TELL

Textbooks in English Language and Linguistics (TELL) 1

Edited by Magnus Huber and Joybrato Mukherjee

Ulrike Gut

Introduction to English Phonetics and Phonology

PETER LANG
Internationaler Verlag der Wissenschaften

Including CD

Introduction to English Phonetics and Phonology

Textbooks in English Language and Linguistics (TELL)

Edited by Magnus Huber and Joybrato Mukherjee

Volume 1

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Wien

Ulrike Gut

Introduction to English Phonetics and Phonology

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at <http://www.d-nb.de>.

ISSN 1862-510X
ISBN 978-3-631-56615-2

Printed in Germany 1 2 3 4 5 7

www.peterlang.de

Acknowledgments

I am indebted to many people for their valuable comments on earlier versions of this book: Lilian Coronel, Patrick Maiwald, Franziska Meltzer, Sandra Mollin, Christina Sanchez, Karin Tausend and Eva-Maria Wunder.

My sincere thanks also go to those who gave me permission to use their photographs, films, icons or voice in the book and on the CD-ROM, and to those who helped with the compilation of the recordings: Rebecca Davies, Christoph Karl, Katarzyna Dziubalska-Kołaczyk, Martin Hermann, Peter James, Leonard Librizzi, Franziska Meltzer, Inger Moen, Kevin Munhall, Joe Perkell, Evgenia Slavianova, Kenneth Stevens, Fiona Tan, Charla Teufel, James Thomas, Holger Voormann, Thomas Wittenberg and the International Phonetic Association <http://www.arts.gla.ac.uk/IPA/ipa.html>.

Table of Contents

1 Introduction: Phonetics and Phonology

This textbook is about the sounds English speakers produce and hear when speaking or listening to speech. Although many are not aware of it, speech and speech sounds play a central role in the life of human beings (except, of course, for the deaf): on an average day, a person produces several thousand spoken words and hears a multiple of that. This means that speakers use several thousand speech sounds every day to communicate their feelings, wishes and intentions and encounter equally many speech sounds when listening to the feelings, wishes and intentions of others. Yet, almost none of the speakers are aware of what they do when they produce or perceive speech. If you ask a speaker who produces several thousand vowels a day which body parts and muscles are active during the articulation of these vowels, he or she will most probably not be able to tell you. Similarly, if you ask native speakers of English how they employ the height of their voice to signal their intentions and attitudes, you will most probably only encounter a puzzled look and a shrug of the shoulders. Neither would most native speakers of English be able to explain why they understand what was said in a recording in which half of the speech sounds are masked by noise. All this shows that the knowledge native speakers have of the pronunciation of their language is not conscious. This textbook is thus about what native speakers of English do not know they know about speaking and understanding spoken English.

For English language teachers and other professionals concerned with pronunciation (such as speech therapists or pronunciation trainers), conscious knowledge of English speech sounds, of their production, properties and perception, is of course essential. They need to know which organs and mental processes are involved in speech production; they need to know which unconscious knowledge native speakers of English have (and learners of English need to acquire) about the sound system of English; they need to know about the organs and mental processes involved in the perception of speech. It is obvious that language teachers will not be able to support language learners adequately until they understand what exactly is entailed in the articulation and perception of speech and also have the means to describe and evaluate differences between the pronunciation of English language learners and that of native speakers. It is the aim of this book to give you the necessary background knowledge and to introduce you to the terminology that is required for a successful pursuance of these professions. Last but not least this textbook is intended for English language learners who wish to be able to identify and improve weaknesses in their own pronunciation of English.

1.1 Phonetics and phonology

There are two fields or subdisciplines in linguistics concerned with pronunciation and sound, namely **phonetics** and **phonology**. Both of them describe and analyse speech from a different perspective. Phoneticians strive to find ways of describing and analysing the sounds humans use in language in an objective way. Three different areas of phonetics can be distinguished: articulatory phonetics, acoustic phonetics and auditory phonetics. **Articulatory phonetics** analyses which organs and muscles are used by speakers to produce speech (see chapter 2). **Acoustic phonetics** is concerned with the physical properties of speech sounds as they travel in the air between a speaker's mouth and a listener's ear (see chapter 5). **Auditory phonetics** focuses on the effect those sounds have when they reach the listener's ear and brain (see chapter 6). Phonetics is thus a linguistic field that draws heavily on other scientific disciplines including anatomy, physiology, neurology and physics.

The beginnings of articulatory phonetics can be traced back to the very detailed descriptions of the pronunciation of Sanskrit, which were made by Indian scholars several centuries BC. The first theory on the function of the vocal folds was proposed by Antoine Ferrein (1693-1769) in the 18th century. Christian Gottlieb Kratzenstein (1723-1795) and Wolfgang von Kempelen (1734-1804) attempted to explain the production of vowels and consonants by building a talking machine. Henry Sweet (1845-1912) and Daniel Jones (1881-1967) were the pioneers in the description and transcription of the articulation of English speech sounds. Nowadays, phoneticians are actively involved in the development of synthetic speech and automatic speech recognition systems, which we encounter almost every day in automatic announcements at railway stations and airports as well as in telephone bookings and interactive computer-based pronunciation training courses.

Questions that phoneticians investigate include: Which speech organs are involved in the production of a particular speech sound or a particular pitch movement, and how do they work together? Which physical properties (i.e. frequency, amplitude) do different speech sounds or pitch movements have? Which body parts are involved in the perception of speech sounds, of pitch and of stress? The methods of investigation in phonetics have profited from many technological advances. As will be described in sections 2.8, 6.5 and 6.6, phoneticians can make use of the sophisticated methods developed in medicine for observing the activity of the speech organs, muscles and the brain during speech production and perception. Direct observation and measurement of nearly all speech organs involved in speech production and speech perception are possible, and computer software enables the exact measurement of the acoustic properties of all aspects of speech.

While phonetics deals with the production, properties and perception of the speech sounds of human languages, phonology is concerned with how these speech sounds form patterns in a particular language. Phonologists investigate, for example, which function a sound has in a language and which sounds can be combined – follow each other – and which cannot. Phonology can be divided into two areas: **segmental** and **suprasegmental phonology**. While segmental phonology deals with speech sounds, suprasegmental phonology is concerned with larger units such as syllables, words and intonation phrases (see chapters 3 and 4). The study of phonology began as early as in the third century BC with the Ancient Greek grammarians describing the sound patterns of Greek and Latin in Europe and scholars in India describing the phonology of Sanskrit. The first descriptions of the sounds of English were published by John Hart (*An Orthographie;* 1569) and William Bullokar (*Booke at large*; 1580), who were concerned with the divergence of spelling and pronunciation in 16th century English. Charles Butler's *The English Grammar*, published in 1634, probably contains the first descriptions of stress patterns of English words.

In the long tradition of phonology many different ideas about language have influenced the methods of phonological analysis. In the later 19th century, for instance, it was popular to treat language as an 'organism' that evolves and dies, modelled on the then revolutionary ideas of Charles Darwin on living organisms. Many German linguists working at that time investigated the 'evolution' of languages, proposed language family trees and provided the first descriptions of sound changes in Indo-European languages. This approach, which was further developed by Ferdinand de Saussure (1857-1913) at the beginning of the 20th century, assumes that language is a system that can be described and analysed by the elements or units it consists of (such as speech sounds) and rules that apply to them (such as rules of sound order). Phonologists working in this framework strive to find units of sound structure and sets of rules that describe patterns and regularities of these units in a particular language. For example, when looking at vowels in unstressed syllables in English (as will be done in section 3.2.3 below), one finds that they have a different quality than vowels in stressed syllables – the technical term for this is that they are reduced. It is possible to describe this vowel reduction as a rule that applies when a syllable is unstressed in English.

In another tradition in phonology, focus is put on the mental representation or knowledge of sounds and sound patterns by speakers. It is assumed that different speakers have the same 'mental idea' of a sound although they might produce and hear this sound in different forms. When you compare the articulation of the /p/ in *pot* with the /p/ in *spot* and in *top* with the methods of phonetic analysis, you will find that they differ distinctly in their articulation and acoustic properties: for the /p/ in *pot*, there is a short but clearly audible burst of air after the speaker opened his or her lips which does not occur in the /p/ in *spot*. For the /p/ in *top*, speakers might not even open their lips and there might be an accompanying stoppage of the airstream in the throat (section 3.1.2. gives you more information on this). However, when you ask speakers about those /p/s, they are usually not aware that they produce or perceive different sounds. This leads phonologists to claim that speakers have just one mental representation (i.e. knowledge stored in the memory) of the speech sound /p/. Different **notation symbols** are used in order to differentiate between speech sounds that form part of the speakers' knowledge and speech sounds that are actually produced and can be measured and perceived. The slashes / / indicate that a speaker's knowledge or mental representation is referred to; this is what phonology deals with. The square brackets [ ] indicate that an actual sound is being talked about, which is what phonetics is concerned with. Figure 1.1 illustrates the usage of the transcription symbols: a speaker's mental representation of "Oh!" is put into slashes, whereas the actual pronunciation of the word is put into square brackets. See chapter 3 for details on this distinction and further transcription conventions.

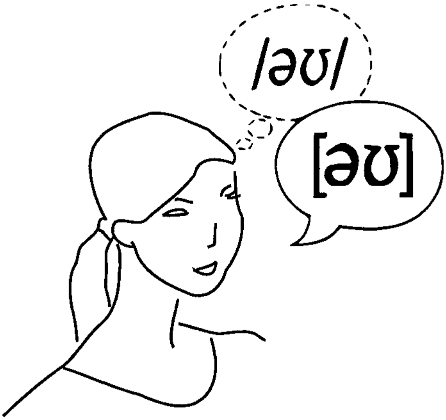


Figure 1.1. Transcription conventions of the mental representation and actual pronunciation of "Oh!".

Recent theories in phonology (e.g. Bybee 2001) put emphasis on the view that language is a tool that is used by speakers for communication (this is called the functional approach). Phonologists working in this tradition claim that language cannot be separated from the speakers who use it. They treat the phonology of a language as the knowledge – represented in his or her brain – a speaker has of phonological units and processes that are necessary in order to produce and understand speech. Questions about English that phonologists investigate include: Which sounds does English have in order to make meaningful distinctions between words? Which pitch movements does English have in order to make meaningful distinctions between different types of utterances such as questions and statements? Which syllables in English words are stressed and which are unstressed? Which sounds can be combined to form an English syllable? Do the same phonological rules apply to the different accents of English (e.g. British English and Australian English) and the different historical forms of English such as Old English and Early Modern English? The methods of phonology are indirect since we do not yet have the means to examine a speaker's phonological knowledge directly. Indirect methods include observing speaker productions and gathering speaker judgments (for example, asking a speaker of English whether *flin* or *lsin* are potentially good English words).

To put it simply, phonology is concerned with what speakers and listeners know and children and second language learners need to learn in order to competently use and understand spoken language. Conversely, phoneticians are interested in the precise activities of the speech organs, the physical properties of speech and the activity of the body parts involved in speech perception. While a phonologist will claim that the vowel in the word *leak* is long and the vowel in the word *lick* is short, a phonetician will measure the exact length of the sound in milliseconds. This is not to say that there is always a clear-cut boundary between phonetic and phonological analysis. Increasingly, linguists combine questions and methods of both fields, for example in laboratory phonology, where phonetic techniques and measurements are used for phonological investigations.

1.2 Sounds and letters

One of the central distinctions that are made in the study of language is that between letters and sounds. In English, more than in other languages, it is obvious that the spelling and pronunciation of words do not match. Take the words *bone*, *done* and *gone* (linguists use italics when they write about words), which are spelled with the same vowel letter but are actually pronounced with three different vowels. *Bone* rhymes with *tone*, *done* rhymes with *fun* and *gone* has the same vowel as *rock*. Or take the words *riff*, *loaf* and *tough*. They all end with the same sound but the spelling varies from <ff> to <f> and <gh> (linguists use the pointed brackets < > when referring to a spelling – the technical term for a spelling symbol is **grapheme**). When talking about the pronunciation of English words it is therefore very confusing and misleading to use the spelling symbols of the Roman alphabet. Chapter 3 describes the symbols of the phonetic alphabet, which enable you to write unambiguously about English pronunciation.

Unfortunately, English spelling is also very deficient in not providing any information about other features of English pronunciation apart from speech sounds, for example features such as stress and intonation. All words in English that are longer than one syllable have one stressed syllable. The word *English*, for example, is stressed on the first syllable, the second one is unstressed. However, no indication of this is given in writing. Equally, the pitch of a speaker's voice moves up and down during speech, which can change the meaning of an utterance. Listeners will understand very different things when somebody says "Yes" with a falling pitch than when he or she says it with a rising pitch. In writing, the only option you have is to use punctuation marks like ! and ?, which, however, is not enough to represent all significant pitch movements produced by speakers. In chapters 3 and 4 you will see which symbols are used in linguistics to describe stress and pitch movement in English.

1.3 The structure of this book

This textbook is intended as a companion for students of English throughout their university studies. Most of the sections were written for absolute beginners and require no previous knowledge of phonetics and phonology. They serve as a first introduction to the sounds and the sound system of English – i.e. the production, perception and physical properties of English speech sounds as well as what we presume to be their organization and structure in the brains of native speakers. The sections marked as 'advanced reading' provide information for especially interested students wishing to specialize in one of the areas of phonetics and phonology.

Chapter 2 describes the structure and function of the body parts involved in speech production and shows how they work together in the articulation of connected speech. Chapters 3 and 4 deal with the phonology of English and describe the sounds of English, English syllables, words and intonation. In chapter 5, the acoustic properties of speech sounds are presented, and it is shown how acoustic phonetics can be employed in language teaching and learning. Chapter 6 provides an introduction to the structure and function of the body parts active during the perception of speech and the phonological knowledge necessary to understand speech.

In order to help you with the reading, all technical terms are printed in **bold** when they first appear. These technical terms form the foundation of phonetic and phonological description and analysis. Sometimes two alternative technical terms are given – as much as one may wish it, in academia there is unfortunately no one-to-one correspondence between scientific concepts and technical terms and different researchers use technical terms with different definitions. All of the basic technical terms are listed in the index at the end of the book so that when you are looking for a definition or description of a technical term, you will find there a reference to the specific page where it is given. Graphs and tables are included in the text wherever they help to illustrate a point. Further illustration is provided by the accompanying CD-ROM. The icons shown in Figure 1.2 indicate video and audio examples included on the CD-ROM.

Figure 1.2 Icons referring to video (left) and audio (right) examples on the CD.

Exercises at the end of each chapter allow you to revise the most important terms and concepts of each chapter. The "Further reading" section points out books, further practice materials and useful websites for a more detailed description of the various topics introduced in each chapter. Full references of the literature listed there are given at the end of the book. When a particular study or book is referred to in the text, they will also be listed in the References section (chapter 7).

1.4 Exercises

1. In what way do phoneticians and phonologists view the sounds and pronunciation of English? What are the major differences?
2. What is the difference between <p>, /p/ and [p]?
3. List all the different spellings of the sound [u] in English words.
4. How many sounds and how many graphemes do the following words have:

   knee, spring, debt, wrapped, fire, stall, key, pint, lamb, print

5. Put the following words into sets beginning with the same sound (not grapheme!)

   knight, cat, king, get, gnat, guest, calm, quay, quest, west, vest, want, vase

6. Ask a native speaker of English how many vowels English has. Compare this to the description of the vowel system of English given in section 3.1.3.

1.5 Further reading

Yavaş (2006, chapter 9) lists all the different spellings of the sounds of English. A history of the theories and methods of phonology and phonetics is given in Clark, Yallop & Fletcher (2007, chapter 11). For especially interested readers, Sampson (1980) describes the history of linguistic thought including phonology and phonetics.

2 Speech production

This chapter describes the anatomy (the structure) and physiology (the function) of the body parts involved in the production of speech. Humans do not have any organs or muscles that are used exclusively for speaking; all of the body parts participating in speech production have other, primary functions, which are more basic and often vital for keeping the organism alive. Producing speech is only a secondary, an 'overlaid function' of the speech organs, one that developed later in evolution. Traditionally, **three 'systems' of speech organs** are differentiated that have different functions in the speech production process (see Figure 2.1). The **respiratory system** described in section 2.1 produces the airstream that is necessary for speaking; the **phonatory system** described in section 2.2 produces voice, and the **articulatory system** described in section 2.3 is responsible for the formation of the different speech sounds. The classification of sounds according to their manner and place of articulation is described in section 2.4. Section 2.5 illustrates how all three systems work together in the production of connected speech. The central role of the brain in speech production is described in section 2.6. Section 2.7 is concerned with speech production in first and second language acquisition, and section 2.8 explains some methods of measuring the activity of the speech organs.

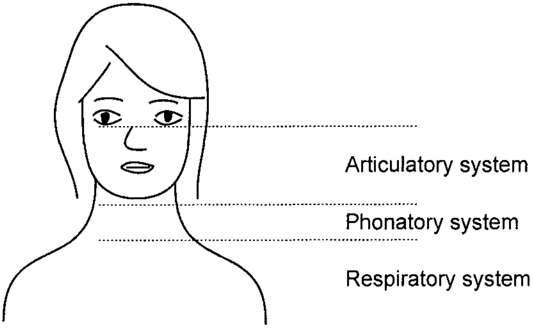


Figure 2.1. The three systems of speech production.

2.1 The respiratory system

The respiratory system provides the airstream on which speech is produced. It consists of the rib cage, the intercostal muscles, the diaphragm, the lungs, the bronchial tubes and the trachea. Figure 2.2 shows the **rib cage**, which is barrel-shaped and contains the **lungs** and the **intercostal muscles**, which run between the ribs. The floor of the rib cage is formed by the **diaphragm**, a dome-shaped muscle. The lungs are a pair of organs consisting of soft sponge-like material that rest on the diaphragm. They are connected to the windpipe, or **trachea**, by two **bronchial tubes**, which divide into increasingly smaller bronchioli that end in millions of tiny air-filled sacs, the alveoli.

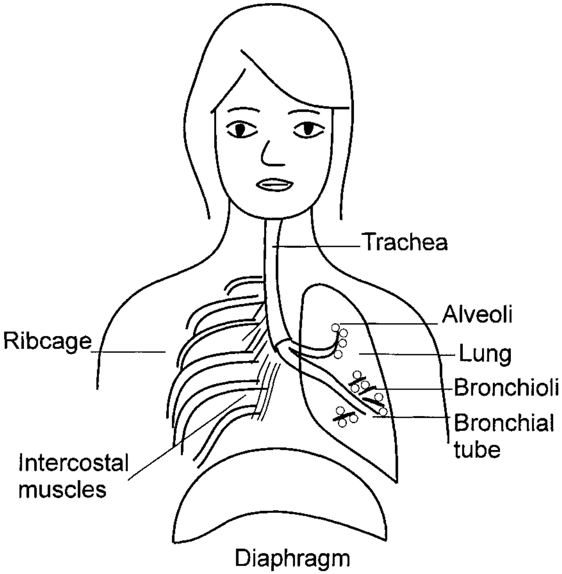


Figure 2.2. The respiratory system.

The primary function of the respiratory system is to breathe and thus to keep the organism alive by providing oxygen for and removing carbon dioxide from the blood. The main function of the respiratory system for producing speech is to provide an **airstream** with which sound is produced. Various types of airstream mechanisms can be used for speaking: the most common one is the **egressive pulmonic** airstream, with air flowing from the lungs (hence pulmonic) through the trachea up into and out of the mouth and nose (hence egressive). Other types of egressive airstreams do not originate in the lungs but further up: an **egressive glottalic** airstream, for example, comes from the larynx (see phonatory system, section 2.2 below). When you hold your breath and then articulate a [t], you are producing it on a glottalic airstream. The resulting sound is called an ejective. Ejectives may occur in English as variants of sounds produced on a pulmonic airstream. They are, however, distinctive speech sounds for example in Hausa, a language spoken in Nigeria, Niger, Sudan and Cameroon. Speech can also be produced on an **ingressive** airstream, for which air is sucked into the body. Some people involuntarily speak on an ingressive airstream when they are out of breath or when they are sobbing. Similarly, the clicking sound that English speakers occasionally make with their tongue in order to signal disapproval is produced with an ingressive airstream. Clicks occur as distinctive speech sounds in South African languages such as Zulu.

Speech in English is usually produced on an egressive pulmonic airstream. In egressive pulmonic speech, the lungs form an air reservoir that provides the energy for sound production. At first, speakers need to breathe in some air. Breathing in, or **inhalation**, does not involve any active sucking of air into the lungs. Rather, by using some of the intercostal muscles connected to the rib cage and the diaphragm muscle, speakers lift their rib cage upwards and outwards and lower their diaphragm. Both actions contribute to expanding the lungs and thus increasing the lung volume. This in turn causes a lowering of the air pressure within the lungs compared to the external atmospheric air pressure, which makes air flow into the lungs via the mouth and the nose so that the internal and the external air pressures are equalized. For breathing out (**exhalation**), the lungs are compressed by the interplay of some of the intercostal muscles and the upward movement of the diaphragm. Both actions reduce the volume of the lungs and press the air out. The total lung capacity of humans varies between 3 and 7 litres, depending on body size, posture and general physical fitness. The lungs, however, never empty completely during exhalation. During the production of pulmonic egressive speech, some of the intercostal muscles and the muscles of the abdomen work together to maintain a relatively consistent level of pressure. Thus speakers can control the speed of exhalation and can vary how long they speak on one breath. When not speaking, during quiet breathing, breathing in and breathing out take roughly the same amount of time. During speech production, however, exhalation is typically about eight times longer than inhalation, a proportion that can easily be increased by trained singers and speakers.

The fact that speech is produced on an airstream and the necessity to breathe in order to refill the lungs determines an important characteristic of speech: it is divided into units, which are often referred to as **breath groups**. Even in perfectly fluent speech, after a certain period of time a speaker needs to interrupt and take a breath. This typically happens at specific linguistic boundaries such as the end of an utterance or between two clauses rather than within syntactic phrases or within words (cf. Yang 2004). Listeners are so accustomed to these breath group breaks that they only notice them when they occur in unusual places. When you record English speakers and measure the length of speech between two breaths you will find that, on average, breath groups are 2 to 3 seconds long. However, they can be as short as 0.7 seconds or as long as 8.29 seconds, depending on factors such as whether a speaker is giving a monologue or participating in a conversation and whether the speech is prepared (like in a presentation learned by heart) or produced spontaneously.

The overall activity of the body parts of the respiratory system involved in speech production also varies with the type of speech that is being produced. For loud shouting greater overall muscular effort is needed, which results in a higher level of air pressure. Similarly, stressed syllables in a word are usually produced with a higher level of air pressure than unstressed ones. Whispering, by contrast, requires very little activity from the body parts of the respiratory system.

In summary, the respiratory system provides the airstream on which speech is produced. It determines the fact that speech is divided into breath groups and contributes to the production of different levels of loudness and stress.

2.2 The phonatory system

The larynx, together with the vocal folds, constitutes the phonatory system. Its main function is to provide voice with the help of the airstream coming from the respiratory system. The airstream passes from the bronchial tubes into the trachea, a tube with a length of about 11 cm and a diameter of 2.5 cm. The trachea has a skeletal frame consisting of rings of cartilage, a material similar to bone but less hard. At the top of the trachea there is the **larynx**, a box-like structure that consists of a series of cartilages (see Figure 2.3).

