrograms might be helpful. Alveolar [s] has lot of energy in the higher frequency range, typically higher than 5 kHz shown by default in Praat. Note that I set the frequency range to 0–8000 Hz in Fig. 6.2, which is seen on the y-axis. Hence, it is useful to increase the frequency range to at least 7–8 kHz when inspecting the place of fricatives (*Spectrum* → *Spectrogram settings* → *View range*). In comparison to alveolar fricatives, the post-alveolar [ʃ] typically displays the loudest noise around 4 kHz. In the non-sibilant group, labio-dentals tend to have the loudest energy concentrated around the lower region (3–4 kHz) than the dentals (around 8 kHz).

If you observe yourself in the mirror, or place the index finger on your lips, you will notice that [ʃ] has a quite salient rounding and protrusion of the lips compared

to [s]. This adds to the feeling of higher pitch in [s] and lower pitch in [ʃ] despite the absence of the vocal cord vibration. You can try to separate these two actions: keep the tongue position stable and say [s] and continually round the lips. Alternatively, say [ʃ] and unround them, keeping the tongue in a fixed position.

By the way, did you instruct your pet robot in Activity 6-3 to round the lips when saying [ʃ] and [ʒ]? Remember from Chapter 1, you are all masters at saying these sounds, the goal of the book is to become aware, and able to describe, how you make them.

The variability in the production of alveolar sibilants is commonly attributed to many social factors such as biological sex, sexual orientation, age or regional dialect. Praat can be useful in projects in which you want to quantify the realization of these fricatives in certain populations. Activity 6-4 describes a pilot approach using the notion of the *centre of gravity* that evaluates the distribution of acoustic energy in aperiodic noise; informally, it tells us where the darkest region in the spectrogram of the fricative noise is.

Activity 6-4: Distinguishing fricatives in Praat

Praat can provide quantitative data regarding the place of articulation of fricatives, or plosive bursts in a similar fashion that the formants provide this data for vowels. We will use the notion of the centre of gravity.

Record several productions of voiceless sibilants [s], [ʃ] and non-sibilants [f], [θ]. Open in Praat and zoom in to the fricatives you want to measure. Place the cursor in the middle of the fricative and ask Praat for spectral slice: *Spectrum* → *View spectral slice*. Praat creates a new object in the Objects window ‘Spectrum X’ where ‘X’ is the name of the file you recorded. With this object highlighted, click *Query* → *Get centre of gravity...* and click *ok* for the default Power value. In the Info window, Praat gives you the frequency value in Hertz that corresponds to the centre of gravity.

Do you see a systematic difference between [s] and [ʃ] fricatives in terms of your centre of gravity values?

The (inter-)dental fricatives, fortis [θ] and lenis [ð] corresponding to the grapheme <th> in English, deserve special attention for two reasons. First, many languages of the world do not contain these sounds in their inventories and thus for many non-native speakers of English these sounds present a challenge due to the interference from the mother language. Replacing them with the alveolar ([s], [z]) or labio-dental fricatives ([f], [v]) or dental/alveolar stops ([t], [d]) is a common strategy of native speakers of many L2 English speakers. Second, despite the limited dialectal variability of consonants in comparison to vowels among native speakers, the (inter-) dental fricatives present rich and salient differences in many dialects. For example, many speakers of African-American Vernacular English (AAVE) or the Brooklyn dialect in the US, or speakers of Irish or Indian English, tend to pronounce these fricatives as dental stops [t], [d], which is referred to as *th-stopping*. Especially for [ð] this variability is compounded with the sheer frequency of this sound in one of the most common English words like *the*, *they*, *that*, etc. and huge variation in their realization.

We have seen in Activity 6-4 that Praat can provide visual and quantitative information regarding the difference between [θ] and [s], which might be useful for evaluating the pronunciation of non-native speakers. In Activity 6-5 we look at how we can also use Praat for visualizing th-stopping.

Activity 6-5: Check the realizations of [θ] and [ð] with Praat

For comparing the fricative and th-stopping realizations in Praat, first review the acoustic differences between stops and fricatives. What do you expect to see when you record ‘day’ or ‘tick’ for the initial sounds as opposed to ‘they’ and ‘thick’? Now record these four words and inspect in Praat verifying your expectations.

Finally, inspect the file ‘th-stopping_cake.wav’ in the book companion and discuss what Praat visualization can tell us about the dental fricative productions of this speaker. This is a longer clip of 8 seconds, so first identify all cases where you would expect dental fricatives; there are five of them. Zoom in sufficiently to inspect the characteristics of the waveform and the spectrogram.

Probably the first, and I admit the intended, impression from the Activity 6-5 is that real speech is hardly as straightforwardly analysable as clearly enunciated recordings. You probably observed that this speaker’s th-productions are variable and while the first three showed more evidence of the fricative component, the last two sounded, and looked, closer to the stop productions. Due to the frequency of dental fricatives in speech, this is another good idea for projects investigating th-stopping in the speech of both native and non-native speakers. Table 6.2 summarizes the characteristics of English fricatives.

Table 6.2 Summary of English fricative consonants

IPA	Manner	Place	Fortis/Lenis	Examples
[f]	Fricative	Labio-dental	Fortis	Fee, affair, photo, fry, flow, few, sphere, forget, coffee, knife, rough, calf, soft
[v]	Fricative	Labio-dental	Lenis	Vice, avenue, heavy, view, revive, sieve
[θ]	Fricative	Dental	Fortis	Thy, thick, cathedral, sympathy, three, thought, math, death
[ð]	Fricative	Dental	Lenis	Then, though, mother, smoothie, breathe, with
[s]	Fricative	Alveolar	Fortis	See, scent, ascertain, sly, ceiling, source, receive, loss, mix
[z]	Fricative	Alveolar	Lenis	Zoom, zebra, deserve, dizzy, anxiety, scissors, ablaze, jazz, tease, cheese
[ʃ]	Fricative	Post-alveolar	Fortis	Shoe, chef, sure, machine, special nation, anxious, fish
[ʒ]	Fricative	Post-alveolar	Lenis	Genre, measure, vision, prestige
[h]	Fricative	Glottal	Fortis	Help, whole, hug

Affricates

Affricates are found in words like ‘church’ or ‘judge’ and are best described as the combination of the previous two manners: they start as stops (with a complete obstruction for the airflow and raised velum) and end as fricatives (the release is slow resulting in distinct fricative noise). In English, there is only one place of articulation at which affricates are produced: *post-alveolar* fortis [tʃ] and lenis [dʒ]. We see that the manner of affricates is reflected in the IPA symbols that combine the symbols for the stop and the fricative. Note, however, that the IPA symbol is also misleading and the affricates are produced at a single place of articulation, not as a shift from the alveolar [t/d] to the post-alveolar [ʃ/ʒ] position. Figure 6.6 illustrates the two phases of English affricates.

The stop-fricative combination in the manner is also reflected in the acoustic signatures of these consonants. They are characterized by a short interval of greatly reduced energy, either as an almost flat waveform for the fortis one or potentially low-amplitude periodicity stemming from the vocal cords vibration for the lenis one, followed by the aperiodic noise component that is in the case of [dʒ] combined with the periodicity of the vocal cord vibration.

Several characteristics mentioned for post-alveolar sibilants [ʃ] and [ʒ], such as the activity of the lips apply to English affricates as well and their place of articulation might correlate with several social variables and can be analysed with the centre of gravity measurement for the fricative component (Table 6.3).

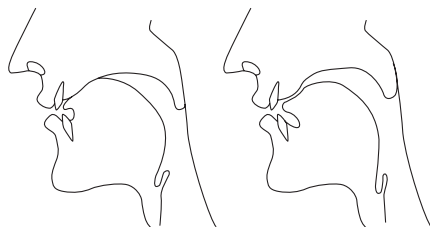


Fig. 6.6 Illustration of the two phases of English affricates: complete palate-alveolar closure on the left, and narrow release resulting in the fricative on the right

Table 6.3 Summary of English affricate consonants

IPA	Manner	Place	Fortis/Lenis	Examples
tʃ	Affricate	Post-alveolar	Fortis	Church, chew, achieve, nature
dʒ	Affricate	Post-alveolar	Lenis	Judge, George, jaw, adjective, adjacent, ajar

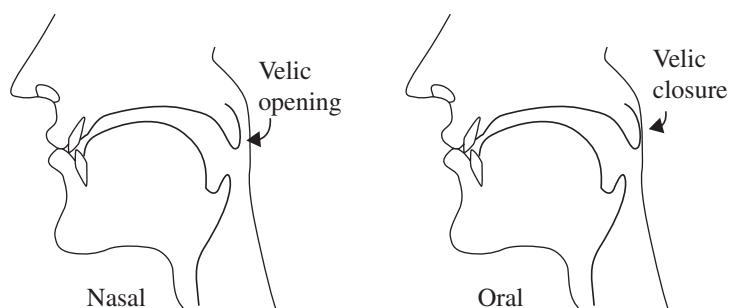


Fig. 6.7 Illustration of bilabial nasal [m] on the left with the velum lowered and the resulting velic opening, and an oral bilabial stop on the right with the velum preventing the passage of air to the nasal cavity

Nasals

Nasals are produced as complete stops in the oral cavity but simultaneous lowering of the soft palate. Therefore, the air is not trapped behind the obstruction but leaves the oral cavity through the **velic opening** into the nasal cavity and out through the nostrils. This can be easily verified by pressing your nostrils shut with your thumb and the index finger and opening them again while saying a sustained nasal sound, for example [m:] or [n:] corresponding to [mmmm] or [nnnn] respectively. Figure 6.7 shows velic opening and how the nasals are otherwise identical articulatorily to the homorganic lenis stops with the complete closure of air in the oral cavity. Recall Activity 3-8 in Chapter 3 in which you became aware of the velic opening and the lowering of velum that is responsible for that.

There are three places of articulation for nasals in English: bilabial [m], alveolar [n] and velar [ŋ], and all of them are voiced.¹ Table 6.4 lists the articulatory characteristics and examples for these three nasals in English. Note that the velar [ŋ] is the only nasal restricted in its distribution in that it never occurs in the word-initial position in English but a more detailed discussion of similar distributional patterns in English can be found in Chapter 8.

Activity 6-6 below explores the visual characteristics of nasals in Praat.

Activity 6-6: Compare oral and nasal stops

To compare the acoustic characteristics of the oral and nasal stops, record several pairs of short words that have the same vowel and the place of articulation and voicing of the consonants and differ only in the manner (oral vs. nasal) of the stop.

¹Nasal consonants that are voiceless (and contrast with the voiced ones or voiceless ones of various places or articulation) are extremely rare among the languages of the world due to their low perceptual discriminating potential.

Table 6.4 Summary of English nasal consonants

IPA	Manner	Place	Voicing	Examples
[m]	Nasal stop	Bilabial	Voiced	Mouth, remind, hammer, plumber, lemon, alarm, volume, condemn
[n]	Nasal stop	Alveolar	Voiced	Knife, narrow, gnat, dinner, minor, ten, align, turn
[ŋ]	Nasal stop	Velar	Voiced	Thing, bang, tongue, language, finger, singing, bingo, link, larynx, ankle

For example, pairs like ‘mom’ - ‘Bob’ (In AE), ‘noon’ - ‘dude’, ‘mean’ - ‘bead’, or ‘Ming’ - ‘Bing’. Save the file, open in Praat with the spectrogram and formant analyses visible and zoom in to see both members of the pair in one window.

Similarly to some previous activities, focus first on the similarities between the nasal and oral stops. Describe them in your own words. Repeat the same with differences. Again, take a minute to strengthen the link between the kin-aesthetic feel when you say these words slowly, auditory sensations when you listen to them, and the visual information in Praat corresponding to these senses.

The articulatory description of the production of the nasals determines their acoustic characteristics observable with Praat. First, the voicing produced at the vocal cords creates the acoustic source that then enters the nasal cavity. There are two acoustic characteristics of this. First, the resonances of the nasal cavity are strongly damped due to its anatomical differences from the oral cavity and start at low frequencies, around 250 Hz. Second, the closure in the oral cavity blocks the regular resonances of the oral cavity, which creates so called anti-resonances, particularly of the frequencies around 1000–2000 Hz, which is visible as a whitish area in the spectrogram around that frequency interval.

The other acoustic marker relates to the leakage of air through the lowered velum. This prevents the accumulation of air behind the obstruction in the oral cavity and thus the release of this obstruction is typically not characterized with a salient burst of energy that is typical for oral stops. Hence, when nasals are compared with lenis oral stops, nasals show much more energy and a clear formant structure during the closure and the absence of a salient burst after the closure. The similarities stem from their shared places of articulation and the formant transitions characterizing them, and partially from the decreased energy during the closure.

When nasals are compared with vowels, the overall energy (i.e. amplitude in the waveform or darkness in the spectrogram) is lower for the nasals and the frequencies between roughly 500 and 2500 Hz tend to show the damped energy.

Approximants

The last group of English consonants includes 4 members: [l], [ɹ], [j] and [w]. Although the first two sounds present immense variability and dialectal differences

in their production, we group them together due to their common stricture degree. All four preceding consonantal categories (stops, fricatives, affricates and nasals) were characterized by either complete or very narrow stricture at the primary place of articulation. All four remaining sounds can be characterized by the stricture degree that is intermediate between that of fricatives and that of vowels. Hence the word *approximant* signals that the active articulator(s) get into the close proximity of the passive or another active articulator. Despite significant differences among the group members themselves, approximation of the articulators that targets a low degree of stricture, is common for all four consonants. Also, while [l] and [ɭ] do involve some contact between the tongue and the roof of the mouth in many contexts and many dialects of English, there are also many speakers and many contexts in which this contact is commonly not made (e.g. [miwk] for ‘milk’ in Cockney).

Similarly to the nasals, all approximants are normally voiced but they might also be devoiced in English in certain context. More on this in the next chapter.

Lateral Approximant [l]

This sound starts words like ‘law’ or ‘Lee’ and ends words like ‘all’ or ‘eel’. We need to explain both characteristics from the heading: lateral, and approximant. A good activity that builds awareness of the first one is to say prolonged [l:] and, without moving anything, reverse the stream of air from going out (egressive) to going in (ingressive) and sucking the air in. Alternate several times and observe where you feel the coldness associated with the passing air. It should be on the sides of the tongue.² Hence, although the blade of the tongue is touching the alveolar ridge (possibly also partly the upper incisors), the primary location of the sound source are the *sides* of the tongue, for which the adjective is *lateral*. You might also try saying [d] immediately followed by [l] and observe how the complete obstruction of [d] is released by the lowering of the sides of the tongue in words like ‘middle’.

The awareness regarding the stricture of approximants might be strengthened by saying a prolonged [l:] and simultaneously attempting to raise the sides of the tongue so that friction is felt. It is somewhat easier to also devoice this fricative. Lateral fricatives, in IPA [ɬ, ɮ], are sounds occurring in several languages, and this exploration shows that English [l] has more open stricture as that of fricatives. Therefore [l] belongs to the group of approximants.

Since the stricture of approximants is even more similar to that of vowels than the previous consonantal manners, the acoustic characteristics of approximants in general and [l] in particular, are very similar to vowels. The lateral has a salient formant structure and the amount of energy in the sound is greater than in nasals but still discerningly lower than in vowels. This can be observed in the vicinity of the consonantal release where the contrast in the energy before and after the release of the tongue blade is usually possible to see in the spectrograms. Due to the contact between the tongue blade and the alveolar ridge, frequencies in

²I first heard about this discovery activity from A. Underhill at the seminar devoted to his book (Underhill 2005).

the range of the second and third formant are commonly attenuated and thus the release of the tongue is best observable as the contrast in the energy of F2-F3. Additionally, a salient signature for /l/ is also F2 lowering. This is because the tongue's muscle anatomy requires the retraction of the tongue body/back in order for the tongue tip to elevate for the contact with the alveolar ridge.

Median Rhotic Approximant [ɹ]

Probably the most difficult English consonant to describe articulatorily is [ɹ]. In all English dialects it occurs in the positions preceding vowels like in 'red' or 'array' and in the rhotic dialects also in the positions not following vowels like in 'car' or 'card'. We will be using the IPA symbol [ɹ] to clearly contrast English [ɹ] of RP or GenAm pronunciations from the trill-type [r] found in other varieties of English (Scottish, Irish) and many other languages.

The term 'median' contrasts [ɹ] from the lateral [l] in that for [l] the sides of the tongue create the primary sound source, and for [ɹ], the midline (technically the mid-sagittal plane) is the location of the primary stricture. But where exactly and how this stricture is created is the subject of enormous variability.

Activity 6-7: Introspection of your [ɹ]

Given the variability of [ɹ] mentioned above and before discussing it in the text below, explore how you yourself produce this sound. Say the first sound of 'red' or 'raw' in a prolonged way and observe first the position of your lips (either with a mirror or by placing a finger on your lips). Is there some systematic target for the lips during your [ɹ]? Next, concentrate on your tongue when saying a prolonged [ɹ] in 'array'. Both the preceding schwa and the following [e] are relatively central vowels and thus suitable for observing the tongue activity when moving from the schwa towards [ɹ] and away from it. Does your tip of the tongue touch anything? Do you feel that the target for [ɹ] is to create stricture with your tip or with another part of your tongue? If you have fellow students or a friend around, discuss together and compare the articulatory strategies for saying this sound.

Figure 6.8 shows MRI images of 22 American speakers producing [ɹ] reported in Tiede et al. (2004).³ In addition to two relatively easily separated variants (*bunched* in which certain part of the tongue is elevated but the tip remains down vs. *retroflex* that involves raising of the tip and a potential secondary constriction in the pharyngeal area), the findings suggest that at least five systematic patterns for American [ɹ] can be identified that are produced by more than a single speaker. These correspond to the five columns in the figure. In your introspection, do you belong to any of these five groups and can you feel these tongue configurations in your [ɹ] production?

³I thank S. Boyce and M. Tiede for permission to use this image.

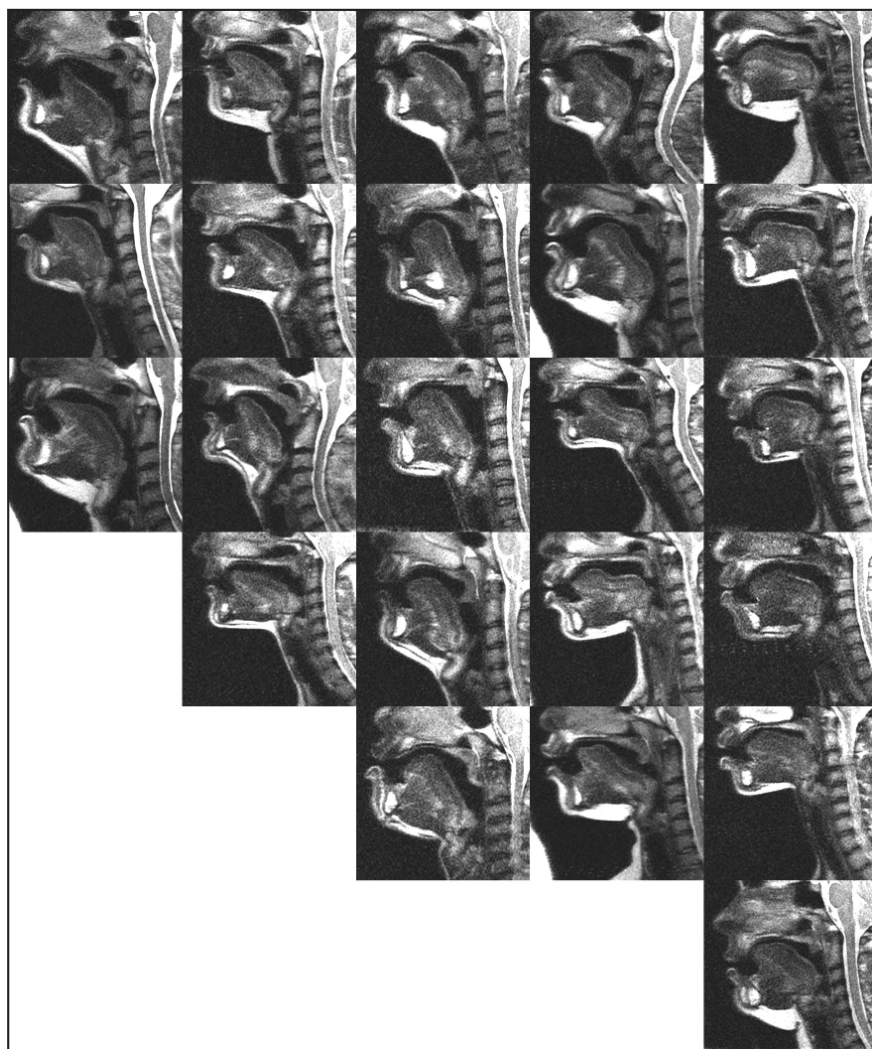


Fig. 6.8 MRI images of the mid-sagittal plane from the sustained production of [ɪ] by 22 speakers of North American English, Tiede et al. (2004)

In addition to the variability of the median constriction location, further differences involve the activity of the lips (many speakers round their lips for [ɪ]) and the dynamic characteristics of the tongue movement (for the retroflex types, whether there is just narrow stricture or also a brief contact between the tongue tip and the alveolar ridge).

Hence, English [ɪ] can be made with several different configurations of the tongue and lip targets while these differences do not make salient acoustic

differences. Hence, there are several articulatory strategies for achieving acoustically similar [ɹ] sounds. The prototypical acoustic signature of English [ɹ] is the distinct lowering of the third formant and partial lowering of the second formant.

Glides (Semivowels) [w] and [j]

The last two English approximants are commonly grouped together as *glides* or *semivowels*. They can only be found in pre-vocalic positions in words like ‘well’, ‘one’, or ‘away’ and ‘yes’ or ‘unique’, respectively. They are the most similar approximants to the vowels both articulatorily and acoustically.

Activity 6-8: Identify the essential difference between glides and vowels

Say a sustained prolonged [j:] at the beginning of ‘yes’. Does it sound like one of the English long vowels? Do the same with a sustained prolonged [w:] at the beginning of ‘why’.

You notice that they sound like high vowels [i:] and [u:], respectively. So the question is how you, in your speech, differentiate between these high vowels and consonantal glides. What is the essential difference? Using our pet robot from Activity 6-3, formulate the articulatory plan, i.e. the precise instructions, for producing glides, and ask the robot to pay particular attention to do X to differentiate glides from the corresponding high vowels. What is the essential X?

While most consonants can be sustained for some time (limited only for stops), glides cannot be sustained because they turn into vowels. Hence, the essential articulatory requirement for glides is the rapid movement towards and away from the obstruction. In other words, the articulatory target for the glides must be specified not only for their location and stricture (degree) but also in terms of their dynamic realization. Note how slightly awkward the pronunciation of words like ‘woo’ is in this respect since the articulators do not really move between the glide and the vowel.

In terms of the place of articulation, the stricture in [j] is produced by the tongue body approximating the hard palate and thus it is described as a *palatal* consonant. In [w], the stricture is produced at two median locations: the back part of the tongue approaches the soft palate (velum), and the lips protrude and approximate each other. Hence, the most common place specification for [w] is *labio-velar*. Some native speakers might produce a voiceless variant of [w] in words spelled with <wh> for which IPA uses the symbol [ɸ]. This applies mostly to Scottish or Southern-US speakers who, for example, might contrast ‘which’ and ‘witch’ with the voiceless [ɸ] and voiced [w], respectively.

The acoustic characteristics of glides correspond to their vowel counterparts: low F1 and high F2 for [j] and low F1 & F2 for [w]. Additionally, due to the rapid creation and release of the stricture, the formants are not stable as for vowels but display clear movement throughout the consonant. Table 6.5 lists the characteristics and examples of English approximants.

Table 6.5 Summary of English approximants

IPA	Manner	Place	Voicing	Examples
l	Lateral approximant	Alveolar	Voiced	Love, light, ceiling, alone, elbow, little, wild, wall, school
ɹ	Median approximant	Alveolar	Voiced	Rate, rule, write, rhythm, area, arrow, mirror
j	Median approximant (glide/semivowel)	Palatal	Voiced	Yes, yawn, use, unit, Europe, queue, pew, cute
w	Median approximant (glide/semivowel)	Labio-velar	Voiced	Weak, what, wave, aware, squeeze, liquid, penguin, suite

6.4 Sound–Spelling Correspondences for English Consonants

Similarly to vowels, we will devote only a brief mention to the notoriously complex relationship between sound and spelling of English consonants. Activity 6-9 asks you for the summary of the patterns known to you or observable from the tables of this chapter, and the text below summarizes the major systematic patterns.

Activity 6-9: Systematic patterns in sound–spelling of English consonants

This chapter included a table for each group of English consonants by their manner of articulation. Try first to list the potential spellings of every single consonant. Then, try to determine if any of the variation in spelling might be systematic. That is, would the spelling of a novel word created with this sound in this environment be predictable?

There are two major groups of patterns that deviate from one-to-one correspondence between one letter and one consonant. The first one is the presence of ‘silent’ consonants. The most known is probably the silent <k> of ‘know’ or ‘knee’ word initially. Its voiced counterpart <g> can also be silent initially in words like ‘gnat’, but also at the end of words such as ‘sign’ or ‘diaphragm’. Other stops are also commonly silent: <p> typically initially in borrowed words starting with ‘psycho-’ or ‘pseudo-’, in word-final positions like ‘bomb’ or ‘debt’, <t> in words like ‘whistle’ or ‘listen’ and rarely also <d> in words like ‘Wednesday’. From fricatives, <s> is exceptionally missing in words like ‘isle’, and <h> is commonly not pronounced in words like ‘ghost’, ‘rhotic’, wh-words like ‘why’ or ‘what’ (but recall [ʌ] in some dialects for these words), and of course <ph> is pronounced as [f] or <ch> as [tʃ] or [ʃ] or [k]; see below. In nasals, <n> is omitted in the final position graphemes <mn> in words like ‘column’. The most common silent letter from the approximants is <l> in words like ‘palm’, ‘talk’ or ‘should’.

The second group presents predictable variation in how a particular spelling combination is pronounced in a particular environment of surrounding sounds. The most common example is the effect of the following vowels on the pronunciation of <g> and <c> letters. In a rule of thumb, the (post-)alveolar [dʒ, s] precede front vowels while the velar [g, k] precede back ones. Another pattern includes word origins. For example, <ch> in French borrowings is pronounced as [ʃ] but as [k] in the technical vocabulary of the Greek origin. Yet another refers to parts of speech and spelling suggesting that word-final fricatives are commonly fortis in nouns whereas they are commonly voiced in similar verbs including <-e> in spelling: 'brea[θ]' but 'brea[ð]e'.

Activity 6-10: English plural and past-tense suffixes

In Chapter 2 in discussing Activity 2-3 I mentioned a systematic pattern for the pronunciation of word-final suffixes <-s> marking the plural in dog[z] but cat[s]. The same pattern applies to all <-s> suffixes (including 3rd person singular in verbs, possessive in nouns) and <-ed> suffixes marking the regular past tense and past participle in verbs. Consider words like 'ribs', 'lips', 'nieces' or 'moved', 'kicked', 'plotted' and try to formulate the pattern yourself before reading below.

The plural or past-tense endings from Activity 6-10, as you probably found out, involve the consideration of voicing and place of articulation. For the lenis [z] and [d] endings in 'ribs' or 'moved', the word-final consonant is lenis, and for the fortis [s] and [t] in 'lips' or 'kicked', the preceding consonant is fortis. This is a special case of *voicing assimilation*, and more on this will be discussed in Chapter 10. Note that it applies only to suffixes and not to <s> that is the integral part of the word since words like 'mince' end in [s] despite the preceding voiced [n]. The third variant with a schwa [əz] and [əd] is used if the preceding consonant is (almost) identical to the suffix, particularly in terms of the manner and place of articulation.

The notes on sound–letter correspondence in this sub-section are only suggestive and far from complete. We mention them since they do represent mental habits and patterns native speakers developed; some of them unconsciously as in the case of plurals or past-tense endings, and some of them consciously when learning spelling in school and practising in reading and writing.

Exercises

- 6-1 Pick one representative from each of the five major consonantal classes covered in this chapter (stops, fricatives, affricates, nasals, and approximants). Find an English word in which the consonant appears between vowels and describe in as much detail as possible all articulatory

- activities necessary for producing this consonant in this environment. Pay attention to detail when talking about stricture, voicing, and the place of articulation. Then describe what you expect to see in the spectrogram of these five words regarding the target consonants. Verify in Praat.
- 6-2 Brainstorm for the phonetic reasons why in several languages [l] becomes vocalized (as in [mɪwk] for ‘milk’ in Cockney or in Serbian/Croatian past-tense suffix /l/ → [o])? Think of both the articulatory and acoustic characteristics of the relevant sounds.

References

- Tiede, Mark K., Susanne E. Boyce, Christy K. Holland, and K. Ann Choe. 2004. A new taxonomy of American English /r/ using MRI and ultrasound. *Journal of the Acoustical Society of America* 115: 2633–2634.
- Underhill, Adrian. 2005. *Sound foundations*. Oxford: Macmillan.



Allophonic Variation in English

7

In this chapter we will

- build awareness of coordinated articulatory habits and their acoustic visualization in combining vowels and consonants into simple English words
- exemplify further the physical and mental aspects of speech in the phonetics-phonology system underlying speech
- outline English allophones as contextual/positional variants and highlight this as an issue for non-native speakers
- explore how syllable affiliation affects the production of consonants
- introduce narrow transcription

7.1 Introduction

The previous two chapters on vowels and consonants have built the foundation for understanding the articulatory processes and their acoustic consequences for the segments at the cost of discussing them in a somewhat unnatural context-free fashion. The current and subsequent chapters will explore expanding layers of context and how our speaking habits are defined by, and integrated in, this context.

This expands on the notions of continuity discussed in Chapter 2 and several pieces of evidence for this notion in the articulatory and acoustic data (e.g. formant transitions). While these characteristics can be construed as automatic, and general of speaking in any language, this chapter focuses more on unconscious speaking habits that are particular for English.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_7) contains supplementary material, which is available to authorized users.

An important note is in order before we proceed. Although most of the words analysed in this chapter will still be short monosyllabic words, we will inevitably start involving also the notions of *syllable* and *stress* in our discussion. These two aspects of speaking behaviour serve as topics for two separate subsequent chapters in which they are discussed in greater detail. For the purposes of this chapter, however, basic informal understanding is sufficient. *Syllable* is a fundamental unit of speech such that longer words have several syllables and speakers can easily count these syllables. Hence, ‘pan’ has one, ‘panda’ two, ‘pandemic’ three syllables and ‘pandemonium’ five.¹ In all these words, [pæn] is the first syllable. Some of these syllables are then more highlighted or prominent than others. For example, all native speakers of English perceive that in ‘panda’, the first syllable is more prominent than the second. Or that the second syllable of ‘pandemic’ is more prominent than the first. This contrast stems from the distribution of *word stress* and the syllables that have it are called stressed ones, while those that don’t are labelled as unstressed.

7.2 English Stops, Phonemes and Allophones

In the previous chapter we identified stops as contrasting in terms of the place of articulation and fortis/lenis dimension. Hence, the primary articulatory gestures required for producing stops involve a complete oral constriction at a particular place in the vocal tract and the appropriate setting of the glottis. But we have not discussed yet how these two discernible subroutines are organized in time. Recall from Chapter 2 that speaking involves coordinations of articulatory activity at multiple levels and skilled performers (= native speakers) developed these coordinations unconsciously while many non-native speakers might struggle with achieving the smoothness of the coordinations performed by the native speakers. Activity 7-1 explores the visual cues to the *temporal coordination* of articulatory gestures.

Activity 7-1: Identify gestural coordinations with Praat

If you are a native speaker, record yourself saying the words ‘a pie’, ‘a bye’ and ‘a spy’. If you are a non-native speaker, record a native speaker or use the recording ‘labial_stop_allophones.wav’ provided in this chapter of the book companion. In any case, introspect the gestures forming the bilabial closure and voicing at the vocal cords and how they are organized in time. This is not easy to do, so do not worry if this coordination feels elusive.

Now let’s add the visual information. Select to view ‘a pie’ and ‘a bye’ in a single window with both the waveform and the spectrogram and a sufficient zoom (cut silent intervals if needed with *Ctrl-X*). Visually identify the beginning and end of the articulatory actions required for:

¹For some speakers or in fast speech, ‘pandemonium’ might also be four syllables.

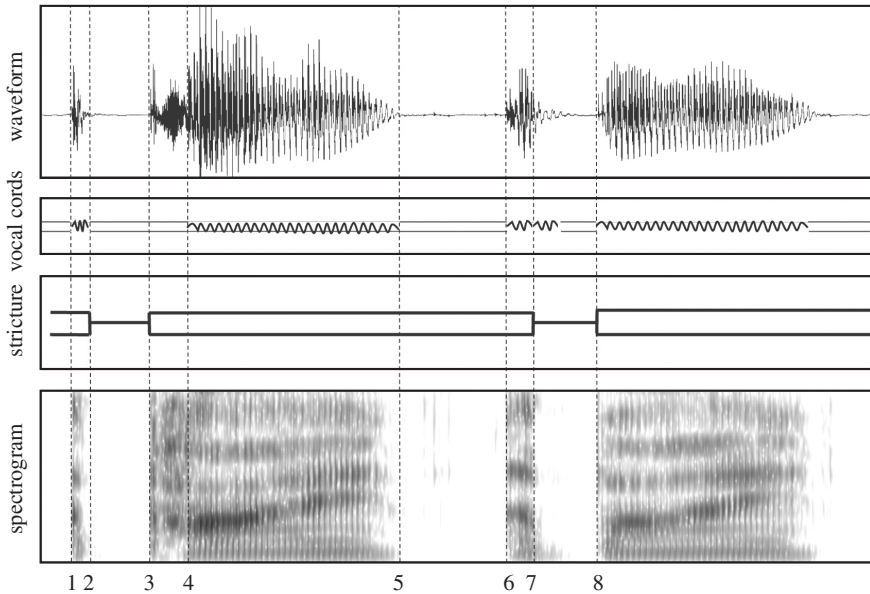


Fig. 7.1 ‘a pie’ and ‘a bye’. See text for explanation of the panels and numbered vertical lines

- the formation of the stricture (in this case the complete closure)
- the vocal cord activity (when the vocal cords are vibrating)

Support your reasoning in as much detail as possible. Finally, how are these two actions organized in time? What follows what and when?

Then consult and compare with the text below and Fig. 7.1.

Figure 7.1 shows the words from Activity 7-1 above (‘a pie’ and ‘a bye’) and their waveform and the spectrogram in the top and bottom panels, respectively. In addition, the middle panels illustrate schematically the temporal development of the articulatory activity at the glottis and the lips for the essential gestures in the stop consonants. The panel below the waveform shows vocal cord vibration: the wavy squiggle indicates adducted and vibrating vocal cords, and the double straight line suggests the vocal cords are apart and not vibrating. The panel below that shows stricture formation: the single line shows closed lips and a complete obstruction while the double line indicates the opened lips. The numbered vertical dotted lines indicate the relevant temporal points corresponding to:

1. The onset of voicing for the vowel,
2. The closure of the lips for [p] of ‘pie’,
3. the opening of the previously closed stricture in [p] of ‘pie’ (lips move apart, open),
4. closing of the previously open vocal cords resulting in the onset of their vibration and voicing for [aɪ],

5. the offset of voicing for the vowel,
6. the onset of voicing for the first vowel in ‘a bye’ associated with already open stricture for vowels,
7. the onset of bilabial lip closure for [b] of ‘bye’; the vocal cords keep vibrating,
8. the offset of the b-closure, i.e. lip opening

Given these temporal points, the intervals for [p] relevant to the discussion and observable in Praat are: an almost flat line, which indicates complete voiceless closure, an interval of aperiodic noise between the third and fourth vertical lines and clear voicing with the formant structure following the fourth point.

Now that we are familiar with these temporal landmarks of articulatory gestures, let us compare [p] and [b] from Fig. 7.1 above to [p] of spy and [b] of bye.

Activity 7-2: Auditory illusion: another ghost segment?

From the same recording as in Activity 7-1 above, select now the word ‘a spy’, zoom to selection so that it fills the entire Praat Edit window. Select the interval starting just before the lip opening and the end of the word as indicated in Fig. 7.2. Click the bar below/above to listen. What do you hear?

Do the same for the identical interval in ‘pie’ and ‘bye’ and listen to the interval starting just before the lip opening and ending with the end of the word. Hence, in Fig. 7.1 interval 3–5 for ‘pie’ and 8–end for ‘bye’. Compare these sounds with the same interval from ‘spy’ shaded in Fig. 7.2. Which of the two (pie or bye) sounds closer and more similar to the interval from ‘spy’? Try to reason about the causes of your perceptions.

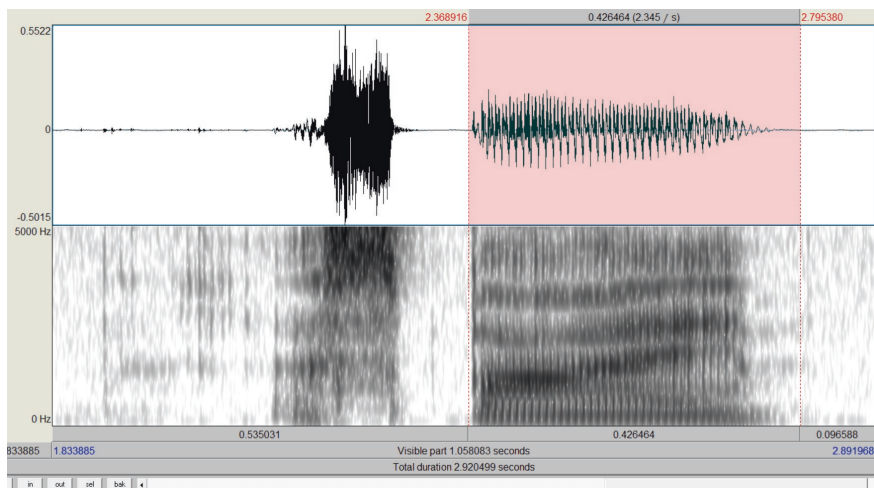


Fig. 7.2 Waveforms and spectrogram of ‘spy’ with the (pink) shaded interval indicating the target interval to listen to in Activity 7-2

There are two questions that require your attention:

- Why do you hear a bilabial closure in the interval shaded in Fig. 7.2 even if the closure itself is excluded from the interval you listen to?
- Why does this bilabial stop in ‘spy’ sound more like the voiced [b] of ‘bye’ rather than the voiceless [p] ‘pie’ and thus why do you hear ‘bye’ if you cut [s] from ‘spy’?

Let’s start with the easier first question from Activity 7-2. The answer stems from the awareness we gained in Activity 6-2 in the previous chapter. Recall that omitting [s] from ‘seed’ sounded like ‘deed’ due to the formant transitions indicating the alveolar place of articulation of the consonant and silence indicated a complete closure for stops. The same applies here and you heard a bilabial stop for the same reasons. However, the observation from the second question, that this bilabial stop sounds like the voiced [b] rather than the voiceless [p], should be a novel, and for some of you also surprising, finding; you did not say ‘sby’ but ‘spy’.

The answer to why you hear a [b] in ‘spy’ involves the understanding of coordination between the oral stricture formation and vocal cord vibration in English stops. In spoken English, there are three stable coordinations of these two actions. The first can be observed with ‘pie’ also shown in Fig. 7.1. The onset of voicing occurs well after the release of the closure. Hence, the *voice onset time (VOT)* is positive, which results in an interval of *aspiration* in which the vocal cords are still open and the air released after the opening of the closure produces aperiodic noise-like sound. Stops produced with this aspiration are transcribed with an upper script [ʰ] and ‘pie’ would thus be [pʰaɪ]. As is clearly shown in Fig. 7.1, the duration of this aspiration interval can be easily measured with Praat.

The second stable coordination results in voice onset time that is significantly shorter, or sometime around zero; that is, the onset of voicing is shortly after the release of the closure. This is the [p] in ‘spy’. Figure 7.2 shows that vocal cords started vibrating right after the short burst in ‘spy’. By comparing ‘spy’ and ‘bye’ in Praat, we see that these two sounds are similar in this coordination and thus [p] is unaspirated while the of ‘bye’ is commonly produced by native speakers in word-initial positions following a silence as a (partially) devoiced [b̥] that sounds somewhat similar to [p]. Hence, this phonetic similarity of voice onset time realization in [b̥] and [p] is the main reason behind the auditory illusion from Activity 7-2.² The unaspirated realization of voiceless stops is common in the majority of Europe’s non-Germanic languages and a handful of the Germanic ones (e.g. Dutch).

The third timing pattern is exemplified in fully voiced stops, which is illustrated in ‘a bye’ in Fig. 7.1. The vocal cords start (or maintain) vibrating well before the

²Another factor that plays the role is that native English words starting with <sb> do not occur. Hence, even though the <p> in ‘spy’ is acoustically similar to in ‘bye’, we still identify it as /p/ of ‘spy’.

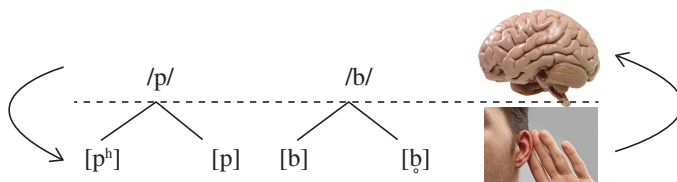


Fig. 7.3 Illustration of phonemes (abstract cognitive units) and allophones (actual realizations of the sounds produced and perceived)

release of the stricture. This is the realization of voiced stops in many languages of Europe even in the word-initial positions preceded by silence.

Figure 7.3 summarizes the situation we have just explored. There is a clear contrast between /b/ and /p/ in English such that replacing one for the other changes the meaning of the word: replacing /p/ in ‘pie’ for /b/ gives ‘bye’. In the minds of native speakers these are different *phonemes*, or *contrastive* sounds. One of the tests for this phonemic contrast is the existence of *minimal pairs*. A minimal pair is defined as two words whose meaning contrast is signalled by a contrast in a sound of the same position while all other sounds remain identical. Phonemes, that signal meaning contrasts in minimal pairs, are thus said to be in *parallel* or *contrastive distribution*. For phonemes /b/ and /p/, minimal pairs in English include for example ‘pie’-‘bye’ or ‘cap’-‘cab’. Note that contrast is in sounds and not necessarily in spelling.

In speaking, when the intention to signal these contrasts is transformed to real physical actions, the realization depends on context (Chapter 2). For example, we saw that the unaspirated [p] occurs in positions following [s] as in ‘spy’ whereas the aspirated [pʰ] occurs in the very initial positions. These phonetic realizations of phonemes are called *allophones* and are *predictable* since they can be often determined from context. Native speakers of English would never mix them up and pronounce *[spʰaɪ] or *[paɪ] and thus we say that these allophones are in *complementary distribution*; a ‘*’ indicates an illicit, or ungrammatical, realization. Phonemes like /p/ and /b/ create meaning contrasts and they can occur in the same positions, such as word-initially as in ‘pie’ vs. ‘bye’, and thus cannot be determined by context.

You might have noticed that in addition to the IPA symbols [p] and [b] for phonemes /p/ and /b/ we have started using also diacritics marking aspiration or devoicing. Primarily, we will be concerned with two types of IPA transcription: broad and narrow. The *broad transcription* uses only symbols representing English phonemes, i.e. the sounds that can be contrastive for meaning. The *narrow transcription* might express more phonetic detail characterizing specific systematic native-like habits of English speakers that non-native speakers might find strange or challenging. For example, it might be a novel observation for less proficient native speakers of Polish, Italian, or Hungarian that English sounds /p, t, k/ in certain positions are aspirated. Hence ‘pick’ would be [pɪk] in broad transcription but it might be [pʰɪk] in narrow transcription to reflect the aspirated realization of /p/ in this word.

Activity 7-3: How is non-initial intervocalic /p/ realized: check your intuition quantitatively

First, record your natural pronunciation of the following four words: ‘copy’, ‘appear’, ‘stepping’ and ‘opaque’ if you are a native speaker or find a native speaker willing to be recorded.

We have identified two allophones of /p/: aspirated [p^h] in word-initial positions like in ‘pie’ and unaspirated [p] in ‘spy’. Without Praat, inspect your p-realizations in the five words you recorded and suggest the allophones in them. Is the binary choice of [p^h] and [p] sufficient or more options would be needed? Write down your suggestions and reasons for them.

Now open the recording in Praat and measure the duration of aspiration, or voice onset time in these words. Zoom sufficiently to the vicinity of the bilabial stop release, identify the time-point of lip opening, left-click and drag the mouse to the point of the onset of vocal cord vibration. Record the duration of the selected interval (in milliseconds). Do the same for /p/ in ‘pie’ and ‘spy’ of the file from Activity 7-1. Do you see any systematic pattern? Check the text against your observations.

You probably noticed auditorily that there is a clear difference in the amount of aspiration between ‘copy’ and ‘stepping’ on the one hand and ‘appear’ and ‘opaque’ on the other hand. It is likely that your measurements of aspiration showed that the former have significantly shorter durations of the aspiration interval than the latter ones and that aspiration in ‘appear’ and ‘opaque’ is comparable to that of ‘pie’. In ‘copy’ and ‘stepping’, there might be some aspiration, probably slightly more than in ‘spy’, but significantly less than in the other group. Words like ‘copy’ and ‘stepping’ are transcribed with an unaspirated [p] in narrow transcription. This discussion nicely shows that narrow transcription is also an abstraction from real speech, it just includes more phonetic details than the broad transcription.

Importantly, now, what is common for ‘copy’ and ‘stepping’ that makes it different from ‘appear’ and ‘opaque’? Note that spelling cannot be the reason because in both groups there are words with a single<p> and a double<pp>. Also, in both groups <p> appears in the middle of the word. If you suspect *word stress* you are on the right track. Aspiration in English is associated primarily with stressed syllables: voiceless stops are aspirated if they begin a stressed syllable.

Returning now to Fig. 7.3, it illustrates the traditional division in the study of speech to *phonology* as the cognitive system of somewhat abstract contrasts in the mind in the top and *phonetics* as the cognitive system of observable physical realizations through studying speech production and speech perception. This further exemplifies the discussion in Chapter 2 when these terms were introduced, but it is important to note that the nature of the boundary indicated with the dotted line, and the arrows, is a fascinating research area.

The allophones, or positional variants, are a crucial concept for non-native speakers. First, it is the case that some allophones do not occur in the native

language of the speaker and s/he needs to create a habit of producing them in a similarly smooth and effort-free nature as the native speakers do. For example, speakers of languages without aspirated stops are either unaware of aspiration and use their L1 unaspirated stops also in their L2 English, or they produce aspirated stops in a stilted way easily recognized by native English speakers.³ Hence, native speakers must develop a new link between the phonological system of contrasts and the phonetic realization of allophones.

A different problem for non-native speakers of English is that sometimes the required allophones do exist in their native language but their distribution based on context is different. Consider data from German and English below. In both languages, the phoneme /b/ may be realized as either voiced [b] or devoiced [b̥] that is very similar to the unaspirated [p]. Moreover, the aspiration of /p,t,k/ is also similar and thus German can also be described with Fig. 7.3. But is the pattern of allophone deployment really the same in the two languages?

German			English	
gab	[g̥ab]	gave	cub	[k ^h ʌb]
Kap	[k ^h ap]	cape	cup	[k ^h ʌp]
Bad	[b̥aɔ̯]	bath	bud	[b̥ʌd]
Pad	[p ^h ɛɔ̯]	(coffee) pad	putt	[p ^h ʌt]

In this data, devoicing can be found in the initial position in English, and in German it takes place both in the initial and the final positions of these monosyllabic words. Due to so called *voicing neutralization* in German and many other languages (Slovak, Polish, Catalan), /b,d,g/ at the end of words, and possibly before a pause, are realized with the saliently devoiced allophones. In English, some devoicing also might take place but it is much less salient than in the other mentioned languages. Hence, German speakers of English must become aware that their L1 word-final devoicing in words like ‘gab’ should not be used in similar English words like ‘cub’ despite the fact that similar devoicing applies in both languages word-initially. In addition to the awareness and skill in terms of purely articulatory habits, all speakers also develop *mental* knowledge to know in which context to deploy which articulatory habit. L2 speakers lower the degree of their foreign accent by learning the L1 pairings of allophones and contexts.

³For example, many non-native speakers after being informed about aspiration produce it as if they were ‘adding a [h]’ after the stop. The problem is that this type of added [h] results from a different coordination of the stricture formation and vocal cord activity. Therefore it is important to become aware that aspiration is just a natural consequence of a particular coordination of articulatory gestures.

Finally, another example of the problems non-native speakers face regarding the mental patterns in phonemes and allophones can be exemplified with Hindi or Thai. Similar to English both languages have aspirated and unaspirated voiceless stops. But consider data from bilabial stops below and try to describe informally how the cognitive system of English is still different from those in Hindi or Thai.