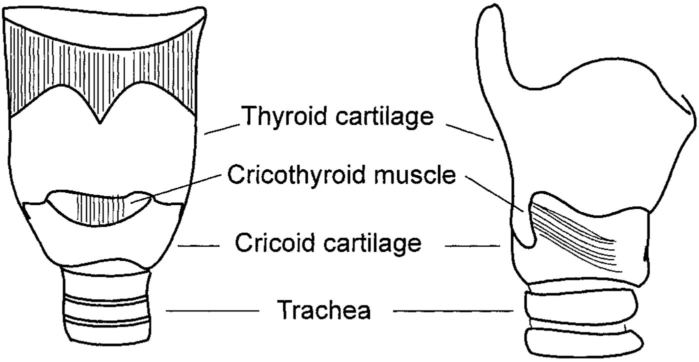


Figure 2.3. Front (left) and side (right) view of the larynx. Adapted from Clark, Yallop & Fletcher(2007: 177) and Ashby & Maidment (2005: 23).

The **cricoid cartilage** functions as the top ring of the trachea and forms the base of the larynx. Two flat plates meeting at an angle at the front of the throat form the **thyroid cartilage**. This thyroid cartilage, which has a more acute angle in males than in females, can be seen and felt as the 'Adam's apple' moving up and down when somebody is speaking or swallowing. It acts as a shield for the vocal folds, which are situated inside the larynx.

Figure 2.4 shows a photograph and a schematised top view of the **vocal folds** (the older terminology is *vocal cords*). The vocal folds basically consist of muscle covered by several layers of fibrous tissue. When looking down on the larynx from above (for example with a laryngoscope mirror; see section 2.8) we can see the two roughly triangular folds of pink tissue with their white thickened edges (the vocal ligaments). In front, they are attached to the thyroid cartilage and at the back to the **arytenoid cartilages**. The two arytenoid cartilages themselves sit on the cricoid cartilage and are attached to it with a joint. This joint allows the arytenoid cartilages to move forward and backward and from side to side, which turns out to be important for speech as it allows a speaker to adjust the position and the tension of the vocal folds. For normal breathing, the back ends of the vocal folds are held apart and air passes in and out silently.

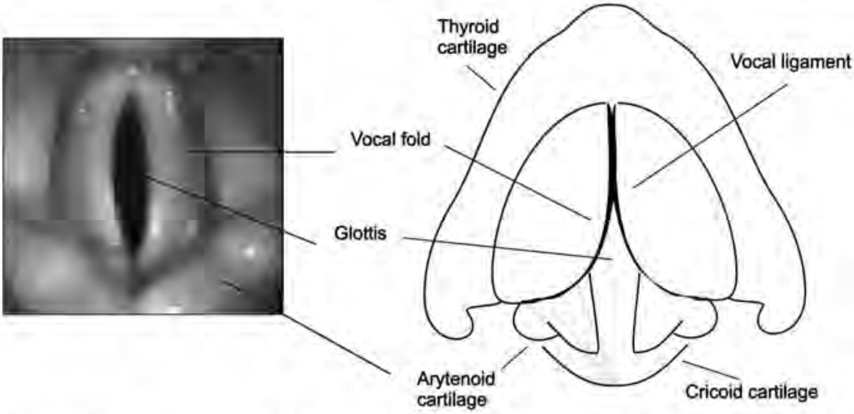


Figure 2.4 Top view of the vocal folds: stroboscopic photograph (left) and schematized view (right).

The primary function of the vocal folds is that of preventing food and drink from falling into the trachea and the lungs and thus saving the organism from choking or suffocation. Further, the possibility to close the vocal folds enables one to hold one's breath during physical exertion, such as lifting weights. The secondary function of the vocal folds and the one important for the production of speech is that of **phonation**. Phonation refers to the vocal folds rapidly opening and closing, which is a process both controlled by the laryngeal muscles and powered by the airstream from the lungs. The process of phonation involves an alternation of forces: it begins with the arytenoid cartilages pressed together so that the gap between the vocal folds is closed and the vocal folds are touching each other lightly. The airstream from the lungs then pushes the vocal folds apart, thus letting some air through. Immediately after that, the vocal folds are pulled together again both because of the elastic properties of their stretched tissue and because of the suction generated by the flow of air through the glottal constriction. This suction, called the Bernoulli force, is similar to the suction that occurs when two trains pass each other at high speed. When the vocal folds are closed again, the air from the lungs forces them apart again and the cycle begins a second time. An adult female speaker's voice is based on an average of 250 openings and closings of the vocal folds per second. Children's voices have a higher rate of vocal fold openings and closings; male voices typically are based on an average of 150 opening and closing movements.

02_vocalfolds.gif on the CD-ROM shows a slow-motion video of the opening and closing cycles of the vocal folds. It was recorded with an endoscope (see section 2.8).

This rapid opening and closing of the vocal folds is called **vibration**. The process of vibration can be likened to the movement your pursed lips make when you blow through them and make them flap. The sound created in the larynx is just a quiet buzz, but after it has passed through the articulatory system (see section 2.3 below) this buzz has changed into a sound we perceive as **voice**. The buzz in the larynx is actually not created by the vibration of the vocal folds itself but by the sequence of puffs of air emitted through the vibrating vocal folds. Not all sounds produced in speech are generated with vocal fold vibration. Speech sounds can therefore be divided into **voiced** ones and **voiceless** ones. Only voiced speech sounds are produced with vocal fold vibration. During the production of voiceless sounds, which are technically speaking noise rather than sounds (see section 5.1.2), the vocal folds are open and do not vibrate. You can test whether a speech sound is voiced – as for example an [a] – or not – as for example a [f] – by putting your hand on your throat above the larynx, where the vocal fold vibration in voiced sounds can be felt.

Smokers who develop laryngeal cancer need to have their larynx removed, which means that they cannot produce voiced sounds anymore. An electrolarynx held to the throat creates electromechanical vibration which functions as an artificial source for articulation. Listen to speech produced with an electrolarynx (02_electrolarynx.wav on the CD-ROM, taken with permission from http://www.webwhispers.org/).

Different speakers have different voices, a quality of sound that is called **timbre**. A voice may sound 'sharp' or 'mellow' or 'squeaky'. These differences in voice quality are partly due to functional differences in the larynx and the vocal folds and partly due to conscious modulation of the voice. Voices can further be high or low, i.e. have a high or low **pitch** (the technical term for height of voice). The limits of a speaker's pitch are determined by the size, length and tenseness of the vocal folds. Roughly speaking, longer vocal folds vibrate more slowly than shorter vocal folds. Equally, tenser and thinner vocal folds cause faster vocal fold vibrations than more relaxed and thicker vocal folds. The faster the vocal fold vibration, the higher the pitch perceived by listeners (see section 6.5 for details).

02_glissando.mov on the CD-ROM shows a video of the vocal folds during a glissando going from a very high to a very low pitch. Note the change in thickness and tension of the vocal folds. Note also that during breathing the vocal folds are held apart.

The length of the vocal folds varies with the speakers' body size and is typically around 3 mm at birth and between 12 and 17 mm for adult females and 17 to 25 mm for adult males. Vocal fold vibration accordingly varies between different speakers. For someone speaking on a very low pitch, the vocal folds perform about 60 to 80 opening-closing cycles per second. Most female voices are produced with a vocal fold vibration of between 150 and 250 cycles per second; male voices are typically based on between 100 and 150 cycles per second. A small child's voice is produced with up to 800 opening-closing cycles per second. The cycles of vocal fold vibration are never perfectly regular, which is not perceptible in a healthy adult voice. It may become noticeable though in older speakers, which is why we can fairly accurately guess the age of a speaker by only listening to his or her voice.

Although the timbre and the highest and lowest pitch a speaker can produce are limited by biological factors, speakers are free to modulate their voices in a lot of ways. Throughout speech, pitch changes continuously and can largely be controlled consciously. Changes in pitch can be employed for linguistic reasons as for example when a speaker produces a rise in pitch across the word *really* and thus signals that the utterance is an incredulous question "Really?!" rather than the statement or rebuke "Really!", which would be produced with a falling pitch. This use of pitch is typical for **intonation languages** such as English and is described in detail in section 4.3. **Tone languages,** such as Yoruba and Igbo spoken in West Africa or the many different dialects of Chinese, employ pitch to signal meaning. In these languages, the pitch height or pitch change on a syllable determines the difference between words. In Mandarin Chinese, for example, [ba] produced with a falling pitch means 'father', whereas it means 'pull' when produced with a rising pitch.

Listen to the difference between "Really" with rising and with falling pitch (02_really.wav on the CD-ROM).

Speakers might also alter the height of their voice for non-linguistic reasons, for example when speaking on a high and squeaky voice in order to imitate someone else or when using a low voice in order to appear serious and strict. In physiological terms, pitch and pitch change are controlled by modifying the length and tenseness of the vocal folds. This happens through a complex interplay of various muscles including the thyroarytenoid, vocalic and cricothyroid muscles. When the muscles in the vocal folds are tensed, vibration is faster and the pitch is higher. When a speaker uses these muscles to increase the tenseness even further, parts of the vocal folds are 'shut down', which causes the vibration to increase even more. Conversely, relaxed vocal folds cause slower vibration and a lower pitch. When you consider singing, it becomes clear which amazing control humans have over their laryngeal muscles. Not only can exact heights of pitch be produced, the pitch range can also be enlarged by switching from chest register to a high falsetto. In chest register, which is used for speaking, the vocal folds are relatively short and thick whereas in falsetto, which underlies high sung notes, they are longer and thinner. One of the challenges of professional singing is to train a middle voice register that smoothes the abrupt change between the two registers. Variation in pitch can also happen unconsciously, for example when shouting. The increased air pressure from the lungs used for producing loud speech and the tension in the laryngeal muscles cause an increased vocal fold vibration – which is why most people's voices are higher when they shout.

If the back ends of the vocal folds are held apart by moving the arytenoid cartilages, a small space opens up between the vocal folds. This space is called the **glottis** (see Figure 2.4). It is the place where the so-called glottal sounds are produced. [h], for example, the sound at the beginning of the word *house*, is created in the glottis with the vocal folds producing a turbulent airflow. The glottal stop [ʔ], which some speakers of English produce at the end of the word *what*, is articulated by closing the vocal folds tightly until the air pressure increases in the trachea and then suddenly opening them and releasing the airstream. Speakers can actively change the shape of the glottis during speech. Through this, the airstream is modulated – which is called **mode of phonation** – and the sound of the voice is altered. Five modes of phonation can be distinguished:

- voice
- voicelessness
- whisper
- breathy voice
- creak

For the production of **voice**, the glottis is closed and the vocal folds are touching each other lightly. The airstream coming from the lungs causes vocal fold vibration as described above (see Figure 2.5 on the left). During the production of **voiceless** sounds the glottis is open, i.e. the vocal folds are wide apart and the airstream passes through without making them vibrate (see Figure 2.5 on the right).

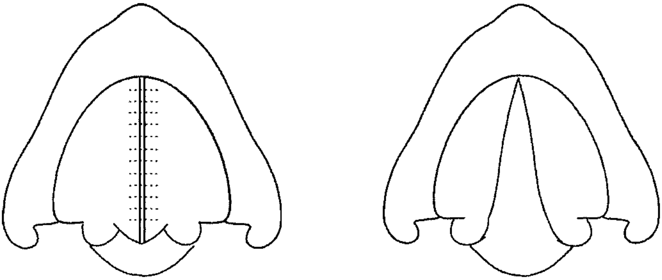


Figure 2.5 Vocal folds during the production of voiced sounds (left) and voiceless sounds (right).

For **whisper**, the glottis is narrowed to about 25% of its maximal opening, the vocal folds are close together but do not vibrate. The arytenoids leave a small gap at the back of the larynx through which air passes with high velocity (see Figure 2.6 on the left). The airflow is strongly turbulent, which produces the characteristic hushing sound.

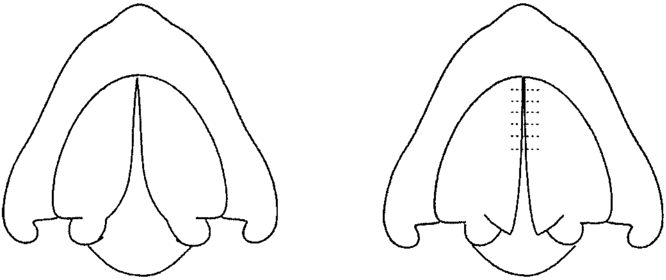


Figure 2.6 Vocal folds during whisper (left) and breathy voice (right). The arytenoids leave a small opening. The vocal folds only vibrate in breathy voice.

Breathy voice is produced by pulling the vocal folds slightly apart but making them vibrate with considerable airflow (see Figure 2.6 on the right). Some speakers of English produce **creaky voice** at the end of utterances with falling pitch. In creaky voice, the arytenoid cartilages are held tightly together so that the vocal folds can only vibrate at the front end (see Figure 2.7). Creaky voice has very low subglottal pressure and vocal fold vibration is reduced to about 30 to 50 cycles per second.

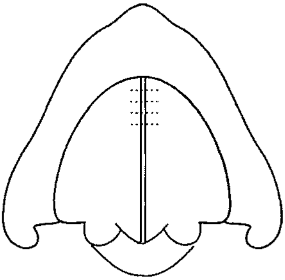


Figure 2.7 Vocal folds during the production of creaky voice. Vocal folds are pressed together. Vibration occurs only in the front.

Listen to an example of breathy voice, creaky voice and whisper (02_phonationmodes.wav on the CD-ROM).

The function of the laryngeal system can be summarized as follows: during silent breathing and for the production of most voiceless sounds the vocal folds are held wide apart and the airstream coming from the lungs passes silently through the open glottis. For glottal sounds such as [h] and the glottal stop [?] the vocal folds move to restrict or interrupt the airstream. For the production of voiced sounds the vocal folds are set into vibration by a complex interplay of various muscles, and the airstream is chopped up into little puffs that enter the pharynx. The thickness and degree of tension of the vocal folds, which can be largely controlled by the speaker via the various laryngeal muscles, determines the pitch of each voiced sound. In English utterances, movements in pitch can be employed to signal linguistic contrasts. Speakers can further produce different modes of phonation that alter the perceptual characteristics of their speech.

2.3 The articulatory system

After passing the larynx the airstream enters the articulatory system either undisturbed or chopped up into little puffs as described above. The articulatory system, sometimes also referred to as the **vocal tract**, consists of three cavities as well as the **active** and **passive articulators**. The organs that function as active articulators are the velum (or soft palate), the tongue, the lips and the mandible (the lower jaw). Passive articulators are the uvula, the palate, the alveolar ridge and the teeth. Just like the other organs of speech, the articulators have other, more vital functions for a person. The mandible, the teeth and the tongue, for example, have the primary function of chewing food and passing it down to the stomach, thus keeping the organism alive. The cavities play an essential role in providing a passage for the incoming and outgoing breath, and they warm and humidify incoming air.

The cavities that make up the articulatory system are the **pharyngeal cavity**, the **oral cavity** and the **nasal cavity** (1, 2 and 3 in Figure 2.8). The pharynx (which is another term for the throat) is a tube of muscle of about 12 cm length and stretches from the top of the larynx to the beginning of the oral and the nasal cavities. At the upper end it divides into two tubes, one leading up to the space inside the nose, the nasal cavity, and the other to the space inside the mouth, the oral cavity. The nasal cavity is about 10 cm long for most adult speakers; the shape of the oral cavity varies widely between speakers. All cavities modulate, i.e. change, the airstream coming from the larynx by acting as **resonating chambers** for the buzz produced by the activity of the vocal folds. This means that some parts of the buzz are amplified and others are dampened, which alters the sound quality (see section 5.1.2 for details). This process is the same as the resonance a musical instrument provides, for example a flute or the body of a violin. The particular shape of the instrument enhances some portions of the airstream blown into it (flute) or produced by the action of the bow setting the strings in motion (violin).

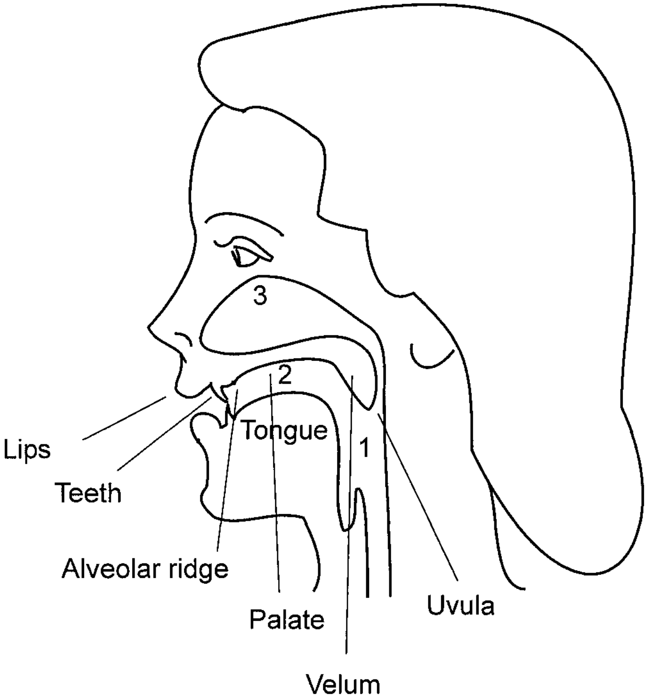


Figure 2.8. The vocal tract. The three cavities and the articulators of the articulatory system. (1 = pharyngeal cavity; 2 = oral cavity; 3 = nasal cavity)

An alteration of the shape of the resonating cavities in a speaker's head leads to the creation of different sound qualities. For example, when the nasal cavity is full of mucus and the tissue is swollen, as is the case when a speaker has the flu, the sound of the speaker's voice is altered considerably. Of course, the sizes and specific shapes of the cavities differ among speakers, which leads to a different volume and different resonating characteristics. Thus, the particular shape of the three cavities in individual speakers contributes to their specific voice quality. Note that the cavities do not provide resonance for an airstream that passed the open glottis and did not set the vocal folds into vibration.

In the oral cavity, the airstream from the larynx can be modified further by various actions of the articulators – the velum, the uvula, the tongue, the palate, the lips, the teeth and the mandible (see Figure 2.8). The **velum** (also called the soft palate) consists of muscular tissue and can be raised and lowered by a set of muscles. The airstream coming from the larynx can only enter the nasal cavity when the velum is lowered. In quiet breathing the velum is lowered all the time so that air can enter through the nose and flow down to the lungs and flow out through the nose again. During speech, the velum is usually raised and the airstream flows through the oral cavity only. Accordingly, sounds produced with a raised velum are called **oral sounds**. Those speech sounds that are produced with a lowered velum and where the airstream flows only through the nasal cavity are called **nasal sounds**. For **nasalized sounds** such as the vowel in the French word *dans*, the air flows through both the oral and the nasal cavity.

The **uvula** is the small fleshy appendage at the lower end of the velum consisting of muscle and tissue. You can see it in a mirror when you open your mouth widely. When the uvula vibrates, much like in gargling, a sound is produced that is common in German (for example at the beginning of the word *Rose*) and other European languages, but not in Standard English.

Although the **hard palate**, the bony structure forming the roof of the mouth, is not mobile and thus does not play an active role in articulation, its importance is shown by studying the articulation of people born with a cleft palate. In about one in 700 newborn babies the two plates that form the palate are not completely joined. Although in most cases surgery is performed immediately to close the gap, the voice quality often remains different. Since the oral cavity does not close completely and parts of the airstream enter the nasal cavity, the speech produced has an overall nasalized quality. The domed roof of the mouth ends in a bony ridge just behind the upper front teeth, which is called the **alveolar ridge**. Many speech sounds of English such as [t] and [s] are articulated at this place.

The **teeth** take part in the articulation of several sounds in English. Although they do not move actively themselves, they play an important role in the production of many sounds, which becomes evident when one considers the difference in sound in the speech of people with false teeth. For interdental sounds, the tip of the tongue is inserted between the teeth in such a fashion that the airstream is pressed through, producing friction. This, for example, happens at the beginning of the words *think* and *that*. For sounds such as [f] at the beginning of the word *fun*, the friction is produced by forcing the airstream through a narrow space created by the lower lip and the upper teeth.

The **lips** consist of two fleshy folds richly supplied with muscles, which control the opening and closing movements. The lips can for example block the airstream coming from the lungs by closing, or they can produce friction by being held close together so that the airstream is forced through and a hissing sound is produced. The lower lip can be held close to the upper teeth as for the production of [f]. Further, the lips can be held rounded and slightly protruded or they can be spread, which makes an important contribution to the production of certain vowels. Humans with a cleft lip, which often occurs together with a cleft palate, have a small gap in the upper lip, which, without surgery, prevents the full closure of the lips for articulation.

The lower jaw or **mandible** is an approximately U-shaped bone structure joined with the skull base. Specific actions of sets of muscles can move the mandible downwards, forward and sideways with a maximal aperture of about 40 mm. Downward movement of the mandible also influences the position of the tongue, which is attached to it with its root. This movement is important for the articulation of many vowels.

The **tongue** is the most important active articulator and the most complex organ of speech. It consists nearly entirely of muscle, is very versatile and can assume a wide variety of shapes. Due to its mobility and flexibility it makes the greatest contribution to changes in the shape of the oral cavity. For descriptive purposes the tongue is usually divided into several functional sections (though these parts cannot move independently of course): the **root**, the **back**, the **front**, the **blade** and the **tip** (see Figure 2.9). The root of the tongue is opposite the back wall of the pharynx. The back of the tongue lies beneath the velum when it is at rest and can be raised towards it with the help of specific muscles. Sounds that are produced with the back of the tongue moving close to the velum or touching it are for example [u] and [g]. The front of the tongue, which is actually fairly in the middle of it and which lies below the hard palate when the tongue is at rest, can also arch upwards and is very active in the production of vowels. The blade of the tongue can move upwards and come near to the hard palate, which forms the roof of the mouth. A sound articulated in such a manner is [j]. The tip of the tongue, the very front end, is its most mobile part and also highly sensitive. It can move between the teeth or curl up and touch the back of the upper teeth. Sounds that involve these kinds of articulation are for example the first sound in *thing* and the [d] in *width*. When the sides of the tongue are lowered and the airstream passes at both sides of the tongue, a sound like [l] is produced. The tongue can further be depressed to make a groove, which is important for the production of the sounds [s] and [z] in *sink* and *zinc*.

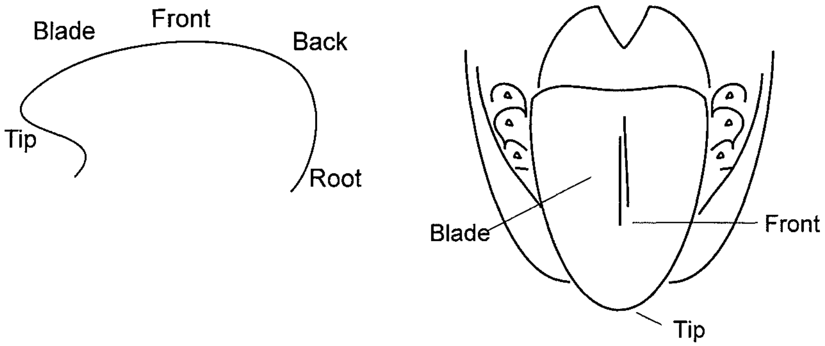


Figure 2.9. The parts of the tongue, side view (left) and top view (right). Right: Adapted from Clark, Yallop & Fletcher (2007: 191).

The articulators of the articulatory system can influence and shape the airstream in many different ways. For example, the airstream can be closed off completely so that pressure is built up behind the blocking. This happens for example when two articulators such as the tip of the tongue and the alveolar ridge touch. After the contact is released, the air bursts out with a popping sound, which is why sounds produced in this manner are referred to as **plosives**. When two articulators, such as the lower lip and the upper teeth, are moved close together but do not block the airstream completely and the air is forced through the small space, turbulence is created, which produces a typical hissing sound. Sounds produced in this way are called **fricatives**. When the two manners of articulation are combined and a blocking phase is followed by a friction phase the resulting sound is called an **affricate**. The tip of the tongue can further make contact with the alveolar ridge repeatedly and in rapid succession, which produces a **trilled sound**. When the tip of the tongue briefly strikes the alveolar ridge only once, the resulting sound is called a **tap**.

To summarize, the articulatory system shapes the airstream, both when it passes the open glottis and after it has set the vocal folds into vibration. This shaping is due to the resonance created by the cavities and the position or movements of the articulators. The term **articulation** refers to the narrowing or constriction of the vocal tract by the movement of an active articulator towards a passive articulator. The articulators can perform many different actions, such as blocking the airstream by making contact or creating friction by forcing the airstream through a small gap between them.

2.4 Classifying speech sounds

Speech sounds can be classified into different categories according to the activity of the organs of speech involved in their production. As already described above, speech sounds can be divided into voiced ones, which involve the vibration of the vocal folds, and voiceless ones, for which the airstream passes the open glottis. Oral speech sounds are produced with an airstream flowing out of the oral cavity only; nasals involve an airstream flowing out of the nasal cavity only. Speech sounds that are produced on both an oral and nasal airflow are called nasalized. In French, nasalized vowels have a distinctive function, which means that a sequence of sounds with a nasalized vowel (e.g. *sans*) means something else than the same sequence of sounds with an oral vowel (e.g. *sa*). In English, nasalized vowels do not have a distinctive function. They are often produced when a vowel is preceded or followed by a nasal consonant due to coarticulatory effects (see section 2.5 below).

Speech sounds are usually divided into **vowels** and **consonants**. Generally speaking, consonants shape the airstream coming from the larynx in a more perceptible manner than vowels. Yet, there are no clear-cut boundaries in articulatory terms between all members of the two groups. The production of a special type of consonant called approximants, for example, involves about the same type of tongue activity than the production of vowels, which is why some classifications of speech sounds propose a third group of sounds called **semi-vowels**. The distinction between vowels and consonants is mainly made on phonological grounds, i.e. the function particular speech sounds have in a syllable (this will be explained in section 3.2.1 below).

All speech sounds can be specified by describing

- the airstream mechanism
- the vocal fold action
- the position of the velum
- the place of articulation
- the manner of articulation

However, usually not all of these features are employed for the classification of vowels and consonants alike. In the description of English **vowels**, airstream mechanism, vocal fold action and the position of the velum are irrelevant since all of them are produced with a pulmonic egressive airstream, all vowels are typically voiced – although in fast speech the neighbouring sounds may cause them to be produced without vibration of the vocal folds (see section 2.5 below) – and are produced with a closed velum (again, coarticulatory effects described in section 2.5 can overrule this). Furthermore, the manner of articulation of vowels is fairly restricted. The airstream passes the oral cavity relatively unhindered; the particular vowel sounds are created by altering the resonating characteristics of the vocal tract through tongue position and the position of the lips. Vowels are usually classified according to three characteristics: **vowel height**, **vowel location** and **lip position** (see Table 2.1). Vowel height refers to the highest point of the tongue in relation to the roof of the mouth. A **high** vowel such as [i] in *bee* is produced with the tongue close to the roof of the mouth; for a **low** vowel such as [a] there is a considerable gap between tongue and roof of the mouth. **Mid** vowels such as [e] in *bed* are articulated with the tongue in mid position between high and low. Vowel location refers to the section of the tongue that is raised during the production of the vowel. In **front** vowels such as [i], the front of the tongue is raised towards the hard palate. In **back** vowels such as [u], the back of the tongue is raised towards the velum. **Central** vowels such as the vowel [ɜ] in *her* are produced with a raised centre part of the tongue. The lips can be either **rounded** as for the production of [u] or **unrounded** (sometimes also spread) for vowels such as [i]. The vowel [i] in *beat* is thus classified as a high, front, unrounded vowel. All vowels that occur in English are described in detail in section 3.1.3.

Table 2.1. Vowel height, vowel location and lip position of vowels.

vowel height		vowel location		lip rounding	
high	[i]	front	[i]	rounded	[u]
mid	[e]	central	[ɜ]	unrounded	[i]
low	[a]	back	[u]		

For English **consonants** it is usual to use a categorisation based on only three features of their production: voicing, their **manner** and their **place of articulation**. The airstream mechanism and the position of the velum are only described when they are not pulmonic egressive and closed, respectively. Thus, a [t] is described as a voiceless alveolar (= place of articulation) plosive (= manner of articulation). Some textbooks claim that English consonants can further be classified into **fortis** (consonants produced with greater articulatory effort) and **lenis** (consonants produced with less articulatory effort). These terms were coined before we had the methods to measure articulatory effort (see section 2.8 below), and it now appears that these categories are misleading: there is no measurable difference in articulatory effort between English consonants. The proposed categories probably stem from a confusion with the categories voiced and voiceless as well as aspirated and unaspirated (see section 3.1.2). All consonants of English can be described and classified exhaustively by their manner and place of articulation alone without reference to 'articulatory effort'.

Table 2.2 lists all the different manners of articulation that underlie consonants. Those consonants that involve a conspicuous obstruction of the airstream are referred to as **obstruents**. Obstruents can be further divided into **fricatives** and **plosives**. For the production of a plosive two articulators block the airstream completely and then suddenly release the pent-up air with an audible popping sound. A [p] is an example of a plosive. For fricatives, two articulators create an obstruction that leaves only a small space through which the airstream is forced. This action creates turbulence, which is perceived as a typical hissing sound, as for example in [f]. Consonants that consist of a plosive followed by a fricative are called **affricates**. The German word *Pflaume*, for example, begins with the affricate [pf]. Obstruents can be either voiced (e.g. [g]) or voiceless (e.g. [s]), and English has both types.

Consonants that are produced with the action of one articulator hitting another repeatedly and thus creating a mini-series of airstream blockages are called **trills**. The 'rolled r' used in Italian or some South-Eastern German dialects is a trill. This sound also occurred in Middle English and Early Modern English, as documented by the playwright Ben Jonson, who wrote in the early 17th century that the 'r' is produced with the "tongue striking the inner palate with a trembling about the teeth" (1640, p. 491). If the tongue tip briefly strikes the alveolar ridge only once the resulting speech sound is called a **tap** (sometimes this term is used synonymously with the term flap, although this refers to a similar but different manner of articulation). A tap, for example, is often produced by American English speakers in the middle of the word *letter*.

Laterals are consonants that are produced with the tongue tip touching the roof of the mouth and an airstream that passes the tongue at both sides like in the speech sound [l]. Laterals are included in the class of the **approximants**. For the production of approximants two articulators move towards each other, but not as closely as to create friction. For example, [j], the first sound in the word *universe*, is an approximant. These sounds are very similar to vowels in terms of their production, but are grouped with the consonants because of their behaviour in the sound system of a language (see section 3.2.1).

For the production of **nasal consonants** the velum is lowered so that the airstream passes through the nasal cavity. Trills, taps, approximants and nasals are usually voiced unless coarticulatory effects (see section 2.5 below) are too strong. Vowels, approximants and nasals are often grouped together as **sonorants**. This is due to the fact that they are usually voiced sounds with little obstruction of the airstream, which gives them a full sonorant sound. All of the manners of articulation listed in Table 2.2 occur in English consonants (see also section 3.1.1).

Table 2.2. Manners of articulation of consonants.

Name	Example	Action of articulators
plosive	[p]	complete blockage of airstream with subsequent sudden release
fricative	[f]	two articulators create obstruction that leaves only a small space through which airstream is forced; turbulence is created
affricate	[pf]	combination of plosive plus fricative
trill	rolled [r]	repeated hitting of one articulator with another
tap	[ɾ]	one articulator hitting another once
lateral approximant	[l]	airstream passes at both sides of tongue
approximant	[j]	two articulators move towards each other; no friction created
nasal	[n]	velum lowered; airstream passes through the nasal cavity

Consonants can further be characterised by their place of articulation. Table 2.3 lists all places of articulation together with the corresponding activities of the articulators. **Bilabial** speech sounds such as [b] are produced with the involvement of both lips. When both the lower lip and the upper teeth participate in articulation – as for example for a [f] – the resulting speech sound is called **labiodental.** Dental or **interdental** speech sounds such as the first sound of *the* are produced with the involvement of the tongue and the teeth. For the production of **alveolar** speech sounds such as a [s], both the tip of the tongue and the alveolar ridge are involved. **Postalveolar** sounds such as the first sound in *shoe* are produced with the blade of the tongue raised just behind the alveolar ridge. For **retroflex** sounds, which many American English speakers produce as the last sound in the word *writer* and which can often be heard in Indian English, the tip of the tongue curls up and back and comes close to the palate. When the blade of the tongue touches or comes close to the palate as for the production of [j], the speech sound is called **palatal**. **Velar** speech sounds such as a [g] involve the back of the tongue and the velum. The first sound in the German word *Rose* when pronounced in the Northern standard way is a **uvular** sound because it involves the uvula. **Pharyngeal** sounds that are produced in the pharynx do not occur in English but for example in Arabic. English has two

glottal sounds, the fricative [h] and the glottal plosive or 'glottal stop' described in section 2.2 above. The first speech sound of the word *was* is classified as **labiovelar** because it involves both the rounding of the lips and the raising of the back of the tongue towards the velum. It is the only English consonant with two places of articulation. All the manners and places of articulation that play a role in the production of English speech sounds are described in detail in chapter 3.

Table 2.3. Places of articulation of consonants.

Name	**Example**	**Articulators involved**
bilabial	[b]	both lips
labiodental	[f]	lower lip and upper teeth
(inter-)dental	[ð]	tongue and teeth
alveolar	[s]	tip of tongue and alveolar ridge
postalveolar	[ʃ]	blade of tongue and place just behind alveolar ridge
retroflex	[ɽ]	tip of tongue, curled up and back, and palate
palatal	[j]	blade of tongue and palate
velar	[g]	back of tongue and velum
uvular	[ʀ]	uvula
pharyngeal	[ħ]	pharynx
glottal	[h]	vocal folds
labiovelar	[w]	lips, back of tongue and velum

Tables 2.2 and 2.3 list examples of typical consonants that are produced with the different manners and places of articulation. However, it needs to be pointed out that speakers seem to differ in their articulations of individual speech sounds. Measurements of the movement of the tongue or the place of contact at the roof of the mouth for particular speech sounds have revealed that speakers use different articulations. For example, some speakers' tongue tip does not meet the alveolar ridge but rather the back of the upper teeth for the production of a [t]. It is assumed that when learning to pronounce a language speakers aim to achieve an acoustic goal rather than an articulatory one. This means that speakers control auditorily whether what they produce sounds correct and then learn to associate the corresponding articulatory movements with the specific speech sound.

Theories of the acquisition of speech production are described in section 2.7 below. Chapter 6 is concerned with the details of speech perception. Methods of measuring articulation are described in section 2.8 below.

2.5 Articulation in connected speech

The above sections showed that the act of speaking involves the complex interplay of body parts such as the muscles connected to the rib cage, the laryngeal muscles and the muscles moving the velum, tongue and lips. Even for the production of a single speech sound the coordinated activities of numerous muscles working in close synchrony are necessary. Yet, speakers do more than produce single sounds. Words consist of sometimes very long sequences of sounds, and utterances consist of sometimes very long sequences of words. In normal conversation, a speaker produces five words per second on average, which amounts to about 15 speech sounds per second. In very fast speech, this rate of articulation can be increased to about 25 speech sounds per second. The articulation of speech is thus an immensely complicated muscular feat.

Let us consider the series of complex and coordinated actions that has to be carried out when pronouncing the short word *pin*. The speaker first needs to breathe in, which requires the concerted action of the diaphragm and the muscles associated to the rib cage and the lungs. Then the airstream needs to be pushed out, which again involves the muscles connected to the rib cage and the diaphragm. The airstream flowing out now needs to be kept constant by using muscles to hold back the collapsing lungs for as long as the utterance lasts. The vocal folds are held apart by the action of the arytenoids. Simultaneously, for the articulation of [p], the speaker needs to close the lips firmly so that no air can escape through the mouth. When the lips are opened, the tongue has already moved into the position needed for the production of [ɪ] and the vocal folds have been brought together so that the voicing of the vowel can begin. During the few milliseconds it takes to produce the vowel, the velum, which was raised for [p] and [ɪ], moves downwards. At the same time, the tip of the tongue moves upwards in the direction of the alveolar ridge. When the tongue tip has made contact there and the velum is lowered, the sound [n] is created.

This simple example illustrates that the tongue is constantly moving during the production of speech and that the articulatory movements required for individual speech sounds are executed parallel to each other. This in turn means that the articulation of each sound is influenced by the articulation of the preceding and following sounds. It is one of the central insights of phonetic research that speaking does not involve producing one sound after another; rather, bundles of sounds are planned together (see section 2.6 below), which means that the articulation of one sound carries characteristics of its neighbouring sounds. If you take a recording of someone saying the word *pin*, it is very difficult to cut it up so that each sound can be isolated. Not only is it impossible to find exact boundaries between the individual sounds, but the three parts of the word *pin* also influence each other in a perceptually distinct way. Parts of the [p] and the [n] always carry the sound of the vowel within them; listening to the cut-out sounds you are able to tell that the [p] is followed by an [ɪ] and that the [n] is preceded by an [ɪ].

02_connectedspeech.mpg on the CD-ROM shows an X-ray recording of a speaker producing several utterances. Watch the continuous movement of the articulators, especially the tongue and the velum: there is no perceptible beginning or end to each of the sounds.

This process of reciprocal influence in articulation is called **coarticulation.** Coarticulation works in both directions: the articulation of a preceding sound can influence the following one, and the following sound can influence the preceding one (see Table 2.4). When the articulation of a speech sound influences the preceding speech sound, the process is referred to as **anticipatory coarticulation**. This happens for example during the articulation of the word *two*. Typically, a speaker's lips are rounded during the production of the [t], although this is not essential for the articulation of this speech sound. Rather, this action anticipates the lip rounding that is necessary for the production of [u]. When pronouncing the word *teeth*, by contrast, the lips are not rounded during the articulation of [t]. **Perseverative coarticulation** occurs when the articulation of a following speech sound is influenced by the preceding one. This happens often when a lateral approximant follows voiceless consonants, as for example in the word *split*. Typically, vocal fold vibration begins only on the vowel, so that the [l] is voiceless.

Table 2.4. Types of coarticulation.

Name	Example	Description
anticipatory	*two*	lips rounded for [t] when followed by rounded vowel
perseverative	*split*	no voicing of lateral approximant [l] after voiceless consonants
anticipatory across words	*have to*	no voicing of [v] preceding a voiceless consonant

Coarticulation of course does not only occur within words but also across words in a breath group. Words are not articulated as separate entities but merge into each other without boundary. This can easily be illustrated with speech analysis software that displays the continuous stream of speech in a breath group (see section 5.4 below). Coarticulation across words occurs for example when in the phrase *have to* the vibration of the vocal folds ends already after the vowel, and the voiced labiodental fricative [v] of *have* is produced as a voiceless [f]. The voicelessness of the [t] is anticipated and the two words are thus realized as [hæftu]. Two neighbouring sounds becoming more similar is a process that is often referred to as **assimilation**. This term has different definitions and is used to describe a variety of phenomena, but we will use it synonymously with the term coarticulation here. Traditionally, three types of assimilation are differentiated. In assimilation of place, one of two adjacent sounds changes its place of articulation in order to make it more similar to the other sound. This is the case with the pronunciation of the alveolar [n] in *unkind* as a velar [ŋ] due to the following velar [k]. Assimilation of voicing is illustrated in the example *have to*, which is pronounced [hæftu] (see Table 2.4). Assimilation of manner refers to two neighbouring sounds becoming similar in their manner of articulation. This, for example, happens in coalescence. When, in connected speech, two adjacent sounds are merged to form a new sound, the process is referred to as **coalescence**. Coalescence can be observed in the way many English speakers pronounce "I get you". Usually, the last sound of *get* and the first sound of *you* merge into the new sound [tʃ] ([t] and the first sound of *shoe*) so that the sequence is not pronounced [getju] but [getʃu].

An important factor influencing the articulation of speech is speaking rate, the speed of articulation. In slow, careful articulation the articulatory organs perform the necessary movements for each speech sound in an ideal way. For instance, in the utterance "This could be true" speakers round their lips for the vowel [u] of *could* and move the back of the tongue close to the velum. Then the tip of the tongue touches the alveolar ridge for the production of [d] before the lips close for the production of [b]. In rapid speech, however, many of the articulatory movements are interrupted before they reach their ideal endpoint, a process which is often described as **undershoot**. In fast speech, the lips are only partially rounded for the pronunciation of the vowel in *could*, making it sound more like an [ə], the sound at the beginning of the word *ahead*, than an [ʊ]. The back of the tongue, similarly, only moves a little way up towards the velum. Most probably, the tip of the tongue never touches the alveolar ridge before the lips close for the [b]. The rapid articulation of "could be" would thus come out as [kəbi]. Sounds not realized at all in fast speech are described as being **deleted**, whereas sounds whose articulation is only partially realised are called **reduced**. In our example the [d] is deleted whereas the [ʊ] is reduced.

Undershot articulation of sounds in English is especially frequent when two consonants follow each other and form a so-called consonant cluster. This occurs for example at the end of the words *just*, *perfect* and *bland*. Sometimes in normal speech, but particularly in fast speech, the second one of these consonants is not articulated so that the words are pronounced [dʒʌs], [pɜfek] and [blæn]. Many studies have shown that there is a systematic variation of such **final consonant cluster deletion** with linguistic factors: it is more frequent in unstressed syllables than in stressed syllables (thus more likely in *government* than in *around*), more frequent in uninflected word forms than in inflected word forms (thus more likely in *last* than in *laughed*) and more frequent in common words such as *just* than in rare words such as *bust* (Neu 1980, Labov 1989, Bybee 2002).

English vowels also often undergo the processes of **vowel reduction** or **vowel deletion**. Whether English vowels have full or undershot articulatory movements varies systematically with stress (see section 3.2.3). For example, English vowels are always reduced in unstressed syllables. This means they are produced without full articulation: the tongue is neither fully front nor back but highest in a middle position of the mouth, it is neither high nor low and the lips are not rounded. This happens for example in the first syllable of the word *around* (which is pronounced [əɹaʊnd]) and the second syllable of the word *comment* (pronounced [kɒmənt] in British English). When these reduced vowels occur between voiceless consonants as in *to come* they are often voiceless, too. In many unstressed syllables, vowels are not articulated in a perceptible way altogether, as for example in the first syllable of the word *police*, which is often realized as [p], so that the word is pronounced [p.lis] (the . stands for a syllable boundary). Vowel deletion also occurs regularly as for example in the second syllable of the word *fashion*, which is usually pronounced [fæ.ʃn].

Section 2.2 showed that some speech sounds in English are voiced and others are not. However, listeners do not perceive gaps in voicing when hearing speech but rather hear a continuous movement of pitch across a breath group. This continuous movement of pitch is referred to as **intonation**; the perception of intonation is described in more detail in section 6.5. Similarly, some speech sounds are produced with greater airstream pressure than others. This varying loudness can be interpreted by listeners as differences in **stress** (see section 3.2.3). In addition, careful, non-reduced articulations of speech sounds that carry the articulatory movement to the ideal target point are interpreted as indications for the presence of stress by listeners (see section 5.5.2 below).

Articulation in connected speech can be summarized as follows: producing a stream of speech is enormously complicated when considering the parallel movements of all muscles concerned. Typically, for the production of a single sound all muscles of the respiratory, phonatory and articulatory systems interact in a specific manner. When sounds are produced in a sequence, the articulatory movements of neighbouring sounds influence each other. Articulatory movements necessary for the production of a preceding or a following sound overlap and cause coarticulatory effects, which can be described as assimilation or coalescence.

2.6 The role of the brain

Sections 2.1 to 2.5 described the function of various muscles and organs in the production of speech. Yet, the most important organ of speech is the brain. Without the brain speech production would be impossible. It controls both the movements of all muscles involved in the articulation of speech sounds and all the other processes necessary for the production of speech, such as the planning of the speech message, the grammatical encoding and the combination of sounds in a way appropriate to convey meaning and to fulfil a communicative purpose.

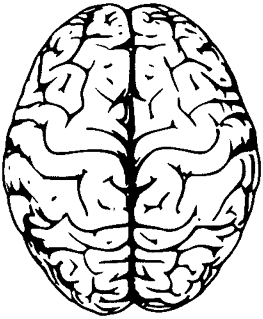


Figure 2.10. Top view of the two hemispheres of the brain (top of the picture = front of the head).

The brain can be divided into the left and right cerebral **hemispheres** (see Figure 2.10), each of which have an outer layer of grey matter called the cerebral cortex and an inner layer of white matter. The two hemispheres are linked among other things by the corpus callosum, which consists of a large bundle of nerve fibres. Each hemisphere contains billions of nerve cells, or neurons, which contribute to its complex functioning. Although, generally speaking, we still know very little about the function of the brain, its different parts have been known to fulfil certain specialized functions since the late 19th century. There are two ways of analysing the function of the brain: early research was limited to observing which dysfunctions occurred when parts of the brain were damaged by a stroke or an accident; more recently, techniques such as EEG and fMRI (see section 2.8 below) are used to measure brain activity during different speech tasks. Studies with split-brain patients, whose hemispheres are disconnected, have shown that the two brain hemispheres fulfil different functions and process information in different ways. This fact is referred to as **lateralization**. For right-handed people and some left-handed people it is the left hemisphere that is generally responsible for speech production and perception, whereas the right hemisphere is primarily involved in the perception of the prosodic aspects of speech (pitch movement and stress, cf. section 6.5). For some left-handed people the functions of the hemispheres are distributed in the opposite way.

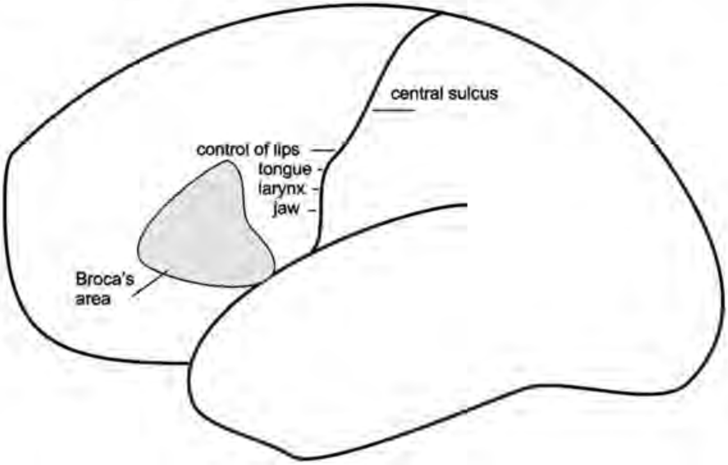


Figure 2.11. The parts of the brain involved in speech production.

Figure 2.11 illustrates that an area around the **central sulcus** is specialized for the control of the muscles in the larynx, tongue, jaw and lips. This area is referred to as the **motor cortex** and is responsible for the planning, control and execution of voluntary movements. Another specialized area of the brain connected with speech production lies in the anterior (front) parts of the left hemisphere for right-handed speakers (an area that was named **Broca's area** after the researcher who identified it). Patients with brain damage in this area suffer from a speech disorder that is referred to as Broca's aphasia. They articulate slowly and laboriously and have little control over pitch and loudness.

Since the articulation of a single sound can involve the coordinated interplay of hundreds of muscles, and since the speed of articulation is so high with a rate of up to 25 speech sounds per second, it is unlikely that each muscular movement of articulation is controlled individually in the motor cortex. Rather, it is assumed that the parallel movements of the different muscles necessary for the production of each speech sound are represented and executed as a coordinated bundle of motor activities. Moreover, as with other motor activities such as walking or riding a bicycle, the execution of articulatory movements is automatized, i.e. the movements are carried out routinely, very fast and without conscious control. Many of the complex patterns of muscular activities that underlie human behaviour involve an automatized motor pattern. That is, after learning the coordinated interplay of the muscles for a specific activity or procedure, the information is stored in the brain as a bundle and the activity is carried out below the level of consciousness. In fact, it becomes very difficult or even downright impossible to carry out the action when thinking consciously about it – try controlling consciously all the muscles involved in making a single step forward!