Hindi		Thai (tone marks omitted)		English	
[pʌ]	nurture	[pa:n]	birthmark	[spʌn]	spun
[pə]	moment	[pa:]	aunt	[spɪ]	spill
[pʰʌ]	knife, blade	[pʰa:n]	belligerent	[pʰʌn]	pun
[pʰə]	fruit	[pʰa:]	cloth	[pʰɪ]	pill
[bʌ]	hair	[ba:n]	to bloom	[bʌn]	bun
[bə]	strength	[ba:]	crazy	[bɪ]	bill

The data show that in Hindi or Thai the realization of aspirated and unaspirated bilabial stops is not predictable as is the case in English. Both [p] and [pʰ] appear in exactly the same environment – word-initially preceding a vowel – and the difference between these two sounds is contrastive since it makes a difference in meaning. Hence, the sounds [p] and [pʰ] in the minds of Hindi or Thai speakers correspond to two different phonemes /p/ and /pʰ/, respectively, which is illustrated in Fig. 7.4. In English they correspond to a single phoneme /p/ that is realized as [pʰ] initiating stressed syllables but as [p] when following /s/, which was shown in Fig. 7.3.

If English speakers learn languages like Thai or Hindi, they have to learn that aspirated stops contrast with the unaspirated ones. Hence, their L1 pattern that [p] and [pʰ] belong to a single mental category, or are ‘essentially the same sound’, has to be changed. Furthermore, they have to develop an ‘awkward’ habit of producing initial unaspirated stops that contrast both with the aspirated voiceless ones as well as the voiced ones. Conversely, Hindi or Thai speakers speaking English have to do the mirror image: treat [p] and [pʰ] as ‘essentially the same thing’ and learn the systematic association between the context (initial vs. following a /s/) and the allophones ([pʰ] vs. [p]).

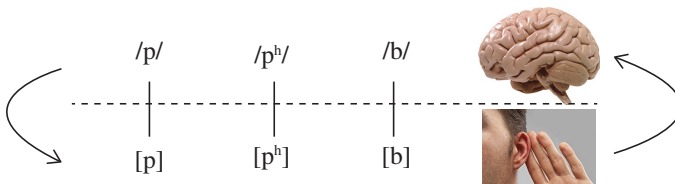


Fig. 7.4 Phonemes and allophones of bilabial stops in Hindi or Thai

We saw above that two allophones in English might be phonemes in another language but two English phonemes might also be allophones in other languages. Consider Spanish data below.

navidad	[naviðað]	Christmas
dos	[dos]	two
madre	[maðre]	mother
dragón	[dragon]	dragon
decir	[desir]	to say
de nada	[denaða]	you're welcome
dando	[dando]	giving

In English, [d] and [ð] signal contrast in meaning, can be found in minimal pairs (e.g. 'den' vs. 'then'), and they are thus phonemes. Their distribution is unpredictable since there is no generalization or a pattern that would tell us which of these two sounds will appear initially followed by [en] and thus rhyme with words like 'pen' or 'when'. In Spanish, however, we do not find minimal pairs with these two sounds and their distribution is predictable and complementary: it is always [ð] and never [d] following a vowel while in the positions not following a vowel we always find [d] and never [ð]. In Spanish these two sounds are thus allophones.

Allophones of a single phoneme are also phonetically similar to each other typically sharing the location, the stricture, or both. This condition is important when deciding on cases like [h] and [ŋ] in English. As will be detailed in the next chapter, the distribution of these two sounds is complementary: [ŋ] could be found word-finally and never initially while [h] is the exact opposite since it could be found word-initially but never word-finally. However, in addition to finding minimal pairs with other phonemes and the fact that both could be found intervocalically ('ahead', 'singer'), [h] and [ŋ] are phonetically radically different and do not share any feature. Furthermore, native speakers readily note the distinctions between phonemes, and thus would never accept [ŋəh] in place of [həŋ] as the realization of 'hang', while they might accept [kɪk^h] as a weird but passable alternative to a more standard [k^hɪk] for 'kick'. Hence, /h/ and /ŋ/ are phonemes in English while [k] and [k^h] are allophones of the same phoneme.

Determining whether two sounds serve as phonemes or allophones in a particular language, and identifying thus the predictable environment(s) for the allophones, belong among the basic aspects of phonological analyses.

Activity 7-4: Identifying phonemes and allophones in languages other than English

If you are a non-native speaker of English, consider your native language. The discussion above, and learning speaking English for many years, should give you a good basis for listing the mental patterns in the distribution of phonemes and allophones in your native language that are different from the English ones

If you are a native English speaker, you most likely spent some time learning another language. Based on this experience, do the same for this language: list the patterns in the distribution of phonemes and allophones that are different from the English ones.

So far we talked about the variability of English stops in the prevocalic positions in words. What about the final postvocalic positions? Activity 7-5 investigates this context.

Activity 7-5: Word-final stops

First, create several minimal pairs so that the phonemic contrast between /b/ and /p/ is at the end of short monosyllabic words like ‘nab’ vs. ‘nap’. Describe informally what you observe in your own speech based on proprioceptive and auditory senses.

Now record some of these minimal pairs in Praat, or explore ‘nap_nab.wav’ file in the companion. Zoom in to inspect the final stops in the ‘nap’-‘nab’ minimal pair. First explore the phonetic characteristics we already discussed: the presence of voicing during the closure and the type of the closure release.

Now, let’s test your friends or family members. Select **only** the interval [næ] in both words, cutting off the closure and release. Play them in succession to your test subjects and ask them if they can tell if [næ] comes from ‘nap’ or ‘nab’. Most likely, even non-expert listeners would be able to discriminate these two words despite missing the crucial discriminating sound. What phonetic cue allows for this discrimination?

You notice that the voiceless [p] tends to have a fully voiceless closure while the vocal cords maintain voicing at least for the initial part of the closure for [b], and maybe longer. In terms of the release, there is quite a lot of contextual variation in native speakers and we might neither see nor hear a salient release although some aperiodic noise might be emitted. Nevertheless, the contrast between [p^h] in the initial positions of stressed syllables and [p] in the word-final position is robust when comparing the intervals of the closure and release.

The crucial observation in activity 7-5 was that even without the cues in the closure and release, the duration of the vowel preceding the stop consonant provides salient cues for the voicing of English stops. Hence, the contrast between two phonemes, i.e. the *phonemic contrast*, is realized differently depending on the context. In word-initial prevocalic positions, the voice onset time is the major phonetic cue signalling the phonemic contrast. But in the word-final postvocalic positions, the duration of the preceding vowel is an important phonetic cue. This is another example of the *redundancy* of phonetic cues in signalling phonemic, or functionally important, contrasts. Previously we discussed, for example, the contrast between [i:] and [ɪ] and how multiple phonetic aspects (duration, lip spreading, tongue peripheral positions, tongue root activity) participate in cuing the phonemic tense vs. lax distinction in high front vowels.

So far, we used bilabial stops to introduce allophonic variability in cuing the phonemic voicing contrast in English. The other two pairs (alveolars /t/, /d/ and velars /k/, /g/) behave in the same fashion. Your skills with Praat allow you to check for yourself by examining minimal pairs like ‘tie’-‘dye’ and ‘sit’-‘Sid’ or ‘cod’-‘God’ and ‘pick’-‘pig’ either in your own speech or in recordings of different speakers. If a certain group of sounds behaves similarly in some systematic pattern of speaking, it is called a *natural class*. It is actually quite rare that some pattern of allophonic variation targets, or is triggered by, only a single sound.

Despite the fact that /t/ and /d/ are fully integral members of the natural class of stops in all aspects discussed above, English displays additional richness in allophones of the alveolar stops in certain contexts. Moreover, the allophonic variation in the English alveolars plays an important role in dialectal variation. Note, however, that the observed systematic behaviour of English stops above applies to most of the English dialects.

Activity 7-6: Check your t/d realization

Record your pronunciation of these words/phrases in English: ‘Adam’, ‘get’, ‘phonetics’, ‘better’, ‘rider’, ‘dentist’, ‘water’, ‘atom’, ‘bottle’, ‘writer’, ‘get out’. Read them as separate words with sufficient pauses in between but as naturally as possible. Then perform three steps.

First, introspect your own pronunciation, read aloud again and make notes about your t/d pronunciation in every word. How many variants do you feel you make? How do you make these allophones articulatorily? How are the words grouped into these variants?

Second, open your recording in Praat and click *Play* in the right-hand menu; that is, use only auditory and no visual cues. Do you need to revise your grouping? How do you hear the differences among the variants?

Finally, employ also the visual modality, open the file in the *Edit* window, zoom to selection for each word, and combine your tactile articulatory introspection, auditory perception and visual inspection of the waveform and spectrogram. Ideally, record another speaker or compare with a fellow student to assess the degree of variability and systematicity in t/d production. Finish with comparing your observations with the text below.

There is no ‘correct’ realization of the words in Activity 7-6 above and great variability among both native and non-native speakers exists (recall the descriptive approach to pronunciation in this book). The selection of the words was meant to elicit your allophones in two main contexts: (1) in between two *sonorant* sounds (vowels, nasals and approximants) at the boundary between a stressed and an unstressed syllable, and (2) at the end of the word preceding a pause. We will discuss these two contexts in turn.

But before we do, we make listening to words and their inspecting in Praat more efficient. Activity 7-6 above asked to record for the first time a greater number of words, eleven in total. The text below will ask you to listen to and compare various pairs or groups of words. With a single file of 11 words, navigating among

them might become laborious. Therefore, this is a good opportunity to introduce the annotating capabilities of Praat that make exploring speech more convenient and are essential for subsequent discussions in the book.

Activity 7-7: Annotate words in Praat recordings

In the Praat Object window, select the recording you made for Activity 7-6 above and then click *Annotate* in the right-hand menu. I recommend familiarizing yourself with a brief Annotation tutorial that concisely describes the available options. When annotating, ask Praat to create a TextGrid, which is a special Praat file that marks phonetic events as either intervals (actions of certain duration with a start and an end) or single temporal points. There are thus two types of tiers: *interval* tiers and *point* tiers. In the *Annotate* option select *To TextGrid...* and Praat then asks you to specify the names of the tiers you wish to make. In the first field you specify all the tier names, and in the second field the subset that should be point tiers (if you want only interval tiers, leave the second field empty). Since now we want to annotate words, which are intervals, let's say 'words' in the first field and leave the second field empty. Praat creates a new TextGrid object with the same name as the originally selected sound file.

Now select both the sound file and the textgrid either by dragging over the two files with your mouse or by clicking one file and then Ctrl-click on the other one, and select *Edit*. Setting both the waveform and the spectrogram visible, place the cursor at the beginning of the first word in either of the two top panels and then click on the (blue) circle at the cursor in the textgrid field shown with a black circle in Fig. 7.5 below. This creates a boundary that could be dragged with a mouse for fine location adjustments or deleted with *Boundary* → *Remove* (Alt-Backspace). Place another boundary at the end of the word. Clicking the newly created interval in the textgrid field you now can type the label for the interval. Continue in the same way of creating intervals and labelling them for all the words as in Fig. 7.5.

Remember: Praat never saves anything automatically so you have to save the TextGrid file manually either from the *Edit* window (*File* → *Save TextGrid as text file...*, shortcut Ctrl-S), or from the *Objects* window by selecting the TextGrid object and then *Save* → *Save as text file*.

Once your annotation is saved, you can easily move to desired intervals simply by Alt+right/left arrow and listen to them either by clicking the bars above/below the interval, or hitting the tab key for the selected interval. Consult other navigating options under the View menu of the Edit window.

With your textgrid annotation file, let's return to the realization of /t/ and /d/. In the first intervocalic context, at least three common varieties for native speakers can be identified. British speakers of the southern regions tend to pronounce /t/ and /d/ with the full alveolar contact. The contrast between /t/ and /d/ relies mainly on voicing during the closure for /d/ and its absence accompanied with weak aspiration of /t/. Note, however, that aspiration of /t/ in words like 'atom' or 'phonetics' is typically weaker than that of words like 'tie' for these speakers. Hence, there is

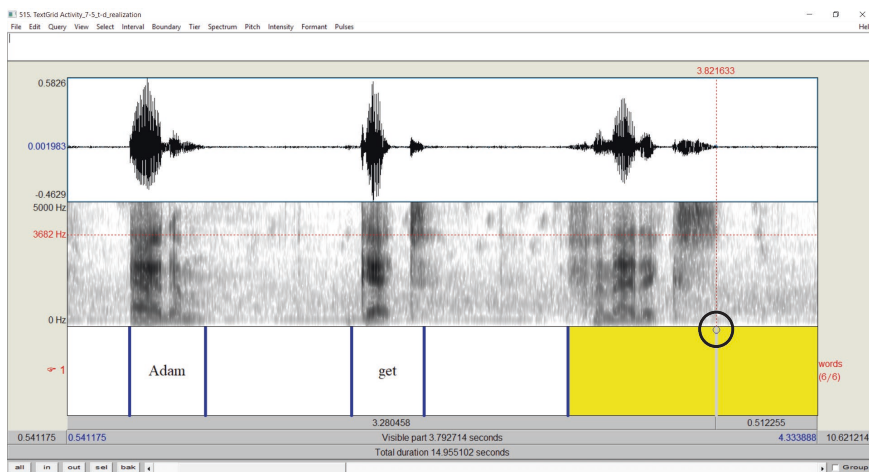


Fig. 7.5 Screenshots of annotating words in Praat. See activity 7-7 for explanation. The black circle marks the dot to be clicked in order to create a boundary

a clear contrast in pairs like *Adam-atom* or *rider-writer*. Speakers of British English in the northern regions have a tendency to significantly shorten the duration of /t/-closure and voice it, sometimes this realization is called a **tap** due to the brief alveolar contact and transcribed in IPA as [t̚]. The third major group, represented by American speakers, have a tendency to pronounce both /t/ and /d/ in this context identically as a very brief tap, but possibly even without a discernible alveolar contact. This is commonly referred to as a **flap** and IPA uses the symbol [ɾ]. Hence, for these speakers there would be no clear difference between ‘atom’ and ‘Adam’ and both would be said as [æɾəm].

The first /t/ of ‘dentist’ is an interesting case. How does your pronunciation here compare to [t] realizations in words like ‘atom’, ‘writer, or ‘phonetics’? Many native speakers exhibit quite rich variation from a fully articulated [t] to a tap or flap, or even a complete omission of /t/ altogether (possibly with a lengthening of the preceding /n/). This is because /t/ here is not flanked by two vowels as in all cases above but follows a nasal, which shares the same place of articulation with the /t/.

The second major group of words, that behave similarly to each other and at the same time typically differ from the first intervocalic context, is represented by three word-final tokens of /t/: ‘get’, ‘dentist’ and ‘get out’. Here, many speakers employ **glottalization** which is a partial or complete substitution of the tongue blade closure of [t] by the glottal stop [ʔ]. Figure 7.6 illustrates the coordination of the two crucial gestures participating in glottalization. At the glottis, vocal cords might be adducted resulting in vibration, which is indicated by the wavy squiggle. The straight single line shows a complete closure at the glottis, and the double line the opened glottis (vocal cord abduction). The activity of the tongue blade against the alveolar ridge is similarly marked: the straight line indicating closure and the double line open airwaves. The left panel of Fig. 7.6 shows a full alveolar

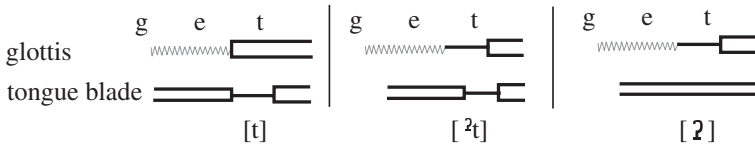


Fig. 7.6 Illustration of the articulatory activity resulting in t-glottalization. On the left, fully released alveolar closure, in the middle (pre-)glottalized [t] and on the right full glottal stop without alveolar closure. See text for detailed description

voiceless closure with the alveolar release. The middle panel shows a very common partial glottalization, or pre-glottalization, in which the glottal closure is produced during the vowel and crucially well before the tongue blade closure against the alveolar ridge is formed. The right panel shows a full glottal stop that replaced completely the alveolar closure and is followed by the glottal release. Activity 7-8 guides you to inspecting glottalization in Praat.

Activity 7-8: Visual cues to glottalization

First, summarize in your own words the crucial articulatory aspects of glottalization.

Second, review what we learned in Chapter 6 about the acoustic signatures of consonants in general and their place of articulation in particular. Brainstorm about what you expect to see in the spectrogram that indicates the glottal rather than the alveolar closure for /t/.

Third, an example from the middle panel of Fig. 7.5 is in the file ‘start_partial_glottalization.wav’ in the book companion at the end of ‘start’. An example of a complete glottal stop is in ‘but_glottal_stop.wav’. Inspect the correspondence of the visual cues to your expectations. Then read the text below and compare.

In the review and brainstorm sections of Activity 7-8, you probably mentioned that a salient closure for the voiceless stops like [t] should show as a flat-like line in the waveform, and a whitish region in the spectrogram. Glottalization involves a partial or full replacement of the alveolar closure made with the blade of the tongue by the closure at the vocal cords. Recall that formant transitions are the major acoustic indicator of the place of articulation of the consonants. Hence, a full glottalization as in the right panel of Fig. 7.6 should show very limited formant transitions at the end of the vowel (since the closure is at the glottis and is not made with the tongue) and a closure typical of other voiceless stops. This is what you can observe in ‘but_glottal_stop.wav’. Additionally, (partial) glottalization is commonly realized with an interval of irregular vocal cord vibrations visible as vertical striations in the spectrogram. These are shown in Fig. 7.7 in the file ‘start_partial_glottalization.wav’ from the book’s companion. The irregular vibration preceding the complete closure results from the stiffening of the vocal cords, which is necessary for the glottal stop.

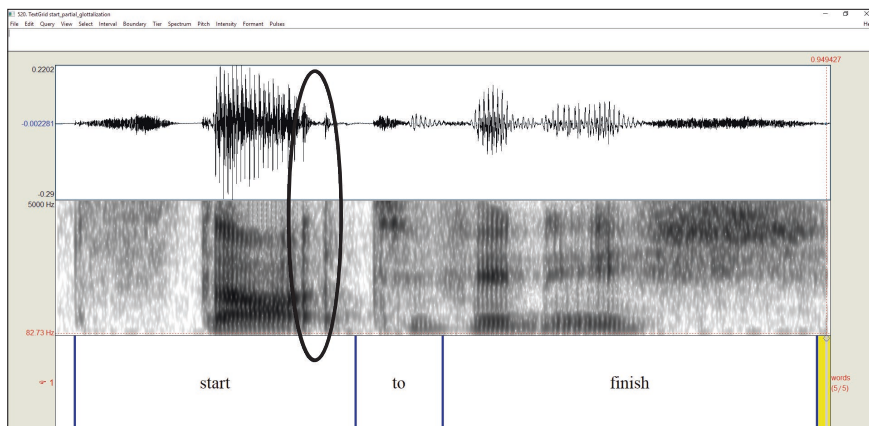


Fig. 7.7 Glottalization of /t/ in ‘start’ marked with the ellipsis

Note how salient allophonic variation regarding word-final /t/ realization results, again, from the systematic pattern, i.e. a speaking habit, of temporal coordination of articulatory gestures. The difference between a ‘regular [t]’ and the glottalized [ʔt] schematized in the left and middle panels of Fig. 7.6. Is due to different timing between when the glottal closure stops modal voicing and when the tongue touches the alveolar ridge.

Check in Praat for any evidence of glottalization in the pronunciation of the words like ‘get’, ‘dentist’, ‘get out’ in your own recording of these words for Activity 7-6? What difference do you see between the first and second /t/ in ‘get out’? If you see any, try to describe it in your own words. What about ‘bottle’? Is your /t/ fully voiceless with a salient alveolar closure, or a tap, or flapped, or combined with partial glottalization, or completely realized as a glottal stop as in the Cockney dialect [bʔl]?

Returning to our discussion of grouping words in Activity 7-6, do you fall into one of these broad categories for these two contexts? You should also check for consistency, which is especially the problem for even proficient non-native speakers. If there is a tap/flap in words like *better* or *atom*, is it also observable in *phonetics* and *writer*? And are you consistent in how you realize the final /t/ in ‘get’ and ‘get out’?

This discussion of allophonic realizations of /t/ above requires a refinement of our characterization of allophones occurring in the complementary distribution. Most likely you have observed that more than one allophone is acceptable for each environment discussed above. Hence, allophones can also be found in **free variation**, in which multiple variants are acceptable in the same context. For example, [ʔ] is in free variation with [t] in ‘bottle’ or ‘get’, since [bʔl], [bʔl], or [get], [geʔ] are all acceptable, but in complementary distribution with [t^h] since the realization of ‘ten’ as [ʔen] is normally not acceptable.

Although (partial) glottalization is most common with /t/, native speakers might employ it also for the other two voiceless stops in pre-pausal positions or in words where one stop is followed by another, as in ‘captain’ or ‘actor’. These stops are also commonly *unreleased* meaning that the closure ends without an audible release of air trapped behind the closure. This is in stark contrast to non-native speakers of languages like Italian or French who have a tendency to release word-final stops quite audibly with a schwa-like vowel. The unreleased stops are marked with the [̚] diacritic in narrow transcription: [k^hæp̚tɪn].

Table 7.1 outlines the major allophones, their contexts and example words for the phonemes /t/ and /d/. A similar table can be constructed for each phoneme. It is important to note several things. First, tables like this give only an illustration of the variability that is dependent on the level of specificity or phonetic detail. In other words, the continuum and multidimensionality of phonetic features may be cut to a chosen set of discrete units. These correspond to discrete-like subroutines that we discussed in Chapter 2 and right now they correspond to allophones. If we explored more contexts, interactions of various phonetic features, degrees of redundancy and other domains, we could come up with more patterns of systematic variation in allophones. Also, there is huge dialectal and even individual variation depending on the register (e.g. formal or not) or plain speech rate. Hence, tables of allophones like 7.1 should definitely not be taken as prescribing how to say these sounds but as capturing only a certain level of phonetic detail that is greater than phonemic distinctions and smaller than physical measurements of articulatory or acoustic events.

The material in this section is crucial for building awareness regarding the mental and physical aspects of the knowledge underlying the speaking habits of native speakers introduced in Chapter 2. We saw examples of how fine coordinations of articulatory activities result in salient and discernible acoustic and perceptual cues. Native speakers know unconsciously how to make them, and importantly, in which contexts to deploy them and how they function in cuing meaningful contrasts. Non-native speakers might need to acquire both the fluidity and automaticity of these physical movements (like sports, music or crafts in Chapter 2) and the knowledge when to deploy them.

7.3 Other Common Allophonic Patterns

In the previous section we talked about the aspiration and voice-onset-time of stops that were always followed by a vowel (‘pie’, ‘spy’, ‘bye’, ‘water’, ...). The specific coordination of the stricture formation and vocal cord vibration results in the aspiration of stressed syllable-initial voiceless stops. While traditionally this is referred to as aspiration in the case of the following vowel, the term used for approximants following these stressed initial stops is partial or complete *devoicing*. Hence, /l, r, j, w/ immediately following these stops are (partially) devoiced in words like ‘clay’, ‘cry’, ‘cue’, or ‘choir’: [kʰleɪ], [kʰɹaɪ], [kʰjuː] and [kʰwɑɪə(ɪ)].

Table 7.1 Summary of alveolar stop allophones

Phoneme	Allophones	Contexts	Example	Dialects
/t/	[t ^h]	Beginning of stressed syllable	<i>tie</i> ,...	all
	[t]/[ɾ]	Between two sonorants starting unstressed syllable	<i>atom</i> ,...	Northern BE, GA
	[ʔ]/[ʔ̚t]/[t̚]	Word-finally or preceding another consonant	<i>cat</i> , <i>bottle</i> , <i>meatball</i> ,...	varied
	[∅] ^a	Following /n/ when starting unstressed syllable	<i>dentist</i> , <i>Atlanta</i>	varied
	[t]	Elsewhere; following [s] in stressed syllables, starting unstressed syllables	<i>stay</i> , <i>today</i> , <i>captain</i> ,...	all
/d/	[d̚]	Word beginnings after a pause	<i>die</i> ,...	all
	[ɾ]	Between two sonorants starting unstressed syllable	<i>Adam</i> ,...	GA
	[d]	Elsewhere	<i>redeem</i> ,...	all

^aThis symbol represents the absence of a sound, a zero realization

This is another use of this IPA diacritic – a small circle below the regularly voiced sounds – in narrow transcription. For many speakers, approximants are also devoiced when they follow voiceless fricatives, and especially the sibilants, in words like ‘slight’, ‘shriek’, ‘Sue’ or ‘sweet’.⁴ Note again how approximants in this devoicing pattern behave together as a natural class.

Another allophonic pattern explored in several experimental studies is the variation in producing /l/. Activity 7-9 investigates this pattern.

Activity 7-9: Dark/clear /l/ in your speech

First, let’s explore your own speech. Say these six words/phrases and self-observe: ‘leap’, ‘peel’, ‘delete’, ‘peel it’, ‘milk’ and ‘peeling’. Inspect both the articulatory production of the six /l/ tokens as well as how they sound to you when you say them. Do you observe any differences? Are there some group(s) within which the /l/s sound more similar and contrast with /l/s in other group(s)?

⁴For some speakers, mainly American English but also some Brits, [j] in words like ‘Sue’ would not be present and it would also not be present in words like ‘tune’, ‘during’, ‘news’, ‘presume’, or ‘luke-warm’. But in words like ‘few’, ‘pew’, ‘cue’ or ‘argue’ [j] would follow the consonant. This is another mental pattern (=habit) that groups the natural class of the alveolar consonants to contrast with other consonants in that the former are not followed by [j] when [u:] follows while the latter always are.

Now record your reading of these words and create a Textgrid identifying the starts and ends of each word in the signal and labelling the intervals (if unsure, go back to Activity 7-7). Don't forget to save both the sound and the Textgrid.

Recall from Chapter 6 that F2 (both its lowering and decrease in its energy) is the major acoustic signature of /l/. Identifying the onset and offset of /l/ might be difficult especially for the onset of 'peel' or 'milk' but make your best educated guess, using both visual and auditory information. Now measure F2 during /l/ in a similar way we measured formants for vowels: Select *Formant* → *Show formants*, adjust if needed for tracking errors (Activity 5-14), and place the cursor around the middle of /l/ where formants are as steady and reliable as possible. Get the F2 value (either from Formant listing or with the shortcut *F2* for *Get second formant*). Keep track of F2 values for all six tokens of /l/.

How do the extracted F2 values correspond to your self-observation? If you observed some grouping of the words, is it reflected in the F2 values?

For many native English speakers, /l/s that are followed by a vowel (and occur in the beginning of a syllable) are referred to as a 'clear' /l/, transcribed as [l] and have a different quality than /l/s followed by another consonant or a word boundary that are referred to as 'dark' or 'velarized' /l/s narrowly transcribed as [ɫ]. Hence, the /l/ in 'lip', 'bleed', or 'believe' is commonly *clear* and the one in 'pill' or 'pilgrim' is *dark*. This issue has received great attention in laboratory phonology studies and there is great variation both dialectally as well as depending on the prosodic context.⁵ For example, what happens when /l/ is word-final but followed by a word starting with a vowel as in 'peel it'. Additionally, many dialects might *vocalize* dark /l/s in which the tongue tip contact against the alveolar ridge is not made and lip rounding is present (we have already mentioned [mɪwk] realization of 'milk' in Chapter 6).

Articulatorily, the dark and clear allophones of /l/ stem again from systematic patterns of coordination between articulatory actions. In Chapter 6, we said that a /l/ requires tongue tip raising and tongue body retraction. If the tongue tip raising is roughly simultaneous with the tongue body retraction, /l/ sounds light and if the tongue tip raising lags behind the tongue body retraction (or the tongue tip gesture is completely omitted for the vocalized variants), /l/ sounds dark. Hence, due to temporarily and spatially more robust retraction of the tongue body, greater F2 lowering indicates a darker or more velarized realization of /l/.

The expectation for F2 from Activity 7-8 would thus be that 'leap' and 'delete' would have the highest values, 'peel' and 'milk' the lowest and 'peel it' and 'peeling' somewhere between. Of course, your dialect if you are a native speaker, and the native language if you are a non-native speaker, might greatly affect the results. For example,

⁵For phonetic studies, see for example a recent paper by Turton (2017) that reviews relevant literature and presents interesting new data. Regarding dialectal variation, Collins and Mees (2013) for example mention that many Welsh or Irish speakers only use clear [l] while many Scottish or American speakers only use dark [ɫ].

Scots or Americans tend to produce all /l/s in general as dark while native speakers of German or Hungarian have a tendency to use the light variant everywhere.

Activity 7-10: Dark/clear /l/ in a recording

To show how the speech of a single speaker might vary, I concatenated several phrases from a longer recording that should all have dark /l/ following our generalization from above. They are in ‘dark-clear_ls.wav’ of the book companion. If you open it together with the associated TextGrid, you see the 6 phrases identified in the TextGrid that have 7 target /l/s since the second one has two (‘initial’ and ‘well’). Investigate this file both auditorily and experimentally (by extracting F2 values from the middle of the /l/s) and summarize your findings. Can you speculate why the speaker varies dark and clear /l/s in these phrases? In other words, is there a systematic explanation for this variation, or is it haphazard? Note that there is no ‘right/correct’ answer here and none will be provided in the text below.

The final salient allophonic variation that we mention here concerns vowels and has already been introduced in the previous section when discussing Activity 7-5 and the contrast ‘nab-nap’. We observed that ‘the duration of the vowel preceding the stop consonant provides salient cues in the voicing of English stops’ so that [æ] of ‘nab’ was significantly longer than [æ] of ‘nap’. This has been called *vowel shortening* or *pre-fortis clipping*. Let’s examine this definition more closely in Activity 7-11.

Activity 7-11: Checking vowel shortening

Design a list of words in which voiced or voiceless stops are followed by either a word boundary or vowels extending the already analysed ‘nap-nab’ minimal pair. For example, ‘phase-face’, ‘phasing-facing’, ‘cab-cap’, ‘cabinet-capital’. Similar to Activity 7-5, record these words and measure the duration of the vowel preceding the target stops. Is the vowel shortening before voiceless consonants supported in each pair roughly equally or do you observe additional systematic grouping of the pairs? If so, how would you append the definition of vowel shortening we gave above? Compare with the text below.

Activity 7-11 was guiding you to discover two improvements over the original definition of allophonic vowel shortening in the previous paragraph. First, you noticed that vowel shortening does not apply to stops only but also to fricatives. Hence, vowel shortening involves all consonants differentiated on the fortis-lenis dimension and voicing. The phonological term for them is *obstruents* and they include fortis /p,t,k,tʃ,f,θ,s,ʃ,h/ and lenis /b,d,g,dʒ,v,ð,z,ʒ/. All other consonants in English are *sonorants*.

Second, you might have also observed that vowel shortening was much more robust in measurements, and salient auditorily, in the ‘phase-face’ and ‘cab-cap’ pairs than in the ‘phasing-facing’ and ‘cabinet-capital’ pairs. The crucial difference between the two groups is that the relevant consonant, the voicing of which

triggers the difference in the preceding vowels, either appears with the vowel in the same syllable, or might begin another syllable. Our definition of allophonic vowel shortening can thus be more precise if it says that the duration of the vowel preceding the stop consonant in the same syllable provides salient cues for the voicing of English obstruents.

Note that the two groups from Activity 7-11 are different on the morphological dimension: while in ‘phasing-facing’ the /z/ and /s/ are at the boundary between the base and the suffix ‘-ing’, in ‘cabinet-capital’ the /b/ and /p/ are not adjacent to such a boundary. The importance of syllables, the affiliation of consonants into syllables, and the effects of morphological structure on various speaking patterns will be explored in subsequent chapters.

► Advanced Section: Relationship between two allophonic processes

Before we move to discussing syllables in more detail, we finish the chapter on allophones with Activity 7-12 exploring the complexity of the cognitive system underlying our speech. We have seen in this chapter that speaking English involves many subconscious habits and systematic generalizations that certain allophones are produced systematically in well-defined contexts. The following activity inspects this cognitive system further and explores the relationship between two allophonic processes.

Activity 7-12: Further characteristics of mental habits

Consider first the realization of the vowels in the first syllables of these two pairs: ‘writer-rider’, ‘seating-seeding’. Are the durations of the vowel pairs comparable? Should they be?

Now, if you are a native speaker of American English, or have access to one, ask them to say these two pairs and record their speech (you might re-order the four words and ask them to read the list several times). Measure the duration of vowels preceding t/d in Praat. Recall the allophonic flapping in American English producing identically sounding ‘atom’ and ‘Adam’. Are the two pairs of this activity also *homophonous*, i.e. do they sound identical? Speculate on the generalizations in the minds of American speakers that would explain their difference or similarity based on what you measured. How is this similar/different from the pairs in Activity 7-11?

Do not worry if this is complex or confusing, any study of real natural phenomena usually is. Recall that the goal is not to provide correct answers but build your awareness and understanding. Consult your speculations with the text below.

My fundamental assumption is that while the alveolar consonants following them are identically realized as flaps [ɾ], the first vowels in ‘rider’ and ‘seeding’ are longer than those in ‘writer’ and ‘seating’ in the speech of most Americans. The brief remarks below provide a potential explanation for this observation. We have two processes here and both of them have been discussed in this chapter. One of them is flapping, in which a /t/ or /d/ in between vowels starting an

unstressed syllable become a voiced flap [ɾ]. It should thus apply in all four target words from Activity 7-12. The other one is vowel shortening, in which a vowel is shortened if followed by a voiceless obstruent in the same syllable. It should thus not apply since we have the voiced flap in all four words. Where does the assumed difference in the pairs come from then?

The most common treatment of this variability in the duration of vowels in pairs like ‘writer – rider’, ‘seating – seeding’ comes from specific **ordering** of the two processes in the minds of the native speakers. If, as assumed in the previous paragraph, flapping applies first, and then pre-fortis clipping, we would not get the different lengths of the vowels. If, however, pre-fortis clipping applies first within the stems, vowels in ‘write’ and ‘seat’ are shortened and those in ‘ride’ and ‘seed’ are not. Adding a vowel-initial suffix results in flapping in all four words. This ordering thus predicts that the first vowels in ‘rider’ and ‘seeding’ remain longer than the vowels in ‘writer’ and ‘seating’.

There are several other factors that might affect the data, including the quality of diphthongs [aɪ] and [aʊ] known as Canadian raising, and morphological levels (stems, stems+ affixes) that we will not discuss here. But the crucial observations for us from Activity 7-12 is that speaking involves complex habitual activity of fine coordination of articulatory actions (how to smoothly flap alveolar stops or shorten vowels), and complex mental habits of knowing in which environments to deploy these articulatory habits, and what their precedence relationships are.

Exercises

- 7-1 Create a Table similar to 7.1 for other phonemes, e.g. /l/, or /æ/.
- 7-2 For a small project, find more examples of fortis/lenis obstruents in the intervocalic positions with or without a morpheme boundary similarly to those in Activity 7-11. For example ‘nab-nap’ & ‘nabber-napper’ with a morpheme boundary compared to ‘lab-lap’ & ‘rabbit-rapid’ and check in Praat if pre-fortis clipping in the speech of native speakers is sensitive also to the presence of this morpheme boundary following the target obstruent.

References

- Collins, Beverley, and Inger M. Mees. 2013. *Practical phonetics and phonology: A resource book for students*. Abingdon: Routledge.
- Turton, Danielle. 2017. Categorical or gradient? An ultrasound investigation of /l/-darkening and vocalization in varieties of English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8 (1): 13. <https://doi.org/10.5334/labphon.35>.



In this chapter we will

- discuss the articulatory, perceptual and psycholinguistic evidence for syllables and what it suggests about the nature of this speech unit in the minds of English speakers
- explore the distributional patterns in English syllables
- investigate how words are divided into syllables

8.1 Introduction

In the previous chapter syllables and word stress were mentioned informally to enable a more detailed exploration of allophonic variation in English. We saw that the position of consonants in the syllables greatly affects their allophonic realization. For example, /t/ is aspirated if initial in stressed syllables, potentially flapped if beginning an unstressed syllable, or very likely glottalized if ending a syllable. In this chapter we explore further how our unconscious habits and mental knowledge related to syllables can be approached through the examination of our speaking patterns. We continue investigating our phonetic and phonological knowledge with the help of hands-on activities to assist in bottom-up awareness building.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_8) contains supplementary material, which is available to authorized users.

8.2 Mental Knowledge of Syllables

In Chapter 2 continuity was presented as a hallmark of speech supported also by difficulties with separating speech into individual sounds. Despite this, we proceeded with extended discussions of the reality of our mental phonological knowledge about these sounds. We investigated actual speaking and how the contrasts and differences among sounds can be observed either through the articulatory or acoustic data. Now the journey continues by putting the consonants and vowels back together again and exploring how *syllables* play a role in our cognitive patterns of speaking.

Activity 8-1: Counting syllables in native language (I)

If you are a native speaker of English, think of longer words like ‘pandemonium’ or ‘responsibility’ and tell the number of syllables as fast as possible. The non-native speakers might in addition to this, also think of several long words in their native language and count the number of syllables in the same way.

How did you determine the syllable count? Did you know right away when reading the word on the page? Did you have to say the word?

Now try to count the number of segments (vowels and consonants) in those words. How does this compare to syllable counting in terms of difficulty and the approach?

Separating speech into syllables is one of the fundamental skills we gain during language acquisition. Activity 8-1 shows that syllable count is not readily available in longer words. You most likely mentally said the word syllable-by-syllable, or even mumbled the word’s articulation silently and simultaneously count the syllables (possibly using your fingers) to reach the correct count of 5 and 6 for ‘pandemonium’ and ‘responsibility’ respectively. The same applied to counting syllables in other languages for native speakers of those languages. Note how for longer words this competence is linked to the actual speech production by either mumbling or silently saying the words ‘in the head’. Nevertheless, the task is relatively easy compared to counting the segments. Obviously, there are more segments than syllables; 11 and 14 for ‘pandemonium’ and ‘responsibility’, respectively. People usually prefer doing this visually and inspect the letters on the page. Using the same approach as with syllables and counting segments ‘in your head’ mentally, or with your fingers, imposes quite heavy cognitive load.

Pre-literate children or illiterate adults easily divide words into syllables by clapping hands and research consistently shows that dividing speech into individual sounds is a skill acquired later, linked to reading skills, and likely to be linguistically more complex than dividing speech into syllables. Additionally, syllables are indispensable in speech production based on data in speech errors or child babbling. In psycholinguistic experiments the perception of syllables is more salient than that of segments. Also, the syllabic writing systems occur more naturally in the history of writing than the writing systems based on segments. Hence, the primacy of syllables in speech over individual segments is supported at many levels. If you

felt counting syllables is easier than counting sounds, as I did, these observations provide explanations for this intuitive mental knowledge that we possess.

A syllable is thus a basic unit of speech, possibly more fundamental than individual vowels and consonants, and evidence comes from speech production, perception, evolution and acquisition.

Following upon our ability to count syllables, Activity 8-2 explores also certain limitations.

Activity 8-2: Counting syllables in Japanese or Berber

In the text we talked about syllable counting in words like ‘pandemonium’ or ‘responsibility’, where most speakers of English, both native and non-native, would provide unified answers. Listen now to the Japanese word すき meaning to like or love. If you have a speaker of Japanese around you, try to ask him/her to say the word in normal speed and if not, listen to the sound file ‘Japanese1_8-2.wav’ in the book companion. How many syllables do you hear?

If you are not proficient in Japanese, you most likely hear a single syllable. However, if you are somewhat proficient in Japanese, your answer is two syllables (some of you might have guessed from the Japanese spelling above with two symbols). You can confirm by listening to ‘Japanese2_8-2.wav’ in which this word is pronounced carefully and slowly. Sound files for Japanese were extracted from this youtube video: <https://www.youtube.com/watch?v=1JtPTTtAYm8>.

For a taste of more complexity, there are words in less known languages like Tashlhiyt Berber or Salish languages like Bella Coola known in phonology literature for very complex syllabification systems and complex consonant clusters. Counting syllables there for speakers not proficient in these languages is an arduous task (e.g. ‘ssrksxt’ ‘I hid him’ in John Coleman’s page <http://www.phon.ox.ac.uk/jcoleman/ssrksxt.wav>).

Hence, the skill of syllabifying words is not completely universal but just like allophonic patterns discussed in the previous chapter, conforms to language-specific systematic patterns and speaking habits that are culturally dependent. More specifically, the reason why we tend to perceive the Japanese two-syllable word as a one-syllable one is the difference in the *phonotactic patterns* between these two languages. Japanese is predominantly a ‘CV language’. Consider any Japanese words or proper names that you can think of. They most likely conform to the CV pattern in which a single consonant is followed by a single vowel (Kawasaki, Murakami, Hirohito,...). Having two adjacent consonants is very rare, they are limited to homorganic nasal+stop combinations (Shinji, kanda, etc.), and these clusters never begin a word. Moreover, Japanese has an allophonic pattern where vowels occurring between voiceless consonants are devoiced. The unconscious knowledge of all these patterns suggest to native Japanese speakers that something that sounds like [ski] has to be, in fact, a two-syllable word with a weak devoiced vowel separating ‘s’ and ‘k’: [sʰki]; [ɯ] is the IPA symbol for a back unrounded high vowel similar to English [ʊ]. And this would apply also to a word they never heard.

On the other hand, /sk/ is a perfectly common consonant cluster in English and there are many words that start with this cluster ('skip', 'school', 'scan', etc.). Hence, for native speakers of English and many languages with similar words beginning with 'sk', this Japanese word would naturally be perceived as a one-syllable word.

Hence, *phonotactics* characterizes permissible combinations and distribution of sounds, and their organization into larger units like syllables and words. Although English phonotactics allows for more complex patterns than Japanese, there are many languages that allow for much more or much less complex patterns than English. Activity 8-2 mentioned the complexity of Berber or Bella Coola, but many European languages, for example Polish or Albanian show more complexity, and Italian or Finnish less complexity, than English.¹

The disagreements in terms of the number of syllables, however, might also arise in English for native English speakers. Activity 8-3 explores.

Activity 8-3: Counting syllables in native language (II)

Say the following words and determine their syllable count: 'feel', 'fire', 'film', 'vegetable', 'actually' and 'brightening'. Ask some other speakers, possibly both native and non-native, to do the same and compare. If you do not have other speakers available, try to hypothesize how English speakers might disagree in terms of syllable counts in these words.

Why do you observe, or hypothesize, variability in these words while the words like 'responsibility' typically produce uniform judgements of the syllable count?

Native speakers of English might disagree on the number of syllables in words like 'feel', 'fire' or 'film'. Depending on their regional accent or register, these might be considered to be one- or two-syllable words. A slightly different case are words like 'vegetable' or 'actually' that might be pronounced with three or four, potentially even two, syllables, depending on style and formality, speed, or whether the speaker is a native English speaker or not. Yet there are other types of words, such as 'brightening', that some people might say has two and other people three syllables irrespective of style or speed.

8.3 Sonority

The observations of cross-linguistic and culturally dependent variation in phonotactic patterns suggest that it might not be possible to identify a universally applicable phonetic basis of a syllable. Nevertheless, phonetic considerations play an important role in understanding how syllables function in the cognitive system of speech. The phonetic dimension commonly employed for describing syllables

¹However, both Italian and Finnish, and also Japanese, have true geminates, or long consonants, that partially balance the lack of complexity in consonantal clusters.

is *sonority*. It can be described as the amount of acoustic energy emitted when sounds are produced in constant loudness. Informally, higher sonority corresponds to sounds carrying far while lower sonority to carrying not so far. Individual sounds vary in their sonority and thus an ordering of sounds along this dimension creates a *sonority hierarchy*. Activity 8-4 invites you to explore this on your own before reading further.

Activity 8-4: Inspecting sonority hierarchy

I used Praat to record my production of individual sounds like [e], [m] or [ʒ] with consonants not released into a schwa, and trying to keep a stable f_0 . I then extracted each sound with the same duration (300 ms) and made Praat scale the intensity so that each sound has identical average intensity (65 dB). Open six such sounds from the book companion into Praat: {e.wav, w.wav, l.wav, m.wav, zh.wav, b.wav}.

Now, please carry out an informal experiment in which you always select a pair from these six sounds in Praat (so that 2 sounds are highlighted in blue), click *Play*, and with headphones listen to this pair. Note which of these two sounds has greater carrying power, or the amount of acoustic energy. Repeat with multiple pairs until you can form a hierarchy, i.e. an ordered list, ordering the sounds from the most sonorous to the least sonorous. Alternatively, you might ask a friend to stand in a fixed distance and play these pairs to him/her and ask him/her to make the judgements.

When finished, compare with the text below.

If experiments similar to the task in Activity 8-4 are run rigorously and with many subjects, the results would most likely point to a continuum between the most sonorous and the least sonorous sounds, or to a sonority hierarchy, along the lines in Fig. 8.1.

Additionally, further division is possible; for example, in vowels, low vowels are more sonorous than high vowels, in approximants, glide s (j, w) would be more sonorous than liquids (r, l), and in obstruents (fricatives and stops), the voiced ones are more sonorous than the voiceless ones. However for our purposes these five major levels are sufficient.

It is likely that the sonority hierarchy in Fig. 8.1 does not correspond completely to your own from Activity 8-4. For example, there were no voiceless stops. Also, fricatives, partly due to the manipulation of intensity, and partly due to the strong noise component, might be perceived as more sonorous than some sounds

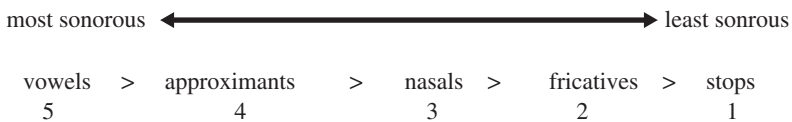


Fig. 8.1 Sonority hierarchy with major classes of sounds. The numbers correspond to sonority levels used in the remainder of the chapter

higher on the hierarchy. Importantly, these loosely defined phonetic scales serve as a basis for phonologically defined hierarchies that explain some of the speaking patterns we discuss below.

Activity 8-5: Sonority in syllables

With the sonority hierarchy in mind, go back to the words inspected for syllable counting in Activities 8-1 to 8-3. Try to use numbers from Fig. 8.1, or chart them in a graph to see what kind of sonority profiles you see. Also, check the sonority profiles in potential disagreements in syllable count explored in Activity 8-3. Does the notion of sonority provide possible explanations for cases when speakers agree vs. those with disagreements?

Charting the sonority of sounds illustrates how our skill for syllable counting, mentioned above, relates to sonority. Figure 8.2 shows that *sonority peaks* have a very close correspondence to the number of syllables.

The top left panel of Fig. 8.2 shows sonority of a prototypical English syllable [kræmp]. We see that the first two consonants rise in sonority from level 1 to 4 and the peak is reached with the vowel. The two consonants following the vowel fall in sonority from level 3 for the nasal to level 1 for the stop. In this ideal example, a single sonority peak corresponds to a single syllable and the sonority profile of the syllable corresponds to a nice ‘hump’.

The top right panel shows the disyllabic word ‘scramble’. The core observation is that if a sonority peak is sufficiently salient, it can represent a separate syllable despite the fact that this syllable does not contain a vowel. Here, the peak

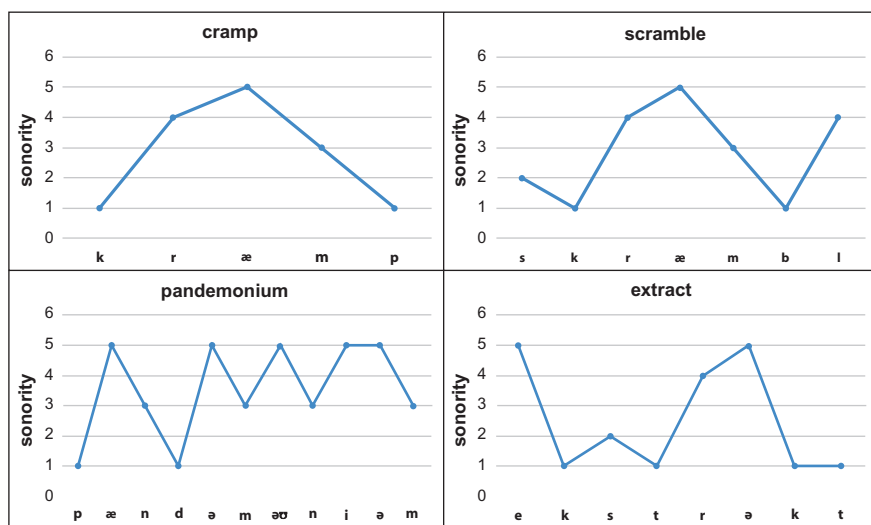


Fig. 8.2 Visualization of sonority in four English words

associated with the word-final [l] is sufficient to constitute the second syllable whereas the first mini-peak associated with the initial [s] is not. If a consonant forms a sonority peak associated with a syllable, we call it a *syllabic consonant*. In English, the most common syllabic consonants are [l, n, m] in words like ‘people’, ‘button’ and ‘rhythm’, respectively, and always in the unstressed syllables (see next chapter). In narrow transcription these words are: [p^{hi}ː.p̩], [bʌ.t̩n̩] and [rɪ.ð̩m̩]; the boundaries between adjacent syllables are marked with a full stop and syllabic consonants with a vertical stroke under the target consonant. There is a lot of variability in the pronunciation of these syllables and commonly a continuum between a full schwa realization and a complete absence of the schwa can be observed.

Alveolar fricatives like [s] are in many respects special in terms of the link between their sonority and their syllable affiliations. For example, the sound eliciting silence in many languages is ‘psst’, which would be considered a syllable with [s] as the sonority peak.