The various muscular activities necessary for the production of speech sounds are believed to be stored as highly automatized motor bundles or **articulatory gestures** in the motor cortex. It appears that humans have evolved specialized neural mechanisms for the automatization of the articulation of speech sounds. For example, it has been found that the respiratory muscles are controlled in a way that their actions are coordinated with the length of an utterance that is being produced. The amount of air a speaker breathes in already reflects the length of the utterance that will be made. Speakers inhale less air for shorter utterances than for longer ones. While speaking, moreover, the muscles of the rib cage work in different ways to uphold the pulmonary air pressure for shorter utterances than they do for longer utterances.

It is assumed that not only the articulatory movements for individual speech sounds are stored in the brain. Given the speed of articulation and the many kinds of reciprocal influence articulatory movements have on neighbouring ones (see coarticulation in section 2.5 above), it is likely that entire sequences of speech sounds are represented in the brain and can be executed automatically. This has been demonstrated in an experiment involving a bite-block, a small object that speakers clench between their teeth during speech production. Two groups of speakers, one with and one without a bite-block between their teeth, were asked to articulate [pi], and the activity in the relevant muscles was measured. It was found that the same muscles were activated for both groups although it is impossible to close one's lips for a [p] with a bite-block between one's teeth. The automatized commands for lip and jaw movements were activated although speakers knew that they could not move them. This experiment was interpreted to demonstrate that speakers have a mental representation of articulatory movements which is executed automatically when planning to produce a certain sequence of sounds.

Speech production involves more than the articulation of sound sequences. Speakers produce speech for a purpose, be it the sharing of some information or the expression of a feeling. This involves further cognitive processes, which are located in various specialized areas in the brain. Psycholinguistic models of speech production (e.g. Levelt 1989) describe the different cognitive processes underlying speaking. Usually, three stages of the speech production process are proposed. In the first, speakers plan what to say by drawing on linguistic knowledge, world knowledge and situational knowledge. At this stage, it is decided what is going to be talked about and how it is going to be said, but as yet on a preverbal level. The elements of the message are still only 'ideas' and are not put into words yet. This preverbal message is turned into a linguistic form in the second stage: first, the appropriate words are chosen and combined according to the syntactic and semantic requirements of the language. Secondly, they are encoded in their appropriate phonological form. We have now reached the stage of 'inner speech' or 'thinking'. In the third stage, articulation is carried out and actual sounds are produced. All models of speech production agree that **auditory feedback** is indispensable for speech production. Auditory feedback refers to the immediate control speakers have over what they are saying by being able to listen to their own speech. This monitoring of speech is used to control the accuracy of speech production. For example, when an error is detected, say a speaker notices that she has pronounced [tɪn] instead of [pɪn], she will interrupt herself and 'repair' what has been said. Speech perception is described in detail in chapter 6.

2.7 Learning the production of speech

The fact that all organs of speech have other, more primary functions shows that speaking is an ability that evolved late in the development of the human species. It appears that the process of evolution of the specialized anatomy and neural mechanisms necessary for speech production covered a period of at least 250,000 years and was probably completed 1.4 million years ago (Lieberman 1984). The specialized anatomy that developed includes a lowered larynx, which creates a longer pharyngeal cavity, leaving room for movements of the back part of the tongue, as well as a smaller tongue in relation to the oral cavity. Newborn humans have, for the first three months, a vocal tract that is shaped much more like that of our ancestors and that of other living primates: the larynx is placed higher, the tongue is larger in relation to the oral cavity and leaves little room for vertical movements and especially movements of the back of the tongue, and the epiglottis is close to the velum. This means that newborns are ill-equipped for the complicated movements underlying the articulation of speech sounds, but it has the advantage of allowing them to drink and breathe at the same time.

2.7.1 Speech production in first language acquisition (advanced reading)

The process of learning how to speak involves the acquisition of patterns of automatized muscular activity and the formation of mental representations of sounds and sound sequences. The acquisition of mental representations is described in section 3.4 below; this section focuses on the learning of the motor activities underlying speech production. Newborns have little control over their speech organs and the only sounds they can produce are cries and involuntary phonation. For cries air is pressed out of the lungs, causing vibration of the tightly held vocal folds. These sounds are usually classified as 'reflexive sounds', which arise as automatic responses to inner states such as hunger and discomfort. Phonation is thought to be involuntary and, due to the immature control over bodily functions and activities, rather occurs as a by-product of other actions.

At around two months of age babies have gained some control over their speech organs. Infants begin to **coo**, i.e. they produce voiced sounds with a resting tongue, sounds that can further be separated by glottal stops. With increasing control over laryngeal and articulatory muscles at between four and seven months of age vocal play begins. Infants try out voicing, changes in pitch and loudness as well as various manners and places of articulation. During the stage of **babbling**, which typically lasts from around six months of age to about one year, first sound sequences are produced that consist of combinations of consonant plus vowel such as [ma] and [ga]. These combinations are often produced in long sequences and, according to some researchers, already show properties of the ambient language/s (i.e. the languages that are to become the infant's native language/s) (e.g. Boysson-Bardies et al. 1984, Boysson-Bardies et al. 1992).

Around their first birthday most infants produce their **first word**, i.e. a sound sequence that is systematically associated with a particular meaning. First word productions show many phonetic differences from the corresponding adult sound sequences: for example, a child may produce [di] for *see* or [na] for *banana*. These early word productions reflect that infants usually produce only one word per breath group (**one-word stage**) and that some manners of articulation (for example approximants) and some places of articulation (for example alveolar) are more difficult to acquire than others. At around 1.5 years of age babies begin to put words into two-word sequences. This usually begins with each of the two words – for example *Mummy sock* – being produced in a separate breath group with an intervening pause. When two words are integrated into one breath group with an uninterrupted pitch movement, the so-called **two-word stage** is reached (see e.g. D'Odorico & Carubbi 2003, Behrens & Gut 2005). In the **multiple word stage** longer productions follow in which more than two words are combined into one breath group. It appears that control over the laryngeal muscles is acquired fairly early and that children can use rising and falling pitch for broadly defined linguistic purposes – e.g. attracting attention – and communicative meaning from the one-word stage on (e.g. Furrow 1984).

However, some of the complex articulatory manoeuvres that underlie the production of particular speech sounds and sequences of speech sounds, such as the lip movements connected with the production of vowels, do not reach adult standard before school age. Large studies with American children have shown that 90% of them produce [p], [m] and [w] correctly in word-initial, word-medial and word-final position at age three. The sounds [b], [k], [d] and [g] are produced correctly by 90% of all children by age four, whereas [t] and [ŋ] have only been acquired by 90% of the children after age six. The last sound to be produced correctly by 90% of all children is [s], which is acquired by age eight (see Menn & Stoel-Gammon 1995). Overall, child speech is slower than that of adults. Due to the comparative lack of practice, articulatory movements by children take longer, coarticulation is far less pronounced (e.g. Kent 1983), and the appropriate reduction and deletion of vowels and consonants in connected speech is only acquired gradually (e.g. Allen & Hawkins 1978).

2.7.2 Speech production in second language acquisition and teaching (advanced reading)

Learning the pronunciation of a second language requires the learning of exactly the same muscular patterns and the formation of the same mental representations as when learning the language as a first language. Yet, the learning processes of first and second language acquisition differ fundamentally. While first language acquisition is inextricably interwoven with the development of muscular control and cognitive abilities, most second language learners have full command over their speech organs and fully developed cognitive competence. In contrast to first language learners, moreover, second language learners have already acquired patterns of muscular activities and have formed the corresponding mental representations for the production of speech in their first language. It is generally assumed that this prior knowledge crucially influences the acquisition of the second language and leads to the 'foreign accent' many language learners exhibit in their speech. In empirical studies, numerous differences between speech produced by language learners and speech produced by native speakers have been found (see Leather & James 1996 for an overview). These include the articulation of individual speech sounds, coarticulatory movements as well as laryngeal activity. For example, German learners of English use different tongue positions for the production of the vowels [e] (in *bed*) and [æ] (in *bad*) than native British English speakers do (Barry 1989). Spanish and French learners of English produce plosives with different articulatory movements than native speakers do. In particular, the time span between the opening of the airstream blockage and the beginning of voicing is different (Flege & Eefting 1987, Laeufer 1996). Yet, a 'foreign accent' is not unavoidable for second language (L2) learners, even those that learned the L2 at an advanced age. A number of studies have demonstrated that some learners can acquire the pronunciation of an L2 so successfully that native speakers cannot identify them as learners (e.g. Bongaerts et al. 1997, Moyer 2004).

In general, speech produced by language learners is slower than that produced by native speakers. The former produce, on average, fewer sounds per second (e.g. Cucchiarini et al. 2002). This is partly due to the fact that the articulatory movements of language learners are often far more extreme than those of native speakers. Russian learners of English, for example, use less coarticulation, i.e. parallel movements of the individual articulators, than native speakers do (Zsiga 2003). Many learners of English further do not delete consonants and do not reduce vowels in the appropriate contexts and to an adequate extent: German learners of English do not often enough delete consonants in consonant clusters at the end of words and they do not reduce vowels in unstressed syllables like native speakers do (e.g. Gut 2007a). The usage of pitch also differs in a distinct way: learners of English produce smaller rises and smaller falls than native speakers, and their overall pitch range, i.e. the difference between the highest and the lowest pitch in one utterance, is much smaller (Mennen 2007, Gut 2007a).

These differences in speech production can in part be explained by the lack of practice language learners have with the articulation of some sounds and sound sequences that do not occur in their native language. It has further been proposed that language learners transfer some articulatory patterns of individual speech sounds and sound sequences from their native language to the second language. This notion of transfer has been integrated in many theories of second language acquisition. For example, the theories based on the theory of Natural Phonology (e.g. Dziubalska-Kołaczyk 1990) and Optimality Theory (e.g. Boersma 1998) claim that second language acquisition means unlearning patterns and processes that have been acquired for the first language. It is often proposed that this unlearning is obstructed by perceptual categories that speakers have formed for their first language (e.g. Flege 1995, Brown 2000). Put simply, it is assumed that speakers cannot hear the differences between sounds or linguistic usage of pitch and loudness in their first and second language and are therefore unable to produce them. The acquisition of perceptual abilities in second language learning is described in more detail in section 6.8.

The teaching of speech production in an L2 has changed over the past decades. While in traditional approaches, the focus of instruction lies on the production of individual speech sounds, recent approaches take into account that speech sounds do not appear in isolation in speech and that their production depends on the neighbouring sounds as well as on prosodic factors, such as whether a syllable is stressed or in which position a word occurs in an utterance (e.g. Trouvain & Gut 2007). Instead of practising individual speech sounds, in those new approaches speech sounds of the L2 are practised in larger units such as syllables, words and utterances (see e.g. Missaglia 2007). Moreover, variations of sounds due to coarticulatory effects or stress (for example, vowel reduction in English) are taught as well.

2.8 Methods of researching speech production (advanced reading)

Our knowledge of the way in which the speech organs are involved in the production of speech sounds is fairly recent. It is only since the 1930s that technical devices have become available for studying the physiology of speech production in detail. A few of the currently most commonly used ones will be described in this section. The functions of the organs of the respiratory system cannot be studied directly. In order to measure the direction of airflow as well as the air pressure in the oral or nasal cavities, a method called **aerometry** is applied. A mask is put over the speaker's nose and mouth, which is equipped with transducers that convert the air pressure into electrical signals, which in turn can be recorded onto a computer.

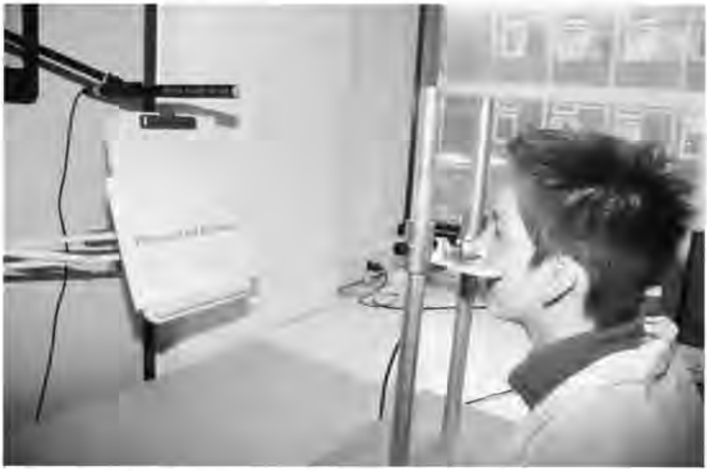


Figure 2.12. A speaker with a nasometer.

To examine nasal airflow a **nasometer** is used (see Figure 2.12). Small microphones separated by a metal plate are put on a series of straps and are either fitted to the speaker's head or attached to poles. The upper microphone records air flowing from the nose, whereas the lower one records air flowing from the oral cavity. This enables researchers to measure the relationship between both types of airflow, which is for example important for the description of speech produced by cleft palate patients (see section 2.3 above).

Phonation, the movement of the vocal folds, can be examined by various means. For an indirect measurement of vocal fold activity a **laryngograph** is used. Two electrodes are put on the speaker's throat on either side of the thyroid cartilage. A weak current is passed between the two electrodes, which measures how often the vocal folds touch. The degree of current reflects the contact between the vocal folds so that vocal fold vibration can be shown in a waveform. This type of measurement gives information about voicing, modes of phonation, pitch and other characteristics of voice quality. Direct observation of the vocal folds is also possible. They can either be seen with the help of a **laryngoscope**, which is an angled rod with a mirror at its end, or by using an **endoscope**, a tube that is fitted with a light source and is connected to a recording device. A very high film speed is needed to capture the rapid vocal fold vibration. Both laryngoscope and endoscope are inserted into the mouth and held directly above the larynx, which of course means that no speech sounds can be produced that involve significant tongue movement. The videos 02_vocalfolds.gif and 02_glissando.mov on the CD-ROM were recorded with an endoscope. The methods of laryngoscopy and endoscopy provide important information about whether a speaker's vocal folds are healthy. Many teachers experience problems with their voice during their career (e.g. Roy et al. 2004); the examination of the vocal folds and of vocal fold activity shows whether the ailment has an organic basis.

The action of the articulators in the articulatory system can be examined in a number of different ways. X-ray techniques, which display the movement of the articulators without hindering the speaker in any way, have not been applied anymore since the harmfulness of the method was discovered. The video 02_connectedspeech.mpg on the CD-ROM shows an X-ray recording that was done in 1974. A new method of measuring the exact movements of the tongue and lips during articulation is called **electromagnetic mid-sagittal articulography,** or EMMA. It allows the analysis of articulatory movements within the oral cavity with high time resolution. For an EMMA study, transducers, which are connected to an amplifier with wires, are placed on a speaker's tongue, lips and nose. The speaker additionally wears a helmet in which the transducers create a magnetic field (see Figure 2.13). The measurement of the alternating currents allows the calculation of the degree and velocity of tongue and lip movements during articulation.

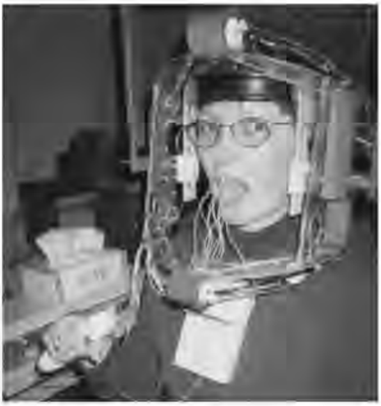


Figure 2.13. A participant in an EMMA experiment.

A device that records the movement of the tongue is the **palatograph**. For this, a thin artificial palate fitted with a large number of electrodes is put onto the speaker's palate. The electrodes fire when they come into contact with the tongue so that the position and degree of tongue contact can be measured with the help of an attached computer. Palatographic studies show that the patterns of articulatory movements can be quite different for various speakers producing the same sound. Figure 2.14 shows the place and strength of contact of the tongue with the roof of the mouth during the production of [t] and [d] for two different speakers. The upper part of the picture illustrates the region of the alveolar ridge, the bottom part the region of hard palate. Black areas signify strong contact; grey areas were only lightly touched by the tongue, and white areas not at all. The left-hand palatogram shows that the speaker's tongue has strong contact with the alveolar ridge and both sides of the palate, whereas the speaker whose palatogram can be seen on the right articulates [t] and [d] mainly with tongue contact at the alveolar ridge.

The activity of the brain during speech production cannot be studied directly. However, with the technique of **electroencephalography** (EEG), the electrical activity of the brain can be recorded in an indirect way. Electrodes placed on the scalp measure voltage differences between different areas of the brain that occur during speech tasks. It is much easier to examine speech perception than speech production with this method since the brain activities involved in articulation complicate the interpretation of the results. **Functional magnetic resonance imaging** (fMRI) is another method for the examination of brain activity during speech tasks. Humans taking part in an fMRI experiment lie on a stretcher-like device and are put into a cylinder-shaped tube where they have to perform different tasks such as listening to language or seeing pictures. The blood flow, which reflects neural activities in the different areas of the brain, is measured by the magnetic resonance reflecting the level of oxygen in the blood.

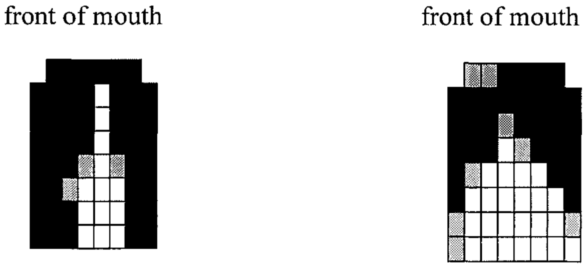


Figure 2.14. Palatograms of [t] and [d] for two speakers.

Finally, the activity of all muscles involved in the production of speech can be examined. In order to measure which muscles are active at which stage of sound production, the technique of **electromyography** (EMG) is applied. With the help of electrodes, which are either inserted into the relevant muscle or placed on the skin above the muscle, it measures the tiny electrical charges that muscles produce when activated. However, raw electromyographic data are difficult to interpret since the activity of a particular muscle must be compared to that of others and the total of muscular patterns. Thus, large numbers of measurements are usually averaged across utterances and across speakers.

2.9 Exercises

1. What is the main function of the respiratory system of speech production?
2. Which organs are involved in phonation and what is their function?
3. List all articulators that make up the articulatory system.
4. Which places and manners of articulation exist for consonants?
5. Why is it difficult to describe the production of a particular speech sound in real connected speech?
6. Which role do neural mechanisms in the brain play in speech production?
7. Why is child speech and the speech produced by second language learners slower than that of adult native speakers?

8. Listen to recording 02_exercise8.wav on the CD-ROM. Mark the breath group boundaries in the text.
9. Ask at least 20 people how they produce the sound [t]. Which articulator does their tip of the tongue touch?

2.10 Further reading

More details on the anatomy and physiology of speech production can be found in Clark, Yallop & Fletcher (2007, chapters 2 and 6) and Lieberman & Blumstein (1988, chapters 2 and 6). For especially interested readers, Laver (1994, chapters 6 to 11) and Ladefoged (2001a, chapters 11 and 12) provide very detailed accounts of the different articulations of English vowels and consonants. Ladefoged and Maddieson (1996) describe the distribution of the different types of articulation in the languages of the world. Lieberman (1984) gives more information on the evolution of speech in humans. The theory of articulatory gestures is explained in Browman & Goldstein (1992). A comprehensive description of the development of speech production abilities in first language acquisition is given in Vihman (1996, chapter 5). Second language speech production is described further in Leather & James (1996) and Zampini (2008).

3 The Phonology of English: Phonemes, Syllables and Words

This chapter and chapter 4 deal with the phonology of English. Chapter 1 explained that phonologists are concerned with **units of sound structure** and sets of **phonological rules** that describe patterns and regularities of the sounds in a particular language. These phonological units and rules are assumed to be stored in a speaker's brain, which enables him or her to produce and understand speech. Which knowledge do speakers have about the phonological units of a language? Speakers will tell you, for example, that words consist of individual speech sounds. Some might even point out that these speech sounds can further be grouped into syllables. Moreover, all speakers would agree that words themselves can be grouped into utterances. The units of speech thus seem to be organized as a hierarchy with larger units (such as words) containing smaller units (such as syllables and speech sounds). Speaker knowledge of phonological units can be illustrated in a **prosodic hierarchy** (e.g. Selkirk 1986, Nespor & Vogel 1986). Figure 3.1 shows the prosodic hierarchy of the (British English pronunciation of the) utterance "This figure". This utterance consists of the two words *this* and *figure*. The word *this* comprises one syllable, whereas the word *figure* contains two syllables (each represented by the symbol σ). Each of the syllables is made up of a number of individual speech sounds. *This* consists of three speech sounds, *fi* and *gure* of two each. The term **prosody** refers to those phonological units that are on a higher level than the speech sounds and is thus another term for **suprasegmental phonology** (see chapter 1). **Segmental phonology** is concerned with the units and phonological rules of the lowest level of the prosodic hierarchy – the speech sounds.

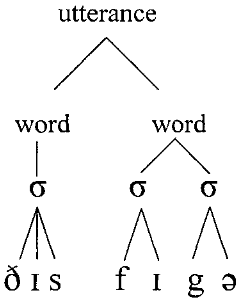


Figure 3.1. Some levels of the prosodic hierarchy: speech sounds (phonemes), syllables and words of the utterance "This figure".

Section 3.1 describes the smallest of the phonological units in English: the phonemes. It explains both their articulatory properties as well as the way they are transcribed in phonological and phonetic analysis. Section 3.2 is concerned with syllables in English and describes their structure and patterns. The phonological properties of English words, including word stress, are presented in section 3.3. Section 3.4 describes how phonological representations are acquired in both first and second language acquisition. The higher levels of the prosodic hierarchy such as phrases, utterances and discourse are dealt with in chapter 4.

3.1 The phonemes of English

Section 2.5 has shown that speech sounds are articulated in different ways depending on their neighbouring sounds (this is called the **phonetic context**) and the general speed of articulation. In English, for example, the voiceless bilabial plosive /p/ in the word *pin* is **aspirated**. This means that between the opening of the closure formed with the lips and the beginning of vocal fold vibration for the vowel /ɪ/ there is a short time interval in which the airstream passes out of the mouth with an audible hiss. In the word *spin*, however, due to the influence of the preceding /s/, the /p/ is not aspirated. Moreover, for the articulation of /p/ at the end of a word (for example in *stamp*) or before another plosive (as in *captain*), the pulmonic airstream ends before the lips are opened – the /p/ is **unreleased**. Similarly, other speech sounds in English can have very different types of articulation depending on the phonetic context. The /d/ in *dish*, for example, is produced with the tip of the tongue touching the alveolar ridge. When a /d/ appears before an interdental sound as in the word *width*, however, it has a place of articulation that is dental, which means that the tip of the tongue touches the back of the teeth.

Speakers of English are normally not aware of the fact that speech sounds can be articulated in very different ways. They do not pay attention to these articulatory differences because they are not functionally relevant in English. Despite producing and hearing many different kinds of voiceless bilabial plosives including unaspirated [p], aspirated [p^h] and unreleased [p˺], speakers have only one 'mental image' of /p/. This mental representation of a specific speech sound is called **phoneme**. The actual speech sounds speakers produce and listeners hear are called **phones**. The phoneme /p/ in English thus includes the phones [p], aspirated [p^h] and unreleased [p˺]. The phoneme /d/ includes the alveolar and the dental voiced plosive. (Remember that the slashes / / are used to indicate the mental representation of a speech sound and the square brackets [ ] are used when we refer to an actual pronunciation). Some phoneticians, in contrast to phonologists, doubt whether speakers really use mental representations of sounds in speech production and speech perception. See section 2.7.2, which discusses whether the phoneme is a useful concept in language teaching.

Not every phone has the function of a phoneme in every language. Phones only have the role of phonemes in a language when they have a **contrastive function**, which means that they cause changes in meaning. A phoneme is thus defined as the smallest meaning-distinguishing unit in a language. Whether a phone has phonemic status in a language or not can be shown with the **minimal pair test**. If you take the voiceless interdental fricative /θ/ in the English word *think* and replace it by the voiceless alveolar fricative /s/, you have changed the meaning of the word to *sink*. *Think* and *sink* thus form a minimal pair in English. The minimal pair test, moreover, shows that /θ/ does not have the function of a phoneme in German. If you replace the /s/ in *Lust* with a /θ/, the meaning of the word does not change – listeners will merely assume that you have a lisp.

Typically, a phoneme of a language can be realized as different phones. As described above, the /p/ in English can be produced as the aspirated [p^{h}]. Alternatively, it can be produced without aspiration as the phone [p] or without audible release as the phone [p˺]. Equally, the /d/ can be produced as an alveolar [d] or as a dental voiced plosive [d̪] depending on the phonetic context. The phones that are realisations of the same phoneme are called **allophones**. Thus, [p], [p^{h}] and [p˺] are allophones of the phoneme /p/, and alveolar [d] and dental [d̪] are allophones of the phoneme /d/ in English. The distribution of the allophones of a phoneme in speech is usually **complementary**. This means that each allophone occurs exclusively in one specific phonetic context. This is, for example, the case for the allophones of /p/ in English. The complementary distribution of the allophones of /p/ can be described by the following phonological rule: [p^{h}] occurs when /p/ is the only consonant at the beginning of a stressed syllable as in the word *picture*. The allophone [p˺] occurs before other plosives – as in the word *captain* – or at the end of an utterance. The allophone [p] occurs in other positions, such as in the word *spin* (see section 3.1.2 below for a list of more allophones of English consonants).

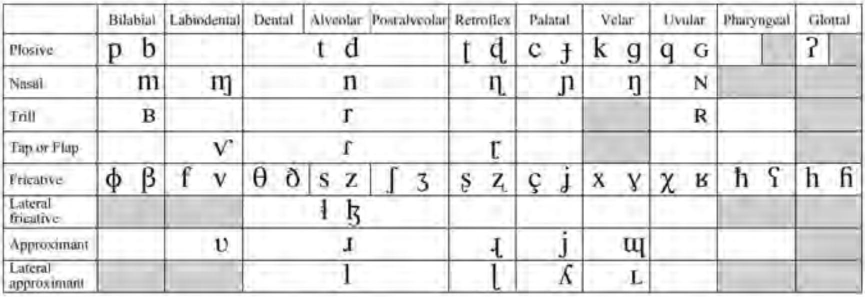
Some allophones have also been claimed to be distributed in **free variation**, i.e. without any systematic distribution according to phonetic context. This seems to be the case for the different realizations of /r/ in German. The alveolar trill (which sounds like the Italian /r/), the uvular trill (a sound like gargling) and the voiced uvular fricative (the /ʁ/ of Standard High German) can occur in every position of German words. Yet, speakers do not produce these different allophones randomly: the choice of allophone seems to be a matter of regional dialect. Bavarian speakers, for example, tend to use only the alveolar trill, whereas speakers from Hannover use the voiced uvular fricative in all places. Many other apparent cases of free allophonic variation can be described as different **stylistic variants** of phonemes. This means that the choice of a particular allophone depends on who the speaker is talking to and on the speaking situation. In British English, for example, word-final /t/ tends to be produced as a glottal stop [ʔ] – for example in the word *what* – by young speakers rather than older speakers, and they do so more in informal situations than in formal situations (Fabricius 2002). Similarly, in slow and careful speech in English, the /t/ at the end of words is usually released, whereas in casual or fast speech it tends to be unreleased.

The phones that function as phonemes in a language make up the phonemic system or **phoneme inventory** of a language. A description of the phoneme inventory of English, however, is far from trivial. English is spoken all over the world and often the different national varieties as well as the different regional (and even social) varieties that are spoken within a country differ widely. Thus, the phoneme inventory of Jamaican English does not match the phoneme inventory of Scottish English, although many phonemes will be shared by the two. For the interested reader, the phoneme inventories of many world-wide varieties of English are described in Kortmann & Schneider (2004). This textbook focuses primarily on the phonemes of Standard British English and Standard American English since these are the most widely taught varieties of English in second language teaching. The term **standard** refers to the variety of English that has the highest prestige in a country, that is the variety described in grammars and dictionaries and whose rules are taught at school. On the British Isles, the standard is variously called **Received Pronunciation** (**RP**), BBC English, Oxford English or Southern British Standard. The standard variety of the U.S.A. is referred to as **General American** (**GA**). Only about 5% of the population on the British Isles speak RP, and the majority of RP speakers originate from or live in the south-east of England, have a middle-class or upper-class background and a high level of education. RP is thus more of a social than a regional accent. General American is more wide-spread in the U.S.A. (approximately half of the speakers there are GA speakers). The term GA refers to a group of accents that does not bear any marked regional characteristics and is the accent most commonly used in radio and television in the U.S.A.

The phoneme inventory of a language is usually divided into **consonants** and **vowels**. Section 2.4 explained that this classification is partly based on the manner of articulation of the different speech sounds. There are furthermore phonological reasons for making this distinction. Consonants and vowels are distributed in a distinct way. Some positions in a syllable can only be filled by vowels, other positions are always occupied by consonants in English (see section 3.2 below). Interestingly, the division of speech sounds into vowels and consonants is not the same across languages. For example, /ɹ/, which is considered a consonant in English, is classified as a vowel in some dialects of Chinese.

3.1.1 The consonants of RP and GA and their transcription

The phoneme inventory of RP and GA comprises 24 phonemes. In order to be able to represent phonemes in writing, the **International Phonetic Alphabet (IPA)** was developed. It contains transcription symbols for all distinctive speech sounds that occur in any language of the world. In addition, it offers transcription symbols for fine phonetic details and prosodic features, the so-called **diacritics**. The IPA has a long history: in 1886, the International Phonetic Association, consisting of language teachers who wanted to devise a phonetic notation in order to support language learning, was founded in Paris. It distributed the first phonetic alphabet in the late 19th century. The symbols employed were chosen with the aim of keeping them as simple as possible for language learners in Western Europe so that most of them were taken from the Roman alphabet. In the past centuries, the IPA has undergone several revisions (the last one was completed in 2005).



	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			r				▨	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		▨	▨

Figure 3.2. The IPA transcription symbols for the pulmonic consonants. Reprinted with permission from The International Phonetic Association. Copyright 2005 by International Phonetic Association.

Figure 3.2 presents the IPA symbols for all pulmonic consonants, i.e. consonants produced with an egressive pulmonic airstream (see section 2.1). The chart can be read in the following way: each row represents a different manner of articulation, ranging from plosive to lateral approximant. Each column refers to a different place of articulation, moving from the lips (bilabial) to the larynx (glottal). Each cell of the chart thus represents a particular combination of a manner and a place of articulation. When two symbols appear in one cell, the one on the left is the symbol for the voiceless consonant and the one on the right is the symbol for its voiced counterpart. Thus, [p] is the symbol for the voiceless bilabial plosive and [b] is the symbol for the voiced bilabial plosive. Cells that are shaded grey signify combinations of place and manner of articulation that are physiologically impossible. Empty cells hold possible combinations of place and manner of articulation, but no such sound has yet been discovered in the languages of the world.

Table 3.1. The consonants of RP and GA.

Phonetic symbol	Examples	
	word-initial	word-final
p	*pin*	*rip*
b	*bin*	*rib*
t	*tin*	*sit*
d	*din*	*rid*
k	*kin*	*Rick*
g	*go*	*rig*
m	*mine*	*rim*
n	*nine*	*Rhine*
ŋ	--	*ring*
f	*fin*	*riff*
v	*vase*	*live*
θ	*thin*	*heath*
ð	*this*	*with*
s	*sin*	*rice*
z	*zoo*	*rise*
ʃ	*shin*	*wish*
ʒ	*genre*	*beige*
h	*him*	--
ɹ	*rim*	*beer* *(GA only)*
j	*year*	--
l	*Lynn*	*bile*
w	*whim*	--
tʃ	*chin*	*hatch*
dʒ	*gin*	*lodge*

Of all the consonants depicted in the IPA chart in Figure 3.2, RP and GA use 24. These are listed in Table 3.1. The phoneme inventory of these two standard varieties of English comprises six plosives: two bilabial (/p/ and /b/), two alveolar (/t/ and /d/) and two velar (/k/ and /g/) ones; one of each voiced and the other voiceless. Furthermore, RP and GA have three nasals, namely the bilabial /m/, the alveolar /n/ and the velar /ŋ/. Of the fricatives, there are nine in RP and GA: a voiceless and a voiced labiodental (/f/ and /v/) fricative, a voiceless (/θ/) and a voiced (/ð/) interdental fricative, the voiceless alveolar /s/ and the voiced alveolar /z/ as well as the voiceless and the voiced postalveolar /ʃ/ and /ʒ/. In addition, there is the phoneme /h/, a voiceless glottal fricative. Both RP and GA have three approximants, the alveolar /ɹ/, the palatal /j/ and the alveolar lateral approximant /l/. In addition, the phoneme /w/ occurs, which is characterized as a labiovelar approximant with two places of articulation (it therefore does not fit into the IPA chart illustrated in Figure 3.2). Lastly, there are two affricates that are sometimes counted as phonemes of RP and GA because they have meaning-distinguishing function: the combination of /t/ and /ʃ/ and the combination of /d/ and /ʒ/.

Table 3.1 lists examples of these phonemes occurring at the beginning of a word (= word-initially) and the end of a word (= word-finally). It is easy to see that not all phonemes can occur at the beginning of words (there are no English words or syllables beginning with /ŋ/) and that not all of them occur at the end of words or syllables (the phonemes /h/, /j/ or /w/ as well as /ɹ/ in RP do not occur in this place). /ʒ/ is a very rare consonant that only occurs in a few words and does so mostly medially, i.e. in the middle of a word as in *plea̲sure*. This can be explained with the history of English. This speech sound only emerged relatively recently, during the Early Modern English period (roughly 1400 to 1650), and only occurs in words that were originally borrowed from French or Latin. The alveolar approximant /ɹ/ occurs in word- and syllable-final position only in GA but not in RP (see section 3.2.1 for further details). Again, this difference is due to a relatively recent development in RP. While the English spoken in Shakespeare's time and consequently that which was transported to the first settlements in Northern America had /ɹ/ in all positions of the word (it was most likely realized as the alveolar trill [r]), /ɹ/ was later dropped in specific positions in most varieties of English spoken on the British Isles, including RP.

Many phonologists claim that /ŋ/ does not constitute a phoneme in its own right but should rather be analysed as an allophone of /n/. Historically speaking, this view is true since up to about the 17th century /ŋ/ was indeed an allophone of /n/ that occurred when /n/ was followed by a /g/ or /k/. In Shakespearean times, thus, the word *sing* was mostly pronounced as /sɪŋg/ and only few speakers began to pronounce it as /sɪŋ/ as we do today. For an analysis of present-day English, however, it can be argued that since minimal pairs like *sin* and *sing* exist, /ŋ/ should be treated as a phoneme in its own right.

3.1.2 The allophonic variation of consonants in RP and GA

All of the phonemes of English listed in Table 3.1 can be realised as several different allophones, whose distribution is either determined by the phonetic context or varies for stylistic reasons. When the variation of allophones is determined by the phonetic context, their distribution can be described by allophonic rules. The major allophonic rules of English are listed in Table 3.2. In order to be able to transcribe, i.e. represent in writing, the articulatory differences between allophones, the IPA includes a list of diacritic symbols (see Figure 3.3). These **diacritics**, when added to a phonetic symbol, indicate additional articulatory details of the production of a particular speech sound. Diacritic symbols can therefore be used to describe the allophonic variation of phonemes.

Table 3.2 shows that, in both RP and GA, the voiceless and voiced plosives have a high number of allophones in complementary distribution. When /p, t, k/ occur as the only consonant before a stressed vowel, they are aspirated and transcribed as [$\mathrm{p^h}$], [$\mathrm{t^h}$] and [$\mathrm{k^h}$]. The voiceless plosives are therefore aspirated in *pea*, *tea* and *key* but not in *spring*, *steam* or *clay*. This allophonic variation can be formulated as a phonological rule. Rule (1) illustrates in which phonological context /p, t, k/ are aspirated:

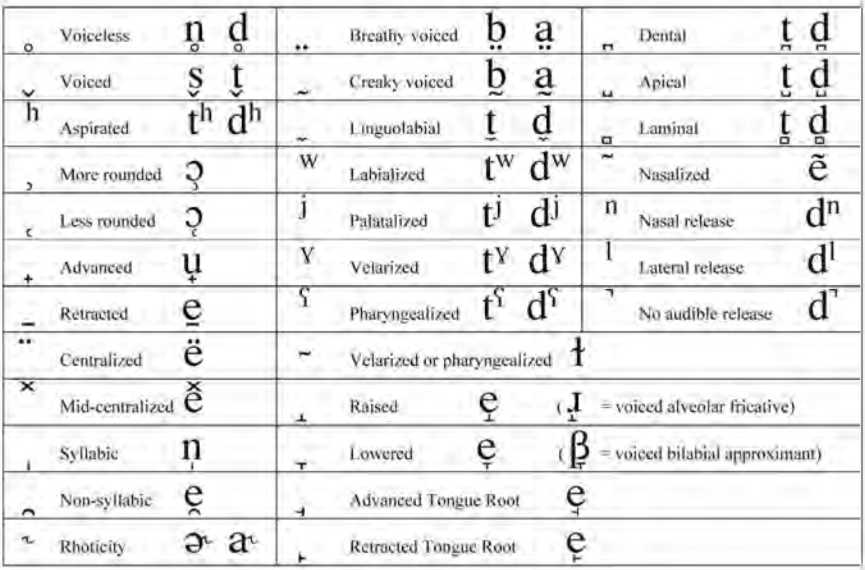
(1)
$$\left.\begin{array}{l} /\mathrm{p}/ \\ /\mathrm{t}/ \\ /\mathrm{k}/ \end{array}\right\} \rightarrow \left\{\begin{array}{l} [\mathrm{p^h}] \\ [\mathrm{t^h}] \\ [\mathrm{k^h}] \end{array}\right\} /.\_$$

It expresses that the phonemes /p, t, k/ are aspirated when they are not preceded by another segment in the syllable. The arrow → indicates a phonological process, the slash (/) gives the condition (when) and the dot . stands for the boundary of a syllable. The _ refers to the position of the phonemes the allophonic rule applies to. When the phonemes /p, t, k/ occur at the end of a syllable after a vowel, they are often **glottalized**, i.e. accompanied or slightly preceded by a glottal stop. Thus, the word *cup* is pronounced [kʌ?p] and *hit* is pronounced [hɪ?t]. When followed by another oral plosive, as for example in the word *captain* or the phrase *good day*, the plosives /p, t, k, b, d, g/ are unreleased. This means that there is no airstream coming from the lungs when the closure formed by the tongue and the other articulator is opened. The diacritic transcription symbol for an unreleased consonant is [˺]. The word *captain* is thus usually pronounced [kæp˺tn] and *good day* is usually pronounced [gʊd˺deɪ]. This phonological process also happens regularly at the end of utterances, for example in "He's a good chap", where *chap* is produced [tʃæp˺]. In all other positions, these plosives have no aspiration, i.e. audible release of air after the opening of the closure and before the onset of vocal fold vibration for the vowel.

Table 3.2. Allophonic variation of some consonants of RP and GA.

/p,t,k/	[pʰ], [tʰ], [kʰ]	as only consonant before stressed vowel
	[ʔp], [ʔt], [ʔk]	syllable-final after vowels
	[p], [t], [k]	elsewhere
/p,t,k,b,d,g/	[p̚], [t̚], [k̚], [b̚], [d̚], [g̚]	before other plosives and at the end of an utterance
	[pʷ], [tʷ], [kʷ], [bʷ], [dʷ], [gʷ]	before rounded segment
/k,g/	[k̟], [g̟]	before front vowel
	[k], [g]	elsewhere
/b,d,g,z,v,ð/	[b̥], [d̥], [g̊], [z̥], [v̥], [ð̥]	next to voiceless sounds, before and after silence
	[b], [d], [g], [z], [v], [ð]	elsewhere
/t,d/	[t̪], [d̪]	before interdental consonant
	[ɾ]	as only consonant at beginning of an unstressed syllable when preceded by a vowel or a sonorant consonant (GA only)
/t/	deleted	in unstressed syllable after /n/ (GA only)
	[ʔ]	in syllable-final position
/n/	[n̪]	before a dental fricative
	[nː]	before voiced obstruent in the same syllable
	[n]	elsewhere
/m/	[ɱ]	before labiodentals
	[m]	elsewhere
/l/	[ɫ]	after vowel
	[l̥]	after voiceless plosive
	[l̪]	after dental consonant
	[l]	elsewhere
/ɹ,w,j/	[ɹ̥], [w̥], [j̊]	after a voiceless plosive
	[ɹ], [w], [j]	elsewhere

The plosives in RP and GA show further allophonic variation (as shown in Table 3.2). When /p, t, k, b, d, g/ occur before a sound produced with rounded lips such as /w/ and /u/, they are produced with lip rounding as well. For example, when articulating words such as *quest* and *cool*, speakers have rounded lips for the production of the first phoneme /k/. This lip rounding or **labialisation** is transcribed as [ʷ]. Similarly, the /g/ in *good* as well as the /t/ in *two* are rounded, which is transcribed as [gʷ] and [tʷ], respectively.



◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized ɫ				
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

Figure 3.3 The IPA transcription symbols for phonetic details: the diacritics.

Furthermore, the velar plosives /g/ and /k/ have different places of articulation depending on the following vowel. The /k/ in *kit*, *cat* and *cot* has three different places of articulation, which you can feel when you articulate these words slowly. When followed by a front vowel, for example in *kit*, the /k/ is fronted as well, which is transcribed as [k̟]. When followed by a central vowel as in *cat*, the place of articulation is palatal. When preceding a back vowel like in *cot*, /k/ is produced as a velar plosive (see section 3.1.3 for a description of the different types of vowels).

The voiced obstruents /b, d, g, z, v, ð/ in RP and GA are sometimes partly or fully devoiced, i.e. produced without vocal fold vibration. This depends on their phonetic environment: when they are preceded and followed by voiced sounds, they are always fully voiced, but when they are preceded or followed by voiceless sounds or silence (at the end or the beginning of an utterance), they are usually partly or fully devoiced. For example, the /b/ in *obtuse* is partly devoiced as well as the /v/ in *dovetail*. These words are transcribed [əb̥tjuːs] (RP), [əb̥tuːs] (GA) and [dʌv̥teɪɫ] respectively. Devoicing is indicated by the subscript circle under the IPA symbol. If you wanted to express this as an allophonic rule you would formulate it as in (2):

(2)

$$/\text{b,d,g,z,v,ð}/ \rightarrow \begin{cases} \text{[b], [d], [g], [z], [v], [ð] / [+voice] \_ [+voice]} \\ \text{[b̥], [d̥], [g̊], [z̥], [v̥], [ð̥] / elsewhere} \end{cases}$$

This rule states that the voiced obstruents /b, d, g, z, v, ð/ are realized as [b], [d], [g], [z], [v] and [ð] only when they are in the position between two voiced segments (the meaning of the symbol [+voice] is explained in section 3.1.5 below); in all other positions they are devoiced. This allophonic rule implies that there is very little difference in the articulation of /k/ and /g/ in *leek* and *league*, between the articulation of /t/ and /d/ in *right* and *ride* and the articulation of /p/ and /b/ in *rope* and *robe*. In fact, the major difference between these minimal pairs lies in the glottalisation of /p, t, k/ and the length of the preceding sonorant (vowel, nasal or lateral). Before a 'voiced' plosive a vowel, nasal or /l/ is considerably longer than when preceding a voiceless plosive. Compare *wide* and *white*, *send* and *sent* and *felt* and *felled.*

The alveolar plosives /t/ and /d/ are produced with the tip of the tongue touching the back of the teeth when a segment that has an interdental place of articulation follows. This is the case in *sit there* and *width*, where the alveolar plosives are produced with a dental place of articulation. This is transcribed as [t̪] and [d̪], respectively. One allophonic rule that only applies to GA concerns the phonemes /t/ and /d/. When they occur as the only consonant at the beginning of an unstressed syllable and have either a vowel or a sonorant consonant preceding them, these phonemes are realized as an alveolar tap [ɾ]. This phonological process, which is referred to as **flapping** (the term tapping would be more correct but is used less often), can be observed in words such as *letter*, *writing* and *startle*, which are pronounced [lɛɾɹ], [ɹaɪɾɪŋ] and [stɑɹɾɫ], respectively. This means that for many GA speakers the words *latter* and *ladder* are homophones, i.e. are pronounced in nearly the same way. This rule also applies across word boundaries as in phrases like *at all* [əɾɔl] and *eat up* [iɾʌʔp].

Another phonological process that occurs only in GA is the deletion of /t/ in unstressed syllables when following a /n/. This phenomenon can be observed in words like *winter* and *rental* which are pronounced [wɪnəɹ] and [ɹɛnəl]. This phonological process is not an allophonic process in its strict sense because the phoneme /t/ is not realized at all. In both RP and GA, the phoneme /t/ is often produced as a glottal stop [ʔ] when it appears at the end of a syllable. This is the case in words such as *batman* [bæʔmən] and *atlas* [æʔləs] as well as across phrases like *hit me* [hɪʔmi]. As mentioned above, this allophonic variation occurs mainly in informal styles, and younger speakers produce it more frequently than older ones.

Table 3.2 further illustrates that the phoneme /n/ also has allophonic variants: when followed by a dental fricative such as [θ] in *tenth*, the place of articulation of the nasal is dental rather than alveolar and is transcribed with the symbol [n̪]. Before a voiced obstruent in the same syllable, such as in the words *lend* or *lens*, the alveolar nasal is lengthened and produced as [n:]. The bilabial nasal /m/ is usually produced as a labiodental nasal [ɱ] when it is followed by a labiodental segment as in the words *Humphrey* [hʌɱfɹɪ] and *emphatic* ([eɱfætɪk] in RP and [ɛɱfætɪk] in GA). This can even be noted across phoneme boundaries in phrases such as *come forward*, which is pronounced [kʌɱfɔ(ɹ)wəd].

In RP, /l/ has four different allophones: when it occurs at the beginning of a syllable as in *leap* or between vowels as in *stealing* and *steal it*, it is realized as the alveolar lateral approximant [l] (sometimes called **'clear l'**). When preceded by a voiceless plosive, however, as in the word *sleep*, it is realized without voice, and is transcribed as [l̥]. When followed by an interdental fricative as in *wealth*, the /l/ is produced with a dental place of articulation – hence the pronunciation [wel̪θ] in RP and [wɛl̪θ] in GA. At the end of a syllable, as in *pill* or *fold*, the lateral approximant is typically realized with the tongue body retracted towards the uvula or raised towards the velum (sometimes the tip of the tongue does not even touch the palate anymore). It is 'velarized' (sometimes called **'dark l'**) and transcribed as [ɫ]. Many American English accents, in contrast to most British English accents, have the allophone 'clear l' in only one position – when it occurs together with another consonant at the beginning of a word as in *clean*. In all other positions, such as between vowels (as in *healing*) or at the end of words (as in *feel*), the /l/ is velarized.