Activity 8-6: Syllabic consonants

In this awareness-building activity, select 4–5 words with potentially syllabic consonants like ‘people’, ‘button’ and ‘rhythm’, ‘nation’ and ‘bottle’. First, say them with and without a schwa in the second syllable; e.g. [p^{hi}ː.p̩] and [p^{hi}ː.pə]. Introspect the different coordination between crucial articulatory actions required for these two realizations. Hence, in the case of ‘people’, what is the coordination between lip opening and the formation of the tongue blade alveolar contact? Is it different in homorganic consonants like in ‘button’?

Second, record the variable realizations of these words as you say them with schwa, without it, and maybe somewhat between. Open in Praat with the spectrogram and search for visual evidence of the absence or presence of the schwa. Measure schwa duration. Strengthen the link between the kinesthetic information from your articulators, and the visual and auditory information provided with Praat.

For example in ‘nation.wav’ in the companion, the first token has a schwa of roughly 60 ms while the second one has just the syllabic [ŋ].

The bottom left panel of Fig. 8.2 shows sonority in a longer word ‘pandemonium’ mentioned previously. Here, all the peaks define syllables. While the first three peaks/syllables are non-problematic, the final two vowels [ɪ ə] are adjacent to each other. Since for most native speakers of English this word has five syllables, these two vowels are associated with different syllables. Note that the English inventory does contain the diphthong [ɪə] that makes a single syllable in words of the NEAR lexical set (Chapter 5). Hence, this panel shows that even non-salient sonority peaks might constitute syllables in certain environments. However, this non-saliency is behind the fact that some speakers might consider ‘pandemonium’ as a four-syllable word.

These ‘weaker’ peaks thus might be in some situations merged and perceived as one. This is probably also the reason why words like ‘actually’ in Activity 8-3

are commonly produced with three syllables by native speakers. This is then one possible strategy if there are adjacent peaks of very similar sonority. Another strategy might be to make the peak even more salient. For example, words like ‘film’, also mentioned in Activity 8-3, might be produced as [fi.ləm] to make a clear peak rather than a sequence of relatively high sonority [ilm] in a single syllable.

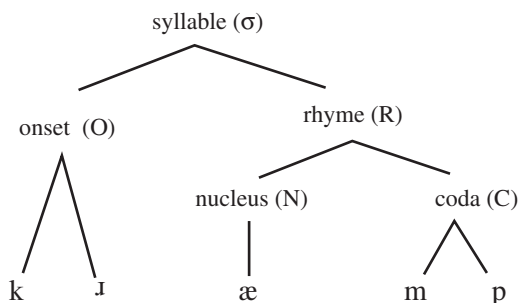
Finally, the bottom right panel of Fig. 8.2 charts sonority in ‘extract’. It is a two-syllable word with clear sonority peaks associated with the two vowels and a minor peak for ‘s’ that does not licence a separate syllable, as we saw with ‘scramble’ above. While the syllable count is not problematic, the precise location of the boundary between the two syllables is less clear. We will return to this in Sect. 8.5.

8.4 English Syllables

Before we tackle the location of syllable boundaries in English in more detail, let’s introduce the basic terminology for syllable structure to facilitate further discussion. The sonority peak is the *nucleus* of the syllable and is the only required element. Hence, there are syllables containing just the nucleus like ‘eye’ [ai] or ‘awe’ [ɔ:]. The consonants preceding the nucleus form an *onset*. In English, there could be up to three onset consonants as in ‘scratch’ or ‘split’. The consonants following the nucleus form a *coda* and English allows up to four consonants in a coda in words like ‘(two) sixths’ [sɪksθs]. The nucleus and the coda together are commonly referred to as *rhyme*, or *rime*, that forms the basis of rhyming in poetry. In terms of sonority, onsets typically rise in sonority while codas fall in sonority. The syllable structure of ‘cramp’ from Fig. 8.2 is displayed in Fig. 8.3.

Activity 8-7 below takes our discussion back to the mental knowledge of phonotactic patterns in English.

Fig. 8.3 Syllable structure in ‘cramp’



Activity 8-7: Creating catchy names for a new product

Imagine you are a creative director of a new international hi-tech start-up company that wants to launch a revolutionary new product. Your task is to come up with a new brand name that will catch on easily in a similar way to relatively recent additions to English like ‘tweet’ or ‘meme’. Here is the final list of 10 suggestions from your international focus group: ‘v_lk’, ‘b_lick’, ‘l_kap’, ‘s_nid’, ‘m_pin’, ‘z_reen’, ‘c_royd’, ‘k_mene’, ‘d_long’ and ‘t_kar’. Which of these would work for markets with native speakers of English?

One way of approaching this task is to group the words into three broad categories: those that would not work (reject), those that are possible but not ideal (backburner) and those that proceed to the final presentation to the board of directors (shortlist). Once you have the 10 words grouped this way, formulate your reasons/justifications for your decisions. Continue reading the main text only after you verbalize your reasons/justifications for the grouping.

Activity 8-7 is framed by the discussion of syllable phonotactics and thus it was not difficult to deduce that the deciding factor for the elicited grouping is the distribution of consonants starting each word. While the codas present no problematic issues, one out of 10 nuclei is problematic: ‘v_lk’. Although we said that ‘l’ is a common syllabic consonant in English, all words have syllabic consonants in unstressed non-initial syllables. Hence, [v_l] is possible in words like ‘oval’ [əv.v_l] but no words starting with a ‘v_lk’ syllable exist in English.

Turning now to onsets as the major grouping factor, my shortlist is {b_lick, s_nid, c_royd}, backburner {z_reen, k_mene, d_long}, and my reject, in addition to ‘v_lk’, includes {l_kap, m_pin, t_kar}. The rationalization is that the shortlist words have onsets commonly occurring in other English words. The backburner words have onsets that rise in sonority, which makes them somewhat acceptable, but would sound a bit strange to native speakers of English since no such onsets occur in other English words. On the other hand, they might be good candidates if the company looks for an exotic-sounding name. The reject list contains names that fail on both counts. They fail to rise in sonority and English does not contain any similar words containing these, or similar, onsets.

Two notes are in order. First, if you are not a native speaker of English, your list might be, sometimes drastically, different. For example, in my native language Slovak, {v_lk, z_reen, k_mene, d_long, t_kar} are perfectly fine and phonotactically well-formed words. Hence, your groupings might be influenced by the phonotactic patterns of your native language. Moreover, some of the reasoning and justifications for your groupings in Activity 8-7 might appeal to the difficulty in pronouncing these onsets. For example, a [tk] onset is considered ‘difficult’ by many English speakers. However, the fact that it is possible in Slovak suggests that the difficulty is not universal but closely linked to the language-specific phonotactics and thus (in)experience in speaking habit formation with these clusters. This is like asking a right-handed tennis player to play left-handed. There are many left-hand players, so inherently it is not more difficult to play left-handed, only the right-handed player has not developed automatic subconscious subroutines.

Second, even for native English speakers, differences are to be expected and I am not claiming that my grouping is the ‘correct’ one. Especially the backburner words might easily be found in many shortlists. Consider for example ‘vlog’, which is now a common English word referring to a video blog. No traditional words of English have ‘vl’ as the onset and yet this word caught on easily. Rising sonority and different places of articulation for the two initial consonants are most likely contributing to the acceptability of this word. Some English speakers might improve the sonority profile by devoicing [v] and produce [fl], which is of course a well-formed English onset. But think of the questionable acceptability of Travel Blog as ‘tlog’ despite the rising sonority profile in the onset. In general, asking native speakers to say things that are unusual or exotic, or in other words throw them off-balance, is a very fruitful method for exploring their mental knowledge and habits.

The most common systematic phonotactic generalizations of English are summarized in Table 8.1, but this is by no means a complete list.

Table 8.1 Summary of the main phonotactic patterns characterizing English syllables

<i>Onsets</i>	
C	no velar nasal [ŋ]
	[ʒ] extremely rare and only in borrowings from French
CC	[s] plus almost any consonant except [h] and [f] (barring rare exceptions like ‘sphinx’ or ‘sphere’)
	Stop or fricative plus an approximant (never a nasal) but excluding some homorganic clusters like [tl] [dl], [pw], [fw], etc.
CCC	Combination of the two CC patterns: s + [p, t, k] + [l, r, j, w]
<i>Nuclei</i>	
N	Any vowel except schwa [ə] in any stressed syllable
	syllabic consonant, most commonly [l, m, n], or schwa in a weak unstressed syllable (Chapter 9)
<i>Codas</i>	
C	No [h] and [ɹ] or glides; [ɹ] in rhotic dialects merges with the vowel and the offglides are part of the diphthongs and not separate consonants
CC	Homorganic nasal + stop (excluding [mb] or [ŋg]), [l] + stops, s + voiceless stops, [p, k] + t
	Any coda C + morphological suffix [s, z, t, d] to agree with the voicing of the coda C and [θ] for ordinal/fraction numbers
CCC	Combination the two CC patterns
CCCC	Extremely rare and often simplified to CCC

Activity 8-8: Compare English phonotactics with another language

After reviewing the generalizations listed in Table 8.1, try to list similar generalizations for a language other than English with which you are most familiar, either as a native or a non-native speaker.

What kind of difficulties can you predict based on the comparison of the two tables for native speakers of this language when they speak English, or for English native speakers when they speak this other language?

In addition to the generalizations in Table 8.1, English features several other phonotactic constraints. For example, while [sl] is a frequent CC onset ('slick', 'sly', etc.), [sr] does not appear in English words. Compare with the possible [ʃr] ('shriek', 'shrine', etc.) but impossible [ʃl]. Or the ubiquity of /s/ in CC and CCC onset clusters but the absence of /z/ in similar function: hypothetical words like 'zmile', 'zbeak' or 'zdrip' are not part of English. This might be related to the general lack of voiced fricatives+sonorant clusters (*[vr] or *[ðr] vs. [fr] or [θr] in 'free' or 'three'). A different pattern involves the distance in consonant clusters in the sonority hierarchy in onsets vs. codas: adjacent sonority levels are generally banned from onsets but allowed in codas. Hence, in stop+fricatives [ps] is not allowed in 'psycho' in onsets while the mirror image [sp] is fine for 'crisp' in coda, or in nasal+approximant [ml] is not allowed in onsets while the mirror image [lm] is fine for 'elm'; but note variability in similar words like 'calm' or 'film'. Both [ps] and [ml] onsets are fine in many languages including Slovak.

8.5 Syllable Boundaries

With the basic overview of sonority and English syllable phonotactics, we are in a good position to touch briefly upon the location of syllable boundaries. This is an extremely complex issue with immense research literature and many remaining controversial issues. Hence, the goal here is not to provide a comprehensive theoretical overview, and certainly not to argue for a 'correct' syllabification model, but rather, make you aware, and possibly curious, about this area of mental speaking habits.

When discussing syllable counting in Sect. 8.2 above, I said that native speakers generally agree on syllable counts in most words, barring a small number of words like 'feel', 'film' or 'vegetable'. Locating where precisely the boundaries between syllables in multisyllable words are is less clear. Let's start with Activity 8-9.

Activity 8-9: Division to syllables

Take a bunch of simple CV-type words like 'many', 'cavity' or 'catamaran'. Say them and count the number of syllables. After this easy task, identify the boundaries precisely; you may transcribe in IPA and divide the syllables marking the boundaries as full stops. **It is important to note that we are interested**

in the mental knowledge of the natural speech patterns of real speakers rather than prescribed rules for end-of-line word hyphenation.

Now take words like ‘taxi’ and ‘hopeless’. Do the same.

Some authors suggest that a good approach is to try to say each syllable of the word twice. Does this suggestion change your original preference?

Ask other people and compare with your preferences.

Some people divide ‘many’ as [me.ni], some as [men.i] and some suggest [men.ni]. Hence, [n] could be the onset of the second syllable, the coda of the first, or be also *ambisyllabic* and ‘split’ between these two syllables. Which one do you feel is the most natural to you? Importantly, note that any preference, i.e. any response other than ‘I don’t know’, must have stemmed from your mental subconscious knowledge regarding the division of words into syllables. This is related to, but not identical with, your mental knowledge regarding syllable counting.

Dividing words into syllables is extremely complex and many factors play a role. For example, the knowledge of spelling or the morphological structure of words unquestionably plays a role. In words like ‘taxi’ and ‘hopeless’, many people might prefer [tæ.ksi] and [həʊp.ləs] over the alternatives [tæk.si] and [həʊ.pləs], respectively. In the case of ‘taxi’ the ‘[ks]’ cluster represents a single letter <x> whose division into separate syllables might not be preferable. Additionally, and perhaps more importantly, English phonotactics disallows words starting with the [ks] cluster. In the case of ‘hopeless’, the suffix ‘-less’ is a separate morpheme and thus making it a separate syllable as well might be preferable.

Besides spelling and morphology, we focus on three aspects affecting the syllabification of words that are directly related to the cognitive system underlying speaking. The first one is the relationship between onset and coda consonants. Let’s start with Activity 8-10.

Activity 8-10: Speeding up CV and VC syllables

First, take any CV syllable, it could be a real English word like ‘pay’ [peɪ] or a nonsense syllable like ‘pi’ [pi] as the 2nd syllable of ‘copy’. Repeat this syllable as if according to a metronome starting slowly and gradually increasing the tempo until you cannot go faster. Now repeat the same with a VC syllable like ‘ape’ [eɪp] or ‘ip’ [ip].

Verbalize what difference you observe in your repetitive productions. You might want to speculate regarding the reasons for the observed difference(s) before continuing with the main text.

I believe that most of you noticed that speeding up with the VC syllable is more difficult to maintain and at some point you had to switch to the CV pattern. Hence, your ‘ape-ape-ape-ape’ switched to ‘pay-pay-pay’. This activity, following the original work of B. Tuller and S. Kelso (1990), shows that, in terms of articulatory dynamics, a CV syllable is more stable than a VC syllable. This difference supports the notion of *onset maximization*, which says that in a VCV sequence, it

is more natural for a consonant to be affiliated as the onset (V.CV) than the coda (VC.V). The fact that onsets are more salient than codas can be observed in many phonetic and phonological aspects of speaking. For example, we already mentioned in Chapter 2 that CVCV sequences are the first to be acquired in the babbling stage. Perceptual experiments also show that people recognize consonants in onsets better than in codas. Additionally, investigating the typology of syllable types among many languages reveals the existence of multiple languages that have CV syllables lacking VC syllables, while showing an absence of any language that would have VC syllables lacking CV syllables. Finally, phonologists documented that phonological alternations that affect coda consonants are much more frequent than those affecting the onsets. Onset maximization is commonly extended to suggest that onsets include as many consonants as permissible and thus in VCCV or VCCCV sequences the syllabification V.CCV or V.CCCV might be sometimes preferred over VC.CV or VC.CCV.

The second factor playing a role in the unconscious mental system of syllable divisions relates to language-specific phonotactic constraints, which is in turn closely linked to the notion of sonority discussed in Sect. 8.3. Hence, syllable divisions normally do not create a syllable that violates the generalizations listed in Table 8.1. For example in words like ‘combine’ or ‘activity’ the application of onset maximization would create onsets not respecting English phonotactic constraints [mb] in *[kə.mbaɪn] or [kt] in *[æ.kti.və.tɪ] (recall ‘*’ indicates an unacceptable form). Since native speakers typically consider these forms much worse than [kəm.baɪn] and [æk.ti.və.tɪ], phonotactics is preferred over onset maximization.

The third and final factor influencing syllable divisions to be mentioned here is the relationship between syllable affiliation and *allophonic patterns* in a language. When discussing these in Chapter 7 we have already mentioned that many allophonic generalizations require the mention of a syllable: consonants are aspirated when in the onset, dark [ɪ] is commonly limited to codas, pre-fortis clipping is triggered by voiceless coda consonants, and so on. This allows us to ‘peek’ into the minds of speakers and explore how they syllabify words in a unique way. Rather than asking them for their metacognitive judgements when we ask them to divide words into syllables, as we did in Activity 8-9, we might be able to infer their division directly from their spontaneous speech. Activity 8-11 provides an example for how exploring acoustics with Praat can inform us about people’s mental habits in syllabification.

Activity 8-11: Syllable division in multisyllable words

Consider the word ‘mistake’. This disyllabic word contains two consonants between the two vowels and their affiliation to syllables presents several options for syllabification. The onset maximization suggests [mɪ.steɪk], the morphology of the word with a prefix ‘mis-’ suggests [mɪs.teɪk] and even [mɪst.eɪk] does not violate English phonotactics. You have several tasks in this activity.

First, review the allophonic patterns discussed in previous Chapter 7 and suggest how these different syllabifications might affect the phonetic realization

of the word. Therefore, what do you expect to see in Praat if a speaker says [mi.steɪk] vs. [mɪs.teɪk]?

Second, ask speakers (preferably both native and non-native) to say this word and record their productions. However, note that a single production of a single word is usually not sufficient to assess a speaking pattern. Recall several words and different environments we recorded in Activity 7-5 when investigating the t/d allophones. Hence, based on the findings from the first task, consider what other words might create nice comparisons to ‘mistake’ and provide you with relevant phonetic data.

Third, ask your speakers after the recording for their metacognitive syllabifications and note down their answers.

Fourth, inspect your data in Praat, measure relevant values and provide now an informed guess about the syllabification of ‘mistake’ and other words for your speakers. Does it correspond to their judgements after the recording?

The main phonetic feature cuing the syllable division in ‘mistake’ is the aspirations of [t] and vowel duration of [ɪ]. If the speakers syllabified [mi.steɪk], [t] should not be aspirated much since it does not start a stressed syllable, and [ɪ] should not be shortened since the voiceless [s] forms the onset of the following syllable and not the coda of the same syllable. If the syllabification is [mɪs.teɪk], some [t] aspiration should take place and [ɪ] should be shortened. The least likely [mɪst.eɪk] should have a shortened [ɪ] but an unaspirated, or possibly glottalized, [t].

The easiest phonetic feature to inspect is the aspiration. If we do not see any aspiration, it does not necessarily mean that the speaker syllabifies [mi.steɪk]. It could be, for example, that this speaker does not aspirate if s/he is a non-native speaker. Hence, we need words like ‘take’ or ‘re-take’ and ‘steak’ or ‘re-state’ to compare aspiration in these words with clear syllable affiliations of the consonants with aspiration in ‘mistake’. The duration of [ɪ] is more difficult to compare, and, in fact, comparing the presence and absence of pre-fortis clipping with [ɪ] is not a good idea because English phonotactics suggests that lax vowels appear in syllables with a coda. Note that English has no words like [pɪ], [gʊ], or [se] but includes many words like [pɪt], [gʊd], or [sed]. For this reason, your Praat visualization might support ambisyllabicity of ‘s’ such that it closes the first syllable and starts the second one: [mɪs.steɪk].

Hence, other aspects such as word stress (next chapter), phonotactics, or morphological structure all affect syllabification and need to be considered in designing words for any phonetic experiment and data collection. For example, the ‘mis’ prefix in ‘mistake’ is not separable nowadays while in words like ‘mistime’ people clearly separate ‘mis-’ from ‘time’, which supports the syllabification [mɪs.tam] and thus predicts more robust and systematic aspiration than in ‘mistake’. You may check with native speakers in Praat if the difference in the separability of the prefix ‘mis-’ can be supported through the measurement of aspiration.

It is important to note that the tendencies mentioned above (onset maximization, phonotactic restrictions, morphological structure and allophonic patterns) combine in various ways in different languages.

8.6 Syllable Weight

We close the chapter with examining syllable rhymes in more detail. Recall that their cognitive reality is established by our subconscious knowing that ‘speak’ rhymes with ‘leak’, ‘peak’, ‘creek’ or ‘streak’ irrespective of the number and quality of onset consonants but not with ‘speck’ or ‘spa’ where onsets are the same but rhymes differ.

In previous chapters we have measured vowel duration either to see the difference between lax (short) vowels like [ɪ] and tense (long) vowels like [i:] in ‘hid’ vs. ‘heed’ and also the effect of coda-consonant voicing on vowel duration in allophonic pre-fortis clipping. We assumed, but never measured, that the duration of diphthongs is longer than that of short vowels and possibly comparable to the long vowels. Activity 8-12 checks this intuition with Praat.

Activity 8-12: Vowel duration revisited

Place short vowels like [ɪ] and [e], long vowels like [i:] or [ɑ:], and diphthongs like [aɪ] or [əʊ]/[oʊ] in an environment as close to minimal pairs as possible. You will find that this is not a very easy job and you might need uncommon or nonsense words!

If you were successful, use your list, my list for inspiration is below; transcription assumes a British non-rhotic dialect

- Dana ([deɪnə], proper name), darner ([dɑ:nə], dragonfly), dinner ([dɪnə], meal), doner ([dəʊnə], e.g. organ doner), diner ([daɪnə], eatery in US, one who dines), deaner ([di:nə], slang for shilling, coin), Denner ([denə], discount supermarket chain in Switzerland)

Record the words, aiming for consistent realization, label with a TextGrid the words in one tier and the stressed vowels flanked by [d] and [n] in another tier. Is the intuition formulated above the activity supported?

In IPA, English [i:] and [ɑ:] of ‘deaner’ and ‘darnar’ are commonly transcribed with the colon indicating their long duration compared to the short vowels. But the comparable duration of long vowels and diphthongs is not overtly marked in transcription. Is this phonetic similarity of long vowels and diphthongs relevant for mental speaking habits that we investigate?

Well, consider the words from Activity 8-12 and try to say just the first syllables omitting [nə]. Which of these syllables are, or could be, well-formed words and which are not? Clearly, all first syllables except [dɪ] and [de] can be separate English words, which supports the phonotactic observation we made in the conclusion of the previous section: long vowels and diphthongs act together in English since they can form monosyllabic words without coda consonants while short lax vowels cannot.

Phonologists use the idea of timing slots in a skeletal tier, ‘x’ in Fig. 8.4, to capture the difference between short and long vowels. Short vowels are represented

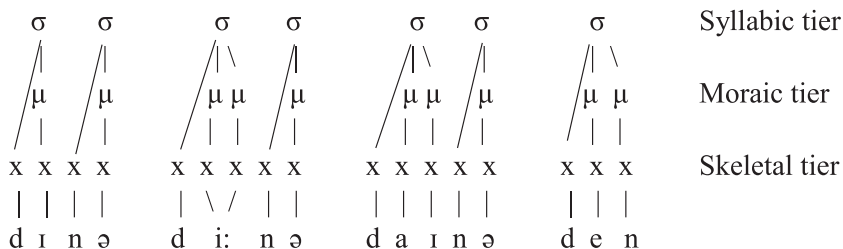


Fig. 8.4 Phonological representations for ‘dinner’, ‘deaner’, ‘diner’ and ‘den’

with a single *x* while long vowels and diphthongs with two *x* slots. We also saw that for rhyming the consonants in the onsets do not ‘count’, and this is represented by *moras*, μ in Fig. 8.4, in the moraic tier where onset consonants are not linked to any mora. We can thus account for our observation that English words cannot have just a short vowel in the rhyme by saying that at least two moras are required for any (content) English word.

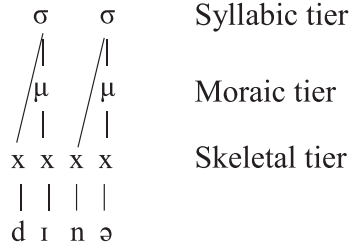
But returning to words from Activity 8-12 we also observe that while [dɪ] and [de] cannot be English words, [dn] and [den] are perfectly fine words. In phonological terms, the coda consonant in English also ‘carries’ a mora, which is shown in the rightmost representation in Fig. 8.4. Hence, syllables with two moras have more *weight*, we call them *heavy syllables*, than those with one mora that are called *light syllables*.

Coda consonants do not carry moras in all languages, multiple coda consonants are not assigned a mora each, and onset consonants never carry a mora. Hence, despite the loose link of moras to phonetic duration, the moraic representation is a phonological construct allowing to express the similar behaviour of rhymes including long vowels, diphthongs and short vowels with coda consonants.

Previewing the next chapter, the difference between heavy and light syllables is very useful for characterising English word stress. It is the case that the stressed syllables are commonly heavy in English. Wells (1990) argues that stressed syllables attract not only onset consonants but also coda consonants and thus words like ‘packet’ should be syllabified as [pæk.tɪ] and not [pæk.kɪt] that would respect onset maximization. This has been called *stress-to-weight* principle. It obviously also suggests [men.i], [kæv.ə.ti] and [kæt.ə.mə.ræn] for words in Activity 8-9, and is used in some dictionaries showing pronunciation. The tiers in Fig. 8.4 also allow for the representation of ambisyllabicity mentioned when discussing Activity 8-9 (e.g. [men.ni] with [n] being split between two syllables) respecting both onset maximization and stress-to-weight. Figure 8.5 shows this for ‘dinner’.

In summary, just like the phoneme, the syllable is a fundamentally phonological concept that is cognitively real, but its phonetic characteristics are not easily defined. While sonority or duration play an important role, carving of these continuous phonetic dimensions into discrete phonological representations provides a useful approach for investigating spoken language.

Fig. 8.5 Phonological representations of ambisyllabicity in ‘dinner’ [dɪ.nə]



Exercises

- 8-1 I googled different types of blogs and came up with interesting blog types such as Music, Homer, Food, Niche, or Rouge blog. Consider the imaginary one-word names for these blogs as ‘mlog’, ‘hlog’, ‘flog’, ‘nlog’ or ‘rlog’, respectively. Following Activity 8-7 and the discussion below it, which one-word blog names have a chance to make it into English and why?
- 8-2 Consider the realization of the first /t/ in words like ‘Atlantic’ vs. ‘attractive’. Inspect the phonetic features, informally and with the help of Praat discuss how the observed differences might illustrate the mental knowledge regarding the location of syllable boundaries and phonotactic patterning.

References

Tuller, Betty, and J.A. Scott Kelso. 1990. Phase transitions in speech production and their perceptual consequences. In *Attention and performance 13: Motor representation and control*, ed. Marc Jannerod, 429–451. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wells, John C. 1990. Syllabification and allophony. In *Studies in the pronunciation of English: A commemorative volume in honour of A.C. Gimson*, ed. Susan Ramsaran, 76–86. London and New York: Routledge.



In this chapter we will

- explore the phonetic aspects of word stress and how they can inform us about the mental knowledge regarding stress
- expand skills in Praat, for example by learning how to manipulate recorded sounds
- review major patterns for stress placement in English

9.1 Introduction

Although the previous chapter investigated syllables, it inevitably forced us to examine them in words of multiple syllables; like when we discussed syllable boundaries and mentioned word stress several times. In this chapter we focus our discussion fully on words consisting of multiple syllables. We will explore our mental knowledge regarding systematic variation in the prominence of individual syllables and again try to use multiple hands-on activities and Praat towards this goal. We will see that word stress, just like other aspects of speaking habits covered in this book, features a complex relationship of rather abstract, intangible, mental patterns that are, however, observable and measurable in the records of our everyday speaking.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_9) contains supplementary material, which is available to authorized users.

9.2 Mental Knowledge of Word Stress

In the previous chapter we counted syllables in longer words like ‘pandemonium’. Children might help themselves by clapping their hands when separating words into syllables. Both counting and clapping may give the impression that all syllables are equal since each receives a clap or adds a digit to the total count. However, closer inspection of individual syllables reveals that they are not equal. Activity 9-1 revisits the word ‘pandemonium’.

Activity 9-1: Two examples of pandemonium

There are several recordings of the word ‘pandemonium’ in the book companion to this chapter. Listen first to the example ‘pandemonium_first.wav’ without inspecting it visually. How does it sound to your ears?

Then, listen to another file with the same word: ‘pandemonium_second.wav’. Now that you have them opened in the Praat Objects window, alternate between listening to the first and second one several times, again without visually inspecting the files and relying on your ears only.

Your first task is to decide which of the two files sounds better. Your second task is to describe the differences between the two recordings. Be detailed, and either explain verbally to a friend the differences using the notions and terminology from this book, or in the absence of a friendly soul, write down your perceptual analysis of the differences. Finally, try to link observations from the two tasks by hypothesizing what in the characteristics of each recording makes you prefer or not prefer that recording.

The exploratory Activity 9-1 is extremely important for the rest of the chapter. Which recording of ‘pandemonium’ did you like more? Both native and proficient non-native speakers should be pretty clear about their preference for the second file (pandemonium_second.wav). You probably intuitively feel that something is wrong with the first file and whatever that is, it sounds better in the second file.¹ The presence of this ‘intuition’ and ‘feel’ is another evidence for your mental knowledge that we pursue in this book.

The second task of Activity 9-1 aimed at pinpointing this previously subconscious knowledge and thus making you consciously aware of the patterns in your speech. I assume your answers might discuss differences in the perceived prominence of individual syllables. While in the ‘preferred’ file the third syllable ‘-mo-’ seemed to be the most highlighted, or stand out the most, in the ‘dispreferred’ file it was probably the absence of this prominence that was noticed, as well as the prominence on the first syllable ‘pan-’. This perceptual prominence of certain syllables is the result of their being *stressed* and these syllables are said to carry, or

¹This preference applies when the words are pronounced in isolation. As we will see in Sect. 11.4, putting words into larger phrases might result in systematic changes in the main stress position.

receive, *word stress*. We will discuss the phonetic aspects of this perceived prominence in more detail in Sect. 9.3 below and I will also reveal the origin of the two pandemonium renditions.

In addition to the clear preference in the two ‘pandemonium’ examples, word stress plays other important functions. You might recall the discussion from Chapter 2 regarding the variation in pronouncing the English word ‘object’. Here word stress participates in signalling the grammatical category of the word: ‘OBJECT’, with the stress on the first syllable, is a noun while ‘obJECT’ with the stress on the second syllable is a verb. We will use capitalization to indicate the stressed syllable for now and introduce IPA symbols later in the chapter. While in the two ‘pandemonium’ examples one stress pattern was preferred and one dispreferred, in the case of ‘object’, both words sound fine only the difference in stress placement is linked to distinct linguistic functions. This grammatical function also participates in the complex mental knowledge English speakers have regarding word stress.

Another characteristic of English stress arises by identifying stressed syllables in words like ‘mother’, ‘deny’, ‘Japanese’ or ‘mathematician’. In English, virtually any syllable might receive stress. In the words above, the main stress falls on the first, second, third and fourth syllable respectively. In many languages, however, word stress is fixed to a particular syllable of each word (e.g. in Hungarian to the first one, Polish to the penultimate (second-last) one, etc.). Hence, for speakers of these languages word stress plays a slightly different role in the cognitive system underlying speech than for English speakers. Yet in some other languages, word stress differences might signal dialectal affiliation, either in a subset of particular words (consider ‘garage’ in the British [gæ.ɪdʒ] and American [gə.ɪɑːʒ] dialects), or in the pronunciation of all words (for example, the East Slovak dialect has word stress on the penultimate syllable while the standard colloquial Slovak stresses the initial syllables). The variability in terms of which syllable receives word stress might thus have multiple motivations: it may signal grammatical differences, regional affiliation, or non-native accent among other things.

Activity 9-2 explores the mental unconscious knowledge of word stress of English speakers further.

Activity 9-2: Saying known and unknown names from an unfamiliar language

The first name ‘Vladimir’ is a very common Slavic name that is popular in Russia but common in many other countries too. You might have heard about politician Putin but also about writer Nabokov, musician Horowitz, or baseball player Guerrero. How do you, as an English speaker, pronounce this name in a regular English sentence like *I wanted Vladimir to come*? In particular, which syllable of the three do you stress disregarding now the uncommon [vl] onset cluster and potential [fl] pronunciation? Try to ask around other English speakers and listen to where they put the stress (of course you have to ask to read the name and not say it since you don’t want them to imitate your pronunciation). Is there an agreement? How sure are you, and them, compared to the certainty with ‘pandemonium’ in the previous activity? Have you said or heard this name pronounced before?

Now try to pronounce the maiden name of my mother ‘Pavljajova’ in regular English; i.e. not trying to think as it ‘should be’ pronounced in Slovak but how would her name be pronounced for example in an English school by the teachers or students (e.g. *Pavljajova is missing today*). I am sure that with this name, you have never heard or said it before. Where do you put the word stress? Again, if possible, ask around other English speakers. Is there an agreement? How sure are you, or your subjects, now about the position of word stress compared to ‘Vladimir’?

What do the data from these mini data-collection experiments say about the mental system of English speakers regarding word stress?

So let us discuss and compare the observations from Activity 9-2. In my experience, speakers of English (both natives and those non-natives who do not speak Slavic languages) commonly disagree regarding word stress in ‘Vladimir’. Some put it on the first syllable (VLAdimir) and some on the third syllable (vladiMIR). Even for a single speaker, there might be differences with different contexts such as with different surnames. We might tentatively conclude that variability is present but it is not the case that any syllable receives stress with equal probability. In other words, there is some systematicity behind this.

You might be surprised to find out that the Russians stress the second syllable of ‘Vladimir’! Hence, the English word stress system is clearly different from the Russian one but the question is how you yourself learned how to say this word. Chances are you have heard this name pronounced by somebody before and since the speakers themselves are not always sure, your level of certainty might be lower than with some other words. But you are unlikely to stress the second syllable like the Russians.

Now, what about the second name ‘Pavljajova’? My own mini-experiment suggests that here the agreement is almost uniform and English speakers put stress on the pre-final, this time the third, syllable: ‘PavlaJOva’. Now think about why most speakers of English would say it this way despite the fact that they have never heard anybody say this name before. A plausible hypothesis might be that English speakers have heard similar names, such as the tennis player Navratilova, who is a native of the Czech Republic and everybody pronounces her name in English with the stress on the pre-final fourth syllable ‘NavratiLOva’. Hence, you say an unknown similar name like ‘Pavljajova’ also with the pre-final stress.

By the way, notice that in ‘Navratilova’ stress is on the fourth syllable while in ‘Pavljajova’ on the third one. Hence, if the previous observation, that English speakers imitate the word stress in Navratilova to put it on ‘Pavljajova’, is correct, we expect the English speakers to count the stressed syllable from the right! Because if they counted from the left, as people not consciously aware of word stress, they would put it on the fourth syllable: *‘PavljajoVA’. Hence, here we are revealing another unconscious mental speaking habit that you as English speakers have: syllables count from the right for the purposes of word stress. I think my bet would be safe that many of the native speakers were not aware of this.

But let’s return to ‘Navratilova’. If you put stress on ‘Pavljajova’ based on your hearing other people say ‘NavratiLOva’ how did these people before you find out

where the stress in Navratilova should be? They certainly did not mimic Czech, Navratilova's native language, in which her name is 'NAvratilova' with the stress on the first syllable.

A plausible conclusion from this discussion is that speakers of every language possess mental knowledge, or a cognitive system, that guides the placement of word stress on individual syllables and that is independent of the segmental make-up of the words themselves. The crucial characteristic of such a system is a certain degree of *predictability* that guides native speakers to assign word stress, rather uniformly, also to words they never heard before. Recall that we similarly discussed predictability when exploring English allophones in Chapter 7 or the wug-test of English plurals in Chapter 2. Different languages, of course, have different degrees of this predictability with languages like French being quite 'easy' with a fixed stress on the last syllable and English much more 'complex', but not as complex as for example Russian. The exploration of this cognitive system guiding stress placement in English is the topic of this chapter.

9.3 Phonetic Aspects of Word Stress

Activity 9-3 begins exploring how the prominence of certain syllables that you perceived in words we discussed above is actually achieved in your mouth. The activity is instrumental for the awareness building regarding word stress.

Activity 9-3: Some nonsense 'fofofo' words

Imagine that a new app is called 'fofofo'. While in some languages this name would allow only a single way of pronouncing, in English we have several variants, at least in theory. By placing the word stress on different syllables, it can be 'FOfofo', 'foFOfo' and also 'fofoFO'. Your first task is to produce these three variants as three different English words clearly differentiating among them and when ready to record your words in a single file in Praat. Since the three words have three identical syllables 'fo', it helps us to observe how you produce word stress in English.

Now, before inspecting acoustics, introspect the articulatory actions and their coordinations. Repeat the three variants and look at your mouth in the mirror. What do you see and feel with your articulators? For example, inspect the pressure of your lower lip on your upper incisors or the degree of lip rounding during the diphthongs. Then, listen to the file you have recorded several times without inspecting it with *View & Edit*. How do you make the stressed syllable stand out from the other two? Consider the features you have at your disposal such as duration, pitch, intensity or vowel quality.

With your hypotheses regarding these features, inspect the recording in Praat. For example, if your hypothesis is that stressed syllables are longer than unstressed ones, actually measure these durations. Always compare the same syllable (first, second, third) and the stressed vs. two unstressed renditions. Is your hypothesis supported? Are the two unstressed syllables always identical?

Pursue this further and inspect the duration of individual sounds. Is it the vowel or the consonant (or both) that primarily change their duration depending on the stress placement? The best way of doing this is to annotate the file with a TextGrid of three interval tiers (words, syllables, sounds) following the skills gained in Chapter 6. Do not forget to save your work!

Visually explore, and measure if possible, all other characteristics that you originally hypothesized such as pitch, intensity, vowel quality, etc. For example, select the sound in the Praat Object window with the TextGrid and click *View*. Make sure that the pitch and intensity tiers are shown by clicking the appropriate boxes in *View* → *Show analyses*. For a very crude inspection, you can select a vowel of the stressed or unstressed syllables in the textgrid and get mean pitch by *Pitch* → *Get pitch*, which will produce mean pitch of the selected interval in the Praat Info window and you might record this in a spreadsheet. The same process can be used for getting mean intensity for a selected interval.

You can also inspect the vowel or consonant quality either visually through formant trajectories for vowels (Chapter 5) or the intensity and characteristics of the noise (such as centre of gravity in Chapter 6) during the consonant.

Activity 9-3 above explored the phonetic realization of word stress. You probably found out that you signal word stress by making the stressed syllables **longer**, **louder** and **higher** in pitch than the unstressed syllables. You may have also observed in the mirror that you open your mouth slightly more and press the lower lip against the upper teeth slightly more in the stressed than in the unstressed syllables. Finally, your stressed syllables might have been diphthongal, and for some more back and rounded, while the unstressed syllables were probably close to a schwa quality. The observations above, that you hopefully made yourselves now, have been known in the phonetics literature since Fry's seminal work (1955, 1958) and form a fruitful area of research ever since.

Let us now see how Praat, in addition to visualization, can be used to quantify these observations. Figure 9.1 shows three 'fofofo' renditions and the accompanying TextGrid file in the book companion.

The important concept in quantifying your observations is the difference between the raw measurements and the analysis based on relative, or derived, values based on those measurements. The raw values depend on individual characteristics, biological differences and many other aspects and conclusions from these values need to acknowledge this. In phonetic research, we are thus often also interested in relative comparisons. For example, we can calculate the difference between the two identical syllables produced by the same speaker as in 'fofofo' words. We can also calculate derived variables such as the ratio of the consonant and the vowel durations in each /fo/ syllable, and compare these ratios for stressed and unstressed syllables. In simple descriptive analyses without statistical modelling, derived measures might provide more robust data for examining our questions.

If you produced the three 'fofofo' tokens in a similar way many native speakers of English would, you may notice that in addition to a clear binary division

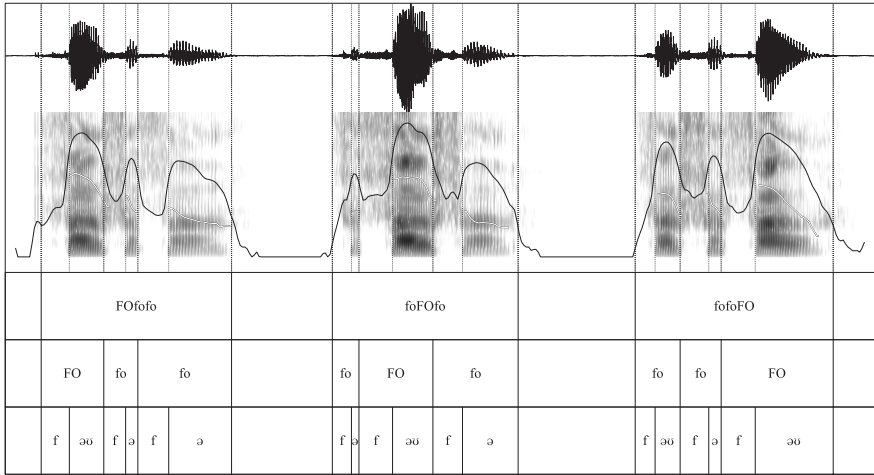


Fig. 9.1 Three examples of non-sense word ‘fofofo’ with varying primary stress annotated in Praat. The solid black curve shows intensity and the greyish line shows pitch

between stressed and unstressed syllables, another level might be identified. For example, in ‘foFOfo’ that is produced similarly to words like ‘concerto’ or ‘tomato’ compare the first and the third syllables. Both are clearly less prominent than the second one but there is also a difference between them. The first one tends to be shorter, the vowel is more centralized and schwa-like than the third one. Similarly, in ‘fofoFO’ that is produced similarly to ‘Japanese’ or ‘overdo’, the first syllable might have been produced slightly more prominently than the following second syllable. This intermediate prominence is related to the *secondary stress*. Hence, English syllables are marked for three levels of stress: primary stress, marked in IPA with a vertical superscripted apostrophe-like line that precedes the most stressed syllable, secondary stress, marked with a vertical subscripted line preceding the syllable, and unstressed. Hence, words like ‘Japanese’ or ‘overdo’ in British English are transcribed as [ˌdʒæpəˈniːz] and [ˌəʊvəˈduː]. The three renditions of ‘fofofo’ would be [ˈfəʊfəʊ], [fəˈfəʊfə] and [ˌfəʊfəˈfəʊ].

Let’s return now to the ‘pandemonium’ examples in Activity 9-1. First, we complete the description of differences between the two original files. In the preferred ‘pandemonium_second.wav’, the third syllable ‘-mo-’ received the primary stress and was longer, higher in pitch and more diphthongal in quality than the same syllable of the dispreferred ‘pandemonium_first.wav’. Additionally, the first syllable ‘pan-’ of the preferred file received the secondary stress while the same syllable of the dispreferred file was made much more prominent (by lengthening it and making it higher in pitch). Hence, the preferred word was the canonical [ˌpændəˈmɒniəm] while the stresses were switched in the dispreferred one [ˈpændəˌmɒniəm].

The dispreferred ‘pandemonium’ file was artificially created in Praat, and we discuss now how Praat can be used to manipulate sounds. Building on cutting,

pastings, or silencing entire intervals of speech with the Edit functions, Activities 9-4 and 9-5 extend your Praat skills in this domain and walk you step-by-step in manipulating pitch and duration of a speech recording.

Activity 9-4: Changing stress perception by manipulating duration and pitch: pitch

Select the original file ‘pandemonium_CDO_BE_10200.wav’ in the Praat *Objects* window. Click *Manipulate* → *To manipulation...* in the right-hand menu and *OK* to the default values for time step and pitch floor/ceiling. Praat creates a new Manipulation object with the same name as the original file, click *View & Edit* and maximize the window to fit your screen. You will see three panels with the waveform and pulses, pitch and an empty duration tier respectively. Select various intervals and play in order to identify (visually and auditorily) approximate boundaries between sounds and syllables. This is a bit challenging since the spectrogram is not easily available in the same window. A way to circumvent this problem is to see both the Manipulation window, and the original sound with its spectrogram in a different window, and group them. If you have these two windows opened on your screen, tick the bottom right field *Group*. When both windows have this field ticked, selecting an interval in one window copies the selection into another window. You can thus select appropriate interval(s) in the Edit window in the spectrogram, and keep the same selection in the Manipulation window.

Returning to the task at hand, you see clearly the pitch raising and the high pitch peak during the stressed third syllable ‘-mo-’. We have several options for manipulating pitch. First, select the approximate interval of the syllable ‘-mo-’. You can now lower the pitch there by clicking either *Shift pitch frequencies...* or *Multiply pitch frequencies...* under *Pitch* menu. In the former you lower pitch by entering a negative number (e.g. -30), which lowers each pitch point by this value (pitch raising is done with positive numbers). In the latter, you lower pitch by entering a value in the (0,1) interval (e.g. 0.75), which multiplies each pitch point by this value; again pitch raising is achieved with factors greater than 1.

You can auditorily inspect your manipulation at each step by playing the resulting file (clicking the relevant bar in the bottom of the window), and comparing with the original sound by Shift-clicking the same bar.

Now select the interval of the first syllable ‘pan-’ and raise the pitch to around 180 Hz. Then select the stressed ‘-mo-’ syllable and lower its pitch. Finally, we can adjust every single pitch point by clicking and dragging. Let’s smooth the resulting contour in this way and eliminate large jumps. The result should look similar to the second tier of the screenshot in Fig. 9.2 with the larger (green) dots marking the new pitch values and smaller (grey) dots the original ones.

Activity 9-5: Changing stress perception by manipulating duration and pitch: duration

After pitch, duration can be manipulated as follows. We want to lengthen the first syllable and shorten the third one. We click in the middle of the aspiration after the initial ‘p’ and add a duration point (*Dur* → *Add duration point at*

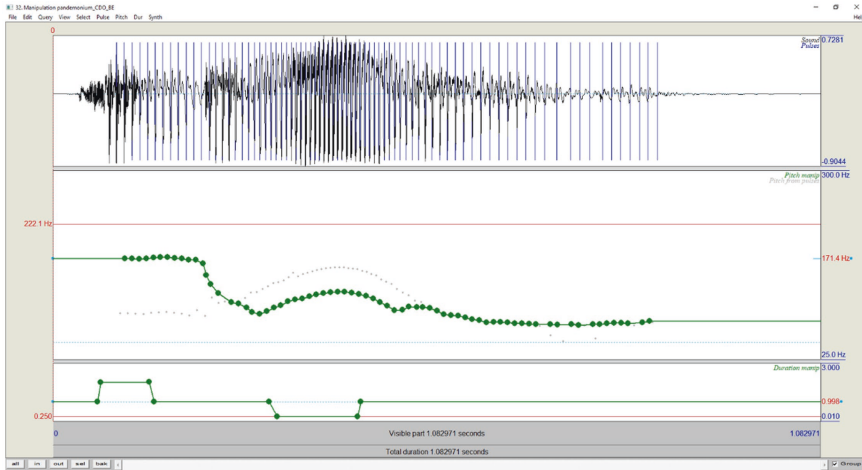


Fig. 9.2 Manipulation of pitch and duration in Praat. The original soundwave with pulses in the top panel, its pitch curve in the middle panel (larger dots represent the new and the smaller the original pitch points), and the duration panel in the bottom (lengthening of the first syllable and shortening of the third one)

cursor) with a shortcut *Ctrl-D*. Duration points are ‘draggable’ so that raising them above the 1.0 level lengthens the duration, while putting them below that level shortens it. If we want to affect the entire sound, a single point is sufficient. But if we want to manipulate only certain intervals, the simplest way is to insert two points close to each other in the vicinity of the interval start (during aspiration) and two points close to the interval end (during the low-amplitude nasal portion). Since in the case of the syllable ‘pan-’ we want to lengthen, we raise the second and third points (e.g. to around 2.0) and leave the first and fourth points at the 1.0 level. This lengthens just the first syllable, which you can check by clicking and shift-clicking the bars. It might be that the ‘n’ sound is too long now, which you might change by dragging the third and fourth points left, closer to the boundary between [æ] and [n]. This results in lengthening primarily the vowel and not the nasal. Now you move to shortening the originally stressed syllable [ˈmæʊ] inserting additional four points and lowering the middle two of those points to 0.25 for example. The range of the manipulation factors (default (3.0, 0.25) can be adjusted with *Dur* → *Set duration range...*). The four-point strategy allows greater control but using only three points and manipulating the middle one is also possible.

If you are happy with your manipulated sound, you can synthesize it with *File* → *Publish resynthesis*, which creates a new Sound object in the Praat Objects window and you may save this newly created file as a regular wav file to your PC. Your manipulation window might look similar to Fig. 9.2.

More details can be found by consulting the Praat manual (*Help* → *Search Praat manual...* and typing ‘manipulation’).

Let me close the discussion of ‘pandemonium’ examples by considering one more manipulation of this word in the file ‘pandemonium_third.wav’. As before, listen to the file first and before reading the text below, try to verbalize how it sounds compared to the previous two examples, and what phonetic features might contribute the most to your perceptual evaluation. In other words, describe what you think I did to the original file so that it sounds this way. How would you transcribe this in IPA?

I expect that you evaluate the file as the worst, or the most dispreferred, of the three. I tried to shift both stresses one syllable to the right: [pænˌdɛməˈniəm]. In addition to manipulating pitch and duration along the lines outlined in Activities 9-4 and 9-5, I also slightly manipulated the third prosodic feature participating in cuing stress: intensity. This is because increased intensity also participates in cuing word stress, as seen in Activity 9-3, and pitch and intensity are physiologically strongly coupled, or related, which we explore in Activity 9-6.

Activity 9-6: Relationship between intensity and pitch

Stand up with your back against a wall, take a deep breath, close your eyes and say continuously some long vowel, such as [i:] or [ɑ:], in a monotonous way with flat level pitch. Ask a friend or a family member to push gently against the lower section of your ribcage in unexpected moments and observe how the production of the vowel is affected. Although having another person is much better for the element of unexpectedness, even if you have nobody around, you can push your ribcage yourself. A similar activity is mentioned in Ladefoged and Johnson (2015).

You observe that the push increases the intensity of your vowel production. This intuitively makes sense since, as we know from Chapters 3 and 4, the intensity of voice is directly related to the amount of air that is emitted when you speak. Importantly, you also observe a rather salient increase in pitch. This is because the increased volume of air emitted from the lungs caused by the push, increases the pressure below the vocal cords (sub-glottal pressure), which makes them vibrate faster, which increases f_0 heard as pitch.