Just as for /l/, devoicing, i.e. voiceless realization when preceded by a voiceless consonant, occurs also for the phonemes /w, j, ɹ/. This is, for example, the case in words like *twin*, *crawl* and *tune* (RP only), which are produced [tw̥ɪn], [kɹ̥ɔl] (GA: [kɹ̥ɑl]) and [tj̊un].

3.1.3 The vowels of RP and GA and their transcription

The vowel systems of RP and GA have many phonemes in common but differ in some respects. In total, RP has 23 phonemic vowels, whereas GA has 16. Traditionally, vowels are divided into **monophthongs** (one vowel, also sometimes referred to as 'pure vowel'), **diphthongs** (two vowels in a sequence) and **triphthongs** (a sequence of three vowels). The word *see* contains the monophthong /i/, the word *say* contains the diphthong /eɪ/, whereas the RP pronunciation of the word *fire* contains a triphthong – the sequence /aɪə/. GA does not have triphthongs, which has to do with the fact that it has coda /ɹ/ (see section 3.2.1 for details). True monophthongs, however, are difficult to find in English – when analysing speech instrumentally as described in chapter 5, one can see that many English vowels do not have stable qualities but are always slightly diphthongized.

Figure 3.4 gives the IPA transcription symbols for vowels. They are usually presented in a **quadrilateral** or **vowel chart** that roughly illustrates the shape of the oral cavity. On the vertical axis, the vertical position of the tongue and lower jaw is represented, ranging from **close** (or high) to **open** (or low). The horizontal axis refers to the part of the tongue that is active during articulation, comprising **front**, **central** and **back**. The location of the vowel /i/, which occurs in *bee*, in the vowel quadrilateral thus shows that it is produced with the tongue in a high position (the mouth is nearly closed) and with the front of the tongue raised in the front of the mouth. For the production of the vowel /o/, by contrast, the mouth is mid-open with a mid-low tongue that arches at the back. Most tongue positions in the IPA vowel chart are filled by two symbols. In these cases, the vowel on the right is produced with rounded lips (for example /u/) and the one on the left is produced with spread lips (for example /i/).

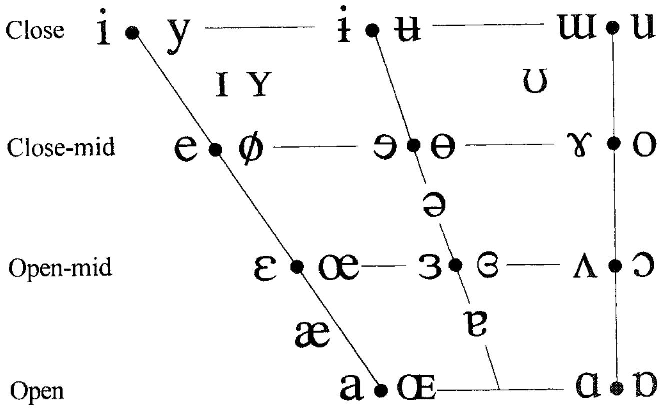


Figure 3.4. The IPA transcription symbols for the cardinal vowels.

Unfortunately, the articulation of vowels cannot be described as neatly as the articulation of consonants (see also sections 2.3 and 2.4). The difference between the front and the central part of the tongue or a mid-closed and a mid-open mouth, for example, is not as absolute and clear-cut as the differences in articulation between a plosive and a nasal. Rather, differences in vowel articulation are gradual and have no clear boundaries. This means that the articulatory descriptions of the vowels in the quadrilateral in Figure 3.4 are idealized. The positions of tongue and jaw that are indicated are extreme positions which are rarely reached in real articulation. These idealized vowels are referred to as **cardinal vowels**. Their primary function is to serve as reference values for phoneticians and phonologists describing the vowel inventory of a language. Thus, it happens often that the descriptions of the same language by different authors do not match entirely. Together with the fact that the vowel inventories of the different varieties of English are continuously changing (cf. Hawkins & Midgley 2005 for RP), slight differences between descriptions of the vowel inventory of a variety of English may always occur.

Table 3.3 lists the vowel inventories of RP and GA as they are described by most phonologists. In terms of monophthongs, both varieties of English have the unrounded high (or close) front vowels /i/ and /ɪ/, the rounded high back vowels /u/ and /ʊ/, the unrounded mid central vowels /ɜ/ and /ə/, the rounded mid-open back vowel /ɔ/, the unrounded mid-open front vowel /æ/, the unrounded mid-open central vowel /ʌ/ and the unrounded low back vowel /ɑ/. GA furthermore has the mid-low front unrounded /ɛ/. RP has the unrounded mid-high front vowel /e/ (although in some descriptions you will find /ɛ/ instead of /e/) and the low back rounded monophthong /ɒ/. Some phonologists claim that the vowel /ɔ/ in GA is increasingly being replaced by the unrounded low back vowel /ɑ/ in words such as *caught*.

As far as the diphthongs are concerned, GA and RP share /eɪ/, /aʊ/, /aɪ/ and /ɔɪ/, which are called **closing diphthongs** because the second vowel /ɪ/ or /ʊ/ is more closed than the first. Where GA speakers use the diphthong /oʊ/, RP speakers have the diphthong /əʊ/, which starts higher and more central than the corresponding diphthong in GA. Only RP has the diphthongs /ʊə/, /eə/ and /iə/, which are referred to as **centering diphthongs** because they end in the central vowel /ə/. Their occurrence in RP has to do with the loss of /ɹ/ in syllable-final position in this variety, as will be explained in section 3.2.1 below. Triphthongs such as /aɪə/ only occur in RP. As with the centering diphthongs, this has to do with the loss of /ɹ/ in syllable-final position in the history of Standard British English. According to Table 3.3, GA has the diphthong /eɪ/, which occurs in words like *fate*. Many GA speakers, however, tend to produce these words with the monophthong /e/ as in [fet]. Similarly, some speakers produce words like *bode* with the monophthong /o/ as in [bod], whereas others will pronounce these words with a diphthong /oʊ/ as in [boʊd].

Listen to the vowels of RP (03_vowelsRP.wav) and GA (03_vowelsGA.wav) on the CD-ROM.

Table 3.3. The vowels of RP and GA and their transcription.

Phonetic symbol	RP	GA
i	*key*	*key*
u	*moon*	*moon*
ɪ	*sit*	*sit*
ʊ	*good*	*good*
e	*bed*	
ɛ		*bed*
ɜ	*bird*	*bird*
ə	*arise*	*arise*
ɔ	*caught*	*(caught)*
æ	*bad*	*bad*
ʌ	*cut*	*cut*
ɒ	*hot*	
ɑ	*laugh*	*hot, (caught)*
eɪ	*fate*	*fate*
aʊ	*mouth*	*mouth*
əʊ	*bode*	
oʊ		*bode*
aɪ	*by*	*by*
ɔɪ	*boy*	*boy*
ʊə	*poor*	
eə	*there*	
iə	*here*	
aɪə	*fire*	
eɪə	*player*	
aʊə	*power*	

Table 3.3 shows the vowel inventories of both RP and GA. However, the fact that RP and GA both have very similar vowel inventories does not mean that they use the same vowels in the same words. In fact, as you know, RP speakers pronounce *laugh* /lɑf/, whereas GA speakers pronounce the word /læf/. Similarly, GA speakers pronounce the word *hot* /hɑt/, while RP speakers pronounce it /hɒt/. These differences can be described with phonological rules. One rule states that RP has /ɑ/ and GA has /æ/ in monosyllabic words where the vowel is followed by a voiceless fricative. Thus, *staff*, *path* and *ask* are pronounced with the vowel /ɑ/ in RP and with the vowel /æ/ in GA. Equally, where RP has the vowel /ɒ/, as in the words *top*, *not* and *spot*, GA speakers produce the vowel as /ɑ/. As you can see by the choice of the phonetic symbol in

Table 3.3, the vowel in the word *bed* is produced with a lower tongue position in GA than in RP.

When looking at Table 3.3, you can see that some of the vowels occur in pairs. In both RP and GA, the vowel pairs /i/ and /ɪ/ as well as /u/ and /ʊ/ have the same vowel height, vowel location and degree of lip rounding. In fact, the greatest difference between these vowels is their length: /i/ and /u/ are long, whereas /ɪ/ and /ʊ/ are short. Furthermore, the long member of the pair can occur in both open and closed syllables (see section 3.2 below), while the short member cannot occur in open syllables, i.e. a consonant has to follow it. The fact that we have minimal pairs such as *beat* and *bit* and *bead* and *bid* in RP and GA shows that **vowel length** is **phonemic** in English. The monophthongs of RP and GA can be divided into long and short ones (see Table 3.4). Many textbooks suggest using the diacritic : to indicate phonemic vowel length, although this is redundant since the phonetic symbols themselves distinguish these vowels clearly.

Table 3.4. Phonemically long and short vowels in RP and GA.

RP		GA	
long vowel	short vowels	long vowels	short vowels
iː	ɪ	iː	ɪ
uː	ʊ	uː	ʊ
ɔː	ʌ	ɔː	ʌ
ɑː	æ	ɑː	æ
ɜː	ə	ɜː	ə
	e		ɛ
	ɒ		

The terms 'long' and 'short' are phonological terms and do not always correlate with phonetic reality. In actual speech, vowel length varies with the type of consonant following it and the type of syllable in which the vowel occurs. Voiced consonants trigger greater length in preceding vowels, voiceless consonants trigger shorter length in preceding vowels. For example, the /ɪ/ followed by the voiceless plosive /t/ in *bit* is always shorter than the /ɪ/ followed by the voiced plosive /d/ in *bid.* Equally, the /i/ in *bead* is longer than the /i/ in *beat.* In addition, vowels are longer before sonorants than before obstruents. This means that the /ɜ/ in *girl* is longer than the /ɜ/ in *gird* in RP. Finally, vowels are longer in syllables that end in vowels – such as *bee* – than in syllables that end in consonants – such as *beat.* Thus, phonemically 'long' vowels can actually be shorter than phonemically 'short' vowels when their real duration is measured.

The pronunciation of words is never stable in languages, and the phoneme inventory of a language can change drastically in the course of a few generations. English is an excellent example of this, as you will find when you have a look at some diachronic descriptions of English phonology. The vowel inventory of languages seems much more susceptible to changes than the consonant inventory, and again, English is no exception. Current changes in RP and GA are well documented (e.g. Upton 2004, Hawkins & Midgley 2005, Gordon 2001, Labov et al. 1972). Processes in RP include the **monophthongization** of the diphthongs /ɔə/, /ʊə/ and /ɪə/, which are increasingly produced as monophthongs now (see Table 3.5). Similarly, the triphthongs /aɪə/ and /aʊə/ are nowadays often realized as the monophthong /a:/ in RP. Conversely, many young speakers now produce the monophthongs /i/ and /u/ in words like *geese* and *goose* as diphthongs – there is a slight **onglide** before the vowel. Finally, the vowel /æ/ in words like *man* is increasingly produced as an /a/.

Table 3.5. Some ongoing vowel changes in RP.

Old vowel	New vowel	Example word
/ɔə/	→ /ɔ:/	*door*
/ʊə/	→ /ɔ:/	*poor*
/ɪə/	→ /i:/	*near*
/aɪə/	→ /a:/	*fire*
/aʊə/	→ /a:/	*power*
/i:/	→ /ɪi/	*geese*
/u:/	→ /ʊu/	*goose*
/æ/	→ /a/	*man*

In the U.S.A, a 'Northern Cities Shift' has been observed, which concerns the pronunciation of some vowels by speakers living in cities in the Great Lakes Region (e.g. Chicago, Detroit and Buffalo). Table 3.6 illustrates that many speakers pronounce *bet* with the vowel /ə/ instead of /ɛ/ now. *Cut* is pronounced with an /ɔ/ rather than an /ʌ/ and words that used to be pronounced with an /ɔ/, such as *caught* now tend to have an /ɑ/. Furthermore, speakers increasingly pronounce words like *cot* with an /æ/ rather than an /ɑ/, whereas words that used to have an /æ/ – like *cat* – are now rather pronounced with a centering diphthong /ɪə/. Finally, the /ɪ/ in words like *kit* can now often be heard as an /ə/.

Table 3.6. Vowel changes in the Northern Cities Shift.

Old vowel	New vowel	Example word
/ɛ/	→ /ə/	*bet*
/ʌ/	→ /ɔ/	*cut*
/ɔ/	→ /ɑ/	*caught*
/ɑ/	→ /æ/	*cot*
/æ/	→ /ɪə/	*cat*
/ɪ/	→ /ə/	*kit*

Just like the consonants, the vowels of RP and GA show allophonic variation. When followed by /n/, /m/ or /ŋ/, vowels are nasalized, which means that at the beginning of the articulation of the vowel the velum is still lowered and some of the airstream passes through the nasal cavity. Nasalization of vowels is transcribed with the diacritic ~ (see Figure 3.3), so that a nasalized /i/ is represented as [ĩ]. Rule (3) expresses this allophonic rule in an abstract way: all vowels (V) are nasalized when followed by a nasal consonant.

$$(3)\ /\mathrm{V}/ \rightarrow [\tilde{\mathrm{V}}]\ /\ \_\left\{\begin{array}{l}/\mathrm{m}/\\ /\mathrm{n}/\\ /\text{ŋ}/\end{array}\right.$$

Other systems for the transcription of speech sounds apart from the IPA have been suggested (e.g. by Gimson in 1962), and there is a tradition in Northern America to replace some IPA symbols with other symbols. Table 3.7 illustrates the differences between the IPA and the Northern American tradition.

Table 3.7. Differences between the IPA and the Northern American tradition of transcription symbols.

IPA	Northern American tradition
ʃ	š
ʒ	ž
tʃ	č
dʒ	j
aɪ, ɔɪ, eɪ	aj, oj, ej
aʊ, oʊ	aw, ow

In this book, only the IPA will be used since it is the most widely used phonetic alphabet and the one that is employed in most dictionaries. Note that although the symbols are usually called phonetic symbols they represent phonemes.

3.1.4 Phonemic and phonetic transcription

With the help of the symbols provided by the IPA, **transcriptions** of sounds, words and utterances can be made. For anyone concerned with the structure and realization of sounds, be it linguists, language teachers, speech therapists or people working in speech technology, transcription is an essential tool. This is due to the fact that the Roman alphabet, which is used as the conventional spelling system for English, has a poor **grapheme-to-phoneme relationship** (see also chapter 1). Sounds and corresponding writing symbols (graphemes) are not always matched in a systematic way. The same sound in English can be represented by different letters: for example, /i/ can be spelled <e> in *me*, <ee> in *see*, <ea> in *sea*, <ie> in *brief* and so forth (the pointed brackets < > are used to indicate spelling). Conversely, the same spelling may refer to different sounds: <th> in *think* corresponds to the phoneme /θ/, whereas it corresponds to the phoneme /ð/ in *this*. Furthermore, there are 'silent' letters in English that are not pronounced at all, as for example in *psychology*, *lamb* and *time*. Thus, a separate spelling system, a phonetic alphabet in which each symbol corresponds to one and only one phoneme, is necessary to describe the pronunciation of English words in a precise manner.

Two types of transcription can be differentiated: phonemic and phonetic transcriptions. In a **phonemic transcription** the presumed underlying representations of sounds (i.e. the speaker's 'mental images' of sounds), the phonemes, are transcribed. It represents all and only the linguistically relevant information about articulation and is the kind of transcription given in dictionaries. The phonemic transcription of the word *pin* is thus /pɪn/. Phonemic transcriptions can be carried out for written texts. For example, the sentence "A tiger and a mouse were walking in a field." can be transcribed phonemically as in example (4).

(4) /ə taɪgə ənd ə maʊs wə wɔkɪŋ ɪn ə fild/

For GA, the transcription would be /ə taɪgəɹ ənd ə maʊs wəɹ wɑkɪŋ ɪn ə fild/ with the difference lying in the last sound of the words *tiger* and *were* as well as the vowel in the word *walking*. (A note on the usage of the symbol /ɹ/: very often in descriptions of English the symbol /r/ is used when referring to the alveolar approximant, although this is the IPA symbol for the alveolar trill (the 'rolled r'). This widespread confusion has many reasons, amongst them the fact that it is easier to use the /r/ symbol in handwriting and on a type-writer. This has to be borne in mind when reading or making transcriptions. In this book, the correct IPA symbol /ɹ/ will be used throughout to refer to the alveolar approximant.)

Phonemic transcriptions can also be made for real speech. Listen to recording 03_transcription.wav on the CD-ROM. A phonemic transcription of the utterance "A tiger and a mouse were walking in a field" produced by this RP speaker is illustrated in (5).

(5) /ə taɪgə ənd ə maʊs wə wɔkɪŋɪn ə fild/

Note that the spaces between the words are only inserted here to make the transcription more reader-friendly. In spoken language, there are obviously no pauses between the last sound of one word and the first one of the next. No differences from the phonemic transcription of the written sentence in (4) exist since transcription (5) too gives the 'ideal' or typical pronunciations of each word as they can be found in a dictionary.

When you are interested in fine-grained articulatory details of an utterance, for example because you would like to inform a language learner of his or her phoneme realizations that are not native-like, you can carry out a **phonetic transcription**. This type of transcription includes information about phonologically not relevant phonetic details – for example, it might specify whether the /p/ in *pin* is aspirated or not. Phonetic transcriptions do not refer to speakers' mental representations but to actual realizations and can therefore only be carried out for real speech. Depending on the amount of phonetic information included in the transcription, the phonetic transcription can be **broad** (including only a few details) or **narrow** (with an exact transcription of all phonetic details). Apart from the IPA symbols for the phonemes, a phonetic transcription includes diacritics to indicate phonetic realizations (see Figure 3.3). Phonetic transcriptions are far more difficult to make than phonemic ones and require a lot of training. A fairly broad phonetic transcription of the utterance "A tiger and a mouse were walking in a field" produced by the RP speaker in 03_transcription.wav on the CD-ROM is (6).

(6) [ət^{h}aɪgəɹəndəmãswəwɔkɪnɪnəfiəɫd]

Several differences from the phonemic transcription can be noted. First, it is difficult to determine word boundaries. Does the 'linking-r' (see section 3.2.1) belong to *tiger* or *and*, and does she say *in a* or rather *inna field*? For this reason, no blank spaces are given in this phonetic transcription. Second, the speaker produces a [ɹ] linking the words *tiger* and *and*. Third, she produces [mas] for *mouse*. Fourth, the last phoneme of the word *walking* is a [n] rather than a [ŋ]. Fifth, the vowel in *field* is a distinct diphthong which is transcribed as [iə]. Moreover, this transcription shows that the first sound of *tiger* is realized as an aspirated [t^{h}], that the vowel in *mouse* is a nasalized [ã] and that the /l/ in *field* is velarized. A narrow phonetic transcription of the same utterance would include many more phonetic details such as the voicing of each segment or the exact place of articulation for each sound.

Table 3.8. IPA symbols and corresponding SAMPA symbols.

IPA	SAMPA	IPA	SAMPA
vowels		consonants	
i	i	p	p
ɪ	I	b	b
e	e	t	t
ɛ	E	d	d
æ	{	k	k
a	a	g	g
u	u	f	f
ʊ	U	v	v
ɔ	O	θ	T
ɒ	Q	ð	D
ɑ	A	s	s
ə	@	z	z
ɜ	3	ʃ	S
ʌ	V	ʒ	Z
		h	h
		m	m
		n	n
		ŋ	N
		l	l
		ɹ	r\
		j	j
		w	w

Nowadays, few transcriptions are carried out by hand or on paper. Typically, especially when transcribing large amounts of speech, phonologists use a computer. Since it is tedious to use the special IPA fonts in text editors and since an exchange of data is difficult when everyone uses different fonts, a machinereadable version of the IPA has been invented, which comprises only symbols that can be found on any computer keyboard. This alphabet is called **SAMPA** (which stands for Speech Assessment Methods Phonetic Alphabet) and was introduced in 1992 by John Wells and other phoneticians (Wells et al. 1992). Table 3.8 lists the IPA symbols relevant for English and their corresponding SAMPA symbols. A phonemic SAMPA transcription of utterance 03_transcription.wav on the CD-ROM would thus look like in (7):

(7) /@ taIg@ @nd @ maUs w@ wOkIN In @ fild/

3.1.5 Phonetic features (advanced reading)

So far we have assumed that phonemes are the smallest units of a speaker's mental representation and the linguistic analysis of speech. In this section, several approaches in phonological theory will be reviewed that claim that phonemes consist of even smaller units. These approaches are motivated by the observation that the phonemes of a language can form groups. For example, those phonemes that involve a complete closure of the oral cavity (e.g. /m, n, ŋ, p, t, k, b, d, g/) can be separated from those that have no such complete closure. The first group is referred to as non-continuants, the second as **continuants**. Equally, in section 2.4 we saw that some consonants are produced with a fairly continuous airstream, whereas other consonants are produced with an obstruction of the airstream. The first group of phonemes, which comprises /m, n, ŋ, l, j, w, ɹ/ as well as all vowels, are called **sonorants**, whereas the other group of phonemes including /p, t, k, b, d, g, f, v, θ, ð, s, z, ʃ, ʒ, h/ are called **obstruents**. Note that this class of phonemes comprises as diverse manners of articulation as plosives and fricatives.

The existence of such **natural classes** of phonemes can be further illustrated with rule (8) introduced in section 3.1.2. It illustrates that the three phonemes /p, t, k/ are aspirated in a particular context:

(8)
$$\left.\begin{array}{l}/p/\\ /t/\\ /k/\end{array}\right\} \rightarrow \left\{\begin{array}{l}{[p^{h}]}\\ {[t^{h}]}\\ {[k^{h}]}\end{array}\right\} /.\_$$

In order to explain why it is only these three phonemes to which the phonological rule applies but not other phonemes of English, one has to refer to the **features** these phonemes share but others do not. /p, t, k/ have in common that they are consonantal, that they are non-continuants, non-sonorants and voiceless. In phonology, **phonetic** or **phonological features** are conceptualised as **binary**, which means that phonemes have one of two possible values: (+) or

(–). A sound is either voiced or voiceless, either continuant or non-continuant – there are no other options. The class of the English voiceless plosives /p, t, k/ can thus be represented as [+consonantal], [–continuant], [–sonorant] and [–voice] (the convention is to use square brackets when referring to features). By the same token, all voiced phonemes in English share the feature [+voice], whereas all voiceless phonemes have the feature [–voice] in common.

Features are abstract, theoretical concepts, but they are assumed to have a phonetic basis. Many different sets of features have been proposed by different researchers, beginning with the classic work by Jakobson & Halle (1956) and Chomsky & Halle (1968). Some of the features refer to articulatory aspects such as the feature [labial], which describes that the lips are involved in the production of a phoneme. The places and manners of consonant articulation listed in section 2.4 in Tables 2.2 and 2.3, for example, are articulatory features. Other features that have been proposed are perceptual or acoustic in nature, such as the feature [low], which describes the property of some vowels in acoustic (physical properties of the sound wave, see section 5.2 below) rather than articulatory terms. In order to be of any scientific value, phonetic features have to be **contrastive** or **distinctive**, i.e. they need to be able to differentiate between individual phonemes and individual phoneme classes.

Phonologists do not only use features in order to describe the phonological behaviour of certain classes of phonemes. It is also assumed that phonemes are represented in a speaker's brain as a bundle of these features. The mental representation of /t/, for example, can be described to be specified as [–voice] which means that it does not have the feature voice, whereas /d/ is specified as [+voice]. The properties of phonemes can be illustrated in a **feature matrix**.

$$\left(\begin{array}{lll}
\text{p} & \text{ɪ} & \text{n} \\
 & & \\
-\text{voice} & +\text{voice} & +\text{voice} \\
+\text{labial} & -\text{labial} & -\text{labial} \\
+\text{stop} & -\text{stop} & -\text{stop} \\
-\text{nasal} & -\text{nasal} & +\text{nasal} \\
-\text{high} & +\text{high} & -\text{high}
\end{array}\right)$$

Figure 3.5. Feature matrix of the word pin.

The feature matrix of the three phonemes of the word *pin* is given in Figure 3.5. Five features are sufficient to describe adequately the difference between the three phonemes: the features [±voice], [±labial], [±stop], [±nasal] and [±high]. The /p/ has the features [–voice], [+labial], [+stop], [–nasal] and [–high]; the phoneme /ɪ/ has the features [+voice], [–labial], [–stop], [–nasal] and [+high], whereas /n/ has the features [+voice], [–labial], [–stop], [+nasal] and [–high].

Recent approaches to describing phonetic features divide them into

- major-class features
- laryngeal features
- manner features
- place features

The **major-class features** are used to distinguish between the major classes of speech sounds: vowels and consonants, sonorants and obstruents (see also section 2.4 above). The major-class features are [±consonantal], [±sonorant] and [±syllabic]. The feature [±consonantal] distinguishes phonemes with a constriction in the vocal tract (oral plosives, fricatives, nasals and the **liquids** /l/ and /ɹ/) from the vowels and the **glides** /w/ and /j/, which are not articulated with a constriction in the vocal tract. The feature [±sonorant] distinguishes obstruents such as /t/ from sonorant consonants such as /l/ and vowels. [±syllabic] describes phonemes that can function as the nucleus of a syllable – which are mainly vowels (see section 3.2 below). The only **laryngeal feature** that plays a role in English is [±voice]. Phonemes that are produced with vocal fold vibration such as /m/ and /a/ have the feature [+voice], whereas all phonemes produced without vocal fold vibration, such as /f/ and /k/, share the feature [–voice].

The **manner features** comprise [±continuant], [±nasal], [±strident] and [±lateral]. As explained above, the continuants are articulated with a free flow of airstream through the oral cavity, which means that an /a/ is [+continuant] while a /m/ is [–continuant]. Phonemes that are [+nasal] are produced with a lowered velum, whereas [–nasal] phonemes are not. Phonemes that have the feature [+strident] are obstruents whose articulation involves a noisy kind of friction as in the case of the phonemes /s/ and /ʃ/. In English, for example, this class of phonemes determines the choice of the **allomorphs** of the plural {s} and the third person singular {s}. When the stem of a word ends in a strident sound, the allomorph is [ɪz], as in *bushes* and *misses*. In other cases, the morpheme is pronounced either [z] (following voiced phonemes except stridents) or [s] (after voiceless phonemes except stridents). Phonemes sharing the feature [+lateral] are produced with tongue contact in the oral cavity and the airstream passing at both sides of the tongue.

There are **monovalent** and binary **place features**. Monovalent features describe the places of articulation [LABIAL], [CORONAL] and [DORSAL]. The binary features specify these places further. [LABIAL] phonemes such as /m, v, b/ are articulated with the lips. In addition, they can be specified for [±round]. English phonemes that have lip rounding are for example /u/ and /o/; phonemes that have no lip-rounding are /i/ and /e/. Phonemes whose mental representation includes the feature [CORONAL] are articulated with the tip, blade or front of the tongue (see Figure 2.9 in chapter 2 above). This applies to a wide range of English phonemes such as /t, θ, s, ʃ, l, j/. These phonemes can be further specified as [±anterior] and [±distributed]. All phonemes specified as [+anterior] are articulated at the alveolar ridge or further forward (this includes labials, labiodentals, interdentals and alveolars) as for example /z/. Phonemes that are [–anterior] are articulated behind the alveolar ridge as for example /ʒ/. Phonemes that are articulated with a constriction that concerns a relatively large part of the vocal tract are [+distributed], whereas phonemes articulated with a constriction that concerns only a small area of the vocal tract are specified as [–distributed]. Because the articulation of /ʃ/ involves a constriction with the blade of the tongue along the post-alveolar area and the palate, it is specified as [+distributed]. /t/, which only involves the tip of the tongue at the alveolar ridge, by contrast, is specified as [–distributed].

The articulation of phonemes that are [DORSAL] involves the dorsum, the back of the tongue (see Figure 2.9), and includes /k, g, ŋ/ and all vowels. These phonemes can be further specified as [±high], [±low], [±back] and [±tense]. [+high] phonemes such as /i, ɪ, u, ʊ/ are articulated with the dorsum close to the palate. /e, o, a/ are correspondingly specified as [–high]. The phonemes specified as [+low] such as /a, ɒ, ɑ/ have a bunched dorsum in a low position in the mouth; [–low] phonemes do not. [+back] phonemes such as /o, ɒ, ɑ, k/ are articulated with a high dorsum in the centre of the mouth or further back. [+tense] vowels are produced with a more peripheral tongue position, as is the case for /i, e, a, u/. This binary feature only exists in languages that have vocalic oppositions such as the ones in English described in section 3.1.2 above. This is why the long vowels listed in Table 3.4 are sometimes referred to as **tense** vowels, whereas the short vowels are called **lax** vowels.

When we assume that the phonemes of English are represented in the minds of English speakers as bundles of features, it becomes obvious that some of the information is redundant. For example, vowels specified as [–high] are also characterized as [+low], which is like saying the same thing in two different ways. Equally, all sounds specified as [+sonorants] in English are also [+voice]. This has led some phonologists to suggest that only distinctive but no redundant features form part of a speaker's linguistic knowledge, an approach that is known as **underspecification theory**. This theory claims that in the underlying mental representation of phonemes those features whose values are predictable are omitted. For example, all front vowels in English are unrounded, so that the feature [–round] can be predicted from the feature [–back] in these cases. These phonemes therefore do not need to be specified for the feature [±round]. The concept of underspecification has been used to explain the difficulties some learners of a second language have with certain phonemes: for example, the claim has been made that when the native language does not have a specified mental representation for some features, it will be difficult to acquire phonemes in a second language that require this feature distinction. The lack of specification of front vowels for the feature [±round] might explain why English speakers find the pronunciation of /y/, the high rounded front vowel in German *Tür* and French *tu*, particularly difficult.

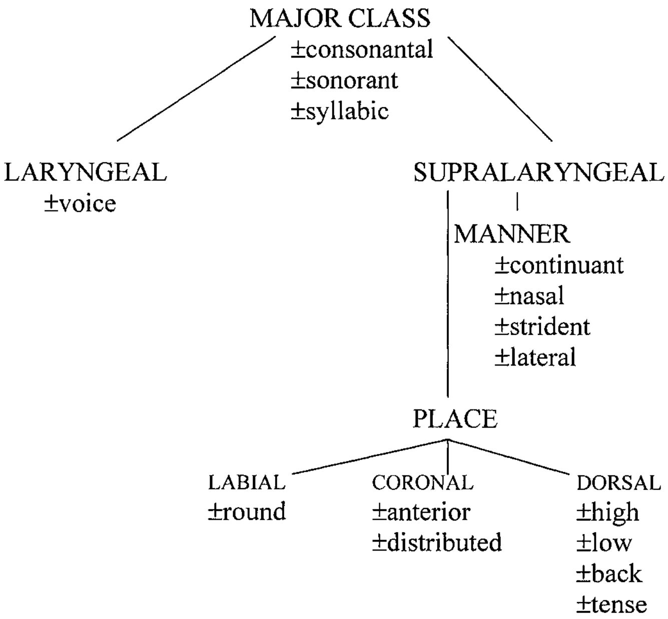


Figure 3.6. A feature tree.

Another way of representing the internal structure of phonemes, apart from drawing a feature matrix, is that of drawing feature trees. This approach is called **feature geometry** and is based on work by Clements (1985) and McCarthy (1988). It suggests that the different features a phoneme comprises are hierarchically ordered. Taking the feature system proposed above, a feature tree would look like the one depicted in Figure 3.6. Each feature grouping in the feature tree is called a **node** (e.g. LARYNGEAL is a node). These nodes form natural classes of phonemes. Beneath each node those features are grouped that make up the subclasses (e.g. ±voice for the node LARYNGEAL). Nodes and features are ranked on **tiers** (the different levels of the hierarchy), which illustrate their hierarchical relationship. The features of each phoneme of English (and any other language, of course) can be specified with the help of such a feature tree.

3.2 The English syllable

Most adult speakers of English agree that the word *dog* consists of one syllable and the word *trisyllabic* of four syllables. Equally, most phonologists agree that phonemes can be grouped into the higher unit of the syllable (see e.g. Blevins 1995). They propose the syllable both as a unit for linguistic analysis and a unit that speakers have a mental representation of in their brains. This conviction is based on the fact that a number of phonological processes have been identified that can best be explained with the unit of a syllable. The distribution of stress in English words, for example, depends on the type of syllables involved (see section 3.2.3 below). Moreover, the motor patterns underlying articulation seem to comprise syllables rather than individual speech sounds, as explained in section 2.5. Similarly, speech errors such as saying *face spood* instead of *space food* can be neatly explained with reference to a particular part of the syllable: the substitution of the speech sounds is not random, but concerns the onset position of each syllable.

The internal structure of a syllable that is thought to be shared by all languages in the world is illustrated in Figure 3.7. A syllable – which is represented by the symbol σ – consists of an **onset** and a **rhyme**, which in turn can be divided into a **nucleus** and a **coda**.

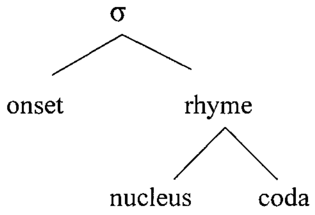


Figure 3.7. Universal structure of a syllable.

Languages vary as to which type of speech sounds are allowed to occur in which position of the syllable. In English, only consonants (C) can appear in the onset and the coda position, whereas the nucleus position is nearly always filled by a vowel (V). The syllable *dog* thus consists of /d/ in the onset position, /ɒ/ as the nucleus and /g/ in the coda position. In English, all parts of a syllable can be filled with more than one speech sound. Figure 3.8 illustrates the internal structure of the syllable *sprint*. It consists of the three consonants /s/, /p/ and /ɹ/ in the onset position, the nucleus /ɪ/ and the consonants /n/ and /t/ in the coda position.

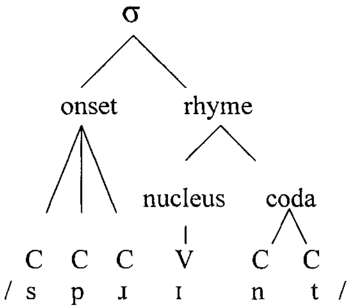


Figure 3.8. The internal structure of the syllable sprint.

In English syllables, only the nucleus is compulsory, that is, every syllable must have a nucleus. Syllable onsets and codas are optional in English: syllables may or may not have an onset and a coda. The syllable *on* does not have an onset, whereas *tea* is a syllable without a coda. The shortest syllables in English thus consist of only a vowel as a nucleus, for example *I, a, oh*. This vowel has to be either phonemically long or a diphthong, as will be explained in the next section. In some special cases, a consonant can function as the syllable nucleus in a word. This is for example the case in the second syllable of the word *button*. The word consists of the two syllables /bʌ.tn̩/ (the dot indicates a syllable boundary, the stroke under the /n/ is the diacritic symbol for a syllabic consonant), with the first syllable comprising the onset /b/ and the nucleus /ʌ/ and the second syllable comprising the onset /t/ and the nucleus /n/. Only nasals and liquids (/n, m, ŋ, l, ɹ/) can function as the syllable nucleus – they are called **syllabic consonants**. Syllables with a syllabic consonant as the nucleus cannot stand on their own in English, but always belong to a word that has another syllable with a vocalic nucleus.

3.2.1 Types of syllables and phonotactic rules of English

Around twenty different types of syllables occur in English. There are, for example, CV syllables such as *do* and *go*, which consist of one consonant in the onset position and one vowel in the nucleus position. Further, there are VC syllables such as *in* or *of* with a vowel as nucleus and one consonant in the coda position. Other types include CVC (*dog*), CCV (*spa*), CCVC (*tram*), CCCVC (*split*), and so forth. Syllables that have no consonant in the coda position are called **open** syllables. **Closed** syllables have at least one consonant in the coda position. Sequences of two or more consonants in the syllable onset and coda, such as /spɹ/ in the onset and /nt/ in the coda of the word *sprint*, are referred to as initial and final **consonant clusters**, respectively. Syllables that have a rhyme consisting of a long vowel (e.g. *bee*), a diphthong (e.g. *bay*) or a short vowel plus consonant (e.g. *beg*) are called **heavy** syllables. Long vowels and diphthongs are sometimes represented as VV so that the syllable structure of *bee* or *bay* can be given as CVV. **Light** syllables, by contrast, have only a short vowel or a syllabic consonant in the rhyme (e.g. *beloved*). The different syllable types do not occur equally often in English speech. Dauer (1983) found that CVC (34%) and CV (30%) syllables are most frequent, followed by VC (15%), V (8%) and CVCC (6%). Gut (2005) found that 43% of all syllables in English are open syllables. Furthermore, their distribution is constrained: stressed syllables are always heavy, light syllables are always unstressed (see also section 3.2.3), and light syllables cannot occur on their own.

The number and type of consonants that can occur in both onset and coda position in English is restricted. The number of consonants in the onset position is limited to three – there are no English syllables with more than three consonants in the onset position. The maximum number of consonants in the coda position is limited to four. If there are four coda consonants, the last one is either an inflectional /s/ as e.g. in *texts* [teksts] or an inflectional /t/ as in *glimpsed* [glɪmpst]. Furthermore, the type of consonants that can occur in English syllable onsets and codas and their order is restricted, which can be formulated in **phonotactic** rules. There are two types of phonotactic rules:

- rules of sound distribution
- rules of sound combinations

Rules of distribution describe in which position of the syllable phonemes can occur. For example, not any kind of vowel can appear in an English syllable without onset and coda: it has to be a long vowel or a diphthong. This is why when the indefinite article *a* is pronounced on its own it is pronounced /eɪ/. (Only in combination with other words is it pronounced /ə/, where it is often considered part of a phonological word; see section 3.3 below). Other phonotactic constraints concern the phoneme /ŋ/, which never occurs in the onset of English syllables. Some phonemes of English never occur in the coda position: these are /j/, /w/ and /h/. In RP, in contrast to GA, /ɹ/ does not occur in the syllable coda: syllables like *car* and *bar* in *barman* are pronounced [kɑ:] and [bɑː]. When phonemes cannot occur in all positions of the syllable, linguists speak of **defective distribution**.

Varieties of English that do not pronounce coda /ɹ/ are called **non-rhotic** varieties (or accents), whereas those varieties that allow coda /ɹ/ are called **rhotic** accents. This phonological difference has historical reasons. Until the Early Modern English period English was a rhotic language, which is still reflected in the present-day spelling of words like *car* and *barman*. This rhotic accent was spoken by the settlers in the early colonies (e.g. in Northern America). During the 17th and 18th centuries, Standard English on the British Isles evolved into a non-rhotic accent and, consequently, many of the settlers of later colonies such as Australia, Singapore and Nigeria, which were founded at that time, spoke non-rhotic English. Present-day Australian English, Singapore English and Nigerian English are still non-rhotic accents. There is one phonological instance when coda /ɹ/ is pronounced even in non-rhotic accents: when a syllable ending in the spelling <r> is followed by a syllable without onset consonant, the /ɹ/ is pronounced. This process, which can be observed in *pairing* /pɛɹɪŋ/ and *far away* /faɹəweɪ/, is called the '**linking /r/**' and reflects resyllabification, a phonological process that will be explained in section 3.2.2 below. When speakers of a non-rhotic accent pronounce a /ɹ/ between two vowels although there is no spelled <r>, the result is called '**intrusive /r/**'. It can be observed in phrases such as *law and order,* which is pronounced /lɔːɹəndɔːdə/ by some RP speakers. It is hypothesized that speakers produce this inserted (**epenthetic**) /ɹ/ to avoid a hiatus, i.e. the separate pronunciation of two adjacent vowels.

Phonotactic constraints of distribution describe the order of consonants in English syllable onsets and codas. In English onsets consisting of a three-consonant cluster, the first phoneme is always a /s/, the second either /p/, /t/ or /k/ and the third /ɹ/, /j/ or – in some cases – /l/ or /w/. Thus, we have in English *splay*, *spray*, *stray*, *squeal* and *stew* (in RP only), but not **pstay* or **wpjay* (impossible forms are marked by the preceding asterisk *). This rule is illustrated in (9). GA does not allow many onset consonant clusters involving /j/, for example *stew* is pronounced /stu/.

(9)

$$\sigma \quad [\mathrm{s} \begin{Bmatrix} \mathrm{p} \\ \mathrm{t} \\ \mathrm{k} \end{Bmatrix} \begin{Bmatrix} \mathrm{(l)} \\ \text{ɹ} \\ \mathrm{(w)} \\ \mathrm{j} \end{Bmatrix}$$

Table 3.9. Permissible two-consonant onset clusters in RP and GA.

RP + GA	RP only	Example
/pl/		*plan*
/pɹ/		*pram*
/pj/		*pure*
/bl/		*blue*
/bɹ/		*brew*
/bj/		*beauty*
/fl/		*flew*
/fɹ/		*friend*
/fj/		*few*
/vj/		*view*
/tɹ/		*tram*
/tw/		*twist*
	/tj/	*tune*
/dɹ/		*drink*
/dw/		*dwell*
	/dj/	*duel*
/θr/		*throw*
/θw/		*thwart*
/sm/		*smell*
/sn/		*snake*
/sl/		*sling*
/sw/		*swing*
	/sj/	*suit*
	/zj/	*Zeus*
/ʃɹ/		*shrink*
/kl/		*claw*
/kɹ/		*crawl*
/kw/		*quest*
/kj/		*cute*
/gl/		*glow*
/gɹ/		*grow*
/gw/		*Gwen*
/gj/		*gules*
/hj/		*huge*
/mj/		*mule*
	/nj/	*new*
	/lj/	*lute*

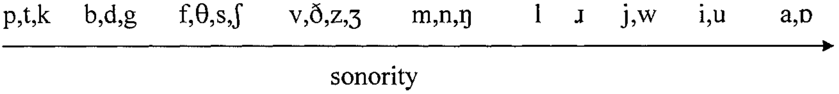
Similarly, not all possible combinations of two consonants in the onset can occur in English. In general, two consonants that share a place of articulation cannot occur together. Thus, while /p, k, b, g, f, s/ can combine with /l/ as in *play*, *clay*, *bleak*, *glimmer*, *flat* and *slip*, */tl/ and */dl/ do not exist as onsets in English. Obstruents (plosives and fricatives) cannot combine with nasals (thus no */bm/ or */fn/), and nasals cannot combine with the liquids /l/ and /ɹ/. Some phonemes such as /ð/, /z/, /ɹ/, /w/ and /j/ cannot appear in the first position of onset clusters at all.

Table 3.9 lists all permissible two-consonant clusters in contemporary RP and GA. Some of the clusters like /gw/, /gj/, /lj/ and /θw/ are in fact very rare, and only a handful of words exist for each of them. You can see that GA lacks some two-consonant onset clusters, in particular, /tj/ as in *tuna*, /dj/ as in *duel*, /sj/ as in *suit*, /zj/ as in *Zeus*, /nj/ as in *new* and /lj/ as in *lute,* which occur only in RP. The cluster /lj/ seems to be dying out in RP; only few speakers still use it in words like *lute* (/lju:t/). Instead, most RP speakers pronounce this word /lu:t/ now. Some words in English have the onset clusters /ʃw/ (*schvartze*), /ʃl/ (*schlep*), /ʃm/ (*shmoo*) and /ʃn/ (*schnapps*), but since all of them occur only in loan words, these onsets are not considered generally permissible in English.

Similar to English syllable onsets, the order of consonants in English syllable codas is restricted by phonotactic rules. Four-consonant coda clusters such as in *texts* (/ksts/) and *glimpsed* (/mpst/) end either in /s/ or in /t/, which have morphemic status. This means that they, for example, consist of the inflectional plural morpheme {s} or the past tense morpheme {ed}. The first consonant of three-consonant coda clusters such as in *elms*, *sprints* and *friends* is always a liquid or a nasal, i.e. either a /l/, /m/, /n/ or /ŋ/ in RP or further a /ɹ/ in GA (as in *carts*). The second consonant in such clusters is one of the obstruents /p, k, t, d, f, θ, s/ or /m/. Three-consonant coda clusters end in either /t/, /s/, /z/ or /θ/ as in *gulped* /gʌlpt/, *links* /lɪŋks/, *friends* RP: /fɹendz/ GA: /fɹɛndz/ and *twelfth* RP: /twelfθ/ GA: /twɛlfθ/. As with the onset clusters, not all phonemes can combine to form two-consonant clusters in English syllable codas. For example, the second consonant of final two-consonant clusters can be any consonant except /h, w, j/ and /ɹ/ in RP.

The phonotactic constraints that apply to English syllables have undergone some changes in the history of English. In Old English, onset clusters such as /hl/, /hw/, /hr/, /wl/, /wr/, /kn/ and /gn/ occurred in many words, but they were lost by or during the Middle English period in most English dialects. Table 3.7 would thus look a bit different if it illustrated the permissible onset clusters of 10th century or 15th century English.

Is there any reason why the order of consonants is restricted in such a way? The phonotactic structure of syllables is often described with reference to the **sonority** of the phonemes involved. The sonority of a phoneme is defined in acoustic and perceptual terms: it refers to the relative loudness of one phoneme compared to other phonemes. The sonority of all phonemes of English can be depicted on a **sonority scale** (Figure 3.9). It can be seen that vowels have the highest sonority of all phonemes in English, with low vowels being even more sonorous than high vowels. The approximants /j/ and /w/ have the closest sonority values to that of the vowels. Voiceless plosives are the least sonorous phonemes. Compared to them, the sonority increases with the voiced plosives, followed by the voiceless and the voiced fricatives, the nasals and the liquids.



p,t,k b,d,g f,θ,s,ʃ v,ð,z,ʒ m,n,ŋ l ɹ j,w i,u a,ɒ

sonority

Figure 3.9. Some phonemes of English on the sonority scale.

When illustrating the sonority of the phonemes of syllables such as *print* and *twist* in a diagram (see Figure 3.10), one can see that the syllables resemble a mountain with a peak. In the onset, the sonority of the first phoneme is lower than that of the second phoneme; the vowel in the nucleus has the highest sonority, whereas the first phoneme in the coda has a higher sonority value than the second one. The order of consonant sequences in the onset and coda position can therefore be explained in the following way: in English onsets, consonants are ordered with increasing sonority, whereas they are ordered with decreasing sonority in the coda position. Not all English syllables conform to this pattern, though. There is one exception to the rule in both onsets and codas, and this concerns the phoneme /s/. Despite having a higher sonority than the plosives, onsets in which a /s/ precedes a plosive, such as in /spl/ and /stɹ/, are permissible in English. In the coda position, equally, /s/ can occur in the last position after phonemes with a lower sonority value, e.g. in /ks/ and /ts/. This is why the /s/ is often treated as a special case and its position in the syllable is called **'extra-syllabic'** or **'appendix'**.

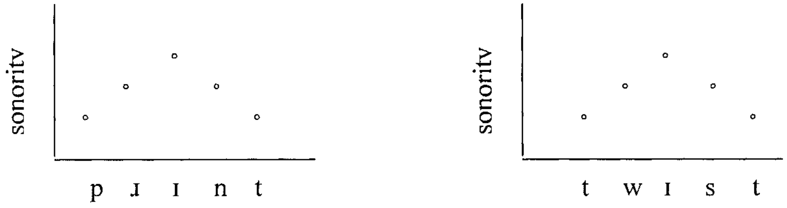


Figure 3.10. The sonority of the phonemes in print *and* twist.