The awareness of the relationship between f_0 and intensity can be further strengthened by trying to mismatch, or de-couple, them in nonsense words like ‘fofofo’ in Activity 9-3. Try producing this word with a stress on various syllables (‘FOfofo’, ‘foFOfo’, ‘fofoFO’) as before, but this time consciously try to increase f_0 but decrease intensity of the stressed syllables at the same time. Similarly, try increasing intensity but lowering f_0 on the stressed syllables. You notice that this is pretty unnatural and actually requires a lot of effort to achieve. Especially if you compare the ease with which you say the stressed syllable when the increased f_0 matches increased intensity.

Hence, manipulating intensity and f_0 in the same direction together is an *ecological*, i.e. physiologically natural, pattern. You can follow the Praat manual (Intro 8.3. Manipulation of Intensity) and adjust intensity of a sound locally. What I did in file ‘pandemonium_third.wav’ was to lower the intensity of the originally stressed vowels [pæn] and [məʊ] and increase the intensity in the originally unstressed [də] and [ni] syllables.

Returning now to your perception of this last pandemonium example, we said that the word sounds weird, or dispreferred compared to the previous two. It is extremely important to discuss the sources of this weirdness. First, the stress pattern is completely altered and has nothing in common with your *mental knowledge of this word*. In the manipulation discussed in Activity 9-3, we took the two stressed syllables (the 1st and 3rd) and only switched their relative prominence in a rather crude way. In ‘pandemonium_third.wav’, however, I attempted to produce phonetic prominence on the originally unstressed 2nd and 4th syllables, which is a much more radical departure from the known stress pattern.

Second, despite employing the concept of ecological manipulation in which both pitch and intensity are correlated and adjusted in the same direction in a given syllable, ‘pandemonium_third.wav’ sounds bad also because it includes other *non-ecological* aspects of manipulation. Word stress is realized in the mouth by an ingeniously coordinated, and unimaginably complex, articulatory activity at several hierarchical levels. For example, I have not modified the quality of vowels. The original vowels in the 2nd and 4th syllables are lax centralized [ə] and [ɪ] and increasing just their duration and pitch and intensity without also changing their quality to a more peripheral, tense, possibly more [i]-like quality in both cases, is very unnatural and un-English-like. The same applies also in reverse, by shortening and lowering the stressed vowels, I am not changing their quality to schwa. Additionally, despite my adjustments of pitch and intensity in tandem and in the same direction, the degree to which I adjusted pitch and duration was not synchronized. Or I may have shortened aspiration (Voice Onset Time) in the first syllable [p^hæn] but left the intensity of the burst unchanged. And I could continue at length in listing the modifications that are not physiologically natural. However, in principle, all these could be modified in Praat and we can adjust formant values, burst intensity and synchronize the pitch/intensity manipulation in a more sophisticated way, but this would take us too far.

Third, the perception of weirdness comes also from the concept of speech resynthesis and the increased number of features that are changed. The more adjustments are made to the original file the more artificial the output would sound. Although it is important to keep this factor in mind, the first two aspects above (the conflict with your mental knowledge and the non-ecological modification) contribute to your perception of dispreference more strongly than this last one.

Let’s finish this section by exploring the links between the mental knowledge of the speaking habits and their phonetic realization in real speech material in Activity 9-7.

Activity 9-7: Continue with the content of your contemplation

First, consider the sentence ‘Continue with the content of your contemplation’ and particularly the words ‘continue’, ‘content’ and ‘contemplation’. You notice they all begin with the same syllable in spelling. Is the pronunciation, and IPA transcription, identical? Keep in mind ‘content’ is a noun here. How does word stress affect the realization of the three syllables? What do you hypothesize happens to these syllables phonetically in this sentence?

Then, if you are a native speaker, or have one around, record the sentence 2–3 times. If you do not have access to native speakers, use the ‘con_examples.wav’ in the book companion. Create an annotation file with two tiers and in the first one label the individual words. In the second tier identify as accurately as possible the various renditions of the syllable ‘con-’ in the recording.

Before reading any further, using the skills with Praat developed so far, compare these syllables in terms of aspiration (duration and intensity), peripherality in the vowel space (through the first and second formant values), syllable duration, pitch and intensity.

How do your phonetic measurements correspond to your hypothesis? Are there differences among the 2–3 renditions of the sentence?

As you likely found out on your own in Activity 9-7, [k^hɒn]/ [k^hʌn] of ‘content’ receives the primary stress and thus should be most prominent regarding all the phonetic dimensions mentioned in the activity: aspiration should be longer and louder, vowel quality more peripheral, duration greater and its pitch and intensity higher. The same syllable of ‘contemplation’ receives the secondary stress and you should thus observe medium values in all these phonetic features. Finally, syllable [kən] in ‘continue’ is unstressed and the values should be the lowest. Naturally, as mentioned in Chapter 1, your values might not necessarily be all in line with these predictions. For example, the speaker in the companion’s example used relatively salient aspiration also with ‘continue’ but other factors (pitch, intensity, peripherality) were clearly indicating the lowest level of stress. Discuss potential reasons for differences, also among the 2–3 renditions of the same sentence in your recording.

9.4 Limitations of the Phonetic Aspects of Word Stress

The previous section discussed how phonetic features like pitch, duration or intensity participate in cuing stressed and unstressed syllables. Our analysis-by-synthesis approach showed that by manipulating these features we can change the perception of stress patterns in words. To fully appreciate, and make use of, the awareness gained in the previous section, however, we have to bear in mind that in real speech the phonetic differences between stressed and unstressed syllables might be less clear and more complex than it might have seemed until now.

To illustrate some of these differences, consider the word ‘mermaid’. It is a two-syllable word with the primary stress on the first syllable. The dictionary British pronunciation of this word is captured in ‘mermaid_CDO_BE.wav’. Inspecting this sound in Praat we see that the first syllable is clearly higher in pitch, has greater average intensity, but is also a bit shorter than the second syllable.

Firstly, we should be cautious with comparing syllables of different segmental makeup. This is not the ‘fofofo’ example in Activity 9-3 or ‘con-’ syllables in 9-7 where all three syllables were identical. It might be the case that the diphthong [eɪ] in ‘-maid’ is on average, or *intrinsically*, longer than the (long) monophthong [ɜ:] of ‘mer-’. Many intrinsic differences have been conclusively shown. For example,

high vowels are intrinsically shorter, but have higher intrinsic pitch, than low vowels. Voiced plosives are shorter than the voiceless ones, and so on. As a result, a syllable with a primary stressed high [i:] vowel might be less loud than a secondarily stressed, but low, [ɑ:]. In short, the segments of a syllable influence the raw phonetic measurements of suprasegmental characteristics of that syllable.

Because of that, it makes little sense to compare directly the raw values such as pitch, duration or intensity among various syllables but we should be aware that listeners perceive *relative* characteristics of syllables and the segments within them. People are aware of the intrinsic phonetic features, i.e. how a default [k] is normally aspirated, how a transition between [k] and the following vowel is usually done, etc., and they are then able to factor out these intrinsic features in mentally ‘calculating’ the prominence of the entire syllable. For example, despite the fact that ‘con-’ in ‘continue’ maybe longer and louder than ‘dic-’ in ‘dictation’ because of the difference in consonants and vowels, people use relative prominence of ‘con-’ with respect to other ‘con-’ syllables and relative prominence of ‘dic-’ compared to other similar syllables, in encoding and decoding the degree of prominence of that syllable in speech.²

Secondly, larger prosodic context of a word affects the realization of individual syllables. It matters a lot if a word is prosodically highlighted relative to other words in a phrase, or if it is in the final or non-final position in that phrase. These aspects will be discussed in greater detail in the following chapters of the book. For now, consider five ‘mermaid’ examples extracted from a speech of a single speaker in Activity 9-8.

Activity 9-8: Mermaids

In Praat, open the file ‘mermaids_chain.wav’ and the associated TextGrid file ‘mermaids_chain.TextGrid’ in the book companion. Select both and click *View & Edit*. Listen to these five tokens and describe the differences among them. The crucial task is to imagine that you do not speak English, do not know the word ‘mermaid’, and you should determine which of the two syllables is more highlighted, stressed, which of them stands out more. One way of approaching this task is to say for yourself the word as if it had the primary stress on the second syllable [mə'meɪd] and consider if any of these five tokens might have been a realization of this different stress pattern. If you have a clear case(s) like that, write those down, if not, try to assign probability (e.g. on a scale from 1, least probable, to 3, most probable) to each of the five tokens.

Now inspect the phonetic characteristics of word stress we have outlined in the previous section. See how these realizations are prosodically different from the dictionary version ‘mermaid_CDO_BE.wav’, and how duration, pitch, intensity and segmental quality among the two syllables of the word are different. Try to imagine the larger prosodic context from which these tokens were extracted. Are they highlighted, (non-)final in a phrase?

²See for example Terken and Hermes (2000) for the notions of syntagmatic and paradigmatic prominence relevant to the notions of ‘relative’ prominence in this paragraph.

Finally, open in Praat the TextGrid file ‘mermaids_contexts_chain.TextGrid’, select this file together with Sound ‘mermaids_chain.wav’ and *View & Edit* them together. You see the transcripts of utterances from which the mermaid tokens were extracted. You notice that ‘mermaid’ has different grammatical functions (e.g. nouns vs. attributes) and different positions (final vs. non-final) among other things. If interested, you can listen to the original utterances in mermaid{1-5}.wav files of the chapter companion.

There are several observations that can be drawn from Activity 9-8. Regarding the response to the first task, the 2nd token might have come from the primary word stress assigned to the second syllable for me. Even if you had different answers, the first observation is that dictionary-based, prescribed, patterns of word stress might have been realized in many complex ways phonetically depending on context.

Secondly, even in the clearest tokens with the perception of the first-syllable stress, like the 5th realization, the comparison of the duration, f_0 and intensity does not yield any clear pattern of differentiation between the two syllables. The relationship between the mental patterns of word stress assignments in the mind and the cognitive system of each speaker on the one hand, and the realization of these mental patterns through observable phonetic characteristics of individual syllables on the other hand, is the topic of active research. For example, some scholars believe that word stress is primarily a cognitive concept reflecting the abstract rhythmical structure and thus, it is possible that the extreme variability in how stress might be realized phonetically, prevents its measurability (e.g. Hayes 1995). On the other hand, native speakers almost always notice when L2 speakers mess up the word stress, which supports the observable phonetic reality of word stress.

All the issues mentioned in this subsection (intrinsic differences, importance of relative values and non-straightforward marking of cognitive patterns in phonetics) suggest that for better understanding of speaking habits involved in word stress it is essential to combine findings from speech production, speech perception, neuro-linguistic studies of speech processing and the phonological analyses.

9.5 Basic Word Stress Patterns in English

Similarly to chapters on syllables, consonants, or vowels, we outline in this section a summary of major factors affecting the placement of word stress in English. The goal again is not to provide a complete phonological analysis; there are entire books or research papers written on this topic that the interested reader is advised to consult (e.g. Chomsky and Halle 1968; Liberman and Prince 1977; Hayes 1981). Additionally, discussing phonological notions of historically transparent but synchronically opaque differences in various prefixes and suffixes, the metrical unit of foot or other technical notions, that would be necessary for a comprehensive treatment of English word stress, would take us too far from the hands-on exploration of speaking patterns in this book.

Moreover, the reader should not expect a list of ‘rules’ for stress placement. Internet searching yields many websites and videos promising the rules of English stress; I found guides that start with 4 rules, or have 5, 6, 8, 10, 12 and reach up to 13 rules, and each rule usually has, sometimes sizeable, exceptions, which the guides might or might not mention.

Rather than providing a list of rules, our goal is to increase your awareness of the type of mental knowledge (the patterns) all native speakers of English acquire subconsciously. The relevance of these patterns, as well as a huge number of exceptions, makes the system of English word stress a *semi-predictable* system. With your skills in visualization and measurements in Praat, you can continue exploring on your own both the predictable as well as exceptional patterns. IPA transcriptions in the rest of this section use Southern British English.

Most treatments of English stress agree that there is a difference between verbs and nouns, with adjectives being similar to nouns in some cases and verbs in others. Verbs, and generally also adjectives, have a tendency to carry stress on the final syllables: consider verbs like ‘obey’ [əˈbeɪ], ‘agree’ [əˈɡri:], ‘amaze’ [əˈmeɪz] or ‘process’ [prəˈses], and adjectives like ‘remote’ [rɪˈməʊt], ‘severe’ [sɪˈvɪə] or ‘divine’ [dɪˈvaɪn]. On the other hand, nouns tend to carry stress away from the final syllable (‘message’ [ˈmesɪdʒ], ‘answer’ [ˈɑːnsə] or ‘concert’ [ˈkɒnsət]). This is clear not only in polysyllabic homographs referring to both verbs and nouns like ‘object’ [ˈɒbdʒɪkt] (noun) vs. ‘object’ [əbˈdʒekt] (verb), mentioned already before, or phrasal verbs vs. nouns like ‘set up’ [ˌsetˈʌp] vs. ‘set-up’ [ˈsetʌp] but in general vocabulary as well. Hence, the first observation is that the speaking habits relevant to word stress are linked to other aspects of grammar like the *part of speech* identification.

Within these categories, *syllable weight* clearly plays a role. Note that all verbs above end in a heavy syllable (i.e. having a long vowel, diphthong, of a short vowel with a coda consonant). In verbs and adjectives with a final light syllable, stress tends to shift to the penultimate syllable: ‘copy’ [ˈkɒpi], ‘carry’ [ˈkæri] or ‘empty’ [ˈempti], ‘heavy’ [ˈhevi], and note that also exceptionally words ending in [əʊ] like ‘follow’ [ˈfɒləʊ] or ‘yellow’ [ˈjeləʊ]. Nouns tend to carry stress on the penultimate syllable if heavy (‘diploma’ [dɪˈpləʊmə], ‘agenda’ [əˈdʒendə]), and if light the stress shifts to the antepenultimate (‘camera’ [ˈkæməɹə], ‘syllable’ [ˈsɪləbəl]).

However, even with these generalizations, there are still systematic patterns not accounted for. For example, what about verbs and adjectives like ‘edit’ [ˈedit], or ‘solid’ [ˈsɒlɪd] that end in a short vowel and a coda consonant that we said form a heavy syllable, but stress moves away to the left on the penultimate syllable? And contrast those words with verbs and adjectives in this group: ‘append’ [əˈpend], ‘adopt’ [əˈdɒpt], ‘collapse’ [kəˈlæps] and ‘direct’ [dɪˈrekt], or ‘intense’ [ɪnˈtens]. Here the words end in two coda consonants whereas those in the previous group ended only in a single coda consonant. Hence, the mental patterns involved in word stress must also include awareness regarding the *quality and number of sounds*. One way of covering these patterns is to treat the very

Table 9.1 Examples of derived forms with suffixes that shift the stress to the syllables preceding the suffix. IPA transcription uses Southern British English

Suffix	Base	Base + suffix	Base	Base + suffix
-ic(s)	symbol ['sɪmbəl]	symbolic [sɪm' bɒlɪk]	alcohol ['ælkəhɒl]	alcoholic [,ælkə'hɒlɪk]
-(t/s)ion	educate ['edʒukeɪt]	education [,edʒu'keɪʃn]	examine [ɪg'zæmɪn]	examination [ɪg,zæmɪ'neɪʃn]
-ian	music ['mju:zɪk]	musician [mju'zɪʃən]	electric [ɪ'lektrɪk]	electrician [ɪlek'trɪʃən]
-ity	human ['hju:mən]	humanity [hju'mænɪti]	personal ['pɜ:snəl]	personality [,pɜ:sə'nælɪti]

final consonant in verbs and adjectives as 'invisible' for stress assignment. Hence, words like 'edit' or 'solid' would end in a light syllable, and 'exist' or 'direct' in a heavy syllable, which would then respect the stress-to-weight principle (heavy syllables attract stress) from the previous paragraph.

It is also plainly clear that the *morphological structure* and the identity of prefixes and suffixes also participate in the system underlying the word stress location. So far we have considered mostly words that are morphologically simple in that they do not contain prefixes and suffixes. However, words with suffixes fall into three basic categories. First, there are suffixes that *carry the stress themselves* like {-ese, -eer, -ee, -aire, -ette, -esque, -ique} in words like Vietnamese [ˌvjetnə'mi:z], 'mountaineer' [ˌmaʊntə'nɪə], 'trainee' [ˌtreɪ'ni:], 'questionnaire' [ˌkwɛstʃə'neə(r)], 'kitchenette' [ˌkɪtʃɪ'net], 'picturesque' [ˌpɪktʃə'resk] or technique [tek'ni:k].

Second, there are those that force the stress to appear on the *syllable preceding the suffix*. These include -ic(s), -(t)ion, -ian, -ity, -ify, -al, -ogy, -graphy among others. The relevance of this mental pattern is best seen in cases where adding a suffix changes the original position of the stress as in the words in Table 9.1. The examples also corroborate the need to 'count' the syllables from the right for the purposes of stress assignment in order to show the commonality among these words.

Finally, the suffixes in the third group do not affect the position of the original stress from the base word. The most frequent ones in this group are -less, -ship, -ment, -hood, -ing, -y, -able, or -like.

Adding prefixes into the considerations brings further complexity. Many words have stress following a prefix (inter#ference [ˌɪntə'fɪərəns] or con#tinue [kən'tɪnju]) but there is also a difference between prefixes that are separable with a clear meaning (like 'un-') and those that people now do not perceive as clearly separable from the word stem (e.g. 'con-'). This area is a small example of the complex relationship between word stress and the morphological makeup of each word participating in the mental knowledge of English speakers. Word stress patterns in general feature prominently in debates on the theoretical models describing the phonology-morphology relationship.

In longer words, the awareness of the *secondary stress* is also important. We have explored the phonetic marking of this intermediate prominence in 'fofofo'

or ‘contemplation’ in the previous sections. Most of these words, barring some borrowed geographical names like ‘Mississippi’ or ‘Manitoba’, are morphologically complex containing affixes. The rule of thumb is that the syllable carrying the secondary stress almost always precedes the one with the primary stress and is commonly not immediately adjacent to it. Adjacent stress syllables, commonly referred to as stress clash, tend to be avoided, which will be discussed further in upcoming chapters. This rule is respected in all the words marked for secondary stress above. Exceptions include ‘trainee’ [ˌtreɪˈniː] above or ‘fifteen’ [ˌfɪfˈtiːn] but note that for ‘princess’ both [ˌprɪnˈses] and [ˈprɪnses] are possible realizations of this word in isolation probably motivated by the avoidance of adjacent stressed syllables.

Secondary stress in words with suffixes also tends to preserve the primary stress of the base without those suffixes. For example, in 6-syllable words with the ‘-tion’ suffix and the primary stress on the syllable preceding it, the secondary stress falls on the initial syllable in ‘justification’ [ˌdʒʌstɪfɪˈkeɪʃn] but on the second syllable in ‘communication’ [kəˌmjuːnɪˈkeɪʃn]. This is to preserve the primary stress of ‘justify’ [ˈdʒʌstɪfaɪ] and ‘communicate’ [kəˈmjuːnɪkət].

The avoidance of adjacent stressed syllables and the preservation of the primary stress can be construed as violable constraints and native speakers are aware not only of their existence but also of their *precedence relationships*, or rankings. For example, in ‘Japanese’ [ˌdʒæpəˈniːz] stemming from ‘Japan’ [dʒəˈpæn], the avoidance of stress clash is more important than the preservation of the original stressed syllable since *[dʒəˌpæˈniːz] is not used by native speakers.

Another type of awareness that native speakers of English subconsciously use in coding word stress relates to *compound* words. Compounding is a very productive word formation process in which two separate words form a novel word (‘bus stop’, ‘old-school’, ‘coffee machine’, ‘blackboard’). Spelling may vary, separated with a white space, hyphen, or adjoined, but crucially for us these words have only a single word stress. This contrasts with two-word phrases that stress both elements: a [ˈblækbɔːd] in schools vs. a [ˈblæk ˈbɔːd] for snowboarding; ‘snowboard’ here being another compound word. There are many such pairs like ‘green house’ (a house made of glass for growing plants vs. a house of green colour) or ‘German teacher’ (person teaching German vs. teacher from Germany). With compounds, nouns and many verbs, are commonly stressing the first word (and very commonly spelled adjoined) whereas adjectives and adverbs might attract the second word stress (e.g. ‘old-fashioned’). Additional tendencies beyond the scope of the book involve consideration like whether the word is a food item, or a geographical name, part of a larger phrase and so on.

Although patterns like these are just tendencies and are not conclusive and exhaustive, they provide some guidelines for non-native speakers of English in trying to determine word stress in a word they are not sure about. Nevertheless, identifying, and becoming aware of, a stress pattern with newly acquired words, either from a dictionary or other sources, is still a necessary part of learning English for L2 speakers.

9.6 IPA Transcription of English Words

Starting with discussing English sounds in Chapter 5, the book includes cursory remarks on the correspondence between English spelling and IPA transcription. I strongly encourage both native and non-native speakers to do transcriptions regularly since it is a great way of increasing awareness of the speaking patterns and proficiency in writing and reading transcriptions is an integral part of that awareness.

Specifically relevant to this chapter, you should be aware that stressed syllables never carry a schwa and the unstressed syllables are by far the most commonly realized as [ə], and sometimes as high lax vowels [ɪ] and [ʊ]. Other vowels are very rare in unstressed syllables. You always mark both primary stress, and secondary if present, and get used to doing this before the syllable itself.

Exercises

- 9-1 Consider the list of words below. Transcribe them in IPA and indicate the primary and secondary stresses based on your intuition. Then, try to provide arguments supporting your stress placement considering the patterns discussed in Sect. 9.5: morphological structure, syllable weight, part of speech, etc. and noting potential exceptions or deviations. Finally, record a native and a non-native speaker saying some of these words and inspect the realization of the phonetic aspects of stress discussed in this chapter.

Example: ‘developmental’ [dɪˌveləpˈmentl̩]. This is clearly a morphologically complex word (‘develop-ment-al’) with the base verb ‘develop’ with the penultimate stress since for verbs the final consonant is ‘invisible’ making the final syllable light, which shifts the stress to the penultimate one. Adding ‘-ment’ does not change the primary stress [dɪˌveləpment] but adding further ‘-al’ does since it is one of the suffixes forcing the primary stress on the syllable preceding the suffix. Regarding the secondary stress, both the avoidance of stress clash and the preservation of the original stressed syllable are respected in [dɪˌveləpˈmentl̩]. Hence, there is no reason to shift the secondary stress to the first syllable; compare with ‘department’ → ‘departmental’.

- Words: comfortable, spaghetti, washing machine, suspect (noun and verb), confederacy, ...

References

- Chomsky, Noam, and Morris Halle. 1968. *The sound patterns of English*. New York: Harper & Row.
- Fry, Dennis B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* 27: 65–768.
- Fry, Dennis B. 1958. Experiments in the perception of stress. *Language and Speech* 1: 120–152.
- Hayes, Bruce. 1981. *A metrical theory of stress rules*. New York: Garland Press.

- Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. Chicago: The University of Chicago Press.
- Ladefoged, Peter, and Keith Johnson. 2015. *A course in phonetics*, 7th ed. Stamford, CT: Cengage Learning.
- Lieberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8 (2): 249–336.
- Terken Jacques, and Dik Hermes. 2000. The perception of prosodic prominence. In *Prosody: theory and experiment*. Text, Speech and Language Technology, vol. 14, ed. Merle Horne, 89–127. Dordrecht: Springer.



In this chapter we will

- explore the speaking habits that underlie smooth linking of adjacent words together
- begin analysing longer phrases and practice further verbal descriptions of articulatory activity
- provide additional guidelines for carrying out projects examining speaking behaviour with Praat

10.1 Introduction

If you ever studied a foreign language, you know that a very important milestone is reached when you are able to recognize individual words in fluent speech. And if you listen to somebody speaking a language you don't know, you might be pretty good at chunking speech into larger phrases but you have no idea where the word boundaries are. The primary reason why individual words are not uttered separately with clear boundaries is the underlying conflict between exerting as little effort as possible (economy) and conveying all information and contrasts necessary for the successful transmission of the message. This was concisely formulated by Bjorn Lindblom in his *Hyper-articulation and Hypo-articulation (H&H)* theory (Lindblom 1990) and applies to many aspects of speaking behaviour both within and across words. For example, try saying any sentence on this page in a style of old-fashioned robots separating each word and you realize how

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_10) contains supplementary material, which is available to authorized users.

effortful this way of speaking is. On the other hand, in environments with lots of noise, or when speaking to a not proficient speaker of a language, speakers have a tendency to exert this effort and signal word boundaries in order to be understood. Since transitions between words that do not signal major boundaries are also hypo-articulated, speakers can afford in many situations to exert little articulatory effort in cuing them. Importantly, studies also show that listeners also expect, and benefit from, this hypo/hyper-articulation both within and across the words.

On the one hand, the magnifying glass with which we zoom to our speaking behaviour in this chapter will be coarser than in the previous chapters since we move from the exploring speaking habits within words, like in segments, syllables or stress, to strategies employed when linking adjacent words together. On the other hand, however, we will explore how the realization of individual sounds is affected by the context of word boundaries. In this we will zoom in to fine articulatory adjustments and their acoustic consequences in the production of individual sounds. We thus extend the understanding of the characteristics of vowels and consonants from Chapters 5–7.

10.2 Preview of the Speaking Habits Under Investigation

Activity 10-1 outlines the types of patterns we will be dealing with in this chapter.

Activity 10-1: In the top right hand corner of the page

Imagine a situation in which you point the attention of your friend where to look in order to find the relevant information and say ‘in the top right hand corner of the page’ in as natural way as possible. Please record this utterance with Praat yourself, or if you are not a native speaker ask one to say this for you and record. Next, record one more time but now say each word separately as if in the dictionary form: ‘in ... the ... top ... right ... hand ... corner ... of ... the ... page’. Now for both files, produce a TextGrid annotation with a single tier ‘words’ in which you label the intervals for each word as accurately as possible. Please indicate also the silences between words in the second recording with a ‘#’.

Your task is to compare the realizations of the words in the two recordings paying specific attention to the boundaries between each pair of adjacent words. Select the interval comprised of two adjacent words with the mouse and listen repeatedly. Hence, inspect ‘in the’ in the first recording and compare with ‘in # the’ in the second recording. Then ‘the top’ with ‘the # top’ and so on. Note down your observations, which might be quite revealing for many of you!

In the text below I analyse a realization of this utterance by an Australian speaker. Compare your observations with the discussion below. What were the similarities? If there were differences, can you speculate as to their reasons? Alternatively, would it sound natural to you to say it like this speaker did?

After doing Activity 10-1, open both the sound and Textgrid files ‘corner_of_the_page’ from the book companion in Praat and click *View & Edit* when both are

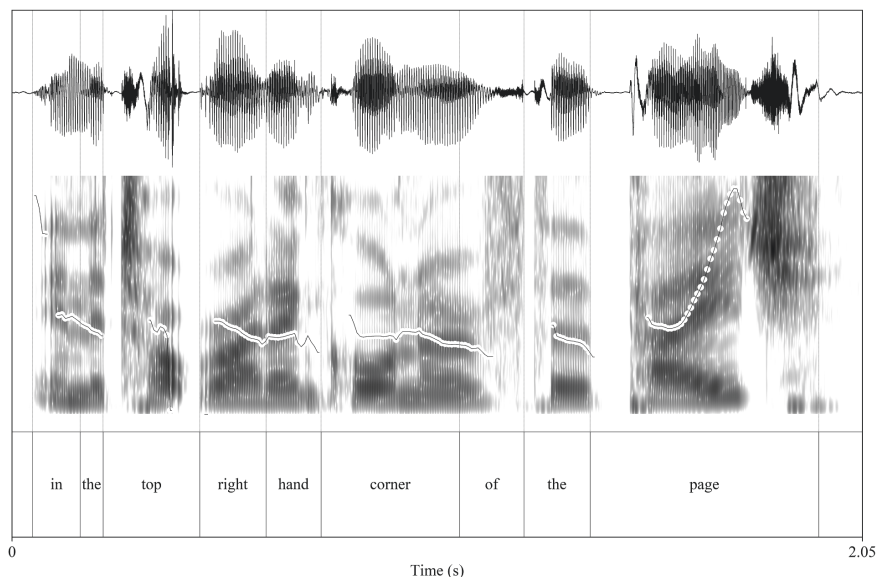


Fig. 10.1 Visualization of the sound wave, spectrogram, f_0 tracking and text-to-speech alignment for the utterance ‘in the top right hand corner of the page’ discussed in the text

highlighted for an image similar to Fig. 10.1. In the extended commentary below I analyse this file, which previews the discussion in the rest of the chapter but also illustrates the level of analysis with which you should be comfortable at this stage in the book.

It is good practice, and a habit you should adopt, to start any investigation of spoken utterances with considering first the global aspects and proceed to more local ones. This utterance was produced as a single chunk but you may realize the speaker slowed down for the last three words, possibly reflecting the cognitive effort in finding the right words relevant for the message. As a result of this slowing down, the last three syllables took roughly 850 ms while the first seven syllables took only slightly longer than that (about 1000 ms). Due to this, the junctures ‘of the’ and ‘the page’ are produced very clearly and may be similar to the realizations when you were producing these words separately ‘of ... the ... page’. In addition to boundaries and tempo, relative prominence of individual words also belongs to the global aspects to consider. Here, ‘page’ is clearly the most prominent word but other content words ‘top’, and possibly also ‘right’ and ‘corner’ are also prominent but these are definitely not as salient as ‘page’.

Moving now to the focus of this discussion, we take each boundary between adjacent words in turn. In ‘in the’ it looks, and sounds, like the speaker deleted [ð] of ‘the’ and realized this juncture as [ɪnə] or alternatively [ɪnnə] replacing [ð] with [n]. Try to say this and note where the tip of your tongue is touching during ‘in the’. Is it the alveolar ridge or partly also the teeth? Most native speakers

would touch the teeth in anticipating the dental [ð] to save the articulatory effort of moving the tongue from the alveolar ridge to the teeth. Some native speakers, and many non-native ones have a tendency to touch the alveolar ridge rather than the teeth. But for both groups, the realization with the tongue tip actually moving from the alveolar ridge to the teeth is only possible in the robot-like ‘in ... the’ when saying each word separately.

Realizations of ‘the top’ and ‘top right’ are not much different from what might be expected if the words were produced separately and both alveolar and bilabial closures in the plosives of ‘top’ are fully realized. The glottalization before and after the bilabial closure is clearly visible with vertical striations in the spectrogram.

The following ‘right hand’ with a smooth transition, however, looks and sounds quite different compared to a full realization ‘right ... hand’. The final /t/ has a very brief, and voiced, closure very similar to the flapped allophone of /t/ [ɾ]. Also, /h/, that is normally voiceless word-initially, is fully voiced, which is another phonetic marker of an extremely weak boundary and tight linking of these two words. You might wonder why this speaker is not glottalizing the [t] in ‘right’ especially since she is glottalizing frequently in her speech and does so also for the [p] in ‘top’. I would say that again the adjacent context plays a role. Note that the following ‘h’ is a glottal fricative requiring a narrow opening of the glottis while glottalization stiffens the vocal cords into a (in)complete closure. The economy of articulatory effort is likely responsible again similarly to the ‘in the’ juncture: producing two different settings of the glottis in quick succession represents great articulatory effort.

In the next boundary ‘hand corner’ we neither hear nor see any evidence for the word-final /d/ and there seems to be only a single velar closure. Try to say ‘hand corner’ imitating the speech and spontaneity of this speaker and note again the placement of your tongue. In this context, many native speakers would not move the tongue from the alveolar place for [n] to the velar closure for [k] but produce the nasal with the already formed velar closure: [hæŋkɔ:nə], which is a rather drastic difference from the canonical realization of ‘hand’.

The final juncture in the fast section of the utterance is ‘corner of’. Here the relevant question is what happens to the final <r> of ‘corner’. This is a non-rhotic speaker of Australian English, which can be seen in the absence of [ɹ] in the first syllable of ‘corner’. Hence, the word-final <r> should not be realized either but in listening to this interval, slight traces of [ɹ] might be heard.

Finally, the last three slowed down words ‘of the page’, as we said before, are produced with full realizations of the segments in the vicinity of these boundaries.

The sample of speaking habits observed above and active in forming smooth transitions between words within the general H&H concept are commonly termed *connected speech aspects* and typically include three basic phenomena: assimilations, elisions and linking (or liaison). Before we discuss them in more detail, note again that we are not talking about ‘rules’ of ‘correct’ pronunciation but rather tendencies of how native speakers adjust to the communicative situation. I doubt that your utterance recorded in Activity 10-1 above displayed precisely those

aspects we noted with the realization just analysed. Connected speech aspects are highly dependent on the style, situation, even the emotional state of the speakers. Important for us is to try to understand the processes and habits rather than specifying any ‘rules’ for their deployment.

A useful approach to connected speech processes is to examine what segments appear in the vicinity of the word boundaries. The three main categories include:

1. **A vowel and a consonant flank the word boundary.** This is when the first word ends with a consonant and the second word starts with a vowel or the first word ends with a vowel and the second word starts with a consonant. This very common situation is represented in our example file with $V_f\#C_i$ junctures ‘the top’ and ‘the page’ and $C_f\#V_i$ junctures like ‘come about’ or ‘he’s unhappy’ in other activities of this chapter.
2. **Consonants flank the word boundary from both sides.** A word-final consonant (C_f) is followed by a word-initial consonant (C_i) in a $C_f\#C_i$ juncture. In Activity 10-1, the junctures ‘in the’, ‘right hand’ or ‘hand corner’ belong to this category. We saw that these flanking consonants might be sometimes left out completely, like /d/ of ‘and’ or ‘hand’, they might be significantly reduced, like /t/ of ‘right’, and might also undergo alterations in which their phonetic characteristics change to become more similar to the adjacent consonants, like /n/ of ‘in’ or ‘hand’.
3. **Vowels flank the word boundary from both sides.** A word-final vowel (V_f) is followed by a word-initial vowel (V_i) in a $V_f\#V_i$ juncture. The only juncture of this type in the examined utterance was ‘corner of’ since in non-rhotic dialects the first word normally ends in a vowel [ˈkɔːn]. In these dialects, the word-final <r> actually surfaces in a smooth transition if the next word starts with a vowel. Other examples for these junctures include ‘the evening’ or ‘to ask’.

In the first case, an adjacent vowel and consonant are simply joined as we would do within the words for the most natural and economic way of producing a smooth transition. Sometimes these junctures display characteristics of word-internal processes. For example, flapping in American English, typical for words like ‘city’ or ‘water’, also takes place in weak junctures like ‘what about’. Also, *re-syllabification* might take place such that C_f , that is a coda of the pre-boundary syllable, might re-syllabify into one syllable with the V_i of the post-boundary word and some people might syllabify in speech (of course not in text) ‘come about’ as [kλ. mə.bəʊt]. On the other hand, allophonic variation, which varies strongly with syllable affiliations, commonly shows the absence of such re-syllabification and actually reinforces the presence of the word and syllable boundary for the listener. For example, in ‘it ends’ people do not re-syllabify the coda /t/ into the onset since the amount of aspiration in non-flapping dialects of English is usually not the same as it would be in the canonical onset /t/ of ‘tends’.

Categories #2 and #3 above require a more detailed discussion that is organized into the following sections. In $C_f\#C_i$ junctures at least two full successive constrictions are required. One or more of these constrictions may be adjusted to be more

similar to the adjacent constriction, (*assimilation*, Sect. 10.3) or the consonantal constriction could be apparently left out completely or significantly reduced (*elision*, Sect. 10.4). In $V_i\#V_i$ junctures English speakers commonly insert *linking* consonants (Sect. 10.5).

Before we proceed to these sections, it is essential to keep in mind, once again, that all the adjustments to the canonical forms are fundamentally continuous and splitting them into certain discrete categories like the elision, assimilation or insertion above is just a convenient help to describe most perceptible speaking habits rather than the precise characterization of all possible situations. Moreover, we should keep in mind the social context and the pragmatic goals that are commonly reflected in the global characteristics of the utterance and interact with the phonetic features characterizing the segments in the vicinity of the word boundaries.

10.3 Assimilations

We talk about assimilations when in the $C_i\#C_i$ environment, one or more characteristics of consonantal constrictions (articulatory place, manner, or voicing) are changed to be more similar to the adjacent consonant. Activity 10-2 sets the stage.

Activity 10-2: Ten billion people

First, say this phrase in a careful formal style ‘ten ... billion ... people’ paying attention to the coordination of articulatory movements in the vicinity of the boundaries. Note how the tongue movement for the alveolar constriction precedes the bilabial closure of the lips. Now think of a situation in which you produce this phrase in casual somewhat faster speech, for example excited about great audience at a concert or a sporting event exaggerating: ‘There were like ten billion people there!’. What is happening to the articulatory movements of the tongue and the lips?

Then, ask a native speaker friend who doesn’t know what you want to study to record the following utterances. If not available, the next best option is to record yourself, and the third best is to inspect the file ‘ten_billion_people.wav’ provided in the book companion.

The recording situation might be something like this:

- A: Imagine. By 2060 there will be ten ... billion ... people on the Earth!
 B: Yeah, but feeding ten billion people is not gonna be easy.

Once recorded produce two interval tiers in your TextGrid: words and closures. Align the intervals for the target words ‘ten’, ‘billion’, ‘people’ in the signal in the first tier with the help of the spectrogram. In the second tier, identify the closures for the word-final nasals and word-initial bilabial stops. You notice that the closures in the first case ‘ten ... billion ... people’ should be relatively easy to identify. For the onset on the nasals, there is a salient decrease in energy, a lighter shade in the spectrogram, particularly around 1500 Hz and

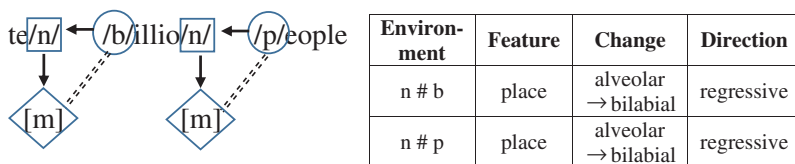


Fig. 10.2 Characterization of nasal place assimilation in ‘ten million people’

most crucially, the disruption of the clear formant structure at the end of the preceding vowel. Is there a release of the nasal closure? The onsets of the bilabial stops might not be clearly visible after silent pauses.

Now do the same for the B’s utterance and inspect the realization of closures in fluent ‘ten billion people’. Describe the difference between the two realizations. Strengthen the link between the visual image, auditory impression and proprioceptive feel of the two realizations.

In ‘ten billion people’, both word boundaries have two flanking consonants and in both, C_f is the nasal alveolar /n/ and C_i is the bilabial /b/ and /p/, respectively. A common unconscious speaking habit of English speakers in casual situations is that the word-final alveolar nasals are not produced with the tip/blade of the tongue but change their place of articulation to be identical to the place of articulation of the following word-initial consonant: ‘te[m]billio[m]people’. This most common /n/-assimilation in English and many other languages falls under *nasal place assimilation* (NPA) and is schematically illustrated in Fig. 10.2.

In this type of change, the C_i /b/ or /p/ is the *trigger* phoneme shown with the circle in Fig. 10.2. In both cases, C_f /n/ is the *target* intended phoneme shown with the square. A phonetic feature of the trigger C_i ‘causes’ the change in that feature of the target C_f : The change is shown with the vertical arrows and the phonetic realization in the square brackets and the diamond. In other words, the speaker anticipates the bilabial place of articulation of C_i and changes the place of C_f to be maximally similar, in this case identical, to the place of the following C_i . This similarity is shown with the double dotted line in Fig. 10.2. As a result, only a single constriction (bilabial) is needed instead of two different constrictions (alveolar, bilabial) and thus the first constriction is not released.

The direction between the trigger and the target of the change is described as *regressive*, or *anticipatory*, and illustrated with the horizontal arrows. These characteristics of nasal place assimilation are summarized in the table of Fig. 10.2.

Activity 10-3: /n/-assimilation

Inspect and describe verbally the processes in phrases ‘ten crowns’ and ‘ten thousand’ using the table in Fig. 10.2 and the associated discussion as a guide. Link to your observations of your own speaking.

In ‘ten crowns’, the following velar plosive [k] triggers the regressive assimilation and the target alveolar nasal changes to the velar nasal [ŋ] so that the adjacent consonants flanking the word boundary share the single velar closure.

In a casual production of ‘ten thousand’, the alveolar nasal changes into a dental nasal [ɲ] to assimilate to the place of articulation of the following dental fricative. Hence, not only plosives with complete obstructions but also fricatives may trigger NPA. Note that some native English speakers, and many non-native speakers might produce the initial [θ] of ‘thousand’ similar to an alveolar [t] and for those speakers assimilation does not take place since the two adjacent sounds are both alveolar. Sometimes, assimilations could result in allophones, like the dental nasal.

Regressive assimilations in which the target is not a /n/ are also common in English and most frequently involve alveolar plosives. For example, /t/ commonly assimilates its place to the place of the following consonant as in ‘tha[p] person’, or ‘go[k̚] that’ in which the target C_f /t/ is realized as a bilabial or a dental stop, respectively.

You may have noticed that so far the targets of regressive assimilations were alveolars with complete oral constrictions. Labials and velars are assimilated much less frequently in English. For example, English /m/, /ŋ/ do not show a strong tendency to assimilate: Compare ‘ten {billion, million, thousand, crowns}’ with ‘some {dude, night, thing, case}’ or ‘sing {billion, million, thousand, decent} people’ and you notice that the bilabial/velar constriction is much more likely to be preserved than the alveolar one. This ‘weakness’, or propensity for change, of alveolar codas compared to labial/velar ones has also been observed in other languages and other processes (e.g. Iverson and Kim 1987 for Korean).

Activity 10-4: Other changes to word-final consonants

Consider the realization of word-final consonants in some realizations of ‘was’, ‘have’, ‘has’, ‘as’ or ‘of’. Produce utterances in which they are followed by a pause. Now say them in junctures like ‘have to’, ‘has to’, ‘was teaching’, ‘of people’ or ‘as soon as possible’. What changes do you observe? Why are they taking place?

In pre-final positions like ‘yes, she was’, ‘We have!’, or ‘such as?’, many native speakers produce the (partially) voiced [z] or [v], which is also a dictionary pronunciation. However in the weak junctures like those in Activity 10-4, you observe the voiceless realization [s] or [f]. The only difference between [s] and [z], or between [f] and [v] is voicing. Hence, another common process is regressive **voicing assimilation**. In English, by far the most frequent is the change from the fricatives /z/ and /v/ to their voiceless pairs [s] and [f] in these functions words.

► Advanced Section: Note for non-native speakers

Speaking habits of many languages include regressive voicing assimilations in which a word-final voiceless consonant changes to a voiced variant. However, in English these assimilations are quite rare and non-native speakers

might want to try to adjust their L1 habits in this way. Hence a phrase like ‘I like that black dog’ might come out as [aɪ laɪk ðæt blæk dɒk] for many L2 speakers. The first three voicing assimilations are common for Slavic, Romance languages or Dutch. In English, however, word-final voiceless (fortis) consonants tend to retain their voicing and word-initial voiced (lenis) consonants might be devoiced to some degree: [aɪ laɪk ðæt blæk dɒk]. On the other hand, the pre-final voicing neutralization in which the voiced (lenis) obstruents are devoiced preceding major prosodic break, like [dɒk] above, is common for Slavic languages, German, or Catalan and these speakers have a tendency to use the voiceless [s] and [f] in the pre-boundary environment in Activity 10-4 (‘yes, she wa[s]’, ‘We ha[f].’, or ‘such a[s]?’). English speakers, again, tend not to fully devoice these consonants, but the voicing of word-final fricatives is extremely variable. I encourage you to record phrases like these from native and non-native speakers and inspect with Praat.

Two less frequent assimilation types in English include the assimilation of *manner* and the *progressive*, or perseverative, assimilation. The former takes place when the place and voicing of the consonant remain the same and only its manner changes. A good example are set phrases like ‘good night’ in which C_f /d/ of ‘good’ is commonly changed to [n] sharing the nasal manner, and the alveolar place, of articulation with the trigger C_i /n/ of ‘night’: [gʊnnat].

Progressive assimilations commonly involve manner as well and are most frequent with dental fricatives in English. How do you say phrases like ‘in the morning’, or ‘all the time’? For many speakers [ð] of ‘the’ is changed to be identical to the preceding C_f : [ɪnnə] and [ɔ:lə]. Hence, the trigger in this case is the word-final consonant, the target is the word-initial consonant and the intended /ð/ is realized as [n] or [l] preserving the alveolar place of articulation but changing the manner into the nasal or lateral. Figure 10.3 schematizes this process. Also very common in these phrases is the combination of the regressive assimilation of place with the progressive assimilation of the manner: [ɪnnə] and [ɔ:lə].

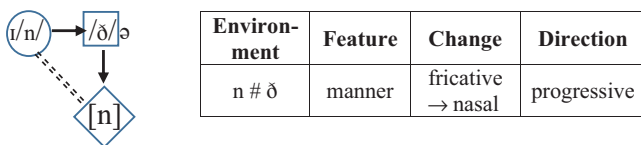


Fig. 10.3 Characterization of manner assimilation in ‘in the’

► Advanced Section: Cognitive patterns underlying ‘in the’ realization

How do people say ‘in the’? There are essentially four options: [ɪn(ɱ)ə], [ɪn(n)ə], [ɪndə] and [ɪnðə]; the brackets indicate the possibility of shortening the nasal. All four are possible and attested. Introspect your own speech, concentrate on the stricture location (dental/alveolar) and the release (burst/not) and ask around also other speakers. What processes lead to these outputs?

- [ɪnðə] arises from the regressive assimilation of place in which the target word-final alveolar /n/ changes to the dental variant [ɱ] to assimilate to the following word-initial trigger [ð].
- [ɪnnə] may result from the progressive assimilation of manner and place (dental fricative [ð] changing to alveolar nasal [n]).
- One way of explaining [ɪnðə] is that after a) above ([ɪnðə]) the progressive assimilation of manner applies. Note that the reverse ordering does not work. With progressive assimilation first, we get b) [ɪnnə], in which both manner and place are assimilated. But now there would be no trigger left to motivate the dental place of articulation for the nasals in [ɪnðə].
- [ɪndə] may result from the progressive assimilation of manner and place ([ð] → [d]) but it might also be the case that the speaker actually pronounces ‘the’ as [də] in th-stopping (Chapter 8).

Hence, both regressive and progressive assimilations might occur separately, resulting in dental [ɪnðə] or alveolar [ɪnnə]/[ɪndə], respectively, but if they occur together, the regressive must precede the progressive in order to produce [ɪnðə]. Similarly to stress clash and preservation in word stress (Chapter 9 discussion of ‘Japanese’) or pre-fortis clipping and flapping (Chapter 7 discussion of ‘writer’-‘rider’), speaking habits at the word boundaries also include not only separate assimilatory processes but also the knowledge of their precedence/temporal relationships.

Finally, we finish this section with a special case of assimilation. We start by brainstorming about potential connected speech aspects in the following phrase: ‘Do you have a favourite thing that you wrote?’ Write down the broad transcription of this phrase (as if each word was said separately) and for each juncture between adjacent words discuss the processes that might apply if this phrase is produced in fast fluent speech. Consider all the word boundaries, but pay particular attention to ‘do you’ and ‘that you’ junctures. Compare your observations with those in Activity 10-5 below.

Activity 10-5: Do you have a favourite thing that you wrote?

For this activity it is essential to work with the sound and textgrid files ‘favourite_thing.wav’ and ‘favourite_thing.TextGrid’ provided in the book companion.

Open the files in Praat and make the spectrogram visible. Now click on the interval in the second ‘words’ tier labelled as ‘do’. Listen just to this interval and transcribe what you hear. Try to explain informally what processes applied for this output to be realized. Now select the interval covering ‘do you’ and assess if your perception has changed and why.

With your informal explanation of this first juncture, consider now the ‘that you’ juncture. Can we expect a similar realization as in ‘do you’? Is this expectation realized? Again, describe informally what is happening in the speaker’s mouth during this juncture.

Finally, apply what we learned about assimilations in this chapter so far and explore these aspects of connected speech in this file, particularly the juncture ‘favourite thing’. Compare your observations with the text below.

So in the first task of Activity 10-5, I would transcribe ‘do’ as [tʃ] and listening to ‘do you’ I would lean to [dʒə], or possibly [dʒjə]. First, the realization of ‘do’ is weakened so much that the vowel is not realized at all. This results in [d] and [j] of ‘you’ being adjacent and the process called *coalescence* creates [dʒ]. Since /d/ is alveolar and /j/ is palatal, the place of articulation of the resulting [dʒ] is post-alveolar; hence, locally ‘in between’ the places of the intended sounds. Therefore, the assimilation does not go in a single direction but both C_f and C_i merge to make a place of articulation that is intermediate between the two. This is a very frequent phenomenon in English with alveolar C_f followed by ‘you’ in cases like ‘did you’ or ‘what you’.

The clear auditory percept of devoicing in [ɖ] is corroborated visually in Praat since there is no periodicity neither in the closure nor in the release within the ‘do’ interval, which can be confirmed by zooming in and is illustrated with the leftmost ellipsis in Fig. 10.4. This is another example of a tendency in English to devoice plosives and affricates following a pause.

The second task of Activity 10-5 focused on the juncture ‘that you’, shown with the rightmost ellipsis in Fig. 10.4. Following the preceding discussion, it would be reasonable to expect a similar case of coalescence [ðætʃjə]. However, we see visually, and hear auditorily, that this is not the case. The speaker glottalized the word-final /t/, seen with vertical striations in the spectrogram, effectively removing the need for the alveolar constriction of the tongue, which in turn obviated the coalescence of the alveolar and palatal constrictions into [tʃ]. Note that the glottalization of the initial /d/ of ‘do’ is not available in this environment. Hence, we see another complexity resulting from the interplay of multiple connected speech processes.

Note also your perception when you play just the word ‘that’. The speaker actually produced something like [ðɪʔ]. This shows that assimilatory processes may also target the phonetic vowel quality since the high front characteristics of /j/ of ‘you’ triggers the realization that is slightly raised and fronted compared to the canonical [ə].

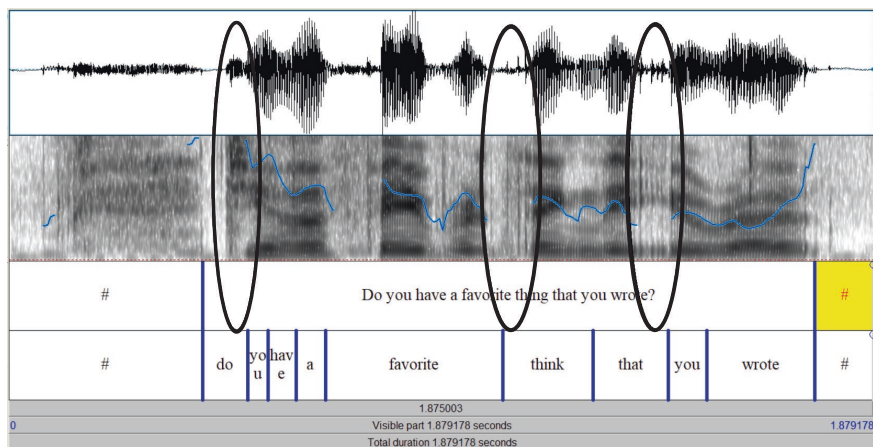


Fig. 10.4 Visualization of three target junctures in Praat. See text for detailed description

Finally, the middle ellipsis in Fig. 10.4 points to the juncture ‘favourite thing’. Despite some modal irregularities during the closure, I would analyse this as involving the regressive assimilation of place in which the C_f /t/ is dentalized [t̪] to anticipate the dental place of the C_i [θ] of ‘think’. Try to produce this juncture and inspect what your unconscious habit in producing junctures like these is. I expect that hardly anybody would produce a full alveolar closure, it would probably be a ‘mouthful’, but both the glottalization as well as the dentalization of the final /t/ are available.

The assimilatory coalescence involving C_f alveolar fricatives may also be observed in phrases like ‘thi[ʃ] year’, ‘I miss [ʃ] you’, or ‘rai[ʒ] your voice’.

10.4 Elisions

In addition to assimilations, another common process active in $C_f\#C_i$ word boundaries in connected speech can be described as weakening, or eliding completely, one of the consonants. We have seen a prototypical example in ‘hand corner’ in Activity 10-1. Similar to assimilations, elisions also target primarily alveolars, and most commonly the plosives /t,d/ when in word-final position and flanked by another consonant from both sides, as in ‘hand corner’. It might also co-occur with assimilation as in [hæŋkɔ:nə]. Additional common elisions involve [v] in ‘of’ in phrases like ‘piece of paper’ or ‘most of the people’.

Consider t/d-deletion as the most widespread case of elision in English. As with all connected speech phenomena, we have to stress that t/d-deletion is an optional process and so far we discussed the importance of the strength of the juncture

following the target word. Additionally, we saw that another optional process of glottalization is strongly linked to various regional accents and might interact with other connected speech processes. In Activity 10-7 we explore what factors affect t/d-deletion systematically, or in other words, we target the conscious awareness of the unconscious habits relating to t/d-deletion. We investigate the habits of native speakers of English but non-native speakers may either compare with natives, if available, or explore how their non-native habits fare compared to the native ones.

Activity 10-6: To delete or not to delete t/d in English?

We will take candidates for t/d-deletion, place them in various environments and your task will be to determine if t/d-deletion is more likely in one or the other environment. Hence, you should try to say the required phrases with and without t/d-deletion, and present a metacognitive judgment what feels preferred and most natural to you. Table 10.1 provides guidance.

First, take ‘most people’ and ‘old people’ (A) and compare with ‘most adults’ and ‘old adults’ (B). Is t/d-deletion more natural in A, B or equally in both A and B? What is the difference between environments A and B and what rationale might explain their potential difference?

Second, take A’s ‘meet friends, ‘odd weekends’ and compare with B’s ‘at least friends’ and ‘cold weekends’. Again, what is the relevant difference between A and B and what rationale might explain their potential difference in cuing t/d-deletion?

Third, consider ‘soft house’, ‘wind hill’ (A) and compare with ‘soft bone’, ‘wind pipe’ (B). Repeat the mental exercise from the previous two tasks.

Finally, compare A ‘(morning) mist picture’ and ‘impact factor’ with B ‘a missed picture’, or ‘packed food’ and try to reason about the potential differences and underlying reasons for them.

Summarize your observations in the table and try formulating informally the mental system that guides subconscious decisions of native speakers whether to delete or not to delete the word-final t/d. Then compare with the text below.

Table 10.1 Environments to consider for t/d realizations. See text of Activity 10-6

Task	Environment A	Environment B	Preference	Rationale
1	Most people, old people	Most adults, old adults		
2	Meet friends, odd weekends	At least friends, cold weekends		
3	Soft house, wind hill	Soft bone, wind pipe		
4	(Morning) mist picture, impact factor	A missed picture, packed food		

The t/d-deletion has been thoroughly studied and found to operate in all major dialects of English. The generalizations from the literature suggest that native speakers prefer the deletion, or delete more frequently, in environments ABBA in the 4 tasks above, respectively. Hence, the following word-initial consonant rather than a vowel (1), a consonant preceding the target word-final t/d (2), oral rather than the glottal following consonant (3) and t/d **not** forming a separate grammatical morpheme of the past tense (4) all induce more deletion than the following vowel or glottal consonant, the absence of a preceding consonant, and t/d marking past tense. The rationale covering all the observations from the first three tasks suggest that the crowding of oral consonantal gestures prompts t/d-deletion, which is in line with the economy of articulatory effort. The fourth task points to the pattern that prefers the retention of the morpheme expressing a grammatical function over its elision and favours conveying important information despite increased effort. Hence, Activity 10-6 and the awareness we gained fit with the H&H theory discussed in Sect. 10.1.

Of course, all of these patterns interact among themselves, with the smoothness of word boundary transitions, distribution of prominences in a phrase and several other constraints. For example, Guy and Boyd (1990) showed that the likelihood of deleting the word-final past tense morphemes decreases with age. Hence, t/d-deletion is an example of an optional, but definitely not random or chaotic, speaking habit that nicely illustrates the complexity and interaction of many participating factors.

Note that in Activity 10-6 I alternated the environments favouring and dis-favouring t/d-deletion to avoid priming your judgement (ABBA instead of AAAA or BBBB). This alternation in the stimuli mitigating possible task effects is a good practice if you are involved in projects exploring speaking habits in which you don't want to analyse and introspect your own speech but want to get a more general picture informed by several observations.

► Advanced Section: Apparent elisions

A closer inspection of the crowding of oral consonantal gestures leading to elisions reveals that some cases are best described as **apparent elisions**. For example, Browman and Goldstein (1990) used the methodology of electromagnetic articulography that traces the movements of little sensors attached to articulators (also discussed in Chapter 2) to investigate the production of juncture clusters like [kt#m] in 'perfect memory'. We saw above that this is the prototypical environment favouring t-deletion. With a hyper-articulated juncture the final /t/ was audibly heard and released whereas in a fluent hypo-articulated phrase the alveolar closure and release were missing in the auditory signal. However, the inspection of the articulatory movement of the tongue blade showed that in both cases the alveolar closure for [t] was attempted, only with the fluent speech it was masked by the overlapping closures for the velar preceding and the labial following /t/. Hence, from the articulatory point of view, it is only apparent elision since the alveolar gesture was produced.

We finish this subsection with exploring $C_f\#C_i$ boundaries in which the consonants flanking the boundary are identical either originally or become such through assimilation. For example, consider phrases like ‘seek closure’, ‘wind tunnel’, or ‘I’ve seen many people’. Say these phrases and inspect your habits producing these $C_f\#C_i$ junctures before reading further.

Most likely you produced a single oral complete constriction: velar [k] in ‘seek closure’, alveolar [t] in ‘wind tunnel’ and bilabial [m] due to assimilation in ‘I’ve seen many people’. Some people tend to describe this as producing both C_f and C_i , only the closure of the first C_f is *unreleased* to smoothly maintain it for the production of the following homorganic C_i : ‘see[k^hk]losure’, ‘win[d^ht]unnel’, ‘I’ve see[m^hm]any people’; as seen in Chapter 7, IPA diacritic [̚] marks an unreleased closure. Other people might feel that C_f is completely elided and only the C_i is produced: ‘see[k]losure’, ‘win[t]unnel’, ‘I’ve see[m]any people’. Which description is closer to your own speaking habit?

With your command of Praat, and understanding of the phonetic patterns of connected speech, how would you design an experiment to test if people generally elide the C_f consonant or maintain it with an unreleased closure? Brainstorm a research plan for such a project before consulting Activity 10-7.

Activity 10-7: Elision or unreleasing in homorganic $C_f\#C_i$ junctures

To determine if people drop out or keep word-final /k/ or /m/ in ‘seek closure’ or ‘team mentality’, we need two basic components. First, we find minimal pairs for which the sole difference is the presence vs. absence of the target C_f consonant. Hence we find pairs like ‘seek closure’ vs. ‘see closure’ or ‘team mentality’ vs. ‘tea mentality’. Second, we determine the phonetic feature, measurable in Praat that can signal this difference. The most natural thing that comes to mind first is the duration of the closure. The unreleasing of the C_f would be supported if the closure is systematically longer in $VC_f\#C_i$ than in $V_f\#C_i$.

This is one potential road map for designing an experiment testing your own intuition about whether English speakers delete completely or just unrelease word-final oral and nasal stops when they are followed by a homorganic stop word-initially.

What other factors should you consider? Think of the frequency of words or the prosodic/syntactic structure of the phrases, spelling, the age of your subjects, etc. There are many potential factors. For a good experiment, you want to control, or consider, as many factors that could affect the closure duration measurement.

I encourage you, either individually, and even better in a group, to carry out a project like this based on your brainstorming and suggestions in Activity 10-7. At this point in the book you have the tools and skills needed for such a project and coming up with your own answers, or possibly more questions, through your own work with data, rather than reading about the answers in a book, is quite rewarding.

10.5 Linking

After discussing $C_f\#C_i$ junctures, we continue with considering how speakers deal with $V_f\#V_i$ boundaries. Adjacent vowels create a so called **vowel hiatus**, which, similar to multiple adjacent consonants, is also a dis-preferred situation that many languages resolve either by deleting one of the two vowels, or breaking the hiatus by inserting a consonant. In English, the second alternative is a productive strategy. Consider, for example, the variation in the indefinite article ‘a/an’. ‘An’ is used when the following word starts with a vowel. Importantly, we consider the presence of the vowel in speech rather than in text since we say ‘a university’ but ‘an umbrella’. If you compare the illicit phrases like *‘a apple’ or *‘a umbrella’ with their licit counterparts with ‘an’, you feel the ease of linking the two words facilitated by this linking ‘n’. This illustrates an active process in English to insert a consonant that would not be normally present in order to break vowel hiatus. The most common **linking consonants** in English are glides, rhotics and glottal stops [j, w, ɹ, ʔ].

Activity 10-8: Vowel hiatus options from recordings

Below I selected three phrases for the analysis from the recordings analysed in Chapter 14. Each has at least one case of vowel hiatus.

- a. ‘texture of the period’
- b. ‘to absolutely capture the moment’
- c. ‘something I was surprised to be asked to do and that’s uh ...’

Transcribe in broad transcription and identify the junctures with $V_f\#V_i$. Then, say these phrases yourself and observe how you link words in these target junctures. Possibly ask other native speakers to say these to you and compare with your own introspection. For each hiatus, discuss possible options.

Each phrase has an associated sound file and the TextGrid file in the book companion named ‘texture’, ‘moment’ and ‘surprised’, respectively. Open them in Praat, compare the realization of the speaker with your observations. Do not forget to listen to the word preceding and following the juncture separately. Then visually inspect the acoustic evidence for the perceived spoken behaviour.

When finished, compare your observations with the text below.

In phrase (a) of Activity 10-8 the target juncture was ‘texture of’. It represents the same environment as ‘corner of’ in Activity 10-1. In non-rhotic dialects, speakers do not normally produce coda ‘r’ sounds but in a word-final position, when present in the spelling, and another vowel follows across a weak boundary, this ‘r’ is used to link the two vowels and thus break the vowel hiatus. This is a very common case of **linking-r**, especially if the two adjacent vowels are non-high and many of you might have produced phrases like ‘corner of’ or ‘texture of’ in this way. However, although the speaker in Activity 10-1 clearly produced this linking-r, the speaker in Activity 10-8 did not. There is visible and perceivable glottalization in

the final syllable of ‘texture’ and a salient interval of non-modal aperiodic noise separating the vowels at the boundary. I would thus analyse this juncture as including an incomplete glottal stop as a strategy to break the vowel hiatus.

The *glottal stop* used as an onset of word-initial onsetless syllables is thus another common strategy, especially with words that are prosodically highlighted or following silent pauses. This is the strategy used by the speaker in the first juncture of the second phrase ‘to absolutely’. Here the glottal stop is very salient. Were you predicting this realization? If not, do you find this strategy natural in your own speech?

Finally, the third phrase contains two junctures relevant to our current discussion: ‘be asked’ and ‘do and’. What did you hear when you listened just to words ‘asked’ and ‘and’ separately? My narrow transcription would include the glides [j] and [w], respectively in these junctures: [bija:st] and [duwən]. Hence, the speaker produces *linking-j* and *linking-w* to break $V_f\#V_i$ hiatus primarily in those junctures where V_f are high vowels. There is also elision at the end of both second words.

As you have seen with ‘texture of’, and in the discussion of assimilations and elisions, these strategies for linking adjacent vowels across word junctures are also optional to a great extent and dependent on context. Another optional pattern of speakers of many English dialects is to link non-high adjacent vowels with an *intrusive-r*. Speakers tend to produce junctures like ‘idea of’ or ‘vodka and juice’ or ‘I saw a man’ as [aɪ'dɪəʊv], [vɒdkə.ɪəndʒu:s], and [aɪsə:ɪəmən], respectively, despite no motivation for a [ɹ] in spelling.

An attentive reader might be wondering that inserting an additional consonantal constriction is not consistent with the idea of simplification linked to smooth transitions discussed in Sect. 10.1. This apparent paradox, however, is seen differently when we consider the syllable structure. The most general and default, in phonological terms *unmarked*, syllables in speech are CV syllables, those having a single onset and a nucleus: CV.CV.CV. Syllables without onsets or with two (or more) consonants between vowels are generally more limited, hence *marked*, than the CV syllables. Some arguments for the unmarked nature of CV syllables have been presented when discussing onset maximization in Chapter 8 and include language development (babbling), speech production and processing and cross-linguistic generalizations regarding phonological processes and the distribution of syllable types. Both $V_f\#V_i$ and $C_f\#C_i$ junctures result in syllables that are more marked than the CV ones: either without an onset or with an coda. Seen from this angle, simplifications of oral consonantal constrictions in adjacent consonants, insertions of consonants to break vowel hiatus or re-syllabifications of coda consonants into the onset of the following syllables are all examples of this single underlying preference for CV syllables and come in play in the favourable contexts.¹

¹Sometimes, however, the linking consonants, especially linking-r might have phonetic characteristics of a coda rather than an onset consonant.

10.6 Processes Within Words

In fast fluent speech, the processes of assimilation, elision and linking can also be observed within words, particularly across various types of morpheme boundaries and commonly at boundaries between two words forming a compound. For example, compounds like ‘hotplate’ or ‘commonplace’ would normally be produced with regressive assimilations of place: [ˈhɒpˌpleɪt], [ˈkɒməˌpleɪs]. Also, elisions of consonants, particularly medial consonants surrounded by other consonants from both sides, are common at boundaries between two elements of a compound like ‘diskdrive’ produced [ˈdɪsdɪˌɑːrv] or ‘waistline’ produced [ˈweɪslɪn].

Elisions of vowels are common as well. For example, schwa vowels between consonants word-initially are frequently dropped in words like ‘support’, ‘particularly’ or ‘connect’. These elisions sometimes result in the loss of entire syllable since ‘support’ might sound like a single syllable word ‘sport’. In the final or word-internal syllables, schwas are sometimes elided, which results in syllabic consonants in words like ‘nation’, ‘national’ or one of possible weakening of ‘and’ to [ŋ] in phrases like ‘rock and roll’. In this case the syllable is not lost and words like ‘nation’ with schwa deletion [ˈneɪʃŋ] are clearly perceived as bisyllabic.

Linking-r in non-rhotic dialects is also observed in cases when vowel-initial suffixes are added to a stem-final /t/ like ‘care’ vs. ‘caring’ or ‘clear’ vs. ‘clearer’. Intrusive-r has been observed in words like ‘drawing’ produced as [drɔːrɪŋ].

In the final activity of this chapter you explore word-internal elisions.

Activity 10-9: Absolutely

In the second file of previous Activity 10-8 you have seen the word ‘absolutely’ and noticed that the speaker produced this word in a very reduced form. In this activity, first, extract just this word into a separate sound file. You can do this easily in Praat by opening the soundfile ‘moment.wav’ together with ‘moment.TextGrid’, selecting the interval labelled as ‘absolutely’, and then *File* → *Save selected sound as WAV file...*

Next, approach several (native) speakers of English and play the sound to them; preferably with headphones. They have two tasks:

1. Identify the word they hear
2. Identify the number of syllables.

You may decide how many times you play the file, it is best to be consistent for each subject. Note their answers. If there are subjects who did not recognize the word correctly, or responded ‘I don’t know’, you can then play the entire ‘moment.wav’ file.

Once done, try to rationalize the results. If there is a systematic pattern in the results, what phonetic features are responsible for the pattern?

I expect some people might not recognize the word out of the context. I myself hear a 3-syllable word [ˈæbsɪʔli]. It is very surprising since the canonical form of

this word [ˌæbsəˈluːtli] has the primary stress on the third syllable and in the realization of the speaker, the second and third syllables coalesce into an unstressed syllable. While a switch between the primary and secondary stresses is commonly observed in English for rhythmical reasons², an elision of a stressed syllable is quite rare. The fact that probably all speakers identify the word correctly in the context is evidence for great redundancy of phonetic features and strong expectations and predictions the context provides.

Exercises

- 10-1 Return to Activity 10-1 and describe all processes there following the approach and vocabulary used in the chapter.
- 10-2 Produce as narrow as possible IPA transcription of any phrases in connected speech. First the dictionary style broad transcriptions, and then the narrow transcription with all connected speech aspects discussed in the chapter represented with IPA. Discuss problematic issues. Remember, the goal is not to seek a fully ‘correct’ transcription but to represent the speaking habits linked to word boundaries in as much phonetic detail as possible.
- 10-3 Analyse and reinforce your awareness of connected speech aspects in the companion’s sound file ‘Three_kids.wav’ and the associated file ‘Three_kids.TextGrid’.

References

- Browman, Catherine, and Louis Goldstein. 1990. Tiers in articulatory phonology with some implications for casual speech. In *Papers in laboratory phonology I: Between the grammar and physics of speech*, ed. John Kingston and Mary E. Beckman, 341–376. Cambridge: Cambridge University Press.
- Guy, Gregory, and Sally Boyd. 1990. The development of a morphological class. *Language Variation and Change* 2: 1–18.
- Iverson, Gregory, and Kee-Ho Kim. 1987. Underspecification and hierarchical feature representation in Korean consonantal phonology. *Papers from the Annual Regional Meeting, Chicago Linguistic Society* 23 (2): 182–198.
- Lindblom, Bjorn. 1990. Explaining phonetic variation: A sketch of H&H theory. In *Speech production and speech modeling*, ed. William J. Hardcastle and Alain Marchal, 403–439. NATO Science Series D: Behavioural and Social Sciences. Dordrecht: Springer.

²This is known as stress clash and will be discussed in Sect. 11.4 of the next chapter.



In this chapter, we will

- engage with speech prosody through a basic investigation of prominences and boundaries
- discuss the realization of weak and strong forms in function words
- explore basic global trends within prosodic units

11.1 Introduction

Previous chapters have guided our exploration of the relationship between how we say something and what meaning/function it conveys in gradually expanding units of analysis. We started with single segments that we later extended to syllables, words, pairs of adjacent words, and for the rest of the book we will tackle this relationship observable in larger phrases. We have been also gradually extending skills for exploring speech in Praat through hands-on activities and facilitating thus deeper awareness of this relationship. At the same time, we have uncovered a number of potential communicative meanings and types of mental knowledge with which the speaking habits are associated.

Now, with expanding our investigations to larger phrases we are confronting ever increasing, and seemingly unlimited, variability in how a particular phrase might be said in various situational contexts and what it conveys regarding the speaker intentions and his/her states. The discussion in the remainder of the book

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_11) contains supplementary material, which is available to authorized users.

provides tools, activities for (self-)exploration, and guidance for navigating this variability in an approachable and systematic way.

In order to do this, we start by introducing prosody as a basic frame of reference for discussing speaking behaviour in larger phrases. Specifically, we will expand the notion of prominence introduced in Chapter 9, and add the notion of disjuncture. The awareness of the central role of prominence and disjuncture in producing larger chunks of speech then sets up our exploration of the variability in the realization of non-prominent function words and global prosodic patterns within a phrase in the remainder of this chapter.

11.2 Primer to Prosodic Analysis: Pitch Accents and Disjunctures

The notion of *prominence* is a crucial concept in the analysis of speech. We have introduced it first in connection to word stress when we compared how individual syllables in a word stand out in relation to other syllables of that word. Here we expand this approach to larger units: we examine how individual words in a phrase stand out in comparison to other words in that phrase and what communicative meanings these patterns of prosodic highlighting might serve.

Activity 11-1: Mind reading (I): I'm a big fan of the program

How would you say the utterance 'I'm a big fan of the program'? Say it out loud now and if possible, record into Praat before continuing to read further.

Which words have you highlighted? Note that I did not provide any context for this utterance, but I suspect many of you have imagined a situation or a context, or possibly your favourite program, before you said the phrase. Try to verbalize what that context/situation for you was when you were saying it the first time.

Now, listen to an actual realization of this utterance in 'BC_Grinch_fan_of_the_program.wav' located in this chapter of the companion. Describe informally the phonetic differences (i.e. in terms of loudness, duration, pitch or voice quality) between your first realization and this one.

What can you infer about the intention of the speaker in the excerpt? Particularly, what might be the context and why the word 'program' was likely more prominent in your rendition than it is in the excerpt? Brainstorm about this and note how you are trying to read the mind of the speaker just based on the prosodic features of his utterance.

When you are done, consider the transcript from an introduction to an interview below. In which sense is the realization of the word 'program' in B's utterance reasonable given this context?

- A: Welcome to the program.
 B: [Thank you for having me.] I'm a big fan of the program.
 A: and I'm a big fan of yours.

Finally, listen to this exchange in the actual opening to the interview in ‘BC_Grinch_opening.wav’ and the associated ‘BC_Grinch_opening.TextGrid’ from the book companion.

The distribution of prosodic prominence to individual words clearly signals different meanings and functions. With no context given, you most likely highlighted the word ‘program’, but it was much less prominent in the particular context shown in Activity 11-1. I will refer to this type of prosodic highlighting as *pitch accent*. As explored below, even though we call it pitch accent, the highlighting is done by other means besides pitch.

Activity 11-2: Phonetic features of pitch accent

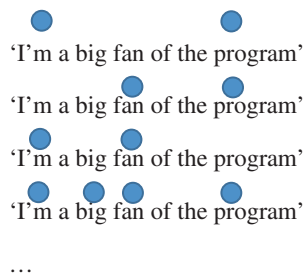
Compare the default realization of ‘I’m a BIG FAN of the PROgram’ with ‘I’m a big FAN of the program’ from Activity 11-1. How do you realize phonetically this intention to accent or de-accent the word ‘program’?

Phonetically, the same features participate in signalling pitch accents as those we reviewed for word stress. If a word is pitch accented, its stressed syllable is typically longer, louder and associated with greater pitch movement than the same syllable of a word that is not pitch accented. In addition, the pitch accented syllable would be produced with more extreme articulation observable through formant frequencies, energy in the burst of the stops or other features we have discussed.

Just like with word stress, where we identified three discrete levels, we will abstract away from this phonetic continuity by assuming a binary distribution of pitch accents: a word is either accented or not. Sometimes, this decision is relatively clear; for example with the first turn of the exchange in 11-1 ‘Welcome to the program’. Accenting both content words is the most natural way of producing this utterance (WELcome to the PROgram) and the female speaker produces it that way in ‘BC_Grinch_opening.wav’. Identifying these two words as accented is easy. However, more often than not, identifying pitch accented words is not a trivial task. Consider the words ‘big’ and ‘fan’ in the last two turns of the excerpt in Activity 11-1. In B’s utterance, ‘big’ tends to have a clearly lower prominence than ‘fan’ but in the subsequent utterance of A, I perceive ‘big’ as slightly more prominent than ‘fan’, and of course both less prominent than ‘yours’.

In these less clear cases, you might find that clapping hands while saying the utterance might be helpful. Frequently, the individual claps correspond to the pitch accents in the chunk. So listen to ‘I’m a big fan of the program’ several times and try imitating the speaker while clapping to the highlighted words. Additionally, clapping is also good for increasing your awareness of pitch accents when you compare alternative renditions of the same utterance with different distributions of the claps. This is very similar to ‘fofofo’ words or distorting the word stress of ‘pandemonium’ in the activities of Chapter 9. Here, try the various possibilities of accenting words and determine, which one is the closest to the way the speaker produced it in the audio. Figure 11.1 illustrates pitch accents with dots.

Fig. 11.1 A single sentence with varying placement of pitch accents marked a dots



Clapping hands helps us identify the metrical structure and melody of the utterances since these two are intricately linked in English.¹

Some descriptions of English prosody use the term *nuclear accent* or *sentence stress* for the last accent in a phrase. Although many people perceive and produce such accent as the most prominent in a phrase, especially in stylized forms of utterances, things are much more complicated in natural dialogues. I invite you to explore this aspect on your own in Chapter 14. At this stage it is sufficient to be aware, and able to identify, the difference between the presence and absence of a pitch accent on a word since this difference cues differences in speakers' intentions.

The second major aspect of prosody, so far not covered in this book, is the notion of *disjuncture* between two words. Activity 11-3 motivates the relevance of this notion.

Activity 11-3: Mind reading (II): Bill doesn't drink because he's unhappy

Record yourself saying to your friend: 'Bill doesn't drink because he's unhappy'. Do this before reading the rest of the text.

Once you recorded, please write down, before reading further, what you actually meant; what message you wanted to convey.

For some, probably a majority, of you, the intention was to say that now Bill does not drink (alcohol) and being unhappy is the reason why he doesn't drink. Under different circumstances he normally does drink. Possibly, he knows he has a tendency to get drunk when he drinks due to unhappiness and thus he drinks only if he's happy. You might be surprised to find out that for other people, and in possibly different contexts, the utterance might convey the fact that Bill now in fact drinks but the reason is NOT that he's unhappy. There's possibly a different reason (this ambiguity is discussed, e.g. in Hirschberg 2006).

Now, say the utterance yourself with these two intentions and compare. What is the major difference? How are the intentions you wanted to convey reflected in the prosody?

¹There are several methods to teaching English prosody that use similar approaches, for example C. Graham's jazz chants (2001).

The sentence ‘Bill doesn’t drink because he’s unhappy’ is an example of *structural ambiguity* when different linguistic structures correspond to different meanings. You can read more about ambiguities like these in Find Out More 11-1.

Find-Out-More 11-1: structural ambiguity

Sometimes ambiguities like ‘Bill doesn’t drink because he’s unhappy’ feel like the famous ambiguous images of a duck/rabbit or the Rubin vase that you can find on the internet, for example here:

- https://en.wikipedia.org/wiki/Ambiguous_image

Other examples from speech include words like ‘unlockable’, or [ˈæɾəm] in American English. ‘Unlockable’ might be referring to a door that you cannot lock, but also a door that you can unlock. The brackets show the grouping of the stem and morphemes in the two meanings that are part of the grammar: {un{lock-able}} vs. {{un-lock}-able}.

As we discussed in Chapter 7, [ˈæɾəm] in American English might refer to ‘atom’ or ‘Adam’ since they differ in the intended structure ([ˈætəm], [ˈædəm]) and both /t/ and /d/ undergo the allophonic process of flapping. Note, however, that this is different from most homophones like ‘son’ - ‘sun’ that reflect the relationship between the sound and the spelling, which is commonly not considered a part of the grammar since the intended structure for both was identical [sʌn].

Most native speakers produce the utterance from Activity 11-3 with the intention that Bill drinks for a reason different from his unhappiness as one unit, phrase or a chunk, whereas the intention that Bill does not drink, and the reason is his unhappiness, is produced with a clear division of this utterance into two smaller phrases/units/chunks: Bill doesn’t drink # because he’s unhappy. This division is typically referred to as a *prosodic break* or a *prosodic boundary*, and signals a strong *disjuncture* between adjacent words ‘drink’ and ‘because’ that is comparatively more salient than a disjuncture between any other pair of adjacent words in this utterance. Also, I will use unit, chunk and intonational phrase interchangeably to refer to the stretches of speech clearly separated by salient prosodic breaks. Furthermore, I will use the terms disjuncture, break or boundary as synonyms referring to this prosodic separation between adjacent words.

Phonetically, such a strong disjuncture is most robustly signalled with a silent pause separating the two adjacent words. Although important and salient, an interval of an actual silence is not a necessary requirement for the perception of a strong disjuncture. Consider Activity 11-4.

Activity 11-4: Realization of prosodic boundaries

Similarly to previous activities, first say this utterance on your own: ‘So tell me how did this come about how did this happen?’ Consider where you would

normally make breaks. Re-write the text and insert # where you would produce a strong prosodic boundary.

Now listen to recording ‘how_did_this_happen.wav’ in the companion. How did this speaker divide her utterance into prosodic units?

Once you identified the units auditorily, inspect the file in Praat both visually and auditorily. You can create a TextGrid annotation for easier navigation. For example, create two-interval tiers called phrases and words (with the sound highlighted *Annotate* → *To TextGrid* → *phrases words* in the first field and the second left empty). Identify the start and end of the prosodic chunks in the first tier and individual words in the second tier with the help of the spectrogram and zooming in and out. Try to match your boundaries as closely aligned with the audio as possible. Don’t forget to save your TextGrid file manually.

Here is how you can compare your alignment of the breaks with mine. In Praat, you have now a TextGrid object called ‘how_did_this_happen’. Select the textgrid file associated with this Activity in a companion website called ‘how_did_this_happen_SB.TextGrid’, save it to your computer and open it in Praat (*Read from file...*). Now highlight both your textgrid and this new one in Praat Objects (Ctrl-click on them) and click *Merge* from the menu on the right. Praat creates a new TextGrid object. Select this new object together with the Sound ‘how_did_this_happen’ and click *View & Edit*. You can now inspect the differences between your alignment and mine. Do not worry too much if the differences are many; manual text-to-speech alignment is a complex skill. It is important to understand why I labelled the way I did. With more practice you will continuously adjust and improve your skills.

Finally, we are ready for the main task of this Activity. Inspect visually and auditorily the phonetic characteristics of the breaks that are identified in the annotations. In addition to silent pauses, what might contribute to the perception of a boundary? Consider the prosodic features we have been discussing: pitch (adjust the pitch setting to the range 90–580 Hz), intensity, duration and voice quality.

In my analysis, the excerpt from Activity 11-4 can be chunked as follows: so tell me # how did this # come about # how did this happen. Two boundaries are produced with a silent pause: ‘this # come’ and ‘happen #’. A part of the silence following ‘this’ stems from the closure phase of the following [k] of ‘come’. But normally, these closures last around 50–80 ms and this silent interval is around 200 ms, which contributes to a clearly perceived break.

The disjuncture between the first and second chunk (‘me # how’) does not have a silence but it has a rather extreme *pitch excursion* with an extreme pitch and intensity fall on ‘me’ that ends with *glottalization*. This is the local creaky voice discussed in Chapter 3 due to the lowering of pitch preceding a boundary. You can visually see two vertical striations in the spectrogram and if you select just this interval, you can clearly hear the creak. Additionally, ‘me’ is also slightly lengthened. You can compare how you would say ‘me’ in this context with ‘me’ in a regular request like ‘Tell me your name’. This local lengthening that precedes

prosodic boundaries is called *pre-boundary lengthening*. All of these phonetic features (lengthening, pitch excursion, intensity lowering, glottalization) contribute to signalling a clear break after ‘me’.

Disjuncture between the second and the third chunk (this # come), on the other hand, does contain silence but does not feature any prominent pitch excursion. Nevertheless, pre-boundary lengthening can be observed by comparing this token of ‘this’ with the same token of the fourth chunk. You can easily see in Praat that ‘this’ preceding the boundary is around 230 ms long while without the following boundary it is only about 130 ms long.

The third major break (‘about # how’) does contain a brief interval of relative silence (about 50 ms) but this belongs to the closure phase of the voiceless /t/. Moreover, this /t/ is glottalized. We can again see the vertical striations preceding the closure. Hence, this boundary is realized in a similar way to the first boundary, only all the phonetic boundary markers (pitch excursion, glottalization, lengthening) are less salient than with the first break.

Finally, the last break is clear since it is followed by an extended silence, and also an exchange of the speakers.

To summarize this primer to prosody, both accents and boundaries signal intentions of the speaker and are very closely interconnected with situational context. Let’s consider ‘Bill drinks because he’s unhappy’ from Activity 11-3 once more. If you wanted to convey that Bill drinks for a reason different from unhappiness, note that you and your hearer(s) already know that Bill is drinking. And you signalled this ‘givenness’ by not accenting ‘drink’ strongly. We will discuss how prosody helps in cuing these meanings and intentions, commonly described as *pragmatic meanings* (e.g. Hirschberg 2006), in the following chapter. This example nicely demonstrates how both accenting and phrasing are harnessed for cuing the pragmatic meanings.

With the basic understanding of the prosody of intonational phrases we now consider how the distribution of pitch accents and boundaries within a phrase affects the realization of English function words like pronouns or prepositions. We then close the chapter by briefly discussing some global trends functioning in this unit of speech.

11.3 Weak Forms

Activity 11-5: for and four

Here we are interested in potential variability in saying the words ‘for’ and ‘four’. If you say them in a dictionary form without context, they would be considered homophones since their pronunciation would be the same: [fɔ:(ɹ)].

Let’s vary now the intentions and produce them with and without prosodic highlighting. In Activity 11-1 we heard ‘Thank you for having me’. Record yourself saying this sentence and a potential retort from your interlocutor when s/he really wants to find out the reason for gratitude:

- (1) A: Thank you for having me!
 B: What are you thanking me for? I should thank you!

Hence, record both A's and B's part. Now employ a similar approach with 'four' and imagine a situation in which we vary the degree of prosodic prominence associated with this word. For example, imagine this dialogue with a friend:

- (2) A: I've got four siblings.
 B: Wow, four sisters, that's crazy!
 A: No, not four sisters, four siblings.

If you have a friend, record together, if not, record just A's part imagining the situation. Now inspect the realizations of 'for' and 'four' in Praat. Identify the tokens (you might annotate words in a single-tier TextGrid file) and compare their phonetic features. What are the differences between the first, non-prominent, and the second, prominent, tokens of 'for' and 'four'? Are the two words different in how the presence or absence of prosodic highlighting affects them? If you do not have access to recording people, use 'thank_you.wav' and 'four_sisters.wav' in the book companion.

With the basic understanding of pitch accents from the previous section, it is clear that the first 'for' in (1) above is not accented. How did you produce the question in (1B)? Was the most prominent word 'thanking', 'me' or 'for'? All might be possible in this context to signal different intentions of the speaker; try to say all three of them. Note how you produced your sentence when I asked you in Activity 11-5 without consciously weighing these three options and deciding for one. Your realization automatically followed your sub-conscious speaking habit and the cognitive selection of your intention. In any case, the first 'for' is definitely shorter, without a major pitch movement, and most likely also less loud than the second one. Importantly, for all native speakers, there is a clear difference in the quality of the vowel: the first token is produced with [ə] whereas the second with [ɔ:].

Now compare these observations with the realizations of 'four' in (2) of Activity 11-5. It is the words 'sisters' and 'siblings' that are the most prominent since they present the new, or additionally, contrastive, information in A's utterances. As a result, these two words are much more prominent than the two 'four' tokens in the second utterance of A. Crucially, the contrast between these two 'four' tokens on the one hand and 'four' in the first A's line on the other hand is not so stark as it was with 'for'. Particularly, the quality of the vowel did not change and all tokens should have the vowel [ɔ:]. For illustration, try to produce 'No, not four sisters, four siblings' with [fə:(ɪ)] and note that this realization is unacceptable for the intention 'four'.

To sum up we saw that some words can be produced in two forms with a change in the quality of segments. In case of 'for' they are [fə:(ɪ)] or [fɔ:(ɪ)]; I put [ɪ] in brackets to include the option for rhotic dialects, and also for the linking-r from previous chapter depending on the following context. In addition to 'for', another word 'you' in (1) of 11-5 presents a similar situation. The first two mentions have a short centralized [jə] or [jʊ] whereas the last one is most likely produced with a long [ju:]. For many speakers without prior conscious exploration of

Table 11.1 Most common weak forms in English

Category	Spelling	Weak form	Strong form
Determiners	the, a, some	ðə(ɪ), ə(n), səm	ði:, eɪ, æn, səm
Prepositions	at, for, to, from, of	ət, fə(ɪ), tə(ʊ), frəm, əv	æt, fɔ:(ɪ), tu, frɒm, ɒv
Conjunctions	and, but, that	ən(d), bət, ðət	ænd, bʌt, ðæt
Pronouns	you, she, he, we your, me, him, her, them, us	jə(ʊ), ʃɪ, (h)ɪ, wɪ jə(ɪ), mi, (h)ɪm, (h)ə(ɪ), ðəm, əs	ju:, ʃi:, hi:, wi: jɔ:(ɪ), mi:, hɪm, hɜ:(ɪ), ðem, əs
Aux. verbs	am, are, is was, were has, have does, do can, could, should would must	əm, ə(ɪ), ɪz wəz, wə(ɪ) (h)əz, (h)əv dəz, də kən, kəd, ʃəd, wəd, məs(t)	əm, ɑ:(ɪ), ɪz wɒz, wɜ:(ɪ) hæz, hæv dʌz, du: kən, kɒd, ʃɒd, wɒd, məs(t)

speech, native and non-native alike, this might come as a novel observation. Other words, like ‘four’ in exchanges like (2), do not allow this variability and the quality of the segments is typically preserved although prosodic characteristics like f_0 , duration, intensity, might be sometimes drastically different.

Words like ‘for’ or ‘you’ can thus be produced in both *strong form* with a full vowel and much more commonly as *weak form* with a vowel reduced to schwa or the centralized lax vowels [ɪ] and [ʊ]. On the other hand, words like ‘four’ only have the strong form. As you might have suspected, the cognitive system of native English speakers provides guidance as to which words might have weak forms and in which contexts they should be used, or rather, they should not be used, since the weak form is the default production of these words. The basic aspects of this cognitive system guiding the automatic sub-conscious realization of strong/weak forms is explored in the rest of this section.

First, consider which words allow weak forms. How is ‘for’ or ‘you’ different from ‘four’ or ‘siblings’? The relevant difference for us is in the part of speech classification. Note that ‘for’ and ‘you’ are the preposition and the pronoun, respectively and both belong among the *function words* of English. These words do not carry much of content or information but rather signal the grammatical relationships among other words in utterances. In that, function words are often (partially) predictable based on the context and thus do not need to be produced with large effort. Nouns, adjectives, numerals or adverbs, on the other hand, belong among *content*, or *lexical, words* and they have high informational content. The content and function words are also sometimes described as *open vs. close set*, respectively. New nouns, verbs or adjectives are added to a language frequently with the changing reality that languages reflect. The set of these words is thus open to new members. However, new prepositions or pronouns are hardly ever created and comprise thus the close set.

Table 11.1 lists the most common function words with both weak and strong forms. Note that verbs ‘to be’ and ‘to have’ might be produced in strong forms

if they represent the main lexical verb meaning ‘to exist’ or ‘to possess’, and typically have weak forms when serving as auxiliaries signalling the grammatical form of the main verb in a verb phrase (e.g. ‘I *was* writing’, ‘she *has* done it’). Additionally, ‘that’ functioning as a demonstrative pronoun (e.g. ‘*that* car is black’) is typically realized in the strong form while the weak forms are possible for conjunctions and relative pronouns (e.g. ‘the car *that* is black’). Finally, note also that the weakening from the vowels in the strong forms to a centralized schwa or [ɪ]/[ʊ] is the most prevalent characteristic. However, the initial consonant of ‘him’, ‘her’ or ‘has’ can be dropped as well in the weak forms.

Now that we identified which words might be realized as weak forms, let’s turn to the basic guidelines for their distribution in speech. Weak forms are a default realization in most cases when these words are **not** carrying a pitch accent or are **not** followed by a salient prosodic boundary. Therefore, the basic generalization from typical realizations is that the strong form of a function word is used if it is accented or followed by a salient prosodic boundary, and the weak form is used elsewhere.

Activity 11-6: Realization of weak forms

Before you continue reading, go back to Activities 11-1 and 11-4 and explore the sound files produced by you and the recordings from the actual interviews checking the application of the generalization above. Describe verbally and/or in writing the phonetic realization and the prosodic context of all the function words.

In 11-1, ‘Welcome to the program’, both ‘to’ and ‘the’ are unaccented and produced in weak forms. In ‘Thank you for having me’, both ‘you’ and ‘for’ are weak. The pronoun ‘me’ is followed in the transcript by a full stop and thus we expect a major prosodic boundary, possibly a pause, and the strong realization [mi:]. Interestingly, however, the actual disjuncture ‘me # I’m’ produced by the speaker is not so strong, although a clearly perceptible break is produced. The vowel in ‘me’ is thus quite short, suggesting a weak form. In ‘I’m a big fan of the program’, listen to the realization of ‘of the’, both in the phrase ‘fan of the program’, and in separate words annotated in the TextGrid. You notice that these two syllables are so contracted that the impression is that only one syllable remains. ‘And’ in ‘And I’m a big fan of yours’ is another interesting case with a potentially present pitch accent and consequently rather full realization of the vowel combined with the elision of the final ‘d’.

In 11-4, the first phrase ‘so tell me’ is followed by a clear prosodic break and a very short silence. Hence, the realization of ‘me’ preceding this break should be in the strong form with the tense quality and the length of the vowel [i:]. This is clearly the case despite the fact that ‘me’ is not pitch accented. Hence, the following major prosodic break is a sufficient environment for the strong form realization. This ‘me’ is in a nice contrast with ‘me’ of 11-1 discussed in the preceding paragraph. One should bear in mind, though, that the pronouns are frequently appearing in the weak form realizations, even if a following prosodic break is present (e.g. tell him! [telɪm]). Note also the two realizations of ‘did’ in the rest of

the utterance. Although we did not list this auxiliary verb in Table 11.1, you notice extremely shortened and centralized vowel realizations that could be transcribed as schwa in the first case, and even as vowel elision in the second case.

Let's close this section with another activity strengthening your awareness of the importance of the weak forms in English below.

Activity 11-7: Perception of strong forms

Have a pen and paper ready. Open file 'Sally_strong.wav' from the companion into Praat *Objects* and click *Play just once*. Write down what was said. Note down your impressions from both the sound and the task of transcribing what was said.

Then, open the file 'Johny_weak.wav' into Praat *Objects* and as before click *Play just once* and write down what was said. Then note down your impressions from both the sound and the task of transcribing what was said.

How do your notes compare? These were comparable utterances in content and length but sufficiently different so that the transcribing of the first file did not facilitate much the same activity in the second file. You can use a different order with your friends and include also Sally_weak and Johny_strong files in the companion.

Compare your observations with the text below.

Most of the students doing Activity 11-7 report that the realizations of the utterances with strong forms only was more difficult to understand than the regular prosody and also sounded very awkward. This is a good lesson for non-native speakers of English, for whom the realizations with weak forms cause problems both in perceiving what was said as well as in their own speaking. Making the relevant words prominent and the function words into weak forms goes a long way towards sounding more like native speakers.

11.4 Global Trends Within Intonational Phrases

In our discussion of prosody in this chapter so far, words that carry important information are pitch accented and thus hyper-articulated while those with little information such as function words are typically hypo-articulated in weak forms. Within pitch accented words it is the stressed syllables that carry out most of this hyper-articulation. Hence, speaking is essentially a constant intention-driven local stretching and lengthening on the one hand, and squeezing and shortening on the other. Moreover, this stretching and squeezing is hierarchical because it can be observed within syllables, words, phrases and even larger units. The characteristics of these local changes at any level are influenced by the intentions at other levels.

Activity 11-8: Phone numbers

In this activity, record yourself, or even better, record a friendly companion when they are asked to provide a phone number including the area code and possibly also the country code. For non-native speakers, it is not important in

this activity to speak English, and phone numbers might be in any native language of the speakers around you.

Your task is to inspect visibly in Praat the prosodic chunking of the phone number provided, and determine which numbers are relatively more prominent than others. What phonetic means did the speaker use to chunk the utterance into smaller units? Additionally, select in Praat the analyses of pitch and intensity over the entire utterance. Do you observe any global patterns?

You can compare your recording with a rendition of a phone number in the book companion: ‘phone_number.wav’ and the associated ‘phone_number.TextGrid’.

By now you are aware of the influence of the prosodic structure (prominences and boundaries) at each unit considered so far (syllables, words, phrases) and also familiar with how the marking at higher levels trickles down to lower levels. For example, pre-boundary lengthening might lengthen unstressed syllables of a word, or segments of a word more prominent at the phrase level are hyper-articulated compared to the same word that is not prominent. Were you lucky to have the same number appearing more than once and in different prosodic contexts? If not, inspect the numbers one, two and four, which all appear multiple times in ‘phone_number.wav’.

In addition to these hierarchical effects on the local changes, global trends also affect the prosodic realization of utterances. The most salient global trend is *declination*. At the physiological level, declination is linked to the breathing cycle and describes the gradual lowering of the sub-glottal pressure as we gradually expel the air with the recoil of our respiratory muscles. This lowering of pressure causes the natural decreasing of f_0 and intensity over the course of an utterance. Recall the tight relationship between the sub-glottal pressure on the one hand and intensity and pitch on the other, which we saw in Activity 9-6: the increase in the pressure by an external push on the chest results in the corresponding increase of pitch and intensity. You probably observed some declinations in your recordings of phone numbers as the pitch and intensity around the start of the utterance was greater than towards the end.

Declination, however, is also under the control of the speaker since individual intonational phrases spoken within a single breath are typically produced with local declination within the phrase and a *reset* at the boundary. Reset is clearly perceivable in the ‘phone_number.wav’ file. To show that it does not take place only in the stylized utterances like phone number listing, consider Fig. 11.2. I took one utterance from the Bake_off interview linked in the companion as ‘Honour.wav’, cut the pauses, extracted pitch, smoothed and interpolated it (all easily done in Praat; more on this in following chapters). The entire utterance consists of five chunks. The grey lines characterize the approximate linear slope of each of the five units as well as of the entire utterance. We can observe the declinations and pitch reset for each unit. Pitch reset is thus another phonetic marker indicating the presence of intonational boundaries.

Inspect the pitch and intensity curves overall and separately in the chunks recorded in phone numbers in Activity 11-8. Examine the declination and resets produced by the speaker and observe how these global trends again trickle down to the realizations at the smaller units of phrases, words and syllables.

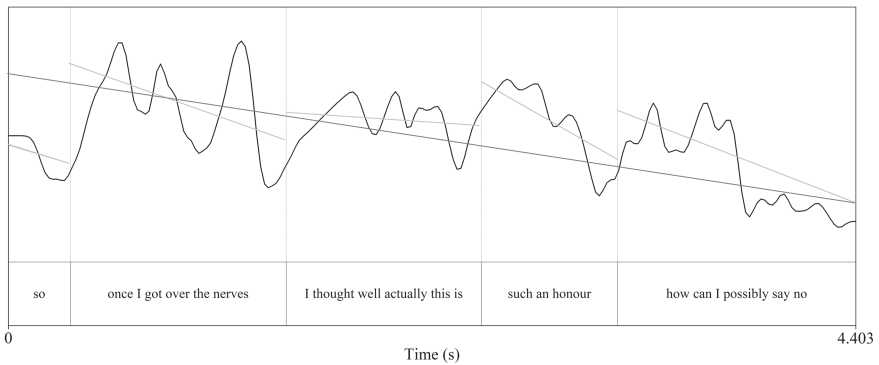


Fig. 11.2 Smoothed and interpolated pitch of an utterance consisting of five intonational units shown with a black curve. Pitch declinations over the units separately as well as over the entire utterance are shown with grey thin lines

Another global trend involves the distribution of prominent syllables within phrases.

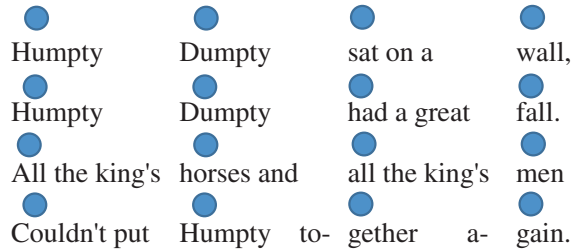
Activity 11-9: Prominence distribution

Consider two target words ‘Japanese’ and ‘overnight’ for this activity. First, think of utterances in which they finish an intonational phrase (e.g. ‘Are you Japanese?’, ‘I plan to stay overnight’). Record them yourself or ask another native speaker of English. Now take the same words but in slightly different utterances: ‘Are you a Japanese student?’ and ‘I’ll take my overnight bag’. Describe informally if you observe any differences in the production of the two target words. Open the recordings in Praat and compare the two realizations in both target words considering all segmental and suprasegmental phonetic features with which you are familiar by now. Do you observe any systematic differences? If so, could you speculate regarding the motivation behind this speaking habit?

Most native English speakers would show a tendency for making systematic differences in the production of the target words in these contexts. One way of transcribing these differences is with word stress. In the pre-final position, the primary stress tends to be in the final syllable and the secondary in the initial one: [ˌdʒæpəˈniːz], [ˌəʊvəˈnaɪt] in the British variety. However, in the other context, the primary stress tends to shift to the syllable with the original secondary stress: [ˈdʒæpəniːz ˈstjuːdənt], [ˈəʊvənaɪt ˈbæg]. This shift might have multiple phonetic correlates in the duration of the syllables, their pitch and intensity, and segmental realizations.

The motivation for this *stress shift* or also termed *rhythm reversal*, in the level of prominence of individual syllables is the tendency to avoid so-called *stress clash*. In speech, similarly to music, the preferred strategy is to alternate more prominent with less prominent units and avoid long stretches of either prominent

Fig. 11.3 Stylized production of a nursery rhyme with dots marking strong stressed syllables



or non-prominent units. The shift observed in Activity 11-9 is an example of this general pattern in English. Keeping the original stresses in ‘Japanese student’ and ‘overnight bag’ would result in two adjacent prominent syllables. Hence, if another ‘stressable’ syllable is available, the word stress, and potentially the pitch accent, can shift to that syllable producing speech that better corresponds to the preferred rhythmical strategy. We saw this avoidance of adjacent prominences also within words when discussing ‘Japan’ vs. ‘Japanese’ in Chapter 9.²

It is important to caution that, as virtually any speaking habit we study, also stress shift is more complex than a simple avoidance of adjacent stressed syllables. Consider phrases like ‘academic tutor’ or ‘old-fashioned music’, in which many native speakers in fluent speech would still reverse the stresses ([ˈækədəmɪk ˈtju:tə], [ˈəʊldfæʃənd ˈmju:zɪk]) despite the fact that a weak unstressed syllables [mɪk] and [ʃənd] separate the original primary stressed syllables ([.ækəˈdemɪk ˈtju:tə], [.əʊldˈfæʃənd ˈmju:zɪk]). Examples like these only scratch the surface of the phonological complexities under the umbrella of English *rhythm*, which could be seen as the stretching and squeezing of speech at the phrasal level resulting in the apparent tendency for semi-regular distribution of prominent syllables. Similarly to other phonological concepts, we will not provide a phonological account of English rhythm here; for that we would need to build a theoretical machinery of feet, syntactic phrase structure in terms of heads and modifiers, and many processes linked to them.

Rather, I invite you to use your skills with Praat for investigating if and where you perceive rhythmic regularity and what phonetic processes might underlie the global perception of rhythm. For example, this regularity may be observed in the production of stylized poetry or nursery rhymes. If you take Humpty Dumpty, it can be nicely produced as if on regular beats of a metronome, as illustrated in Fig. 11.3.

The distribution of stressed and unstressed syllables in the first two lines is /-.-.-./ and /-.-.-.-./ in the third and fourth lines. Hence, the stressed syllables occur regularly (on a beat) despite the fact that the number of unstressed syllables

²Recall also the discussion of ‘pandemonium’ tokens in isolation from Chapter 9. We said that [ˌpændəˈmɔːniəm] is preferred while [ˌpændəˈmɔːniəm] dispreferred. But in phrases like ‘pandemonium Fortress’, the latter might be more preferred rhythmically than the former.

between the beats varies (one or two). The unstressed syllables are thus squeezed and the stressed ones stretched both in time and in the extent of other prominence lending characteristics. This characteristic is traditionally invoked when describing English, together with Russian or Arabic and other languages, as a *stress-timed language* that is different from *syllable-timed languages* like many Romance languages in which the duration of syllables is more stable and the stretching and squeezing is less salient. However, I have to caution again since this seeming binary categorization of languages is in fact a continuum and it is greatly affected by multiple differences both across but also within languages.

In other forms of speaking, fully spontaneous or less stylized, this type of rhythmical regularity is usually not very obvious. Yet, several patterns of spontaneous English speaking that we have already mentioned, such as shortening and centralization of unstressed syllables, reductions in weak forms, stress shift or clapping to facilitate the identification of pitch accents, suggest that the global rhythmic patterns influence the realization of speech. The next two chapters will take us further on this path. I encourage you to use Praat and the extended material in the interviews linked in Chapter 14, or other natural speech of your choosing, to test your intuitions regarding the visible and measurable aspects of rhythm in speech.

Exercises

- 11-1 The companion includes the file ‘Her.wav’ and the associated ‘Her.TextGrid’ file extracted from the Bake off interview. The excerpt has four tokens of the pronoun ‘her’. Employ the concepts from the weak and strong forms discussion and describe the phonetic characteristics of these tokens. Discuss if all tendencies mentioned in the text are respected in this file and if not, how they can be amended.
- 11-2 Consider again the file ‘Three_kids.wav’ and the associated file Three_kids.TextGrid. from the exercise in the previous chapter. Analyse now weak forms and basic prosody in terms of prominences and boundaries, and explore how these prosodic aspects are interconnected with the connected speech aspects happening at the word boundaries.

References

- Graham, Carolyn, and Marilyn Rosenthal. 2001. *Jazz chants old and new*. New York: Oxford University Press.
- Hirschberg, Julia. 2006. Pragmatics and intonation. In *The handbook of pragmatics*, ed. Laurence R. Horn and Gregory Ward, 515–527. Malden, MA: Blackwell.



In this chapter, we will

- discuss ways of describing intonational contours expanding on the characteristics of boundaries and accents introduced in Chapter 11
- examine the difference between fundamental frequency and pitch
- become more proficient in reading and manipulating Praat's pitch tracking
- explore major factors affecting the relationship between intonation and meaning in speaking habits

12.1 Introduction

The previous chapter introduced the notion of basic prosodic analysis by examining the phonetic characteristics and pragmatic meanings of prosodic boundaries and pitch accents. In this chapter, we will expand on that by considering not only the presence vs. absence of boundaries and pitch accents but also the intonational characteristics associated with both boundaries and pitch accents. For that purpose, we will present a discrete system for describing the naturally continuous intonational characteristics, following the similar approach of using IPA for transcribing continuous segmental characteristics of sounds. And similarly to the approach in other chapters, we will investigate how the changes in the intonational characteristics affect the communicative and pragmatic meanings of the resulting utterances.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_12) contains supplementary material, which is available to authorized users.

12.2 Fundamental Frequency and Pitch in Praat

The focus of the entire chapter is on intonation, or the melody, of spoken utterances. The physical correlate of intonation/melody in speech production is the *fundamental frequency* typically marked as f_0 and introduced in Chapter 4. It corresponds to the rate of vibration of the vocal cords and that in turn is controlled voluntarily by varying the tension of the vocal cords by the muscles attached to the cartilages in the larynx (Chapter 3). Activity 12-1 reinforces the awareness of this relationship.

Activity 12-1: Siren

Imagine a warning siren. Pick any vowel, take a deep breath, and imitate two cycles of this siren. Utilize the limits of your voice: start as low as possible, go as high as possible and finish as low as possible. Introspect the feelings in your larynx, particularly tension. The sensations in your larynx as well as auditory information through your hearing are two types of sensory information linked to the fundamental frequency.

Now do the same and record yourself in Praat. Save the sound and click *View & Edit*. Make Praat show you f_0 by clicking *Pitch* → *Show pitch*, and adjust the pitch range to see a smooth curve. If you see any sharp jumps, you know the pitch tracker in Praat cannot be trusted since your siren was smooth and not jerky. These jumps result from so-called pitch halving or pitch doubling and typically can be reasonably well resolved with adjustments to the pitch range (*Pitch* → *Pitch settings*). For example, Fig. 12.1 shows the Praat pitch track of my recording of a siren, in the companion as ‘warning_siren_a.wav’, with pitch range 90–250 Hz and then the adjusted 90–300 Hz. The curve (blue in Praat interface and black in Fig. 12.1) is another representation, this time visual, of f_0 produced by your vocal cords.

Finally, if you have a regular air-balloon available, try to reproduce your siren with it. Blow a sufficient amount of air into it. Then pinch the sides of the balloon’s ‘neck’ in between the thumb and the index finger of both hands, and adjust the tension by stretching the sides moving the hands away from each other and towards each other until you master producing a sound siren. This is yet another way of becoming aware of the process changing f_0 in your larynx where the balloon’s neck represents your vocal cords and the force of your hands stretching and relaxing the tension represents the activity of the muscles attached to the cartilages.

The correlate of intonation/melody in speech perception is *pitch*. An attentive reader notices that terminology used in Praat is slightly confusing and refers to pitch and not fundamental frequency. We have seen the blue line corresponding to f_0 in several figures and Praat activities in previous chapters but now is the time to look at it more closely. There are several aspects of discrepancy between what Praat shows in the blue line and what our auditory mechanism perceives as pitch. It is important to be aware of these in order for Praat to be our tool and helper and not a master enslaving our perception.

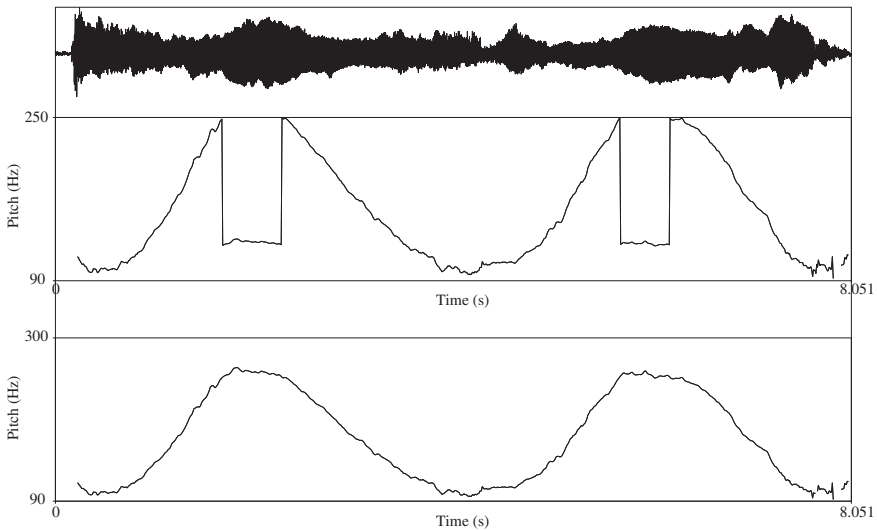


Fig. 12.1 Praat's pitch track of 'warning_siren.wav' showing pitch halving with 90–250 Hz range and the adjusted 90–300 Hz pitch range

First, there are things that we hear as pitch but they do not show as the blue line. Activity 12-2 describes the most relevant cases.

Activity 12-2: pitch perceived with no f_0 contour (I)

Open again the file DSE.wav that was discussed also in Chapter 4 and is located in that chapter of the companion. Listen to the sound with the *Play* button without inspecting it visually. Try to imitate the melody of this utterance by humming it.

Now visualize the sound in Praat and make only the pitch visible (in *View* → *Show analyses...* only pitch is ticked) and adjust the pitch range to 60–200 Hz. You see that the blue line is not continuous but disjointed. Why are there gaps in the blue line? Do you perceive the pitch in this utterance as disjoint corresponding to the blue line? Did you hum the melody in this disjointed way?

To appreciate Praat's ability to fool you, adjust now pitch ceiling to 100 Hz instead of 200 Hz (*Pitch* → *Pitch settings*). Describe how the f_0 curve has changed and why it should not be trusted.

Now, in the Praat Objects window, highlight the DSE sound again, and then extract just the fundamental frequency (pitch) by *Analyse periodicity* → *To pitch...* and set the floor and the ceiling values to 60 and 200 Hz, respectively (since we verified this as my real pitch range) and click *OK*. Praat creates a new Pitch object called DSE; click on that and click *View & Edit*. You see a curve made of pink dots corresponding to the blue curve before. Listen to the file by clicking the bar Visible window or Total duration. Strange, isn't it! Praat imitates just the disjointed source vocal fold vibration without any filtering by

articulators and cavities above the larynx explored in Chapter 4. How is your perception of the disjoint f_0 curve now?

Next, we can use Praat functionality to fill the empty intervals of the originally disjoint f_0 curve by interpolating: with the pitch object selected, click *Convert* → *Interpolate*. Praat made another Pitch object that you can rename by clicking the *Rename...* button in the bottom of the Objects window and name it ‘DSE_Interpolated’. Click *View & Edit*, and listen to it, compare this continuous version with the un-interpolated original one.

Finally, alternate listening to the three objects in your Objects window: Sound DSE, Pitch DSE and Pitch DSE_Interpolated; for the Pitch objects click *Sound* and then select either *Hum*, which is the same as what you heard before, or *Play pulses*, which gives a slightly different perception. Which of the two Pitch objects sounds more similar to your perception of melody/intonation in the DSE file, and probably closer to your humming at the start of the activity?

Activity 12-2 above explored the visual gaps in f_0 tracking provided by Praat as the pitch contour. In terms of production, they represent the intervals in which the vocal cords were not vibrating in a modal voice and these intervals primarily correspond to voiceless sounds like [s], [k], [p] or [ʃ], or partially devoiced sounds like the initial [d̥] in my realization of ‘discovering spoken English’.

As always, Praat can only show what the acoustic signal produced by our mouths contains and not what we hear with our ears. Hence, in the rest of the activity’s task you probably found out that the disjoint f_0 curve that you see does not correspond to a more continuous perception of pitch and melody that you

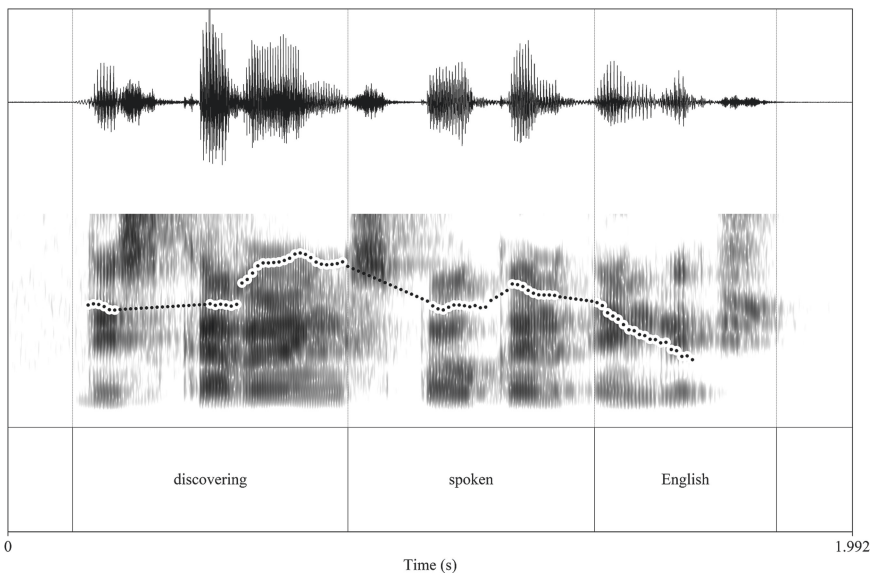


Fig. 12.2 File ‘DSE.wav’ with the original f_0 (white larger dots) and the interpolated one (black smaller dots)

experience when listening to the sentence. Figure 12.2 illustrates both the original pitch contour in Praat as well as the interpolated one.

Activity 12-3: pitch perceived with no f_0 contour (II)

For this activity, record any short phrase in loud whispered speech highlighting a single word. For example, whisper ‘Discovering SPOken English’ with the intention to contrast ‘spoken’ with ‘written’. Open the file in Praat and listen to it. Do you perceive the melody of this utterance? Inspect it now visually and note the absence of f_0 since the vocal cords are not vibrating in whispering. Do you see any visual evidence in Praat that corresponds to your perception of melody? If you cannot record your own file, you may use ‘DSE_whispered.wav’ in the book companion.

As you might have experienced yourself, speakers commonly compensate for the absence of f_0 , particularly when they intend to signal a specific intention; in our case the contrast between ‘spoken’ and ‘written’ English. Most likely, information present in the intensity contour (recall the relationship between f_0 and intensity we explored with word stress in Activity 9-6 of Chapter 9), and the formant frequency shifts can be decoded by our ears as pitch even in the total absence of f_0 in the acoustic signal.

Hence, Activities 12-2 and 12-3 show that pitch might be perceived even in the absence of f_0 contour in Praat. The next activity looks at the other side of the discrepancy between the visual and auditory senses when changes in f_0 shown with Praat are not perceived as pitch changes.

Activity 12-4: changes in f_0 contour seen but not perceived

In this activity we return to a file suggested for exercises in Chapters 10 and 11: ‘Three_kids.wav’. Open the sound file together with the associated TextGrid in Praat, maximize the window, select the spectrogram and pitch analyses and set the pitch range to 80–400 Hz.

Listen to the file, paying special attention to its intonation. Again, try to hum it on your own. Now, try humming slowly but tracing the blue f_0 contour as closely as possible expressing in your pitch every upward or downward motion indicated by the curve.

You notice that this ‘exact’ tracing is also not conforming to your perception of pitch and melody in this utterance. Try to identify which peaks and valleys represent really produced peaks/valleys and which do not. Select various intervals, listen to pitch, compare with what you see. Can you generalize the basic approach to which f_0 tracings are ‘real’ and which should be disregarded?

There are multiple ways, in addition to pitch halving and doubling and very common creakiness of the voice in which Praat and other automatic software might

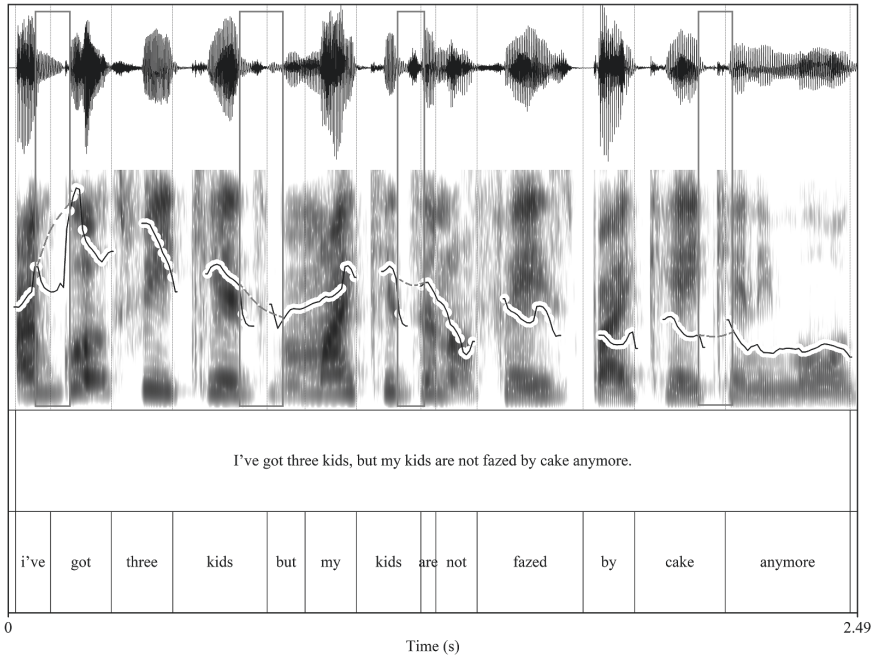


Fig. 12.3 Examples of unreliable f_0 contours in the vicinity of consonants. The grey box marks the interval and the dashed line through this box traces the idealized contour

show unreliable f_0 . Activity 12-4 illustrated the unreliability stemming from the effect of consonants on f_0 . Figure 12.3 illustrates four such cases in this particular utterance. A useful general approach is to be very cautious, and rather disregard, f_0 in the vicinity of most obstruents (stops and fricatives). The voiced (lenis) ones, especially labials, commonly lower f_0 during the closure while the voiceless (fortis) consonants have a tendency to increase f_0 or produce artefacts at their releases. Also, f_0 during obstruents is much less salient than f_0 during vowels or sonorant consonants. Hence, these very local variations in f_0 tend not to be perceived in the overall pitch contour of an utterance and thus should be disregarded in most cases. Our perception tends to reflect the dashed lines in Fig. 12.3 interpolating through these unreliable regions.

Intonation is traditionally described as an *autosegmental* or *suprasegmental* feature of speech meaning that its primary characteristics can be overlaid on any segmental material and thus the effects of segments on pitch perception can be disregarded for pragmatic meanings. A useful approach to studying overall f_0 contours and their relationship to communicative functions is to take a step back, as you commonly do to appreciate a picture on the wall, and inspect the overall contours and intentions rather than focus on every single turn of the visible pitch contour provided by Praat.

Finally, I have to briefly mention the important issue of the *non-linear relationship* between produced f_0 and perceived pitch. Informally, this means that a certain change in f_0 measured in the acoustic signal in terms of Hz does not

correspond to precisely the same change in pitch perceived by the listeners. Therefore, simple differences in Hz cannot in fact be interpreted, without a reference value. Various scales derived from frequencies measurable in the signal were proposed that reflect human perception of pitch better than Hz. The most common are *Mel* or *semitone* logarithmic scales. Hence, in investigations or project targeting the perception of pitch, it is advisable to convert the measured Hz values to Mel or semitone values using conversion formulas readily available either directly in Praat or by searching the internet.

12.3 Approaches to Marking Intonation

Up to now IPA was sufficient for putting on the paper everything crucial related to speaking habits. These transcriptions accompanied the exploration of these characteristics in the speech signal with Praat and the introspection of your own speaking. In addition to IPA symbols for segments we were gradually adding diacritics, syllable boundaries or word stress.

For intonation we also need a system for transcribing the crucial and most salient aspects that are important in communicating various intentions. We need such a system for the same reasons, outlined in Sect. 2.4 of the book, for which we introduced IPA. Reading and writing in IPA facilitates the awareness of the unconscious aspects of speaking. A system of discrete labels transcribing salient aspects of intonation serves the same purpose. Just like IPA, both producing the transcription based on speech signal and producing intonational contrasts based on (any) discrete notation go a long way towards uncovering the systematic relationship between the form and functions of intonation.

Similarly to IPA symbols, when we abstracted away from precise characteristics of lip rounding or tongue positions during the b-closures of ‘beet’ or ‘boot’, and simply transcribed [b], we will abstract away from multiple continuous intonational characteristics in working with any discrete labels. Crucially, just like with IPA, we will use the discrete labels while understanding this is just a convenient abstraction focusing our attention on the most salient functions and meanings that can be cued with intonation.

Activity 12-5: informal descriptions of intonation

In Praat, open the sound file and TextGrid ‘welcome_to_the_program_triplet.wav’ found in the chapter’s companion. A part of this was already discussed in Activity 10-1. Make the pitch visible and adjust the pitch range to 60–400 Hz.

Try to describe intonation in as much detail as possible. We can again use the metaphor for our pet humanoid robot that you should instruct to imitate the melody of the three utterances in the file.

Consider, which aspects of the visible f_0 contour are the most relevant and important for a good imitation of the speakers’ intonation and thus, previewing our discussion, which aspects should receive a discrete label.

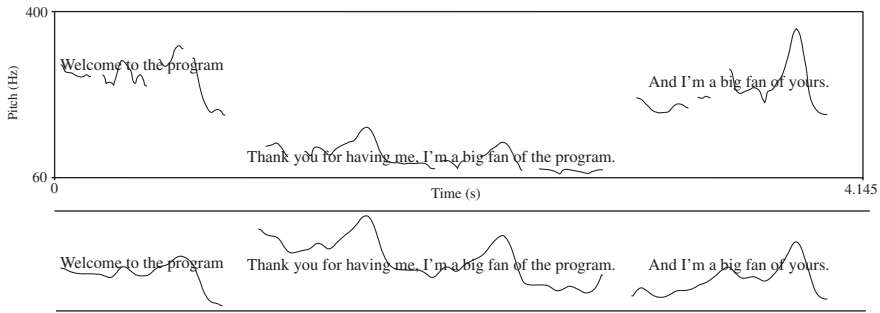


Fig. 12.4 Alignment of pitch contours with text for ‘welcome_to_the_program_triplet.wav’ analysed in Activity 12-5. The top panel traces the raw f_0 in Hz and the bottom shows interpolated, smoothed pitch contours in semitone (based 100 Hz) scale that corresponds more to the perception of melody

What words did you use in your informal descriptions? Maybe you recalled our discussion in Sect. 11.2 and identified prominent words, like ‘program’, ‘having’, ‘fan’ or ‘yours’, and potentially a weak boundary after ‘me’ in the second utterance. Maybe you also mentioned adjectives like falling or rising pitch, or flat pitch range. You might have also perceived speakers as engaged, animated or enthusiastic.

The notational systems for marking potentially meaningful variation in intonation display various levels of abstraction mentioned above. The top panel of Fig. 12.4 shows the f_0 contour in Hz as shown in Praat aligned to the text. This is a true representation of pitch for the speakers within the 60–400 Hz pitch range, but also seems to indicate that the male voice is flatter than the female voice. It also includes gaps and artefacts not corresponding to our perception of pitch discussed in the previous section. In the bottom panel, pitch is visualized in semitones and is more relevant to each speaker’s pitch range, which is schematically indicated with the straight horizontal lines. We see that the male’s voice is close to its upper levels in ‘having’ whereas the peaks of the female voice are somewhat lower in her range. Does this correspond to your perceptions from Activity 12-5?

Contours in the bottom panel were also interpolated and smoothed. I could have manually removed certain spurious pitch points in Praat but refrained from that. This display shows the first abstraction between the raw measurements in Hz and melody perception.

The next step is to identify the candidates for the second level of abstraction. Recall from Sect. 11.2 that we are interested primarily in pitch accents, i.e. how pitch cues the prominence of particular words, and boundaries, i.e. how pitch signals the chunking intended by the speakers. In this particular example, the prominent words, and particularly their stressed syllables, are nicely correlating with the major pitch movements and the ‘peaks’ in the contour. All four major ‘humps’ correspond to prominent words ‘program’, ‘having’, ‘fan’ or ‘yours’. What about the words like ‘welcome’, ‘thank’, ‘and’ or ‘big’? You probably feel they are also somewhat prominent, although less than the first four words and their ‘humps’ are certainly less robust.

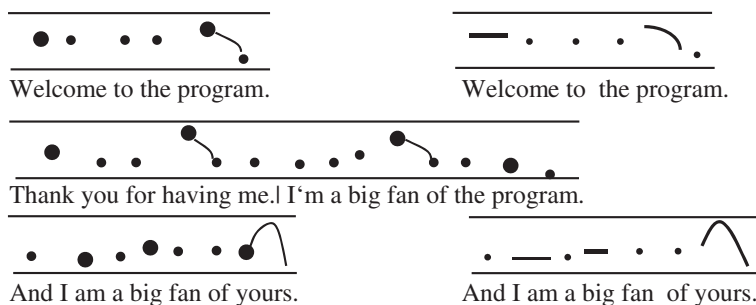


Fig. 12.5 Examples of interlinear systems traditional in the British school of intonation analysis with dots/dashes corresponding to syllables, their thickness or size to the word stress and prominence, horizontal lines to pitch range, and the major contour movement described with stylized tails or curves

One further step in the abstraction from the continuous signal to discrete labels, commonly employed in intonation marking systems, is to indicate syllables with dots or dashes. This allows for marking the contrast between the stressed or unstressed syllables with the thickness of the lines or the size of the dots. Stylized curves are then used to indicate the most salient pitch movement, which is in English commonly occurring towards the end of each prosodic unit. An example of this is shown in Fig. 12.5.

The notational systems arising from the British tradition of intonational analysis abstracts further and identifies the *tonic syllable* (or *intonational nucleus*) containing a *nuclear tone*. This is the most prominent syllable of a unit and might carry most commonly one of five or seven nuclear tones: (*high/low*) *fall*, (*high/low*) *rise*, *fall-rise*, *rise-fall*, and *mid (or flat) tone*. Additionally, the first prominent syllable that is not a nucleus is the *head* (that could be high or low), non-prominent syllables preceding it are the *pre-head*, and non-prominent syllables following the nucleus are the *tail*. Hence, the intonational unit //welcome to the program// starts with a high head in the first syllable, has a high fall nuclear tone on the nucleus 'pro' and the tail with 'gram'. Various suggestions can be found for marking this information in text; for example: //°welcome to the `pro-gram// or //and °I'm a big fan of ^yours//. Some treatments of English intonation, for example an influential discourse-based model of Brazil (1997), include five tones but additional primitives such as key or tonicity.

This approach to intonation is well documented in multiple publications on this area. The 'contour-based' system of intonational description above can be compared to the 'target-based' system arising from the American tradition of intonational research. The most widespread framework within this approach is the system of Tones and Break Indices (ToBI) based on the work by J. Pierrehumbert (1980), elaborated in terms of communicative meanings of these tonal targets in Pierrehumbert and Hirschberg (1990), and further extended in Beckman et al. (2005). In this framework, the salient aspects of intonation are not construed as contours (fall or fall-rise), but as pitch targets that are either H(igh) or L(ow) and

align with both prominent words as pitch accents marked with a ‘*’, and with prosodic boundaries as boundary tones marked with ‘%’ and ‘-’. The pitch accents can signal simple targets, or bi-tonal combinations of H and L. Hence, ‘Welcome to the program’ described above would be analysed as having two prosodically prominent words, both associated with a high target pitch accents and a boundary associated with a low target.

H* H* L-L%.
Welcome to the program

In this book we will devote more space to this approach for several reasons. First, although ToBI was originally based on the intonation of American English, active research in describing other languages resulted in the descriptions of a wide range of languages by extending or adjusting the basic ToBI; e.g. Jun (2005, 2014). Hence, chances are that the intonation of the native language of many readers has been described with a ToBI-like system.

Second, ToBI is easily adaptable to the tasks at hand and can be simplified to cover only some basic functions of intonation, for example identifying only prominent words, or appended to include meaningful intonational pitch targets in other languages.

Third, it corresponds to the fundamental functions of prosody described in 11.2 in terms of marking prominent words and signalling the chunking of speech into units in a widely *used autosegmental-metrical* approach to intonation and prosody in both theoretical research (e.g. Ladd 1996) and applications to speech processing, for example making speech synthesizers speak with more natural prosody.

Fourth, there exist excellent online sources, already using Praat, for readers interested in becoming more proficient at transcribing intonation in ToBI. Particularly the online course in MIT opencourseware (Veilleux et al. 2006) linked in the Find Out More box below.

Find-out-more 12-1

There is an excellent self-study material developed by N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos for a course at MIT ‘Transcribing Prosodic Structure of Spoken Utterances with ToBI’ available through MIT’s open courseware initiative (<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/index.htm>).

The course has approachable lecture notes linked to sound files that you can hear, and uses Praat for visualizing intonation and its marking (more below). It also has multiple exercises for practicing ToBI annotation at your own pace.

Finally, and most importantly, I believe that listening to, and visualizing, the continuous aspects of intonation in Praat appended with using ToBI for discrete marking of intonational contrasts offers a most meaningful *hybrid* approach to studying intonation. Hybrid in the sense that any discrete marking system is

commonly quite distanced from the actual phonetic continuous and dynamically developing features present in the speech signal. However, the option to constantly visualize and possibly also measure and quantify, various aspects of intonation offers a tool very useful for any subsequent investigation of intonation on your own or in projects motivated by the material in this book. Just as we saw with phonetic variability referring to segments, intonational variability is at least as huge, and apart from very stylized utterances, typically covered in introductory books, intonation in real dialogue interactions is enormously complex. Without a discrete-like framework it would be difficult to search for contrastive meanings, but without constantly being aware of the continuity and variability in the speech signal, we might be reaching unfounded or plainly wrong conclusions.

The primary motivation for using a discrete set of symbols for marking continuous speech is to make the contrasts functioning in language more tangible (recall, for example the concepts of minimal pairs or allophones raised for the same reasons). We will outline this relationship between the form of the contours and their major pragmatic meanings in the next section. Before we do, however, we have to also be aware of the disadvantages and general limitations inherent in the use of discrete symbol marking system like ToBI.

These are nicely described, for example in Ward (2019, pp. 69–72). Primarily, the reality of any symbolic representation is questionable since the level of abstraction is commonly too great and symbols do not straightforwardly correspond either to real speech signal or to pragmatic meanings. For example, the inventory of labels might predict contours that are not perceptually salient. Also, any labelling system is essentially theoretically based and ToBI is not an exception. There are assumptions about what kind of pragmatic meanings are signalled with prosody and marked with ToBI, but many other salient meanings are not covered.

Bearing these limitations in mind, intonation falls squarely with other speaking patterns covered so far in that we investigate how phonetic continuous characteristics of articulation/acoustics cue the presence or absence, or a degree, of communicative functions or intentions.

12.4 Simplified ToBI

There are three dimensions of intonational variation that primarily signal discrete-like aspects of pragmatic meanings. Two of them – assigning prominence to particular words and signalling the division of continuous speech into units or chunks – have been introduced. The third one is the quality of pitch targets. These three form the basis of a slightly simplified version of ToBI that is useful for awareness building and self-exploratory goals of this book. As before, we present the basics of the framework along with the exploration in activities.

ToBI captures linking words into chunks and separating chunks from each other with five levels of *break index* (BI, 0-4). This index is associated with the right edge of every word and indicates the strength of the (dis)juncture between two words. The basic smooth linking of adjacent words is marked with BI #1. Connected speech processes of assimilations, elisions, coalescence or linking

discussed in Chapter 10 tend to occur in the vicinity of very weak boundaries. With clearly perceivable processes like these and an extremely tight juncture between adjacent words, BI #0 is commonly used. Break index #2 labels hesitations, disjunctures associated with speech errors, and mismatches between the phonetic cues for indices #1 and #3. These weaker junctures (#0-#2) are not expected to be associated with tonal targets.

Break indices #3 and #4 are the most relevant for this chapter and delimit the right edge of the (*minor*) **intermediate phrase** marked with ‘-’, and the (*major*) **full intonational phrase** marked with ‘%’, respectively. Both are associated with pitch targets. The right edge tone marking for intermediate phrases are called **phrase accents** and are represented by tones L- or H- and BI #3. The full intonational phrases contain one or more intermediate phrase and the right edge **boundary tone** represented as L% or H% and BI #4. Since any full intonational phrase must contain at least one intermediate phrase, every BI #4 contains both the phrase accent and the boundary tone: L-H%, H-H%, etc.

As mentioned before, we can easily adjust ToBI for our needs. For example, if interested only in crude chunking, we might lump indices #3 and #4 under #4 and disregard the rest. If, however, we want to study connected speech aspects, identifying breaks #0 and #1 makes sense. Similarly, studying phonetic characteristics of hesitations and speech errors invites the annotator to use #2 index.

In sum, the greater degree of disjuncture, the higher break index and the more robust and salient the phonetic cues signalling the boundary. Pre-boundary lengthening, exemplified in the previous chapter, i.e. with the phone number, applies to all breaks and greater lengthening indicates a higher break index. The phrase accent (‘-’) of BI #3 induces some lengthening and a weaker tonal target, and the presence of the additional boundary tone of BI #4 (%) is realized as greater lengthening, more pronounced tonal marking that is commonly, but not necessarily, followed by a silent pause.

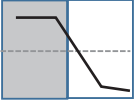
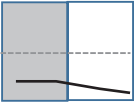
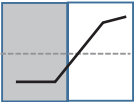
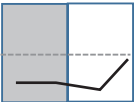
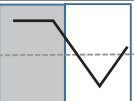
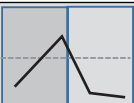
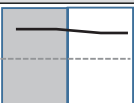
In addition to tonal targets associated with boundaries, prominent words are highlighted with salient tones called **pitch accents** and marked with the ‘*’. Just like the nucleus (tonic syllable) is required for each tone unit in the contour-based annotation, the basic rule in ToBI is that every (minor) intermediate phrase has to contain at least one pitch accent.

Hence, we now have the basic inventory of primitives used in our simplified ToBI framework that we may start using for exemplifying the system in real speech:

- pitch accents: H*, L*
- phrase accents: H-, L-
- boundary tones: H%, L%

With this inventory, Table 12.1 lists possibilities of the ToBI annotation for six of the seven nuclear tones of the British tradition. The schematized f_0 contours in the third column of the box represent a two-syllable word with the stress on the first syllable (the shaded box) and the speaker’s pitch range as the dotted midline. The fourth column is left intentionally blank; see also Activity 12-6

Table 12.1 Traditional contour-based labels, ToBI annotation, and schematized f_0 for basic nuclear tones

Nuclear tone	ToBI	Schematized representation	Context/Intention
High fall	H* L-L%		
Low fall	L* L-L%		
High rise	L* H-H%		
Low rise	L* L-H%		
Fall-rise	H* L-H%		
Rise-fall			
Mid (flat)	H* H-L%		

The first five ToBI annotations are fairly intuitive and the H/L targets correspond well with the mnemonics (high/low) rise or (high/low) fall. The flat contour, called also a *plateau* in ToBI, is marked with H-L% that does not depict a fall that the succession of H- and L% might evoke. Recall the essentially arbitrary character of the symbols in any discrete system of labels. You can compare these schematic representations with your productions in Activity 12-6.

Activity 12-6: Basic tunes

Select several two-syllable words with the initial stress, for example {‘Mary’, ‘sister’, brother’, ‘Cathy’, ‘really’,...}. Say each of them with the intonation described in the six unshaded boxes of Table 12.1. Think of the potential contexts in which you would produce these tunes. Can you use any contour for any word or are there restrictions?

After practicing the production of these schematized contours several times until you are comfortable, record yourself saying them, save and open in Praat. Adjust the pitch range and examine the contours. Can you identify the basic

shape of the schematized contours despite the effects of consonants and their voicing on the pitch tracks? Inspect the pitch contours further and strengthen the link between the auditory and visual information provided by Praat.

Now prepare a similar list of monosyllabic words; e.g. {'John', 'yes', 'no', 'Pat', 'great', 'wow',...}. Again, can you use any contour for any word or are there restrictions?

Finally, look back to what you were thinking when saying these utterances. Try to write down the context, what you wanted to express when you were saying these disyllabic and monosyllabic words with particular contours. In fact, try to fill the fourth column of Table 12.1. When done, compare with a companion/group who was doing the same activity and discuss similarities and differences.

Observations from Activity 12-6 include at least three main points. First, essentially any contour can be produced with any word or a phrase. Hence, intonation is truly auto- or supra-segmental since it is independent of the words or sounds with which it is realized. Second, in order to produce these tunes, you most likely, consciously or sub-consciously, imagined a situation and intentions relevant for that situation. This context and intention is indispensable in any discussion of intonation and also facilitates the natural production of these tunes. Hence, the tunes are said to express various situational, contextual, and pragmatic meanings, or functions. Third, if you compared your contexts and intentions for each tune with a friend, you most likely come to the conclusion that a single tune can express more meanings/intentions and sometimes a similar intention might be realized with different tunes. This final observation is responsible for the complexity, challenge, and beauty of studying intonation since natural speech exhibits *many-to-many relationship* between the *form* of intonation, whether represented by continuous phonetic features or discretized symbols, and its *functional meanings* that speakers and hearers assign in natural spoken interactions.

With the basic ToBI system illustrated, now is the time to actually start producing your first annotations. Before we do that, however, we have to mention briefly some formal requirements, and then three important aspects to be aware of when ToBI labelling. Regarding the formal requirements, standard ToBI annotation has four layers, or tiers: *tones* identifying the tonal targets (pitch accents, phrase accents, and boundary tones), *words* showing the alignments of every word to the signal, *breaks* identifying the break index at the right edge of each word, and *miscellaneous* that might be used for labellings of laughter, speech errors or other miscellaneous aspects of speech. The pitch accents are marked anywhere within the stressed syllable of the word, and boundary tones at the right edge of the word.

Now the three points to keep in mind. First, skills and awareness gained in the process of producing your annotations is the real goal. The output of reasonable ToBI annotations is a by-product that will improve gradually with more time, energy, and engagement you devote to this skill. In this, ToBI is exactly the same as IPA transcription, especially the approach when IPA transcribing entire phrases and not individual words.

The accuracy of labelling concerns the second point. Apart from clear formal mistakes (like having a phrase without a pitch accent, or producing labels like L*H% instead of L-H%), there is no gold standard or ‘correct’ ToBI annotation. Some contours are straightforward, some are more challenging, and with some, even expert ToBI labellers disagree. This is natural and you should remember the main goal from the first point above is to gain appreciation of the variability.

The final point concerns the selection of material for labelling. Starting with Chapter 10, most analyzed sound files come from real natural speech. This provides a great challenge in analyzing speech patterns compared to artificially recorded phrases that clearly exemplify a relevant point of interest. An approach that describes first clean, stylized utterances and only then goes on to naturally occurring speech makes a lot of sense, and is employed in many textbooks on this topic. Getting your hands ‘dirty’ with real speech data from varied situational contexts right from the start is also a relevant approach that highlights the complexity and multi-dimensionality of speech. We take an intermediate approach and use exclusively real speech data in this chapter selected to illustrate intonational meanings with limited context, and extend it to greater situational context in the rest of the book.

With all these in mind, Activity 12-7 guides you to your first ToBI labelling in Praat.

Activity 12-7: First ToBI annotations in Praat

Let’s start with the phrase already analyzed in the previous section: ‘welcome to the program’. First, we need to prepare a TextGrid file for ToBI annotations. Open the TextGrid ‘welcome_to_the_program_GN_empty.TextGrid’ from the book companion and click *View & Edit alone*. You should be able to create an identical TextGrid on your own. Go ahead and try, review activities in previous chapters if needed. Only if you get stuck, check below.

The easiest way to proceed is to open the sound file ‘welcome_to_the_program_GN.wav’ from the book companion, then with the file highlighted in the Objects window *Annotate → To TextGrid...* and specify the 4 tiers we need to create (tones, words, breaks, misc) and indicate that tones, breaks, and misc should be point tiers. Open the textgrid with the sound file, set to view the spectrogram, and align the four words with the sound by creating and labelling the intervals in the second tier. Select the right boundary of a word in the 2nd tier, and click the circle at the cursor in the 3rd tier, which creates an empty point for labelling the break index. Compare your TextGrid with the empty one in the book companion and if everything is the same, save your work.

Finally, make the pitch track visible, adjust the pitch range, and enter the relevant tones in the 1st tier and breaks in the 3rd tier. The tones were already discussed in the previous sections, and breaks are all #1 with BI #4 after ‘program’.

The book companion contains two more files from the same speaker: 'daughter_of_imigrants_GN.wav' and 'reached_her_earlier_GN.wav'. Proceed by opening the sound file only in Praat, creating a ToBI textgrid file, and then filling the annotation, and saving your work. Finally, compare with the associated TextGrid files with the '-full' suffix in the companion.

How did you proceed when annotating the two files in Activity 12-7? It is important to develop your own approach. I usually start with determining the phrases/chunks; hence where the #3 and #4 breaks are. Then I select just the relevant chunk, listen to it several times, and try to determine which words are pitch accented. I might use hand-clapping described in Sect. 11.2. Then I consider how to best capture f_0 targets and my perception of the intention of the speaker within the inventory of tones provided by ToBI. I finish with entering the missing break indices.

Using real speech and reinforcing points made earlier, *uncertainty* about ToBI annotation is very natural, common, and to be expected. In fact, labelling something as uncertain, and reasonably explain the uncertainty in non-trivial terms, qualifies as understanding in the approach promoted in this book. The ToBI annotation scheme includes several ways of expressing uncertainty in labelling. If you opened 'daughter_of_imigrants_GN_full.TextGrid' from Activity 12-7 you saw that I put '*?' on 'she', which indicates that I am not sure if this word receives a pitch accent. Did you have a pitch accent on this word?

To explain this uncertainty in non-trivial terms, I would say that the metrical structure of the utterance conforms to having a pitch accent on 'she' reasonably well. Informally, when I clap the accents with my hands, clapping on 'she' feels a bit closer to what the speaker produced (or meant to produce) than not clapping. On the other hand, the pitch is mid-level, which is expected at the onset of a phrase, 'she' is relatively short, and I also feel less prominent than 'daughter'. Hence, following this, or similar, non-trivial explanation of doubt, labelling '*?' is justified.

Other uncertainty labelled in ToBI include 'X' if not sure what type of pitch target was intended; e.g. X* for a pitch accent or X-X% for a boundary tone. Hence, for X* you are confident that a word is accented, only the type of the pitch accent is not clear. A great way of dealing with uncertainty, which I strongly recommend, is to include one more point tier in ToBI TextGrid called 'alt' for *alternative labelling*. This is also supported in the online ToBI course linked in the Find Out More box 12-1. Here, if you cannot decide between two alternative annotations, this second-best alternative might be included. In all cases of uncertainty overtly indicated in ToBI annotation, always make sure you can argue reasonably why this uncertainty is used and not to overuse it. Naturally, the goal of any discrete annotation is to capture the relevant aspects of speech in a concise and consistent way. However, admitting reasonable uncertainty, leads to better understanding and awareness of the speech patterns, and better appreciation of the strengths and weaknesses of ToBI and discrete annotation in general. Both of these fall squarely within the goals of this book.

Our system of 6 primitives (H^* , L^* , $H-$, $L-$, $H\%$, $L\%$) captures a lot of variability in naturally occurring speech. However, it still lacks means for describing some of the very common patterns in English. For example, return to the file ‘welcome_to_the_program_triplet.wav’, particularly to the third utterance ‘and I’m a big fan of yours’. The accent on ‘yours’ informally corresponds to a rise-fall, which in the current ToBI inventory cannot be handled. The following passage describes two elements of the full ToBI system that we will add to our simplified system.

Both of them involve *bi-tonal pitch accents*. In addition to simple pitch accents H^* or L^* , the two most common bi-tonal pitch accents are rising $L+H^*$, and a ‘stepped-down’ $H+!H^*$. Both of these accents are very common in English but as you can see also represent an enlargement of our set of primitives. First, we are adding the plus sign as a new symbol to the inventory. This indicates that the two tones are associated to a single stressed syllable but the pitch targets are not equal. The tone followed by the star described the more prominent target and $L+H^*$ is thus perceived more as an H target than an L target. The $L+H^*$ accent is commonly perceived as a ‘scooped’ accent in that pitch falls slightly, typically before the stressed syllable (L), and then rises during the major portion of the stressed syllable (H^*). This addition allows us to describe the rise-fall nuclear tone as $L+H^* L-L\%$ in ToBI and fill the empty shaded row in Table 12.1.

The second bi-tonal accent, $H+!H^*$, includes the notion of *downstep*, marked with an exclamation mark. It captures the intention of the speaker to reach a pitch target that is lower, or stepped down, from the preceding H target. $!H^*$ is commonly found either as a separate pitch accent, stepping down from a preceding H^* target, or as the stronger second element of a bi-tonal accent $H+!H^*$ in which the H-target commonly aligns with unaccented material before the accented syllable. In addition to the already mentioned general sequential restriction that every ‘-’ phrase accent must be preceded by a ‘*’ pitch accent (i.e. every unit must have a prominent syllable), $!H$ is restricted to follow an H target in the same intermediate phrase (i.e. f_0 can step down only from a preceding H target). A $!H^*$ can also follow another $!H^*$ if the latter steps down further from the former. There are only these two sequential restrictions relevant for us and none of the other tones are restricted in their occurrence by preceding or surrounding tones.

Both $L+H^*$ and $!H^*$ pitch accents are exemplified through Praat manipulation functionality in Activity 12-8.

Activity 12-8: Manipulating stylized pitch in Praat

Recall that in Activity 12-6 I asked you to mimic the stylized f_0 contours in Table 12.1 to build the awareness regarding the contrast among pitch targets. Here we further strengthen this awareness. Instead of changing your own pitch, you can make gross alterations directly in Praat. We build here also on pitch manipulation in Activity 9-4.

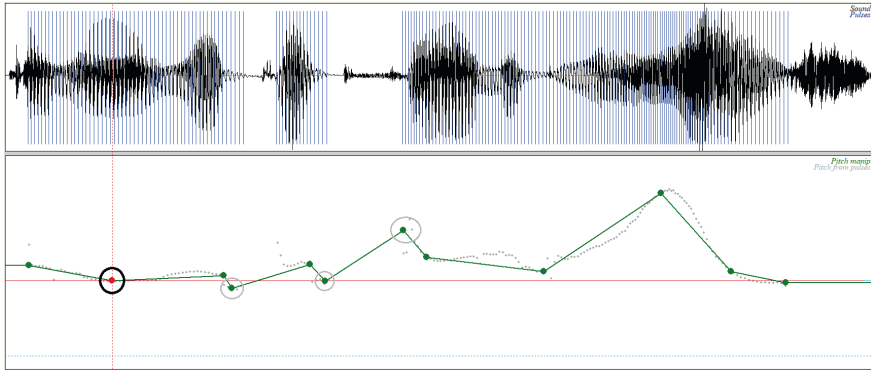


Fig. 12.6 Screenshot from manipulating pitch in ‘fan_of_yours.wav’. See Activity 11-8 for a detailed description

Open the last turn of our triplet example in the companion: ‘fan_of_yours.wav’. With the sound highlighted, create a Manipulation object: *Manipulate* → *To Manipulation...* and set the floor and ceiling to 80 and 500 Hz, respectively. With the manipulation object highlighted, click *Extract pitch tier* and then create a stylized version of pitch with *Modify* → *Stylize...* and set Frequency resolution to 2 semitones. You can *View & Edit* the stylized Pitch Tier. Finally, select both the manipulation and the pitch tier objects and click *Replace pitch tier*. Select the Manipulation object only and click *View & Edit*. Figure 12.6 shows the first two panels of the manipulation object that you should see on your screen.

The straight (green) line represents the stylized pitch and you can check how it traces the original pitch contour represented with grey dots. Employing our understanding of pitch errors and artefacts, we should delete the three turning points marked with grey circles in Fig. 12.6, since they represent unreliable pitch in the vicinity of the consonants. It would be also good to add one more pitch turning point for ‘I’m’ marked with a black solid circle (*Pitch* → *Add pitch point at cursor*).

Now, you are ready to alter the pitch contour by dragging the pitch points. Start with the last two points changing the boundary tone, for example from the original L-L% to L-H% or H-H%, as approximately indicated in the left panel of Fig. 12.7. Listen to the created utterance by pressing the bar below the sound and compare it with the original by clicking the bar while holding the shift key. Consider context and intentions of the speaker when producing these new contours.

Finally, manipulate pitch accents on ‘yours’ by lowering it (top right panel of Fig. 12.7) to H* and compare it with the original L+H*. Then, increase pitch on ‘big’ to create a downstep on ‘yours’ (middle right panel in Fig. 12.7) and compare with a typical ‘hat pattern’ of H* H* L-L% of the bottom right panel. Combine the manipulation of pitch accents in the right of Fig. 12.7 with that of boundary tones in the left. Again, consider context and intentions of the speaker when producing these new contours and observe the types of contrasts conveyed by these manipulations of the pitch contour.

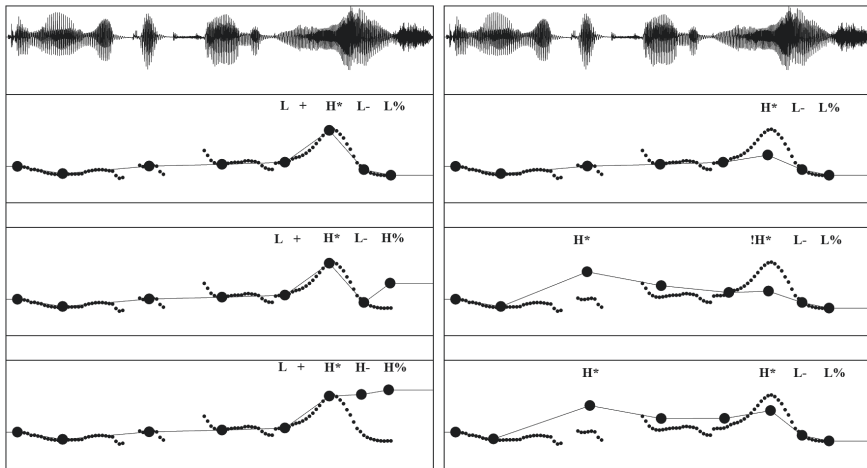


Fig. 12.7 Manipulation of stylized pitch in Praat for generating various ToBI annotations

Gross manipulation of pitch contours with Praat described in Activity 12-8 is both fun and provides important insights through basic analysis-by-synthesis approach (which changes make perceptually salient differences). Naturally, the synthesized utterances sound somewhat unnatural (already discussed with Praat manipulation in Chapter 9). On the other hand, my manipulation of H*L-H% or H*!H*L-L% sounded very reasonable and the contrast with the other contours was very clear. I encourage you to try the manipulation of stylized contour with utterances that you record yourself.

Table 12.2 illustrates schematic f_0 contours that our simplified system of primitives can capture. Try producing them with single words or short phrases.

We will close this section with several important remarks. First, I'd like to re-iterate the primary goal is to provide the tools for describing English intonation that would facilitate your own hands-on engagement with intonation by exploring recordings with Praat. We do not have space to examine every primitive and every contour in Table 12.2 or every contrast in the break indices. Nor can we outline a complete ToBI scheme, which includes, for example L*+H pitch accents, !H-H% boundary tones, 'p' label for marking hesitations in the break tier, and others. The online course listed in Find Out More box 12-1 is an ideal reference for those interested.

Second, both IPA and ToBI are essentially phonological frameworks since they abstract from the continuity of speech in order to capture meaningful contrasts. IPA describes relevant contrasts using discrete symbols for phonemes, allophones, stress marks, etc. ToBI also uses a fixed inventory of primitives to capture functional meanings (more on this in the following section). ToBI describes intonational contours as arising through *interpolation* between successive tonal targets. It is important to bear in mind that the H and L labels never correspond to absolute f_0 values and should be considered as *relative* describing the intention of the

Table 12.2 Summary of the schematized f_0 contours predicted in the adapted ToBI framework described in this section. As with Table 12.1, the box represents a two-syllable word with the stress on the first syllable (the shaded part) and the speaker's pitch range as the dotted midline

	L-L%	L-H%	H-H%	H-L%	L-	H-
H*						
L*						
L+H*						
(H+)!H*						

speaker to reach certain tonal target within a given pitch range and in the context of other material in the utterance. Furthermore, just like with the aspects of connected speech, the tonal targets can also influence each other or be constrained by the duration of the segmental material. In other words, the same ToBI labelling, for example H* L-H%, may be realized in actual f_0 tracing variably depending on the length of the phrase in terms of syllables, the voicing of the consonants, how far the pitch accent is from the boundary, and many other aspects.

Finally, there are many continuous prosodic dimensions of speech that ToBI does not capture. For example, there are no labels for changes in speech tempo or intensity. But these interact closely with the three basic dimensions (breaks, accent location and type) in non-trivial ways and convey various emotional and attitudinal aspects of the speaker. Additionally, we also abstract away from some natural characteristics of pitch such as declination mentioned in the previous chapter.

Despite limitations like these, I believe that the concepts of adaptability for particular needs or domains and marking uncertainty described in this section make ToBI a useful hybrid framework for our main goals.

12.5 Basic Functional Meanings of Intonation

The previous sections were primarily concerned with characterizing the intonational contours. Activities always encouraged you to consider the situational context and intentions of the speakers when producing such tunes. However, I did not want to cloud the breadth of your varied observations of the links between the form and meaning of intonation with either my own perceptions or various expectations for these relationship found in the immense literature on this topic. The aim

Table 12.3 Sample tunes and meanings for ‘It’s nine o’clock’

It’s nine o’clock	Context/Intention
H* H* L-L%	Default response to A’s question ‘What’s the time?’; assertive. Possible also without the question with various pragmatic meanings (e.g. please hurry up, or, slow down there is plenty of time)
H* L-L% H* L-L%	You are ready but your partner is not and was supposed to be ready long time ago; possibly slightly angry or annoyed
(L*) L+H* L-L%	Your partner, still not ready, asks ‘Is it quarter to nine already?’; you want to correct the partner
L* L* H-H%	You are not ready and thought it was 8.30 and you had plenty of time. But your partner enters the room ready. You are surprised
H* !H* L-L%	After a lengthy discussion of when to leave for the party; this is the end of the discussion, my final word
L+H* L-L%	A asks: is the party at 8 o’clock?; you want to correct that.
H* H* L-H%	Darling hurry up we don’t want to be late
...	...

was to incite your curiosity and increase your awareness and appreciation of the complexity of this relationship. Activity 12-9 expands on this approach.

Activity 12-9: It’s 9 o’clock

Imagine a dialogue between A and B in which A says something, to which B responds with ‘It’s nine o’clock’. Your first task is to brainstorm and list as many situational contexts and intentions underlying B’s response as you can think of. Write them down. This is a good activity to share with a friend or a group of people.

The next task is to consider how these different functions might be cued with variation in intonation. Try to say the phrases with the intended meaning, and consider their description in ToBI. Of course, the best approach is always to record and visually inspect with Praat.

One way of approaching the tasks of Activity 12-9 is to imagine a particular context and try different variations in the intention of the speaker. For example, suppose that you (B) and your partner (A) are getting ready for a party and A enters the room. Table 12.3 lists several examples of possible tunes and their intentions.

Your brainstorming during Activity 12-9, alone or with companions, produced without a doubt links between the tune and the intention that possibly, but not necessarily, overlap with those listed in the table. The variability reached from Activities like 12-9 for various groups of English speakers corroborates the already mentioned complexity of many-to-many relationship between pitch contours and their conversational meanings. Yet, despite this enormous complexity, people

typically communicate smoothly and these intentions are easily produced and perceived. Hence, the link between meaning and form of intonational tunes surely belongs among the unconscious speaking habits we want to investigate in this book. Of course, occasional mis-understandings in spoken interactions take place, but the very expectation of something that turns out to be different than expected, is evidence for an underlying system of these expectations.

Our approach to intonational meanings in English will be similar to the approach to English word stress patterns in Chapter 9. Our goal is not to provide a complete phonological analysis of English intonation; there are entire books or research papers written on this topic that the interested reader is advised to consult (e.g. Ward 2019; Sczepak-Reed 2010; Gussenhoven 2004; Cruttenden 1997; Brazil 1997; Ladd 1996). Rather, we will limit the presentation to the major types of mental knowledge that is generally accepted to be active in the unconscious speaking habits.

The central notion is the *common ground*, or a system of *mutual beliefs*. The speaker has to constantly infer what the hearer believes, which ideas or concepts the two interlocutors share, and which might consider new information to be added to this common ground or mutual belief system. As speakers, we all conceptualize intonational meanings within this frame, i.e. how an utterance contributes to our interlocutor's beliefs, what the goals are, and what action is intended next. This view is based on seeing speech as a tool for *joint action* (Clark 1996), where asking, informing or expressing disagreement are similar to other common joint activities such as dancing, team sports or playing music in a group. *Coordination* at the *physical* as well as *mental* levels is the essential key to all such activities. For example, a late pass in any team sport can be seen as a problem of linking the physical coordination of two players to the mental model of what each player is going to do. This discrepancy will then manifest itself in temporal or spatial discrepancy and in problems with the continuity of the game. Speech communication is the same as team sport or, for example the effort of two people to move a heavy piece of furniture. They need to understand where and how they move, coordinate with each other during the move, and continually adapt their physical manifestations as well as the mental models of this activity. Intonation can thus be seen as one of the means for achieving this mutual mental coordination.

Activity 12-10: Welcome to the program again

Listen again to the opening of an interview explored in the chapter 'welcome_to_the_program_triplet.wav'. Consider how the notions of coordination, joint action and common ground might be playing a role in the smoothness of this spoken interaction.

To include several examples, a simple (H*) H*L-L% suggests that the speaker considers entities with H* pitch accents as new information to the hearer while L-L% signals the utterance is self-contained and nothing needs to be added

necessarily. This is the case of the first row in Table 12.3. The most common use of L+H* is to mark extra emphasis, or in some cases contrast with what was said or assumed before. Hence, the accented item, rather than some alternative, should be added to the common ground.

Both of these accents are closely linked to the notion of *focus*, that is, what is currently being talked about. If we present some new information without assuming any prior context or common ground, the utterance is said in *broad focus*, and H* is the most common pitch accent. If the focus is *narrow*, or *contrastive* such that some elements have been recently invoked, L+H* is commonly used as in row 6 of the table. The L+H* on ‘fan’, and the absence of a pitch accent on ‘program’ (also called *de-accenting*) due to its previous mention and forming an old, given information, are both indicators of narrow focus in ‘I’m a big fan of the program’.

The downstep tonal target (!H*) is commonly used in lists, calling contours or to signal greater cohesiveness between the pitch accented elements. For example, the notion of ‘nine o’clockness’ is much better invoked with H* !H* accents of ‘nine’ and ‘o’clock’ than with alternative H* H* in which the two elements are more separable.

The pitch targets associated with boundaries are most relevant to the expectations of what is to follow. High targets typically express a *forward-looking* meaning that invites the hearer to search for the connection between what was just said and what is to come. For example, in many polar (yes/no) or tag questions an answer is expected to follow, or in a prototypical continuation rise (L-H%), the speaker her/himself plans to continue, or implies something linked to the current utterance, as in the 8th row of Table 12.3. The low boundaries typically express finality, independence of the two phrases separated with a low boundary, or expected confirmation after a tag ending in low boundary.

In addition to these general remarks, some (stylized) contours might also have well-defined particular meanings, like the calling contour, or the surprise-redundancy contour. The ‘calling_surprise-redundancy.wav’ in the companion has examples of both. Can you identify what is surprising and what redundant in the utterance ‘Intonation isn’t easy’?

For the sake of completeness, descriptions of these meanings and functions in the British tradition is commonly structured under the headings of *accentual (focus)*, *attitudinal*, *grammatical* and *discourse* functions of intonation. In some cases the listed (attitudinal) functions like ‘expressing doubt’, ‘mocking’, ‘being impressed’, ‘friendly’ feel subjective and situation dependent. Or, the most salient grammatical function is commonly illustrated with the falling wh-questions and rising polar (yes/no) questions. This does apply in some contexts and some intentions but there are many cases in real data in which these do not apply. Even with the discourse function of accenting new information and de-accenting old/given information, there are numerous instances not respecting this pattern in natural conversations. Hence, in the spirit of the book, I invite you to explore the material and the commentary in Chapter 14, but ultimately explore and investigate on your own with the skills, tools and awareness gained by using this book.

Exercises

- 12-1 One of the general examples used by Pierrehumbert and Hirschberg (1990) is the utterance ‘The train leaves at seven’. First, discuss the distribution of H* accents with a fixed L-L% boundary. How does accenting one, two or three of the relevant elements (‘train’, ‘leaving’, ‘seven’) signal what the speaker believes about the common ground s/he shares with the hearer? Additionally, inspect how H% boundary (possibly in L-H% and H-H%) reflects the forward-looking meaning of the utterance. Test how a reasonable continuation, like ‘or nine thirty’, compares with apparently non-sensical ‘there’s a full moon tonight’ if the latter follows a H%.

References

- Beckman, Mary, Julia Hirshberg, and Stefanie Shattuck-Hufnagel. 2005. The original ToBI system and the evolution of the ToBI framework. In *Prosodic typology: The phonology of intonation and phrasing*, ed. Sun-Ah Jun, 9–54. Oxford: Oxford University Press.
- Brazil, David. 1997. *The communicative value of intonation in English*. Cambridge: Cambridge University Press.
- Clark, Herbert H. 1996. *Using language*. Cambridge: Cambridge University Press.
- Cruttenden, Alan. 1997. *Intonation*, 2nd ed. Cambridge: Cambridge University Press.
- Ladd, Robert D. 1996. *Intonational phonology*. Cambridge: Cambridge University Press.
- Gussenhoven, Carlos. 2004. *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
- Jun, Sun-Ah. 2005. *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press.
- Jun, Sun-Ah. 2014. *Prosodic typology II*. Oxford: Oxford University Press.
- Pierrehumbert, Janet. 1980. The phonology and phonetics of English intonation. Ph.D. dissertation, Massachusetts Institute of Technology.
- Pierrehumbert, Janet, and Julia Hirschberg. 1990. The meaning of intonation contours in the interpretation of discourse. In *Intentions in communication*, ed. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, 271–311. Cambridge, MA: MIT Press.
- Szczepek-Reed, Beatrice. 2010. *Analysing conversation: An introduction to prosody*. London: Red Globe Press.
- Veilleux, Nanette, Stefanie Shattuck-Hufnagel, and Alejna Brugos. 2006. *6.911 Transcribing prosodic structure of spoken utterances with ToBI*. January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare. <https://ocw.mit.edu>.
- Ward, Nigel. 2019. *Prosodic patterns in English conversation*. Cambridge: Cambridge University Press.



Prosody III: Beyond Intonational Phrase

13

In this chapter, we will

- discuss how prosody signals meaningful structuring of speech
 - through the deployment of words like ‘and’, ‘um’, ‘okay’ and others
 - through relating individual intonation units into larger meaningful units of speaking
 - through participating in turn-taking organization
- explore prosodic alignment as a potential means for managing alliances in social communication
- extend Praat skills with basic scripting and feature extraction

13.1 Introduction

This chapter concludes our investigation of speaking habits in gradually enlarging units of analysis. In Chapters 5–12 we explored sounds, syllables, words and intonational phrases, and this chapter focuses on a selection of speaking patterns observable more globally in speech segments spanning several intonational phrases. We started with a magnifying glass, studying how particular movements are made, and gradually zoomed out to consider how these movements are combined into larger functional routines, and how these sub-routines combine into smooth actions influenced by the context and intentions of the speaker. We looked at how our mental models of conveying different types of messages and information can be fruitfully investigated through studying continuous phonetic

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_13) contains supplementary material, which is available to authorized users.

characteristics of speech that are observable in the articulatory, visual, auditory or kinaesthetic domains. Invoking the similarity between speaking and sports or musical performance from Chapter 2, we are now aware of the underlying mechanisms and types of knowledge going into saying a phrase, which might be compared to routines and knowledge involved in R. Federer's backhand stroke or M. Schiffrin's single slalom turn. But there are other speaking habits active in terms of strategizing, planning and adjustments that we effortlessly and unconsciously perform at a more global level. This can be compared to strategizing, planning and adjustments needed to win an entire match, a complete slalom run, or convey emotions of a piece of music to the audience.

The chapter discusses various notions of how the prosodic realization, particularly in terms of timing and intonation, of sample utterances might function in larger chunks of speech. All the example sound files are extracted from the three NPR interviews described in Chapter 1 and analysed in detail in Chapter 14. After the presentation of these concepts, I advise you to listen to the entire excerpts and examine the pattern under discussion in this larger speech material that is linked in the companion with Chapter 14 material. For example, this chapter starts with investigating 'and' in one interview. In addition to the two tokens discussed here, there are 13 other tokens in the entire excerpt of a little over a minute long and you should explore the similarities, differences and additional observations on top of the ones discussed in the main text.

13.2 Discourse Markers

In the previous chapter we have mentioned *discourse* when outlining functions of intonation and how the distribution of pitch accents may signal what the speaker considers new or given information to the hearer, or how the tonal targets of the boundaries may help orienting the listeners to what to expect next. In the narrow sense relevant for us, spoken discourse can be construed as a collection of utterances viewed as a cohesive social interaction. For example, a lecture, an interview, a re-told story or a chat over coffee are all examples of spoken discourse.¹

These larger chunks of speaking typically have internal structure. Just like a piece of larger text on paper is commonly organized into paragraphs, sections, and chapters, any spoken discourse has a similar hierarchical organization. Think back to a recent lecture, a told story or a chat you participated in and you should easily identify the main message or thesis, possibly some opening, body, and closing, how a certain story or an idea supported the main message, how individual sub-components were linked together, or other characteristics of this hierarchical

¹Discourse is a multi-faceted term and the object of study in a broad area of discourse analysis. Therefore, discourse in a broader sense may refer to both linguistics but also to other political, social and institutional domains, educational or media discourse, or even larger aspects of culture or genres.

organization. There are several means in the repertoire of our speaking habits that help with conveying and decoding such organization.

Activity 13-1: What does ‘and’ mean? (I)

First, spend a moment reviewing the meaning of one of the most frequent words - ‘and’. What concepts, functions or (grammatical) relations does it signal?

Now, consider the text transcript of the opening of an interview explored in the previous chapter containing the initial triplet analysed before, plus one more utterance:

GN: Welcome to the program

BC: Thank you for having me I’m a big fan of the program

GN: And I’m a big fan of yours and I’m a big fan of “The Grinch” which is of course a Christmas classic

BC: Yes

The second turn of the host GN, contains two tokens of ‘and’. What functions/relations do they signal? Is there a difference between them?

Now, listen to the above exchange that appears in the companion as ‘welcome_triplet_and.wav’. Is your perception from the text regarding two ‘and’ tokens supported or rebutted after listening? You can now open the associated TextGrid and inspect the two ‘and’ tokens in Praat and compare them phonetically along the dimensions we have been exploring in the book: duration, pitch, intensity, quality of the segments, weak/strong form, pitch accents, boundary strength, etc.

What unconscious speaking habit is speaker GN employing? In other words, why is she producing them differently?

The two ‘and’ tokens can be thought of as a special case of a minimal pair, similar to ‘fofofo’ or ‘mermaid’ words in previous chapters, which have identical segmental material but potentially different meanings or functions cued by their prosodic characteristics. It is no doubt that both ‘and’ tokens signal some form of coordination and continuation between the two utterances flanked from both sides building the cohesion of the larger unit. Furthermore, there is the phrase ‘I’m a big fan of’ produced three times, which further supports the idea that the three utterances containing identical lexical items, syntactic constructions, and joined with ‘and’ belong together despite the fact that they come in separate turns of the two speakers. In addition to these, pronouns, ellipsis, and other *cohesive devices* by the speaker also signal what the listener should already know from previous material in the current discourse unit.

However, despite all this, the prosody of the two ‘and’ tokens is clearly different. The first one is a weak form, unaccented, linked smoothly with the following word, shorter and not carrying a pitch accent. The second one is a strong form, pitch accented, separated from the adjacent word with a strong boundary. In short, the speaker made the second ‘and’ prosodically much more salient than the first one. One might expect that this greater prosodic saliency intensifies the default functions of addition/coordination, showing even greater cohesion. However, my

analysis is that the intention of GN was actually to signal something like ‘hey, enough of the opening pleasantries, I am going to pivot to the main topic of the interview, the Grinch’. She did it very ingeniously by including ‘the Grinch’ with the re-used construction ‘I’m a big fan of’, hence the use of ‘and’ is fully fitting. However, the prosody also signals a degree of separation from the initial opening.

Activity 13-2: What does ‘and’ mean? (II)

Given the text above, what other words could she have used to convey this intention? What would be the prosody? Could a different prosody of ‘and’ convey the same function?

I have switched the two ‘and’ tokens in ‘welcome_triplet_and_switched.wav’. Listen to this file and consider if it disrupts the flow and cohesion of the interaction for you. You can play further by making both tokens either the strong version or the weak version using your Praat skills and investigating the effect on perceived cohesion.

‘And’ in the above interaction functions as a *discourse marker*, or a *cue word*, helping to code and decode the organization of discourse and the intentions of the speakers. Other discourse markers include ‘okay’, ‘so’, ‘well’, ‘now’ and many others. These are words with little information content but rich functional load: in addition to their literal meaning, they can signal a wide array of communicative discourse functions. Importantly for us, the prosody participates in cuing these different discourse meanings and intentions.

While engaged with this opening sequence, consider also the sharp f_0 rises of GN in the third turn and ‘yes’ from BC. Despite ‘yes’ and the rising pitch contours, GN is not really asking anything (she would come across as foolish), and ‘yes’ is not an answer. Consider the intentions of the speakers, and why ‘yes’ is a low fall L*L-L% rather than H*L-L% in this context in the light of Sect. 12.4 of the previous chapter.

13.3 Conversational Fillers

The above discussion of ‘and’ and discourse markers brings to light how the prosody of very frequent and seemingly unambiguous words is harnessed for helping the hearer to orient herself in the discourse. Activity 13-3 explores another, maybe surprising, candidate.

Activity 13-3: What does ‘um’ mean?

Let’s continue exploring the BC interview. This time, please listen to the file ‘welcome_triplet_and_um.wav’ linked in the companion; it includes the opening from the previous two activities plus the continuation of GN’s turn. Your task is to speculate why the speaker used the ‘um’ preceding the question ‘Why did you wanna play the Grinch?’ How does the prosody of ‘um’ help the listener to decode the speaker’s intention?

‘Um’ in Activity 13-3 is commonly referred to as a *filled pause*, or a *conversational filler*. Two main variants of these fillers are the nasalized one [ə(:)m], transcribed as ‘um’ or ‘ehm’ in text, and non-nasalized [ə(:)] transcribed ‘uh’ or ‘er(r)’. Fillers are considered lexically empty, which in turn makes them a good candidate for conveying multiple functions in spoken interactions. Traditionally, fillers were treated as imperfections, speaking manuals were suggesting their avoidance, and official transcripts, like those from NPR for the interview in Activity 13-3, do not include them.

The most salient function for fillers is to signal hesitation, thinking, planning what to say, or searching for a particular word. However, there is a much richer inventory of functions they serve. For example, in Activity 13-3, ‘um’ is not likely to signal hesitation or planning. It’s too short and also a bit higher in pitch for a filler with this function. I suggest you stop here and read the previous sentence again. Note how my sentence assumes a systematic link between the prosody of the filler and its function in communication. Do you agree with the statement in that sentence?

My analysis is that the speaker uses this ‘um’ again in a pivot function, signalling to focus the attention from the introductory remarks to the first real question of the interview, and thus help in organizing the discourse. This falls within a general function of fillers to enhance comprehension by focusing attention, and providing extra time that is useful for producing and perceiving unpredictable or critical information. Fillers have also been shown to signal boundaries of discourse units also in monologues, indicate speech errors and upcoming repairs, or signal whether the speaker knows the answer to a question. Fillers are usually flat in pitch but their relative height can signal attitude or confidence of the speaker.

In short, fillers and their prosody play multiple functions in spoken discourse despite their lack of lexical content. The functions fillers play are so unconscious that people are very good at filtering them out of speech and notice them only in cases when they are severely over-used, typically in semi-formal settings like public speaking or interviews. On the other hand, being too conscious about one’s speaking might increase their frequency.

13.4 Turn-Taking

Another prosodically conditioned family of speaking patterns is the rich system of *turn-taking management*. For example, both fillers and discourse markers like ‘um’ or ‘and’, especially when prolonged, followed by a silent pause, and with a flat pitch, may indicate the intention of the speaker to continue in her turn. Turn-organization is a complex interplay between how we say things on the one hand, and if we wish to continue talking, to yield the floor, or assume speaking on the other hand. It is amazing that in conversations people typically start and finish speaking with surprising smoothness and ease. This suggests we all have culturally dependent mental knowledge regarding who speaks when and the system

relating the speech prosody and this turn-taking management system. Additionally, the smoothness of turn-taking exemplifies a need for the coordination of mental models among interlocutors.

One of the important continuous features characterizing turn-taking is the temporal lag or *gap*, between the end of the turn by the current speaker and the onset of the turn from the upcoming speaker. It is positive if the next turn starts after the current one ended, and negative in case of *overlap*, or cross-talk, when the following turn starts before the current one ended. The gap values thus represent the continuous phonetic measure that participates in cuing communicative functions. Just like in other speaking patterns explored in this book, we can use Praat to extract and visualize the gap values in order to become more aware how they participate in our unconscious speaking behaviour.

In Activity 13-4 we engage in another hands-on exploration of turn-taking with Praat that adds to our repertoire of Praat skills - the *scripting* functionality. This book does not require any programming from you. This activity just gives a glimpse of this very powerful option in Praat that allows the user to label things of interest and then extract relevant values, and possibly then analyse the data statistically with R or other software, for which Johnson (2008) or Harrington (2010) provide useful guidance. Although I find scripting in Praat quite intuitive, trained coders might find Praat cumbersome or strange at times, but one of the reasons for its wide community of users is that even novices in programming might relatively easily create customized scripts.

Activity 13-4: Label gaps and extract values with a script (I)

We continue our exploration of the BC_Grinch interview. Open the sound file 'BC_Grinch.wav' linked in next Chapter 14 of the companion and the textgrid 'BC_Grinch_phrases.TextGrid', linked in this chapter. Click *View & Edit*. Our goal is to assess the turn-taking management by investigating the gaps, but we will not measure by hand. Rather, we label the gaps and then extract their durations with a simple script provided in the companion.

The first task is to label gaps. These are temporal intervals, so add another interval tier to BC_Grinch_phrases with *Tier* → *Add interval tier...* and name it 'turn-taking'. Since the first two tiers have starts and ends of each turn already labelled, you create intervals representing each gap by clicking at such turn-end boundary, creating a new boundary at the same time in the third tier by clicking the circle as in the first labelling in Activity 7-6, and repeating the same with the following turn-start. The script will then go through each interval of your new tier and extract the duration of those intervals consistently labelled as gaps.

In this activity you should ignore intervals labelled as <laugh>. But there are multiple decisions that you will have to make. First, some gaps are overlaps, and the value is negative, and some are switches with a silence and the gap value is positive. How to label so that a script could get this information? Then, there are two cases of 'mm' from speaker GN. You have to decide whether this is a regular turn exchange we are interested in and if yes, then how to label the gap. There is also an issue with BC's speech that overlaps GN's turn 'It's the

classic evil character the British accent, you know the whole thing’ both when it starts and when it finishes. What are the gaps here that correspond to our goal of investigating turn-organization and management? Support all your decisions verbally or in writing. Save the textgrid.

Finally, open the ‘BC_Grinch_turn-taking_SB.TextGrid’, click *View & Edit* with the sound and compare my labelling with yours. Discuss, disagree, evaluate,...

Labelling is essentially a first analysis and many times requires making informed decisions and assumptions that always affect the results. My decisions are outlined below. First, I decided to have both positive and negative gaps as intervals and label ‘+’ or ‘-’, respectively. Normally, gap values would be extracted directly from turn labels like in the first two tiers, but the script for that would be much more complicated.

Second, I decided not to include GN’s ‘mm’ as regular turns, and thus disregard them for the gap investigation. My reasoning is that they are not turns in which the speaker wished to take the floor, or contribute some content. Rather, she just expressed engagement and positive evaluation of what BC was saying. Of course, we might study if the temporal location of these ‘mm’ backchannel tokens within BC’s turn is systematic, which it likely is to some extent, but that would be a different question.

Third, the attempted turn by BC ‘Because that would...’ was unsuccessful, I labelled it as BI for butting-in (Beattie 1982) as in Fig. 13.1. But regarding the gaps I labelled it as a relatively long positive value since that was the intention of BC, he took a breath and planned to take the floor with this gap despite the fact that GN actually started the continuation of her turn earlier.

Finally, the extended overlap at the end of this turn, also in Fig. 13.1, I decided to label as a positive gap rather than a long negative overlap. This is because GN made a relatively salient silent pause after the low-pitched ‘you know’, which are both turn-ending signals. I took the addition of ‘the whole thing’ as an after-thought from GN not relevant for BC’s decision of when to start speaking. We will return to these issues below.

Now that we have the labels, we take the script that extracts all values and calculates mean gap duration. The script is very basic, and measuring by hand only 8 values might be faster than writing a script. However, in your projects you might need to extract tens or hundreds of values objectively, without the need to extract all over again manually if you want to change anything. Hence, scripting saves time and adds objectivity. Essentially everything we have done in Praat regarding value extraction (durations, formants, pitch, intensity, centre of gravity, etc.) and much more is scriptable. Praat includes its own tutorial for scripting and other online resources for scripting could also be consulted.

Activity 13-5: Label gaps and extract values with a script (II)

Now, open the script ‘extract_gaps.praat’ that is linked in the book companion: in Praat Objects *Praat* → *Open Praat script...* and navigate to the script location on your computer. The script expects a single TextGrid highlighted in the

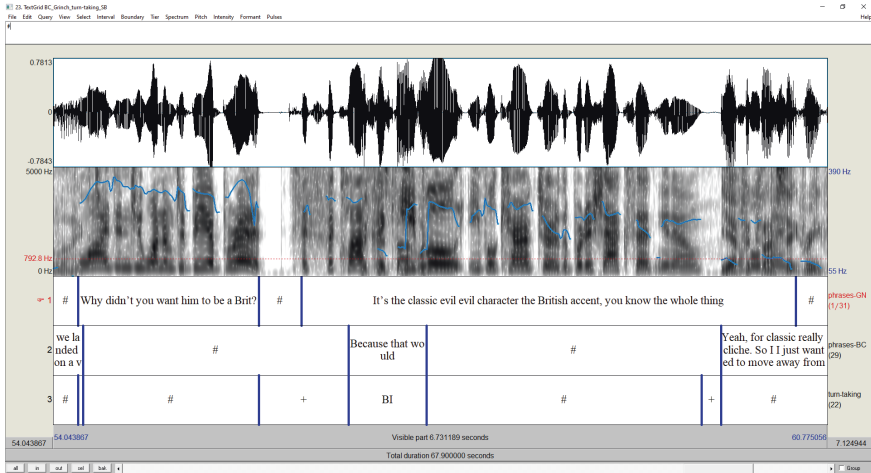


Fig. 13.1 Illustration of gap labelling for Activity 13-4 with the problematic overlaps discussed in the text

Praat objects and the gap information in the 3rd tier of that textgrid with intervals labelled as ‘+’ and ‘-’ as illustrated in Fig. 13.1.

The script loops through the intervals of the third tier, checks the labels, and if it finds a ‘+’ or ‘-’, it appends the gap count, total gap duration, and makes the interval value negative for intervals labelled with a minus. If you run the script with the textgrid selected (in the script window *Run* → *Run* or *Ctrl-R*), it outputs a list of relevant gap intervals with their labels and durations rounded to 3 decimal points (milliseconds) and also the mean gap duration into the Praat Info window. This is shown in Fig. 13.2 with the script on the right, selected textgrid in the Objects window, extracted values in the top middle Praat Info window and the soundfile with the textgrid in bottom left window. It is good practice to check several values manually if they correspond to what you expect the script to do.

In my labelling, six values were positive and two negative with a mean gap duration of 0.212 seconds. This corresponds very well to gap intervals observed in large corpora of conversational speech (e.g. Heldner and Edlund 2010). Research in psycholinguistics shows, however, that about 0.8–1.0 seconds is needed to formulate even the simplest speech responses (e.g. Levinson 2016). To reconcile these very robust observations (turn switches around 200 ms but planning speech needing 1000 ms), we have to conclude that *we normally start planning what we say much earlier than at the end of the interlocutor’s turn*. For example, relatively soon we predict the intent of the speaker and when s/he will probably finish speaking. Recall our discussion of all intricate habituated calculations of sensory information that underlies both catching a ball and speaking. Being able to smoothly switch turns in conversation is a similar type of learned behaviour that requires mutual coordination and active listening.

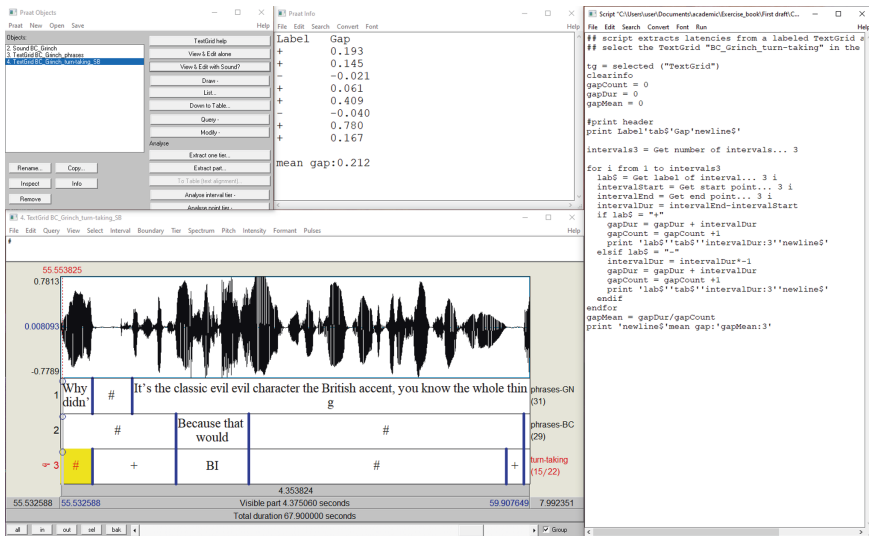


Fig. 13.2 The input and output of the script ‘extract_gaps.praat’; see Activity 13-5 for details

While gap durations provide information about the coordination between interlocutors, the cognitive system underlying smooth turn-taking involves the consideration of both syntactic and semantic completion as well as prosody. Regarding prosody, studies show that, for example, the duration and pitch of turn-initial fillers, the pitch range and declination in the first half of the utterance, or regions of low pitch and low intensity about 100–150 ms before the turn-end all provide useful prosodic cues to the listener regarding when the speaker is about to finish.

Let’s briefly return to the two final gaps in Fig. 13.1. Note how an unusually long gap from BC (0.780 seconds) results in an extended overlap of ‘because that would’ and his unsuccessful taking of the floor. In this light, consider the following, rather a quick initiation of the BC’s turn with a gap of only 0.167 seconds. A potential explanation is that the speaker is adjusting his timing of turn-initiation to avoid another overlap and problems in assuming the floor. These observations support constant monitoring, adjustments and negotiation of the interlocutors in terms of who speaks when.

13.5 Discourse Organization

In addition to fillers and discourse markers and their prosody, the hierarchical structure of discourse might be signalled with other prosodic characteristics. The most researched are the pitch range, silent pause duration, and variation in intensity and speaking rate. When you next attend a lecture or listen to a story, it might

be revealing to pay attention to these characteristics. Activity 13-6 shows how you could use Praat to test your intuitions or subjective observations.

Activity 13-6: Prosodic patterns in discourse structuring

In this activity we explore a roughly half a minute long monologue in which the radio host GN introduces her interview in ‘bake_off_intro.wav’ of the companion. First, listen to this monologue and characterize her stylized way of catching the attention of the listeners and building the motivation to listen further. What is the logical structure of information? In other words, which chunks belong together, and where can you identify junctures, or minor topic changes, in this information structure?

Here we want to investigate how the prosodic features of pitch, intensity, speech rate and pause duration participate in signalling the discourse organization of this monologue. I created the textgrid ‘bake_off_intro_bgs_words_sylls.TextGrid’ in which I labelled the monologue to units separated by in-breaths in the first tier and words within each such unit in the second tier. The syllable count for each word is in the third tier.

Open the script ‘extract_prosody_bake_off.praat’ located in the chapter’s companion from Praat (*Praat* → *Open Praat script...*) and run it (Ctrl-R). The initial form will ask for the path to the folder where you placed the sound and the textgrid. The script then extracts median pitch in semitones, which I take as a very crude proxy for pitch range here, intensity in dB, speech rate in syllables per second for each unit, and the duration of the silent pause preceding the unit. Feel free to inspect the script if interested.

If everything works, the script outputs the information in Table 13.1 into the Info window. Is there a systematic way in which these prosodic features are utilized by the speaker for signalling the discourse structure of the monologue and the relationship among individual units?

Literature suggests that those units that introduce new discourse (sub-) topics tend to follow longer pauses and are produced with an increased pitch range, or a higher initial pitch accent. Also, in units towards the end of the current topics the loudness tends to lower and speaking rate tends to slow down. I have to stress that the script in Activity 13-6 is meant mainly to illustrate a basic approach and Praat functionality in extracting quantitative values. A more sophisticated approach would also consider intonation phrases as units, pitch range calculated through minima and maxima, speech rate measures that consider also intrinsic characteristics of segments, etc.

How did the extracted values correspond to your perceived discourse structure of the monologue? The core observation is that the values show some tendencies in correlating with discourse structure, but it is certainly not the case that they provide a clean straightforward pattern. We have seen this before and it is good to remind ourselves that speech is complex, which makes it both challenging but also enticing to study.

Table 13.1 Prosodic features characterizing breath-delimited units extracted with Praat. See text and Activity 13-6 for details

Unit	Pitch (st)	Intensity (dB)	Rate (syl/s)	Prec Pause (s)	Words
1	12.1	63.9	4.15		There's a big birthday in Britain today
2	11.4	62.2	4.50	0.231	Queen Elizabeth turns 90. And like many of us do on our birthdays, she'll be celebrating with a bit of cake
3	10.9	61.2	4.28	0.420	This year, the task of coming up with that cake fit for a queen falls to Nadiya Hussain
4	9.3	64.5	5.88	0.385	She was the winner of the most recent season of the wildly popular TV show
5	9.0	55.4	4.92	0.249	"The Great British Bake Off"
6	10.1	62.3	5.83	0.319	Her story captivated audiences in the U.K
7	7.4	61.8	6.39	0.282	She's the daughter of immigrants from Bangladesh
8	12.1	57.5	4.66	0.310	and she personally delivered her orange drizzle birthday cake with orange curd and butter cream and purple frosting
9	10.4	58.6	5.35	0.380	to the queen this morning
10	13.0	61.3	3.74	0.423	We reached her in Milton Keynes

Inspecting the values, some unit boundaries are very clearly prosodically marked. For example, unit #6 has greater pitch range, intensity, and is faster than the previous unit, and also follows a salient silent pause. Similarly, unit #10 is prosodically separated from the previous material, although speech rate is slower. Looking at the transcript in the last column, both #6 and #10 can be reasonably analysed as representing a new segment in the logical development of information in the monologue, and these thus corroborate the findings in the literature. We also see some partial patterns. For example, #7 is prosodically linked with the preceding unit, which corresponds to discourse structure, but #8 is less so. Or that units #1–5 gradually decrease in pitch range, which follows our hypothesis, but the other three features do not offer such a clear trend.

The activity and the subsequent discussion of the data show an approach in which the hypotheses regarding the intentions and discourse meanings can be tested, or investigated, using phonetic data labelled and extracted with Praat.

13.6 Speech Entrainment

Consider how in body language we might unconsciously signal our positive feelings towards other people by leaning towards them, and negative ones by leaning away from them. Spoken behaviour shows a similar potential. *Speech*

entrainment, also called alignment, accommodation or others terms, is the tendency of conversational partners to become similar to each other along various aspects of speaking. In addition to the lexical or syntactic dimensions where identical words or constructions are re-used (e.g. ‘I’m a big fan of’ in the Grinch interview), here we are mostly interested in the prosodic characteristics of speaking. An interesting hypothesis is that speech entrainment is similar to ‘leaning’ and we make our speaking more similar to those interlocutors about whom we feel positively.

Speech entrainment is a ubiquitous speaking habit operating beyond single intonational phrases. It is commonly found with crude phonetic features extractable directly from the acoustic signal. For example, interviewees converged towards their interviewers in mean pitch, intensity, speech rate or gap durations. Additionally, more complex and derived features show entrainment. For example, entrainment in ToBI labels, rhythm or backchannel inviting cues – that is the prosodic pattern of eliciting mhm-type feedback response – have also been observed in multiple studies and corpora. All of us also adjust various regional and identity-based features of speech based on our interlocutor as we talk differently to our friends, family members or co-workers. Politicians or TV hosts have been shown to adjust their accents to their audience or interviewees. Hence, the degree of speech entrainment can thus provide a mean for managing social distance among interlocutors and assessing strategic aspects of inter-personal communication.

We have already seen several examples of behaviour that can be studied through entrainment. Recall the smooth turn-taking in the opening sequence, or BC’s adjustments in gaps in Fig. 13.1, of the BC_Grinch interview. Also the type of features extracted with Praat in Table 13.1 can be used for assessing entrainment between dialogue partners. Activity 13-7 expands this pool with an example of a perception experiment using manipulated speech prepared in Praat.

Activity 13-7: Muffled speech

This activity asks you to run a small perception experiment with your friends or relatives. I prepared four excerpts in the companion. Two of them include muffled speech of about 12 seconds long and comparable intensity (‘GN_in_bakeoff_360_70.wav’ and ‘GN_in_grinch_360_70.wav’). Another two include just the initial portions of 3.5 seconds extracted from the first two excerpts (‘GN_in_bakeoff_360_70_short.wav’ and ‘GN_in_grinch_360_70_short.wav’) and these have also comparable mean pitch values. I describe in the text below the activity how the muffled speech was created.

Your task is to elicit subjective assessments of speaking behaviour in two comparable files. Hence, pick one of these pairs, either two short sounds or two long sounds, and play them to your subjects. Ask them to describe how the speaker comes across. I admit, this is not a very well-defined task for your subjects, and they will probably comment on how strange the sound files seem, but stick with it. Note the answers, and ask how your subjects would characterize the difference(s) between the two excerpts, elicit their observations.

Is there a systematic pattern in how you and your subjects describe the differences between the two files?

You might have guessed from listening to previous sound files (and their file names), but your subjects will not know, that both excerpts come from the speech of a single person - interviewer GN - from the two interviews we have analysed in this book before. The muffled speech, technically low-pass filtered, is used to mask the lexical content, i.e. what the speaker is saying, but preserve the prosodic features such as pitch or rhythm.

I prepared the files in Activity 13-7 in the following way. I first concatenated speech from GN from about a minute of speech starting the dialogue (the full audio files for the Grinch and Bakeoff interviews are linked in Chapter 14). I manually excluded all cross-talk and then low-pass filtered (with sound selected *Filter* → *Filter (pass Hann band...)* and used 0–360 Hz and smoothing left at 100 Hz). This removes higher (formant) frequencies muffling the identity of segments but preserves many prosodic characteristics. Since low-pass filtering lowers intensity, I then unified the loudness of the two files with *Modify* → *Scale intensity...* and kept 70 dB, which makes mean intensity of the entire file 70 dB.

The responses from your subjects probably vary wildly for this strange, and not very clearly defined task. However, I believe that some differences between the two files will be observed and commented on. Since both files come from the same speaker, and have comparable duration, mean pitch, and mean intensity, the hypothesis is that the observed differences come from the fact that GN talks to different interviewees, and adjusts her speaking habits slightly.

For example, examine the opening sequence in the two interviews:

GN: Good morning

NH: Good morning

GN: So how did this happen how did this come about

GN: Welcome to the program

BC: Thank you for having me I'm a big fan of the program

GN: And I'm a big fan of yours and I'm a big fan of the Grinch (...)

In Bakeoff, GN starts with a cautious 'Good morning' but after NH's spirited and enthusiastic 'Good morning' with an increased pitch range, GN's subsequent turn uses also quite varied pitch. One phonetic feature that captures this variability is the standard deviation of the pitch, which is 4.8 semitones in the short low-pass filtered file. To get this value I extracted the pitch from the sound file (*To Pitch...* 85 and 500 Hz for floor and ceiling) and asked for standard deviation (*Query* → *Get standard deviation...* and specified *semitones*). My impression from the Grinch interview is slightly different. GN starts with a very enthusiastic greeting but BC's pitch range is lower, also because he's a male, and the subsequent GN's turn is flatter than in the Bakeoff; this time standard deviation is around 3 semitones. Hence, it is reasonable to hypothesize that GN has a tendency to entrain her pitch variability to that of her interviewees.

You can inspect the entire interviews, consider also higher pitch register in certain turns used by both BC and GN in the Grinch interview. There are multiple ways how we can study entrainment. We can compare the same person when speaking with other interlocutors, observe gradual changes when the interlocutors might become more similar, for example by studying convergence with comparing the first and the second half of the interview, or consider the development of speakers' similarity in the vicinity of their turn-exchanges.

Since by now you are well aware of the complexity in the speaking behaviour, you know better than to expect clean patterns of entrainment in any dialogue. Studies also observe prevalent dis-entrainment of speakers and the potential interaction of speech entrainment with many higher-level aspects of dialogues like dominance, or dialogue intentions. There is also a debate whether entrainment is used sub-consciously, yet under some control of the speaker, for negotiating social distance (e.g. Communication Accommodation Theory, Giles et al. 1991), or occurs more or less automatically as priming of the speaking patterns that are activated in the listener from the speaker's speech (e.g. Pickering and Garrod 2004, 2013) and might be linked to strong biological relationship between the perception and production.

Unquestionably though, speech entrainment is a powerful sub-conscious speaking habit that, along with all other habits examined in the book, might provide insights into the social cognition of humans by examining the phonetically measurable and quantifiable aspects of speaking.

Exercises

- 13-1 Compare the monologue of host GN in the Bakeoff interview with her speech in dialogues with both NH in Bakeoff and BC in Grinch. Consider, for example how boundary signals and speech rate might participate in these two styles of speaking.

References

- Beatie, Geoffrey. 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica* 39 (1–2): 93–114.
- Giles, Howard, Justine Coupland, and Nicolas Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, ed. Howard Giles, Justine Coupland, and Nicolas Coupland, 1–68. Cambridge: Cambridge University Press.
- Harrington, Jonathan. 2010. *Phonetic analysis of speech corpora*. Malden, MA: Wiley-Blackwell.
- Heldner, Mattias, and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 30 (4): 555–568.
- Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell.

-
- Levinson, Steven C. 2016. Turn-taking in human communication, origins, and implications for language processing. *Topics in Cognitive Science* 20: 6–14.
- Pickering, Martin, and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27 (2): 169–226.
- Pickering, Martin, and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36 (4): 329–347.



Exemplifying the Book Material in Real Interviews

14

In this chapter, we will

- engage in a comprehensive analysis of three roughly half-a-minute excerpts from the radio interviews
- exemplify various speaking habits covered in the preceding chapters in situated dialogues and further contextualize your understanding and awareness of these habits

14.1 Introduction

The concluding chapter of the book brings all the concepts and notions together with commentaries to the transcription of several short excerpts of dialogues. The excerpts come from sections of three interviews aired in NPR and described and acknowledged in Chapter 1. We will call them *Bake_off*, *Grinch* and *Jeeves_Wooster*. The first one is further split to three sound files (*Bake_off_intro*, *Bake_off_dialogue*, *Bake_off_conclusion*). All five sound files have the associated Praat Textgrid files with aligned phrases and words to the signal.

In two roughly one-minute long files (*Bake_off_dialogue*, *Grinch*) the texgrids include additionally full IPA and ToBI annotations, and in *Jeeves_and_Wooster* the first half-minute is annotated this way. Finally, the text in three sections below provides my commentary to approximately half-a minute of each of the three files (*Bake_off_dialogue*, *Grinch*, *Jeeves_and_Wooster*).

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-54349-5_14) contains supplementary material, which is available to authorized users.

One way of approaching this material is to treat it as a general exercise in which you are asked to engage with speech through Praat, attempt to analyse what speakers are doing and why they might be doing it using the understanding gained in the book chapters. Hence, you might proceed by chunking the speech, making a basic prosody analysis, possibly ToBI annotation, then IPA annotation of individual words in a chunk, and considering the connected speech aspects. I would encourage you to do this first and on your own without consulting neither the textgrids nor the commentary. I suggest you start with the first half-minutes of Grinch, Bake_off_dialogue and Jeeves_and_Wooster.

The entering of IPA symbols into Praat textgrid might be cumbersome at first, but with the help of three tables for consonants, vowels and diacritics provided in the help menu (Help → Phonetic symbols) the learning curve is quite fast. You can alternatively copy/paste from other text editor, or transcribe with IPA on the paper. In any case, when visually inspecting, zoom in sufficiently to see the complete IPA transcription of each interval. Finally, I am attempting a transcription that is narrower than the broad transcription, indicating allophones and connected speech processes discussed in the book. I also try to be consistent in the level of phonetic detail I use in the IPA transcription, but inevitably some aspects might be omitted.

Once you attempt to analyse on your own, consult with my annotations and commentary. It is absolutely essential to listen to the units, separate words or any other relevant chunks in Praat while looking at the ToBI and IPA annotation provided in the textgrid. The textgrids with ‘full_ToBI_IPA_final’ appended include 12 tiers: each speaker has phrases, words, IPA, and ToBI-tones, ToBI-breaks and ToBI-misc annotated with the last one containing occasional alternative ToBI labelling.

These **should not** be taken as an ‘answer key’ to the exercise of analysing speech. In the spirit of the descriptive rather than prescriptive approach mentioned in Chapter 1, my commentaries and annotations should be taken as opening a dialogue, in which you assess, grapple with, extend or disagree with what I am saying or what my annotations indicate. The important thing is that you understand why I am saying it and what my annotations represent, and that you can reasonably argue for alternative or additional analyses or observations that you make.

I comment only on a subset of issues that are in some aspects relevant given the issues we covered in the book and thus the commentary is by no means complete.

Once you have done the analysis yourself and engaged in comparing yours and mine analyses of these half-minutes, I suggest you try the remainder of Bake_off_dialogue and Grinch. You can consult with my IPA and ToBI but this time without any commentary. Finally, the remaining excerpts (the rest of Jeeves_and_Wooster, Bake_off_intro, and Bake_off_conclusion) are only provided with the word alignment and the rest of the analysis is entirely for you to do. For further material, context and background, you can listen to the entire interview and/or read the transcripts here:

- Bake_off: <https://www.npr.org/sections/thesalt/2016/04/21/475007057/british-bake-off-winner-takes-on-the-toughest-judge-of-all-the-queen>

- Grinch: <https://www.npr.org/2018/11/11/665873897/its-lovely-being-mean-benedict-cumberbatch-gets-into-character-as-the-grinch>
- Jeeves_Wooster: <https://www.npr.org/2018/12/02/672391198/jeeves-and-wooster-but-make-it-a-modern-spy-novel>

14.2 The Grinch

GN: Welcome to the program

In this opening turn consisting of a single Intonational Phrase (IP), the host GN uses increased pitch range in a regular ‘hat pattern’ intonation accenting both content words and a low boundary (H*H*L-L%). Both function words are weak forms. The first syllable of ‘program’ is, interestingly, also somewhat weakened despite the presence of a salient pitch accent. This weakening is realized as the monophthongization; in the enunciated dictionary pronunciation it has a diphthong [ˈpɹoʊ.ɡræm] but here it is pronounced with a monophthong [o]. This is supported by steady formant trajectories during the vowel. Also, there is not much devoicing of [ɹ] that would be normally expected in the stressed syllable. Since this is the first speech we analyse here from this speaker, we can keep these observations as tentative to compare with some other relevant environment later to see if they should be treated as general patterns of the speaker or if they reflect some contextual influence.

BC: Thank you for having me I’m a big fan of the program

This turn has two units. The phrase accent (L-) explains why intonation goes down after ‘having’ despite very weak pre-boundary lengthening. Note that analysing this as a single phrase would not explain this pitch lowering. After H* on ‘thank’, the two L+H* might be alternatively analysed as H* since the ‘scoop’ in pitch is not very salient. I opted for L+H* since I perceive a clear pitch rise during, and beyond, the accented syllable and the intention of the speaker to add extra enthusiasm, and maybe surprise. When I produced this with ‘regular’ H* accents, my perception was different from the one I had from the speaker’s utterance.

Several aspects of connected speech take place. First, ‘you for’ link is rather strange. If you listen to them separately, you would hardly recognize these words. I am not certain but some form of overlap of the initial [j] and [f] occurs, which might be the reason for clear devoicing of [j] and voicing of /f/ as [v]. Second, there is a nasal place assimilation /ŋ/ → [m] in ‘having’ and linking [j] connecting ‘me I’m’. Finally, ‘of the’ in the second unit are both also extremely weakened. Note also the diphthong realization of ‘program’, which is different from GN’s realization before.

GN: And I’m a big fan of yours and I’m a big fan of the Grinch which is of course a Christmas classic

This turn is divided into four IPs. The first one, ‘And I’m a big fan of yours’ has a very clear L+H*L-L% phrase-final intonation with L+H* expressing the contrast

between being a fan of the program (BC) and a friend of BC (GN). As you can see from the textgrid, I am much less sure about other pitch accents in this phrase. Especially deciding whether ‘big’, ‘fan’ or none are accented is very difficult. On the one hand, they are quite high and one should account for high pitch targets through ToBI labelling since the next accent on ‘yours’ is L+H*. On the other hand, metrically (with the help of clapping), there would be too many accents, and I perceive either none or one on ‘big’ only.

In connected speech, the elision of /d/ in ‘and’ despite the full vowel, and final devoicing of /z/ in ‘yours’ can be observed. For non-native speakers, the word ‘big’ by both speakers might be instructive since the initial /b/ is quite saliently devoiced, the quality of the vowel is the lax [ɪ], and the final [g] is, at least partially, voiced. All three habits go against common realizations in many languages whose L2 speakers might have a tendency not to say [bɪg] but something like [bɪk].

Note also the release of the phrase-final word, which I transcribed here as ‘uh-’ and [əʔ] in IPA. This, I think, is not a broken filler. Both the audible, and visible, release, and the subsequent glottal stop, corroborated by the vertical striations at the onset of the following ‘and’, are strategies for producing a strong disjuncture. It might be revealing since I assume very few of you actually ‘hear’ the release and the glottal stop. Normally I would include the release with the interval of ‘yours’ but wanted to illustrate this point here.

The following phrase consists of only strong form ‘and’ with a plateau intonation contour (H*H-L%), strong boundary signals in terms of pre-boundary lengthening, the release of /d/, and the subsequent glottal stop. The communicative functions of this prosody with this discourse marker were discussed in Sect. 13.2.

The phrase ‘I’m a big fan of the Grinch’ has a downstepped pattern of three pitch accents followed by a high boundary tone. The high boundary conveys continuation and the speaker orients the listener to interpret this phrase as linked to the following one. The other segmental and suprasegmental aspects are as expected.

The next intonational phrase is ‘which is of course a Christmas classic’. I perceive an intermediate boundary following ‘Christmas’, and a somewhat weaker disjuncture, still labelled as a #1 break index, after ‘is’. This makes the labelling of pitch targets problematic since without this break, the pitch accent on ‘classic’ would clearly be H+!H*. But with the high pitch target and the plateau on ‘of course a Christmas’ already assumed with the phrase accent H-, I opted for L* on ‘classic’. This is not without problems, since the pitch is relatively high, but I decided on L* due to a rather salient rise to the final H-H% boundary.

I found the realization of the stressed vowel in ‘classic’ interesting. I would expect the American pronunciation of [æ], also used in ‘class’, but GN used a retracted [ɑ], which appears in the British pronunciation of ‘class’. I can only speculate that GN might in this way be entraining, or accommodating, to her British guest. Note also salient aspiration and devoicing of the segments following [k] in ‘course’ and ‘classic’, respectively.

BC: Yes

The rising intonation of the preceding turn-end and the low fall of this one was discussed in Sect. 13.2. Note the creaky voice associated with low pitch targets, and the negative gap, which corroborates the analysis from 13.2 that ‘yes’ is not an answer (for which BC might require some time) but rather just affirming the obvious.

GN: And like “A Christmas Carol” the main character is not in the Christmas spirit.

ToBI labelling of the phrasing is quite challenging. I decided for the intonational phrase boundary following ‘character’ due to significant pre-boundary lengthening of the final syllable and the additional intermediate boundaries as seen in the textgrid due to minor lengthening and glottalizations. The H* and !H* pitch accents are fairly clear.

Noteworthy aspects of connected speech include the expected nasal place assimilation (/n/→ [ŋ]) in ‘main character’ but I fail to hear clearly how the speaker produced this and would lean to not annotating this boundary with such assimilation. The juncture ‘in the’ shows the combined regressive and progressive assimilations discussed in Chapter 10.

GN: Um why did you wanna play the Grinch

The discourse function of the first filler, separated from the rest with a #4 boundary was discussed in Chapter 13. Phonetically, although I transcribed as [əm], the final sound might actually be [w], which would be a special case of assimilation. The #4 break after ‘play’ and labelling of the tonal targets posed no major challenges in this utterance. Note the decreased pitch range in the final IP. In connected speech, there is a coalescence in ‘did you’, which makes [dɪdʒu] difficult to divide into separate words. Note also the likely th-stopping in ‘the Grinch’, which can be attributed to the initial boundary strengthening.

BC: I think it’s a time-old classic and I think it was time as well for a for maybe a reimagining of it and

Both chunking and determining pitch targets is quite a challenge here. The speaker uses a relatively compressed pitch register, hence deciding between H and L targets is difficult, and his chunking reflects his thinking process more than the syntactic structure of his sentences. Hence, while H targets might be surprising here, I concluded that his L in the bottom of his register is around 75 Hz and I thus made a decision to consistently label targets higher than 80 Hz as H targets barring other evidence for the opposite.

In ‘I think’ the reduction is so extreme that ‘I’ is a creaky [ə] and ‘think’ might be missing the nasal if you listen to it separately. Compare this with the following full realization of ‘I think’ in the next phrase. The reduction is also surprising given the H* accent. In ‘old classic’ the weak unaccented realization of ‘old’ results in monophthongization and a very clear case of d-elision. The quality of the stressed vowel in ‘classic’ was also a challenge and it is less clear than the GN’s realization discussed above.

In ‘time as well for a for maybe a reimagining of it’, there is a prototypical example of continuation rise with a L-H% boundary. Note also two cases of linking: linking-r ‘for a’ and its absence in the corrected ‘for maybe’, and also a linking-j in ‘maybe a’.

BC: I think it’s such an iconic American role

This utterance is divided into two intonational phrases with the break following ‘American’ supported by very salient lengthening of the pre-boundary [n] with a plateau intonation. You see that in the onset of the first phrase BC uses very breathy voice. I would speculate that he does it intentionally to convey the earnestness of his proposition. I wasn’t sure whether the accent is on ‘I’ or ‘think’ rhythmically, but because of the absence of f_0 on ‘I’ opted for ‘think’. Also the accent on ‘role’ was not clear due to the overlap. Segmentally, note the almost absent closure for the affricate [tʃ] in ‘such’.

GN: mm

Here there are two interesting aspects. The first is the communicative function. Note how ‘mm’ could mean ‘tasty’ (in fact used this way in the Bakeoff interview), express hesitation or affirm, and positively evaluate, the previous proposition. Both context and the prosody of ‘mm’ play a role in disambiguating among these functions and I would analyse this as the third option above. The second interesting aspect is the temporal placement of this phrase within the interlocutor’s speech. We see that ‘mm’ is fully overlapping with BC’s speech but smoothly aligns with the strong intonational boundary following ‘American’. In our discussion of turn-taking, the prosodic cue of pre-boundary lengthening might be a ‘go ahead’ signal for the speaker to deploy her semantically empty discourse marker intended to convey positive engagement and evaluation despite the incompleteness of her interlocutor’s intended message.

BC: and something I was surprised to be asked to do and that’s uh that’s a good thing in an actor’s life when you’re surprised by an offer

Let’s take the filler ‘uh’ first in the middle of this utterance. Its function is to convey thinking or planning and phonetic features of its duration, flat pitch, and the following silent pause all support this communicative meaning. In the speech before the filler, the intonation is quite unproblematic. Note multiple examples of linking adjacent words with unreleased word-final [tʰ] in ‘surprised to’, linking-j in ‘be asked’, linking-w in ‘do and’ already discussed in Chapter 10, or nasal place assimilation in ‘and that’s’.

There are also two interesting segmental aspects. First, note the realization of ‘something’ and how you might perceive a [p] instead, or in addition to, the intended /m/. This arises from glitches in fine coordination between the laryngeal and supralaryngeal articulation in which vocal cord opening for [θ] starts a bit earlier, while the lips are still closed for [m] resulting in the perception of a voiceless bilabial closure. Second, the realization of ‘asked to’ seems to be simplified by eliding [k]. This is possibly due to alleviating articulatory effort to form the

alveolar-velar-alveolar sequence of constrictions and make just the alveolar one. Another strategy for simplifying this sequence would be to metathesize, that is, to reverse the order of consonants and say [ɑ:kst]. This also reduces the number of changes in the place of articulation, and thus articulatory effort, and is a common pattern, for example in African American vernacular English.

In the speech following the filler, there are challenges in ToBI labelling. In addition to the boundary following ‘life’, I perceive minor breaks after ‘thing’ and ‘surprised’. They could have been labelled as #3 breaks with H- phrasal tones, which would then require L* pitch accents on ‘life’ and ‘offer’ as indicated in alternative labelling since each intermediate phrase has to have an accent. I decided to go with a simpler analysis and the primary reason was that the accents of ‘thing’ and ‘surprised’ are so salient that the material following them feels de-accented.

Consider also the regressive assimilation of place in ‘good thing’ and also the [æksəz] realization of ‘actor’s’, which I would analyse as a speech error.

BC: And then they said we love your accent

This was produced as a single intonational phrase with a very prominent L+H* accent on ‘accent’, which previews the contrast between the Grinch being American and BC having a British accent. Note two cases of nasal place assimilation with the first one following the elision of /d/ of ‘and’.

14.3 Bakeoff_Dialogue

GN: Good morning

NH: Good morning

GN: So tell me how did this come about how did this happen

The prosody of this initial sequence was discussed in detail in previous chapters. In connected speech, NH’s juncture in ‘good morning’ might have some degree of assimilation. Although I hear a much less clear [d] than in the GN’s greeting, I do not hear a clear [b]. In the last turn, there are multiple reductions, particularly diphthongs in ‘so’ and the second ‘how’ are clipped, or truncated, such that only a monophthong remains, word-final /z/ in ‘this’ is devoiced in voicing assimilation.

NH: Uh the queen’s cake or just life in general

There were multiple issues with transcribing this turn. In ToBI, I was considering a #3 break after the initial filler but decided the lengthening is not sufficiently salient for that. In addition to t-deletion in ‘just life’ and some indication for th-stopping in ‘the’, the great issue was the realization of ‘general’; especially the number of syllables. Since I do not visually see clear consonantal releases for schwas but still perceive it as a three-syllable word, I made both [ɹ] and [ɪ] syllabic.

GN: Let’s start with the queen’s cake

Intonationally, this phrase presents a clear example of a downstepping contour. This is possible in part by having the concepts of queen's cake fully in the common ground of both speakers and thus no need to foreground it with H* accents.

I found the realization of 'cake' interesting. Despite transcribing in IPA with aspiration, note how voice-onset time is relatively short, about 30 ms. This contrasts with the aspirations of about 60 ms in the GN's opening monologue, and even more so with her very first mention of this word in 'Bake_off_intro' in an extremely hyper-articulated fashion with VOT of about 150 ms. I would speculate that the degree of GN's aspiration in 'cake' in this turn might be an indication of her entrainment to her guest who did not aspirate her 'cake' a lot (about 35 ms.). Otherwise, why would she not keep her regular aspiration of about 60 ms from her two previous mentions of this word also in this highlighted realization due to the pitch accent and lengthening in the pre-boundary position?

Another noteworthy aspect is the t-deletion in 'start'. From Chapter 7 we know that this environment, a /t/ not preceded by a consonant and followed by an approximant, is in general not favouring deletion; yet it very clearly takes place here.

NH: Well oh my goodness so um a couple of weeks ago I just got a phone call to say

The response starts with three intonational phrases with discourse markers. Both 'well' and 'so um' are clearly chunked as separate units and all three chunks provide a pivot from the previous questions but all three with slightly different meanings. 'Well' might primarily acknowledge the question and signal she will respond. 'Oh my goodness' expresses her emotional stance, and 'so um' signals stalling for time. Note how 'well' and 'oh my goodness' have the same intonation pattern of H*L-L% but 'so um' is flat. Try for yourself alternating and use H*L-L% for 'so um' and flat H*H-L% for 'well' and 'oh my goodness'. You realize how the intonation is critical in cuing these intentions of the speaker.

The rest of the utterance is intonationally quite straightforward. In connected speech there are nice examples of unreleased [t'] in 'just' and a complete replacement of the alveolar closure for /t/ by glottalization in 'got'. The realization of 'phone call' sounds somewhat unusual but I am not confident to pinpoint the reason. It seems that /l/ in 'call' is distinctly the clear allophone of /l/, which might relate to her dialect, and /n/ of 'phone' might be affected through some form of anticipatory realization similar to that /l/.

NH: Hey would you like to do this and my initial reaction was well no

You see in the ToBI annotation that some decisions in the first unit were difficult to make and both alternatives present a plausible analysis. I point your attention first to the low boundary after 'this'. Here intonation corroborates the discourse meaning of this reported speech with no justification for rising question intonation. I also found the clear strong releases of word-final consonants in 'this' and 'and' quite interesting. Strong articulatory releases might serve as markers of

intonational boundaries, which we have not mentioned in the previous chapters. Both of the releases are realized with noticeable glottalization further strengthening the saliency of the disjuncture.

In the second half of the utterance, the IPA transcription of the pre-boundary realization of ‘was’ presents a challenge. I opted for the full vowel but the continuity of all patterns we discussed is exemplified nicely here again since it is neither a clear schwa nor the vowel [ɒ].

GN: Really

Unfortunately, this phrase overlaps completely with NH’s speech and thus a phonetic analysis is not really possible. We can only rely on the perception. Cross-talks and overlaps are much better studied if both speakers wear separate microphones, which is the case in many spoken dialogue corpora but not in the recordings available from the NPR. Despite the absence of a clear f_0 contour to inspect visually, the intention of the speaker was reconstructed in my ToBI annotation.

NH: Because I was so afraid of getting it horribly wrong

The IPA transcription and the location of word boundaries in ‘because I was’ is only approximate due to the overlap but it is clear that all three words are radically reduced and weakened. The smooth linking of words in this phrase involves various strategies discussed in the chapters from almost completely elided tongue-tip closure for /d/ in ‘afraid’, the glottalization of ‘it’, but also a bit non-standard velar release in ‘wrong’. We can also nicely see natural declination of f_0 , for example in how the final H* on ‘wrong’ is about 30 Hz lower than the H* of ‘so’.

NH: um and then it dawned on me I can’t say no to the queen so once I got over the nerves I thought well actually this is d- such an honour how can I possibly say no

In this almost seven seconds of speech I did not identify an in-breath so we analyse it as a single utterance. The division into intonational phrases is relatively straightforward with the exception of the minor speech error following ‘d-’ On the one hand, there are neither temporal nor clear tonal markers signalling a boundary. On the other hand, there is a brief silent pause and also the presence of pitch reset, both supporting the presence of a boundary, which I indicate with the BI #3.

Regarding ToBI tonal targets, I would like to comment on two boundary tones. First, consider the L-H% following ‘nerves’. Strictly speaking, the pitch contour shows a clear fall to the L- phrasal tone but the subsequent rise for H% cannot be seen in the contour. My assumption, open for discussion, is that this H% could not be fully realized due to the partial devoicing of the final fricative, but that the intention of the speaker was to produce a continuation rise (L-H% in ToBI). So I am taking a clearly ‘phonological’ position here trying to capture the pragmatic meaning whereas those who wish to strictly describe the pitch contour phonetically, might prefer L-L%. As I said in the previous chapter, both ToBI and IPA are

trying to find a balance in capturing both phonetic and phonological aspects of the signal, but sometime one has to pick a side. Crucially, the most important feature of labelling is the consistency so that the systematic relationship between prosody and meaning could be effectively studied based on the annotations.

The second comment I want to make concerns the final boundary following ‘possibly say no’. By deciding not to label a pitch accent on ‘no’, of which I am not sure as indicated with the *? label, we would expect the pitch contour to gradually fall between the H* on ‘possibly’ and the low boundary target aligned at the end of the phrase. However, we see a clear, and intended, deviation from this since pitch falls sharply right after the accented syllable of ‘possibly’ and stays low during the last three syllables of the phrase. This is a rather frequent pattern for H*L-L% (and also H*L-H%) contours in phrases in which the pitch accented syllable is separated from the right edge of the phrase by more than one syllable. The analysis within the ToBI framework captures this by saying that the phrase accent L- aligns directly following the final pitch accent while the boundary tone aligns with the edge of the phrase. Independent evidence for this phrase accent alignment analysis was provided from English and other languages. Alternative treatments suggest that the pitch accent is actually a fall, e.g. H*L, which effectively dispenses with the need for phrase accents that might be redundant in many situations. Irrespective of the theoretical treatment, the phonetic and phonological aspects of the alignment of pitch targets with the segmental material is an exciting research area to be aware of for anybody who engages with labelling intonation.

In the connected speech aspects, besides the already discussed patterns in t/d-deletion and glottalizations captured in the labelling, the juncture ‘on me I’ provides interesting material. First, there is no nasal place assimilation, which suggests a somewhat hyper-articulated production. In the next juncture ‘me I’ we see and hear an indication for linking-j, which would suggest a rather weak break #1. However, we also see glottalization at the beginning of ‘I’ indicating a stronger break. This nicely illustrates the interplay and interactions among many local speaking habits when they are employed for marking intentions in larger prosodic chunks.

GN: Well how did you even decide what kind of cake to bake for the queen

The chunking of this utterance into units does not present major issues. I analysed the contour with downstepped pitch accents, which is unproblematic in the second unit, but the last accents of the third unit on ‘bake’ and ‘queen’ fall apparently quite low and one might consider the alternative of L* accents. However there are three arguments against this analysis. First, the pitch range of GN is a bit lower than NH and her low reaches close to 120 Hz; for example in the low boundary of ‘Let’s start with the queen’s cake’ before. Second, multiple downstepped accents, together with f_0 declination, result in narrowing the pitch range so that even H-targets appear to be relatively low for the speaker. Finally, all content words ‘cake’, ‘bake’ and ‘queen’ have the same discourse status as fully active and given in the common ground and thus there is no specific reason to have a high accent on ‘cake’ and a low one on the other two words.

14.4 Jeeves and Wooster

GN: And I wanna talk about the language

Intonationally, we see again the decreased pitch range, and consequently, the absence of the fall between !H* and L-L%. However, the low boundary tone is supported by strong glottalization in the last syllable of ‘language’; recall that creak is commonly associated with low pitch targets. After the previous two excerpts, in which we started from the beginning of the interview, here we start roughly after two minutes. Despite not knowing the preceding context of the interview, we can safely deduce that GN has talked about something else before, and that ‘language’ was most likely mentioned, or understood in the context, since it has !H* and not H*. Therefore, the most important element marked with clear prosodic highlighting is the ‘wanting’ to talk about the language.

The juncture ‘about the’ is worth mentioning for two reasons. One is the reduction in the stressed second syllable of ‘about’ resulting in a clear monophthong. The other is the realization of ‘the’ that corresponds to th-stopping, very likely due to the progressive assimilation from the word-final /t/ of ‘about’.

GN: In your book you use words like something is described as dreadfully spoony people say hello with a what-ho crumpets uh that’s not how we speak today obviously

I decided to lump together three units delimited by two in-breaths, particularly to highlight how prosodic chunking might in natural spontaneous speech differ from divisions we would expect based on just the text or in read speech. For example, ‘described # as dreadfully spoony’ seems more natural than the realized ‘described as # dreadfully spoony’. Also, an in-breath separating the indefinite article from the following word in ‘with a # what-ho crumpets’ is not expected. All #3 and #4 boundaries, however, have clear phonetic markers in pre-boundary lengthening, glottalization or silent pauses. The speaker’s prosodic chunking thus clearly reflects her speech planning indicating her search for these unpredictable expressions and simultaneously orienting the listener’s attention to them. In pitch targets, an important issue was to differentiate the plateaus H-L% with a forward-looking expectation of the continuation from the low targets L-L%.

There are regular realizations of weak forms and linking between adjacent words with one challenging issue of word-final /t/ realizations in ‘that’s not how’. I transcribed with an elision of the first one and flapping of the second one. I am not very confident regarding the latter one especially since flapping occurs most commonly between vowels and here /h/ is following. But the closure is extremely brief and voiced.

I would also invite you to inspect and measure the aspirations of stops in this utterance. For example comparing /p/ in ‘people’, ‘spoony’ and ‘crumpets’ and checking the expectation from the word stress chapter that the initial stops in stressed syllables are significantly more aspirated than those preceded by /s/ or in unstressed syllables.

ST: It's not and most the pity I would say

ToBI annotation in this utterance is really challenging for me since there are several reasonable options. As the primary labelling I put the simplest option with a single intonational phrase, downstepping contour and the L- aligned right after the last accent on 'pity', as discussed in the previous section. However, all of these decisions had reasonable alternatives: the breaks after 'not' and 'pity' could be #3, and pitch accents might be rising L+H* ones, and there was also a possibility of accents on 'I' and 'say'.

In connected speech we have familiar cases of t/d-deletion and glottalizations. Segmentally, I would point your attention to the vowels of 'pity'. Recall from the pattern of 'happy tensing' from the end of Sect. 5.3 describing the tenser variant of short /i/ at the end of words like 'happy' or 'coffee'. You can test if this speaker conforms to this pattern, and to my transcription, by measuring formants in 'pity'.

GN: I agree

Here the ToBI transcription is unreliable due to the overlap but note again the smooth alignment of GN's agreement with the end of 'I would say' from ST. Moreover, ST gave the impression of ending his turn with a clear L-L% boundary tone. This is another example of mutual coordination of the speakers on the one hand, and the non-deterministic system of turn-taking, in which overlaps are very frequent despite this mutual coordination. I think that the word 'so' initiating the following utterance is lengthened, and followed by a silent pause not because the speaker needed more time to plan, but to resolve the overlap, make sure he has the floor, and then continued speaking.

ST: So the words are the secrets of Wodehouse

Intonationally, the lengthening of 'words' might suggest a #3 break following, but the absence of resets and the presence of general downstepping argues for the primary labelling in the textgrid. In both 'the' tokens the speaker employs th-stopping, which can be visually seen in the clear bursts following the complete closures represented as vertical bars of energy in the spectrogram.

ST: There's obviously plots and there's obviously characters but every single word every syllable has to ring true

This continuation of ST's turn presents several challenges for labelling and nicely exemplifies variability in speech production. First, In ToBI chunking, the question of the association of 'but' arises. Again, this is a very common pattern in which conjunctions such as 'and' or 'but' align either with the preceding or the following clause they link. Second, it is not clear if 'has to ring true' is a separate intermediate phrase and if the accents are H* in severely narrowed pitch range or if they are L*. You see my preference in the textgrid. I took the salient fall, and minor lengthening, on 'syllable' as strong indications of the presence of the L- phrase accent, and assumed a narrowed pitch range due to the upcoming in-breath and previous 4.3 seconds of speaking.

In IPA transcription there are clear boundary-related initial glottal stops preceding both ‘every’ tokens, which I decided not to mark with a [ʔ] to be consistent with the rest of the IPA transcription. Then there is the realization of the two tokens of ‘obviously’. There is a radical weakening of the vowel hiatus between [v] and [s] resulting in the three-syllable realization compared to the dictionary [ˈɒbvɪəsli]. Moreover, I neither saw nor heard the closure for /b/, and in the second case for /l/. Since this was a radical departure from the dictionary pronunciation, my IPA transcription tries to capture it.

Finally, the realization of ‘and there’s’ juncture is also interesting. Despite what we said regarding the elision vs. unreleasing of the adjacent homorganic consonants in Chapter 10, I perceive here extreme reduction that fits elision better than unreleasing: [ənez]. Neither the sound nor the visual information provides support for the place of articulation of the nasal and thus whether the nasal place assimilation from alveolar to dental took place or not.

ST: And it’s an opportunity to make a mistake and be anachronistic

The analysis of intonation in this utterance does not present major challenges. The most interesting aspect is the realization of four weak forms: ‘and’, ‘an’, ‘a’, ‘and’. The first ‘and’ is phrase initial and probably due to strengthening associated with this position the vowel is not a weak form schwa but [æ]. Despite the full vowel, the final /d/ is evidently elided despite the following vowel, which presents the least conducive environment for d-deletion. Hence, continuity of speech with [æɪ] as ‘and’ is at the display here as well.

I transcribed the indefinite article ‘an’ as [ə²] since I hear and see clear glottalization and I do not detect any traces of /n/. Here I can only speculate that the speaker mispronounced and planned to say ‘a’ instead of ‘an’, or hypo-articulated so much that any alveolar constriction is replaced with the glottalization, not only /t/. The other indefinite article ‘a’ is also extremely reduced into just a voiceless release of the preceding [k]. The extreme reduction and weakening is also present with the second ‘and’ despite being initial in an intermediate phrase. All aspects discussed in Chapter 10 take place: schwa deletion, d-deletion and the nasal place assimilation triggered by the following bilabial stop.

ST: But it’s also an opportunity just to absolutely capture the moment and the time and the texture of the period

This is a very challenging utterance for ToBI annotation. The key issue is the nature of the boundaries that follow the main concepts of this utterance: ‘opportunity’, ‘moment’, ‘time’ and ‘texture’. Moreover, the speaker’s voice is frequently creaky, which complicates the analysis. I finally opted for #3 breaks after the first three words of the list primarily due to evidence in the lengthening preceding the boundary and the pitch reset following it. There was also a slight but perceptible hesitation on ‘the’ preceding ‘texture’, which I deemed not salient enough for a #3 break. The capturing of the tonal targets was less problematic with the already discussed issues of narrowed pitch range in the last intermediate phrase.

The utterance also presents many cases of connected speech aspects. There is another case of flapping in the initial ‘but’, which is noteworthy for British speakers. In the juncture ‘an opportunity’ I hear a word-final [m], which probably stems from the nasal place assimilation triggered by the bilabial [p] of ‘opportunity’. However, the /p/ is not adjacent to the target /n/ as they are separated by a vowel. This is a nice example of continuity and overlap in the articulatory speaking actions and what we said in Chapter 2 that sometimes also non-adjacent sounds might influence each other. Other aspects such as the glottalization of word-initial vowels as boundary signals, d-deletions and nasal place assimilations in ‘and’, were discussed before, and the absence of linking-r in ‘texture of’ was analysed in Chapter 10.

GN: Do you have a favourite thing that you wrote like a favourite phrase that you were just like oh nailed it

In the first phrase ‘do you have a favourite thing that you wrote’ we can mention the initial coalescence in ‘do you’. The affricate is voiceless [tʃ] but since this initial devoicing was not transcribed in IPA I left [dʒ] to preserve consistency. It is worth listening separately to ‘do’ and then to ‘do you’ and note how the clear voiceless perception of the ‘do’ alone changes to the voiced one in ‘do you’. As seen in the alternative tier, I had difficulties with labelling pitch targets in the second part of the utterance, especially the association of ‘favourite phrase’ into units. The primary labelling is closest to my perception, but shows very different chunking from the one expected just based on text.

Segmentally, it is interesting to compare the two tokens of ‘favourite’ and observe how the second mention is still reduced despite being pitch accented. Furthermore, we can also observe the pragmatic use of breathy-creaky voice quality in ‘oh’ conveying emotional excitement.

This concludes my commentary. As mentioned in the Introduction, you might want to proceed with the remainder of *Bake_off_dialogue* and *Grinch* that are labelled, the rest of the three interviews that are not labelled, or to any other spoken material that you might be curious about. I hope your investigation and engagement with speaking English exemplified in the book and these commentaries have been, and will be in the future, intellectually challenging, practically useful and rewarding.

Further Reading

As described in the Introduction, the coverage of the material overlaps greatly with other introductory books on English Phonetics & Phonology but offers a unique and novel approach of discovery activities and guided hands-on explorations of speech patterns with Praat. These are followed by conceptualizing just experienced observations through core aspects of English sound patterns in the phonetics and phonology domains. These are further exemplified by commenting their deployment in spontaneous speech. Below is a brief, and very incomplete, selection of suggested reading for those who wish to take their studies further or seek alternative and complementing approaches to English phonetics and phonology. A fuller treatment of specific issues can be found in the main academic *journal*s of the area like *Journal of Phonetics*, *Laboratory Phonology*, *Language and Speech*, *Journal of the Acoustical Society of America*, *Phonology*, and others. Representative reviews of the major topics and the issues and questions within them by leading researchers of the field can be found in *handbooks* such as *The Handbook of Phonetic Sciences*, *The Handbook of Phonological Theory*, or *The Handbook of International Phonetic Association*.

General Phonetics and Phonology Textbooks

Colins, Beverley S., Mees, Inger M. (2013). *Practical Phonetics and Phonology: A Resource Book for Students*. 3rd Edition. Routledge.

- This book offers a very useful division of progressively advanced blocks (Introduction, Development, Exploration, Extension) allowing for a modular approach. It complements the current book by deeper discussions of accents, dialectal differences, and learning foreign languages (e.g. Spanish, Japanese, Italian, or German).

Ladefoged, Peter, Johnson, Keith. (2015). *A Course in Phonetics*. 7th Edition. Cengage Learning.

- This is a classic textbook with nice illustrations and many useful new performance exercises and covering slightly broader area than the current book.

Roach, Peter. (2009). *English Phonetics and Phonology: A Practical Course*. 4th Edition. Cambridge University Press.

- A very approachable book geared mostly towards non-native speakers of English and focusing primarily on the articulatory patterns rather than on acoustics.

Zsiga, Elizabeth, C. (2013). *The Sounds of Language*. Wiley-Blackwell.

- While the current book contains a one-semester material focused on phonetics more than on phonology, Zsiga's book covers more ground for a full two-semester course nicely balancing phonetics and phonology.

English Prosody

Szczepek-Reed, Beatrice (2010). *Analysing Conversation: An Introduction to Prosody*. Palgrave Macmillan.

- In addition to discussing some prosodic aspects of speech in context and presenting relatively rich audio material for the readers with a brief description of Praat, the book aims at exemplifying various patterns uncovered within the Conversational Analysis framework and the sound examples are picked to illustrate these patterns.

Ward, Nigel. (2018). *Prosodic Patterns in English Conversation*. Cambridge University Press.

- This book presents the experimentally grounded bottom-up approach to how measurable acoustic-prosodic characteristics of pitch intensity, timing and voice quality participate in systematically cuing various communicative and pragmatic functions.

Other Useful More Advanced References

- Crutenden, Alan. (1997). *Intonation*. Cambridge University Press.
- Johnson, Keith. (1997). *Acoustic and Auditory Phonetics*. Blackwell.
- Kenstowicz, Michael. (1998). *Phonology in Generative Grammar*. Blackwell.
- Ladd, Robert, D. (1996). *Intonational Phonology*. Cambridge University Press.
- Ladefoged, Peter, Madsen, Ian. (1996). *Sounds of the World's Languages*. Wiley-Blackwell.
- Laver, John. (1994). *Principles of Phonetics*. Cambridge University Press.
- Stevens, Kenneth. (2000). *Acoustic Phonetics*. MIT Press.
- Zemlin, Willard. (1998). *Speech and Hearing Science, Anatomy and Physiology*. Prentice-Hall.

Index

A

abstraction, 17, 66, 217–218
acoustics, 45, 49
action, 8
active articulators, 36–39
age, as social factor, 22
airstream mechanism
 egressive, 29
 glottalic, 30
Albanian, 140
allophone, 120
 variation, 126
allophonic pattern, 19, 132, 149
allophonic process
 ordering, 136
alveolar
 closure, 180–189
 ridge, 36–39, 65
ambiguity, structural, 199
ambisyllabicity, 148, 150, 152
amplitude, 51–53, 55
analysis-by-synthesis, 229
anti-resonance, 106
aperiodic waves, 52–54, 95
Arabic, 209
aspiration, 119, 131, 150, 165
assimilation, 180, 221
 manner, 183
 nasal place, 181
 progressive, 183
 regressive, 181
 voicing, 112, 182
autosegmental, 216
autosegmental-metrical, of intonation, 220
awareness, building, 14, 25, 137

B

babbling, 9
backchannel, 241, 246
Beckman, M., 219
Bella Coola, 140
Berber, 139
Berko-Gleason, J., 10
body language, 245
Boersma, P., 46
boundary
 between sounds, 16
 between syllables, 144
 between words, 175, 179
 morpheme, 192
 prosodic, 21, 196–199
 tone, 222–230
Boyd, S., 188
Brazil, D., 219
break
 index, 221–230
 prosodic, 199
bronchi, 30
burst, 30, 52, 95, 99, 106

C

Canadian raising, 136
cardinal vowels, 72, 74
cartilages, in the larynx, 30–33
Catalan, 122
category
 grammatical, 20, 157
cavity
 as filter, 57

nasal, 106
 pharyngeal, 59
 centre of gravity, 102
 chimpanzee, 40–42
 click, sound, 30
 coalescence, 185, 221
 coarticulation, 18
 cognitive science, 16
 cohesive device, 237
 common ground, 232
 communication, interpersonal, 8
 Communication Accommodation Theory, 248
 communicative situation, 21
 competence, 18
 compound, 171, 192
 compression, 50
 continuum, in context, 18
 contrast
 phonemic, 125
 coordination, 8
 between speakers, 232, 242
 temporal, 116
 Croatian, 113
 cross-talk, 240, 247
 cue word. *See* discourse, marker
 cycle, 50, 52
 Czech, 159

D

dark/clear-l, 132–136, 149
 de-accenting, 233
 declination, 206, 230
 derived, measurements, 160
 devoicing
 of approximants, 131
 of stops, 119, 122
 dialect
 regional, 21
 diaphragm, 28, 33
 diphthong, 72, 87–91
 centring, 90
 closing, 90
 falling, 90
 discourse, 236
 marker, 238
 structuring, 244
 discreteness
 of sounds, 17
 discreteness, in intonation, 220
 disjuncture, 196–198
 distribution
 complementary, 120, 124, 130

parallel, contrastive, 120
 downstep, 227
 Dutch, 119

E

elasticity, 29, 33
 electromagnetometry, 13, 71, 188
 elision, 180, 221
 apparent, 188
 English varieties
 African-American Vernacular, 102
 American (AE), 66, 81, 128
 British (BE), 66, 81
 British, north, 128
 Brooklyn, 102
 California, 77
 Cockney, 34, 107, 113, 130
 Dublin, 77
 Indian, 102
 Irish, 102, 108
 New England, 83
 non-rhotic, 83
 r-full, 22
 rhotic, 83
 r-less, 22
 Scottish, 83, 90, 108, 110
 Southern US, 90, 110
 epiglottis, 40–43

F

F2 lowering, 108, 110
 F3 lowering, 84, 110
 filled pause, 239
 filler, conversational, 239
 filter
 acoustic, 54–58
 low-pass, 247
 Finnish, 23, 65, 140
 flap, flapping, 128, 135, 179, 184
 fluidity, of movements, 14
 focus
 broad, 233
 narrow, contrastive, 233
 formant. *See* frequency, formant
 movement, 88, 96
 tracking errors, 82
 transition, 98, 106, 129
 fortis/lenis consonants, 96
 Fourier Transform, 57
 free variation, 130
 French, 77, 131, 159

frequency, 52–54, 60
 formant, 58, 64–67
 fundamental, 52–55, 212–217
 harmonics, 54–59, 61
 fricative, 100
 (inter-)dental, 102
 Fry, D.B., 160
 functional load, 238
 function word, 201–203

G

gap, in turn-taking, 240
 General American (GA), 64
 German, 23, 65, 122, 134
 gestures
 articulatory coordination, 130
 crowding, 188
 glide, 110, 141, 190
 glottalization, 128, 130, 178, 185, 200
 glottal stop, 34, 100, 129, 130, 190
 glottis, 34–35, 39, 52, 55, 99, 117, 128, 178
 goal, articulatory, 16
 Guy, G., 188

H

habits, speaking, 9, 18
 hands clapping
 pitch accents, 197
 syllables, 138
 hard palate, 36–38, 65
 head, in tone unit, 219
 hesitation, 222, 239
 Hindi, 123
 Hirschberg, J., 219
 homograph, 169
 homorganic, consonant, 189
 Hungarian, 77, 120, 134, 157
 Hyper/Hypo-articulation theory (H&H), 175

I

identity, speaker's, 21, 22
 information
 given, 233
 new, 202, 232
 inspiration, 29
 intensity, 33, 164
 intentions, 21, 231
 International Phonetic Alphabet (IPA), 7, 25,
 217–221
 interpolation, 229
 intonation

American tradition, 219
 British tradition, 219, 233
 form of, 224
 functional meaning, 224
 functions, 233
 intrinsic, difference, 166
 introspection, 8, 15, 63, 80, 87, 108
 intrusive-r, 191
 Italian, 23, 77, 120, 131, 140

J

Japanese, 139
 jaw, 36–39
 movement of, 19
 Ju'hoansi, 24
 juncture. *See* boundary; break

K

Kelso, S., 148
 knowledge
 mental, 9, 10, 17, 20, 122, 159, 165
 phonological, 18
 subconscious, 8, 17
 Korean, 182

L

L1
 speakers, 10
 L2
 accent, 10
 language, 10
 speakers, 10, 74, 102, 171
 Labov, W., 22
 language, 16
 larynx, 30–35, 40, 41, 94
 lowering, 40–43
 lateral
 approximant, 107–108
 fricative, 107
 lexical set, 73, 143
 Lieberman, Ph., 42
 Lindblom, B., 175
 linking, 180
 j/w, 191
 r, 190
 lips, 36–39
 closure, 13, 15, 16, 18, 19, 117
 liquid, 141
 loudness. *See* amplitude
 lungs, 28–30, 33

M

mandible. *See* jaw
 Mel, scale, 217
 merger
 cot-caught, 81, 91
 north-force, 84
 strut-foot, 87
 mind–body, dichotomy, 16, 17
 minimal pair, 84, 120, 221, 237
 monophthong, 72
 mora, 152
 morphological structure, 170
 muscles
 extrinsic, of the tongue, 36–38
 intercostal, 29, 33
 intrinsic, of the tongue, 36–38
 orbicularis oris, 37
 mutual beliefs, 232

N

natural class, 126
 non-linearity, pitch vs. f_0 , 216
 nuclear accent, 198
 nucleus, syllable, 144–147

O

obstruent, 134–136, 141, 216
 onset maximization, 148, 152
 oscillogram, 13
 overlap, 98
 in time, 14, 18
 turn-taking, 240
 overtones. *See* frequency, harmonics

P

palato-alveolar. *See* place of articulation,
 post-alveolar
 part of speech, 169, 203
 passive articulators, 36–39
 perception, 8
 experiment, 246
 period, 52
 periodicity, 51, 52
 periodic waves, 52–54, 95
 pharynx, 38, 41
 pharyngeal wall, 36–38
 philosophy of mind, 16
 phonation, 30
 phoneme, 120
 phonetics, 17, 121
 phonology, 17, 121
 framework, 229

 phonotactics, 140, 145–147, 149
 phrase
 accent, 222–230
 intonational, 199
 Pierrehumbert, J., 219
 pitch, 55, 160, 212. *See also* frequency,
 fundamental
 accent, 196–197, 222–230
 accent, bi-tonal, 227
 excursion, 200
 intensity link, 164
 intrinsic, 167
 manipulation, 162, 227
 range, 218, 243
 reset, 206
 standard deviation, 247
 target, 219, 221–230
 place of articulation, 96, 110
 alveolar, 99, 100
 bilabial, 99
 glottal, 101
 (inter-)dental, 100
 labio-dental, 100
 labio-velar, 110
 palatal, 110
 post-alveolar, 100, 104
 velar, 99
 plateau, contour, 223
 Polish, 120, 122, 140, 157
 Praat
 annotating functionality, 127
 scripting, 240, 241
 textgrid, 127
 tutorial, 46
 pragmatic meaning, 201
 pre-boundary lengthening, 201, 222
 predictability
 allophones, 120
 word stress, 159
 pre-fortis clipping, 134, 136, 149, 150, 184
 pre-head, 219
 primates, higher. *See* chimpanzee
 prominence, 177, 196
 distribution, 207
 prosodic
 structure, 206
 puffs, in recording, 47

R
 rarefaction, 50
 r-colouring. *See* rhoticity
 Received Pronunciation (RP), 64
 redundancy, 125
 register, 131

- relative, features, 167
 representation, mental, 17
 resonator, 59, 67. *See also* cavity; filter
 re-syllabification, 179
 rhotic approximant, 108–110, 190
 rhoticity, 83, 91
 IPA transcription, 84
 rhythm, 208
 entrainment, 246
 reversal. *See* stress shift
 Romance languages, 209
 routine, 11, 12
 Russian, 158, 209
- S**
- schwa, 66, 69, 86, 90, 143, 192
 secondary stress, 161, 166, 170
 semitone, scale, 217
 semivowel. *See* glide
 sentence stress, 198
 Serbian, 113
 set, open/close, 203
 shortcut, mental, 17
 sibilant, 101, 132
 sine wave, 50–51, 54–58
 skills
 acquired, habitual, 17
 Slovak, 23, 122, 145, 147, 157
 socio-economic status, 21, 22, 26
 sociolinguistic factors, 21, 102
 soft palate, 36–42, 105, 110
 sonorant, 126, 134
 sonority
 hierarchy, 141–144
 profile, 142
 sound
 adjacent, 14, 19
 source, 54–58
 velaric, 30
 source-filter theory, 58, 59
 Spanish, 16, 77, 90, 124
 speaking
 continuous, 12, 17
 ecology, 164
 rate, 243
 spectrogram, 13, 59–62
 broadband, 61
 narrowband, 61
 spectrum, 54–59
 speech
 entrainment, 246
 error, 222, 239
 visualization, 12
 spelling-pronunciation link, 23, 79, 85, 111
 silent consonant, 111
 stress clash, 171, 184, 207
 stress shift, 207
 stress-timed, 209
 stress-to-weight, 152, 170
 stricture
 degree of, 94, 95, 107
 strong form, 201–205
 subroutine, 9, 116, 131, 145, 235
 suprasegmental, 216
 syllabic consonant, 143, 145
 syllable, 9, 116, 135, 191
 coda, 144–147, 179
 counting, 138, 140
 heavy/light, 152
 onset, 144–147, 179
 rhyme, 144–147, 151
 stressed, 121
 tonic, 219
 weight, 169
 syllable-timed, 209
- T**
- t/d
 deletion, 186–189
 realization, 126
 tail, 219
 tap, alveolar, 128
 target, of assimilation, 181
 Thai, 123
 th-stopping, 102, 103, 184
 tier
 moraic, 151–152
 skeletal, 151–152
 tone, nuclear, 219
 Tones and Break Indices (ToBI), 25, 219
 tongue, 36–38, 41
 bunched, 108
 retraction, 64, 133
 retroflex, 108, 109
 trachea, 30
 transcription
 broad, phonemic, 120
 narrow, 120, 121, 131, 132, 143
 trigger, 181
 trill, 108
 Tuller, B., 148
 tuning fork, 49–51
 Turkish, 65
 tut-tuting, 29

U

uncertainty, in annotation, 226
unreleased, stop, 131, 189
uvula, 38, 43

V

velarized-l. *See* dark-l
velic opening, 105
velum. *See* soft palate
vibrations
 in the larynx, 30, 31
 source of, 49–52
vocal cords, 30–35, 52–59, 94, 117, 212
 abduction, 33, 128
 adduction, 33, 94
vocal folds. *See* vocal cords
vocal fry, 35
vocalization, dark-l, 133
voice
 breathy, 34, 35
 creaky, 34, 35, 81, 200
 modal, 34, 35, 214

 quality, 34–35
voice onset time (VOT), 119
vowel
 2D model, 71
 duration, 75, 151
 hiatus, 190
 quality, 71
 shortening, 134, 136
 tense/lax contrast, 76, 80, 96, 125

W

Ward, N., 221
Weenink, D., 46
Wells, J.C., 73, 79, 84, 152
whispering, 34, 35
word stress, 116, 121
wug test, 10

X

x-ray, 11–23