3.2.2 Syllabification in English

When words consist of more than one syllable, when they are **multisyllabic** or **polysyllabic**, the problem of **syllabification** arises. Should the consonants /kstɹ/ in the word *extreme* /ɪkstɹi:m/ be analysed as coda consonants of the first syllable, as onset consonants of the second syllable or should they be split up somehow to fill both positions? Phonologists have proposed that these consonants are distributed according to the **maximum onset principle**. This principle states that intervocalic consonants are syllabified as the onset of the following syllable as far as the phonotactic constraints of the language allow it. This means that the two syllables of *extreme* are split up into /ɪk.stɹi:m/ as illustrated in Figure 3.11. /stɹ/ is syllabified as the onset of the second syllable, since this conforms to the phonotactic rules of onset clusters. /kstɹ/ violates these rules, so that /k/ has to be syllabified as the coda of the first syllable.

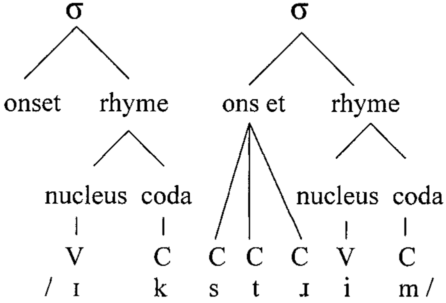


Figure 3.11. The syllable structure of the word extreme.

In some cases, the maximum onset rule leads to questionable results. For example, the words *apple* and *epic* would have to be syllabified into /æ.pl/ and /e.pɪk/ (GA: /ɛ.pɪk/), which violates the rule described above, namely that syllables have to consist of at least a long vowel or a short vowel plus consonant. /æ/ and /e/ are short vowels and, according to this rule, cannot stand on their own in a syllable. A way around this problem is to propose that the /p/ in both cases is an **ambisyllabic** consonant, i.e. that it belongs simultaneously to the coda of the first and the onset position of the second syllable (see Figure 3.12). There seems to be good reason to assume that speakers indeed have a mental representation of ambisyllabicity for these consonants – when you ask native speakers to name the syllables of *apple* and *epic*, they will usually say *ap.pl* and *ep.pic*.

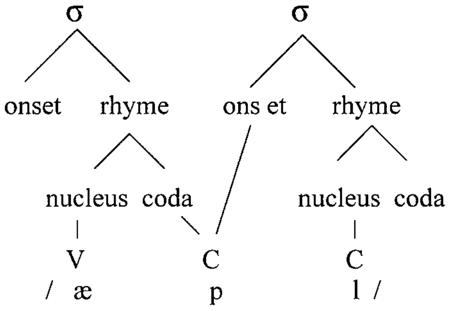


Figure 3.12. The syllable structure of the word apple *with /p/ as an ambisyllabic consonant.*

In connected speech, the problem of syllabification arises also across word boundaries. Speakers often produce speech that contains **resyllabification,** where one or more consonants of one word are attached to the following word. Resyllabification also follows the maximum onset principle. This can be seen for example in the phrase *have it*, which is usually resyllabified into [hæ.vɪt]. The last consonant of the word *have* is produced as the onset of the second syllable. As mentioned in section 3.2.1 above, the 'linking /r/' in non-rhotic accents can also be described as resyllabification. In a phrase such as *Peter and Mary*, which is pronounced /piː.tə.ɹənd.mæ.ɹɪ/, the /ɹ/ in *Peter* is resyllabified as the onset of the following syllable. In non-rhotic accents of English, coda /ɹ/ is not pronounced. The /ɹ/ in phrases like *Peter and Mary,* however, is resyllabified as an onset consonant and is therefore pronounced.

3.2.3 Stress and speech rhythm in English

Syllables in English can be either **stressed** or **unstressed**. In the word *English*, for example, all native speakers of RP and GA would agree that the first syllable is stressed, whereas the second syllable is unstressed. The phenomenon of stress can be explained both with reference to speech production mechanisms and to acoustic and perceptual properties. Sections 2.1 and 2.2 showed that at least three factors lead to a perceptual impression of stress. Speakers produce greater air pressure by increased muscular effort in the respiratory system and they produce higher pitch by tensioning the vocal folds on stressed syllables. Moreover, the articulatory movements for the production of speech sounds in stressed syllables are carried out to a larger extent than for speech sounds in an unstressed syllable. Section 5.5.2 explains that in acoustic terms this can be measured in an increased loudness or energy in the speech signal as well as an increased duration of individual segments and syllables. The perceptual impression of stress is one of **prominence** – a stressed syllable appears more prominent (louder) than an unstressed syllable. Similarly, the speech sounds in stressed syllables have a different quality. These vowels and consonant are fully articulated rather than reduced. The physiological mechanisms underlying the production of stress determine that stress is a relative rather than an absolute property of syllables. One syllable can be more or less stressed than another syllable, but it is very difficult to determine the degree of stress on a single syllable if this is the only one we hear.

Speakers of English know which syllable or syllables in a word are stressed. They have a mental representation of word stress, which will be described in section 3.3.1 below. Thus, stress is an abstract property of a syllable and part of the linguistic knowledge of a speaker. The degree of stress on a particular syllable, however, is also a real physical event that can be measured, as explained in section 5.5.2 below. In order to keep the two things – the phonological representation and the phonetic event – separate, two different terms should be used. These terms are **accent** and **stress**. Unfortunately, different researchers use the two terms with the opposite meaning. Laver (1994) regards 'accent' as the abstract category and calls the actual physical occurrence 'stress'. Jassem & Gibbon (1980), on the other hand, call the mental representation 'stress' and refer to the observable manifestation of stress as 'accent'. In this book, the term 'stress' is used to refer to the abstract property of syllables and the mental representation, whereas the term 'accent' is employed for the measurable phonetic event. Thus, a stressed syllable is the one a speaker knows to be stressed, whereas accents can be measured in real speech.

Not every syllable in English can be stressed. Stress is correlated with the phonological properties of a syllable. In English, syllables consisting of a short vowel only and CV syllables with a short vowel are always unstressed. This means that the phonological distribution of vowels in stressed and unstressed syllables differs widely. While in stressed syllables all vowels listed in Table 3.3 above except /ə/ can occur, the only vowels that appear in unstressed syllables in both RP and GA are /ə/ and /ɪ/. This can be seen in the unstressed syllables of the words *sofa* /'səʊfə/ (GA: /'soʊfə/), *about* /ə'baʊt/, *police* /pə'li:s/ and *family* /'fæmɪlɪ/. (The stressed syllable is marked by a preceding '). This restricted occurrence of the **schwa** /ə/ in both RP and GA has led many phonologists to propose that the schwa does not constitute a phoneme in its own right. Apart from a few minimal pairs with /ɪ/, as in RP *sofa* (/'səʊfə/) and *Sophie* (/'səʊfɪ/), /ə/ does not contrast with any other vowel of English. For many speakers, furthermore, /ə/ and /ɪ/ are interchangeable in pre-stress position so that *extreme* can be pronounced both /ək'stɹi:m/ and /ɪk'stɹi:m/. The rule that unstressed syllables can only contain the reduced vowels /ə/ and /ɪ/ is a phonological description. In section 2.5 a phonetic explanation of this phenomenon was given: in unstressed syllables, due to the rapid articulation, the tongue height and horizontal position of the tongue in vowels is far removed from the ideal articulatory target. The tongue only briefly reaches a central position – neither high nor low, neither front nor back – before the articulators move on to form the next consonant, which means that the vowel appears 'reduced'.

Unstressed syllables in English are often also called **weak** syllables (whereas stressed syllables are referred to as **strong** syllables). By the same token, the alternative pronunciations of a number of English function words are called **weak** and **strong forms**. Table 3.10 lists the different pronunciations of some of these words (for a full list of 40 function words see Roach 1991, chapter 12) and gives examples in RP. The weak form is the one that is usually used in connected speech. There, the processes of reduction and coarticulation described in section 2.5 apply and determine the articulation of the vowel as well as the articulation of consonant clusters. Typically, the vowel is reduced (i.e. not fully articulated) and consonant clusters are reduced (i.e. one or more consonants are not realized at all). Only when such a function word is produced on its own or for purposes of contrast or when it appears at the end of an utterance do speakers use the strong form. Thus, in "You and I" *and* is usually pronounced in its weak form [ən] (/juːənaɪ/), whereas in "You AND I" with emphasis on *and* it is pronounced in its strong form /juː'ændaɪ/. *Of* appears in its weak form in "One of you" /wʌnəvjuː/ but in the strong form /ɒv/ in "The woman I saw a photo of" /ðəwʊmənaɪsɔɹəfəʊtəʊ'ɒv/ (GA: /ðəwʊmənaɪsɑːəfoʊtoʊ'ɑːv/).

03_weakandstrongformsRP.wav on the CD-ROM is a recording of the strong and weak forms of English function words read by an RP speaker.

Table 3.10. Weak and strong forms of some function words in English (RP).

Word	Weak form	Example	Strong form	Example
and	ənd/ən/n	*fish and chips* /fɪʃəntʃɪps/	ænd	*it was me AND him* /ɪtwəzmi'ændhɪm/
the	ðə	*open the window* /əʊpənðəwɪndəʊ/	ði	*it was THE woman* /ɪtwəz'ði:wʊmən/
a	ə	*take a look* /teɪkəlʊk/	eɪ	*not A coat but THIS one* /nɒt'eɪkəʊtbət'ðɪswʌn/
of	əv	*one of you* /wʌnəvju:/	ɒv	*not that I know of* /nɒtðətaɪnəʊɒv/
to	tə	*time to go* /taɪmtəgəʊ/	tu	*where has he gone to* /weəɹəzhɪgɒntu/
from	fɹəm	*back from work* /bækfɹəmwɜ:k/	fɹɒm	*where did you get that from* /weədɪdjʊgetðætfɹɒm/
for	RP: fə (GA:fəɹ)	*time for a break* /taɪmfəɹəbɹeɪk/	RP: fɔ: (GA: foɹ)	*what was that for* /wɒtwəzðætfɔ:/
he	ɪ	*did he come* /dɪdɪkʌm/	hi	*it was he* /ɪtwəzhi/
his	ɪz	*take his dog* /teɪkɪzdɒg/	hɪz	*that was HIS dog* /ðætwəzhɪzdɒg/
her	ə	*give her time* /gɪvətaɪm/	RP: hɜ GA: hɜɹ	*I gave it to her* /aɪgeɪvɪtəhɜ/
them	ðəm/əm	*let them go* /letəmgəʊ/	ðem	*give it to them* /gɪvɪtəðem/
was	wəz	*I was there* /aɪwəzðeə/	wɒz	*I don't know who it was* /aɪdəʊntnəʊhuɪtwɒz/
but	bət	*small but cheap* /smɔ:lbətʃi:p/	bʌt	*I'd say so but* /aɪdseɪsəʊbʌt/

The particular interplay of stressed and unstressed syllables in English is referred to as **speech rhythm**. English speech rhythm is said to be characterized by a regular occurrence of stress beats. Read utterance (10) and snap your fingers in synchrony with the stress beats occurring on the stressed syllables: the first on the syllable *ti*, the second on the syllable *mouse*, the third on the syllable *wal* and the last on the syllable *field*. You will probably snap your fingers in a fairly regular rhythm. This is why English has traditionally been called a '**stress-timed**' language (Abercrombie 1967). The time interval between two stress beats in stress-timed languages is claimed to be **isochronous**, i.e. roughly equal in time. Other languages such as French, by contrast, are supposed to be syllable-timed with each syllable occurring at regular intervals.

(10) A '**ti**ger and a '**mouse** were '**wal**king in a '**field**

In example (10), the total number of syllables between two stress beats varies: there are three unstressed syllables between *ti* and *mouse* but only one unstressed syllable between *mouse* and *wal*. If speakers aim to produce stress beats at regular, isochronous intervals, they will have to adjust the duration of the intervening unstressed syllables. The three syllables between *ti* and *mouse* should have the same duration as the single syllable *were* between the syllables *mouse* and *wal*. In order to achieve this, they have to be crushed together and be spoken much faster than the syllable *were*. This variable length of syllables has been taken as a characteristic property of stress-timed languages.

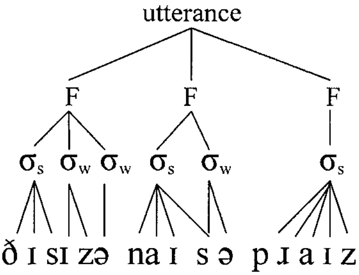


Figure 3.13. The utterance "This is a nice surprise" (in RP) divided into syllables and feet.

The time interval comprising a stressed syllable plus all following unstressed syllables up to but not including the next stressed syllable is called the **foot**. It is therefore possible to insert another level into the prosodic hierarchy introduced in Figure 3.1 – the level of the foot above syllables and below words. Figure 3.13 shows how the phonemes of the utterance "This is a nice surprise" can be grouped into syllables, which themselves can be grouped into feet (F). Traditionally, stressed or strong syllables are indicated by a subscript $_s$, whereas unstressed or weak syllables are transcribed by a subscript $_w$. The utterance consists of three feet, the first one comprising the strong syllable /ðɪs/ and the weak syllables /ɪz/ and /ə/. The second foot consists of one strong and one weak syllable, and the third of the strong syllable /pɹaɪz/ only.

The concepts of stress-timing and isochrony are very popular among researchers and language teachers. However, when you actually measure the time interval between stressed syllables in English, you will find that they do not have the same length (e.g. Roach 1982, Dauer 1983). Recent research on English speech rhythm therefore claims that isochrony is a perceptual phenomenon: speakers perceive regular beats although they do not really exist in speech (Couper-Kuhlen 1993). In recent approaches, speech rhythm is measured using other phonological domains than the foot. For example, the relative vowel length and the duration of consonants (e.g. Ramus et al. 1999) or the relative duration of stressed and unstressed syllables is compared (Bolinger 1981, Gut 2003).

3.3 The phonological word in English

All speakers of English agree that English utterances consist of words strung together. It is, however, quite controversial what a word actually is. In writing, a word is defined as the letters between two blanks – the sentence "What is a word?" thus consists of four **orthographic words**. Yet, even in writing, there are some difficult cases such as *body-paint*, *flower pot*, *What's* and *I've*. Should these cases be considered one or two words? In speaking, the problem is even greater. There are no pauses between words in an utterance, all phonemes are strung together. How can the boundaries of individual words be determined? In fact, this is a major challenge for language learners who initially find it impossible to segment the speech stream into words. Phonologists have proposed the unit of the **phonological word (pword)**. This is based on the reasoning that this is a unit in which various phonological processes apply. In particular, word stress (see section 3.3.1 below) happens within the string of speech sounds defined as the phonological word. Other processes that occur within the unit of pwords such as aspiration and velarization are described below in section 3.3.2.

Consider structures like *What's* and *I've* again. The /s/ and the /v/ in each example cannot stand on their own phonologically. They do not form a permissible syllable but are incorporated into the syllable structure of the preceding /wɒt/ and /aɪ/, respectively. These structures that are dependent on phonologically adjacent words are called **clitics**. If the word they are attached to, the **host**, precedes the clitic, they are called **enclitics**; if the clitic precedes the host it is referred to as a **proclitic**. Although they may seem similar, clitics have a different phonological behaviour than bound morphemes. With affixes, there may be a change in the pronunciation of the root. For example, *do* changes into *don't* and *will* changes into *won't*. When clitics attach, conversely, there is rarely a change in the pronunciation of the root. For example, you do not have *I have* changing to *[iv] for *I've*.

Many phonologists consider phrases such a *the house*, *a game* or *have it* phonological words. This is based on the observation that the stress pattern in these syntactic phrases can be described with the same rules as those for polysyllabic words such as *arouse*, *again* and *habit*. In fact, there is no difference in either pronunciation or stress between *abridge* and *a bridge* (both are produced /ə'bɹɪdʒ/), which is why both are considered phonological words in English.

3.3.1 Word stress in English

The languages of the world can be divided into two broad classes regarding word stress. Some languages have fixed stress, where it is always a particular syllable of a word that is stressed. This is for example the case in Czech, where stress always falls on the first syllable of every word. Similarly, in Kiswahili, it is always the penultimate (last but one) syllable in a word that receives stress. Other languages have free word stress, which means that stress can be on any syllable of a word and cannot be predicted by rule. In English, word stress is neither entirely free nor fixed. Due to historical reasons such as the Norman Conquest in 1066 and the subsequent rule of French speakers in England, English is a language that has a particular richness of loan words in its vocabulary. French words tend to have stress on the last syllable, whereas Germanic ones tend to have initial stress, which contributes to the complexity of the rules describing English word stress.

Since in English word stress is a property of each content (or lexical) word, it is referred to as **lexical stress**. Other word classes in English, such as determiners, conjunctions and prepositions, which are categorized as function words or grammatical words, do not have stressed syllables. Typically, these function words are very short consisting of only one syllable and, as explained in section 3.2.3 above, stress is a relative term and cannot be applied meaningfully to monosyllabic words.

English content words consisting of three or more syllables usually have more than one stressed syllable. Consider for example the words *entertainment* and *refrigerator*. In each of them, two syllables have stress, but the strength or degree of stress is different. In *ˌenter'tainment* the syllable *'tain* carries the main or **primary stress,** which is marked by the preceding apostrophe '. The syllable *ˌen* carries less, **secondary**, **stress** and is marked by a preceding ˌ. In the word *re'frigeˌrator*, primary stress lies on the second syllable *'fri* and secondary stress lies on the syllable *ˌra*. Phonologists thus assume that English has at least two different levels of stress. Some approaches propose even more levels of stress (e.g. Chomsky & Halle 1968), which are, however, generally difficult to distinguish by listeners.

Native speakers of English know without thinking which syllable of a word is stressed, i.e. which can receive an accent in speech. Even when encountering a completely unfamiliar word, native speakers usually agree on the syllable that can be accentuated. In contrast, learners of English find word stress very unpredictable and often hesitate which syllable to accent. (This is why good dictionaries indicate the stress pattern of each word.) What are these rules English speakers know? Several phonologists representing different theoretical traditions have worked on formulating rules that describe word stress in English. Most theoretical approaches agree that the following factors influence stress and accent placement in English words:

- the number and phonological shape of the syllables in a word (light and heavy syllables)
- the morphological complexity of the word (morphologically simple or complex words, type of affix)
- the grammatical category the word belongs to (noun, verb etc.)

In general, only heavy syllables can bear stress. Thus, if a word has only one heavy syllable, this is the one that is stressed. Examples for this are *decay*, *attract*, *agenda* and *entertain*. In *de'cay*, the second syllable is heavy with a diphthong as the nucleus, in *a'ttract*, the second syllable is heavy with two consonants in the coda position. Equally, *a'genda* has a heavy syllable as the second syllable and in *enter'tain* the last syllable is heavy. As with almost every linguistic rule, there are some exceptions, such as *co'mmittee* and *'essay*, where a light syllable is stressed. Very often, words with exceptional stress patterns are loan words for which the original stress pattern has been kept.

In morphologically complex words that consist of a stem and one or more affixes, stress patterns can be influenced by the type of affix. Some affixes are **stress-attracting**, i.e. they are always stressed. The prefix *semi-* and the suffixes

-ain,
-ee

-eer
-ality
-ese
-ette (in RP)
-esque
-ique

are always stressed in words such as *'semicircle*, *enter'tain*, *absen'tee*, *mountai'neer*, *nor'mality*, *Journa'lese*, *ciga'rette* (RP only, in GA it is *'cigarette*, see Table 3.9), *gro'tesque* and *u'nique*. Other suffixes determine that stress in the word stem (i.e. the word without any affixes) shifts one syllable to the right, as for example in

-graphy in *pho'tography*	(but *'photograph*)
-eous in *advan'tageous*	(but *ad'vantage*)
-ial in *ad'verbial*	(but *'adverb*)
-ic in *a'tomic*	(but *'atom*)
-ion in *inhi'bition*	(but *in'hibit*)
-ty in *tran'quillity*	(but *'tranquil*)
-ive in *re'flexive*	(but *'reflex*)

These suffixes are called **stress-shifting**. Suffixes that are **stress-neutral**, by contrast, do not have any influence on the stress pattern of the stem. Examples for stress-neutral suffixes are *-ness, -able, -ance/ -ence, -al, -like, -ful, -ing, -ish, -less, -ment, -ly* and *-wise*.

The grammatical category of a word also influences stress placement in English. In general, English nouns with stress on the final syllable are rare. Most English nouns are not stressed on the final syllable and examples such as *cham'pagne*, *maga'zine* (GA: *'magazine*) and *ca'noe* are loan words in which the original stress patterns have been preserved. Verbs and adjectives with final stress, conversely, are very common in English. The general rule for word stress in English nouns is that stress falls on the penultimate (the last but one) syllable if it is heavy and on the antepenultimate (the syllable preceding the penultimate syllable) if it is not. Thus, we have *u'tensil* where the second syllable is heavy and *'discipline* where it is not and the preceding syllable is stressed. Should the antepenultimate syllable also be a light one, the onset consonant of the penultimate syllable becomes ambisyllabic as in /'dɪs.sɪ.plɪn/.

Compounds in English normally receive stress on the first element. Thus, it is *'flower pot*, *'greenhouse*, *'White House*, *'redhead* and *'spoonfeed*. In fact, stress is what makes the difference between the compounds *'greenhouse* (a house in which you grow vegetables) and *'blackbird* (a particular type of bird) and the corresponding phrases *green 'house* (a house that is green) and *black* *'bird* (a bird that is black). Some exceptional compounds that can be accented on the second element include

- compounds that have a number as the first element
- compounds that end in *-ed* and function as adjectives
- proper names of people, roads, institutions and public places
- compounds that have a place, time, material or ingredient as the first element
- compounds that function as adverbs

An example of a compound with stress on the second element that has a number as the first element is *second-'class*. A compound ending in *-ed* and functioning as an adjective with stress on the second element is *bad-'tempered*. Proper names of people, places and institutions such as *Mary 'Jones*, *Charing Cross 'Road*, *Queen's 'College* and *Royal Albert 'Hall* also have stress on the second element. When the first part of a compound describes a particular place, time, material or ingredient, stress falls on the second element. This applies to English words such as *school 'dinner*, *summer 'fruit*, *olive 'oil* and *leather 'jacket*. An example of a compound functioning as an adverb is *South-'East*. However, when these compounds occur in connected speech, stress can shift to the left, especially when the word following the compound begins with a stressed syllable. Thus, *half-'timbered* becomes "a 'half-timbered house", *second-'class* turns into "a 'second-class ticket" and *after'noon* appears as *'afternoon 'tea*.

To a limited extent, stress is **phonemic** in English. This means that a small number of minimal word pairs exist that are differentiated only by stress. Table 3.11 lists some of these minimal pairs. These words are spelled in exactly the same way and have a prefix-plus-stem structure. It can be seen that these words, when functioning as verbs, have stress on the second syllable, whereas they have stress on the first syllable when they function as nouns or adjectives.

It was shown above that some compounds have **variable stress** depending on the stress pattern of the following word. This is true for a number of English words: *Heath'row* for example changes the stress pattern in *'Heathrow 'airport* and *thir'teen* does the same in *'thirteen 'girls*. Equally, *ho'tel* turns into *'hotel* in *'hotel 'management*, it is *cham'pagne* but *'champagne 'breakfast* and *can'teen* but *'canteen 'cook*. In general, four types of words can be identified in English that often change their stress pattern when they are produced in combination with other words. These are

- adjectives ending in *-ese*
- adjectives starting with *un-*
- adjectives that can be used adverbially
- numerals between 12 and 99, excepting 20, 30, 40, 50, 60, 70, 80 and 90

For example, *Japa'nese* changes into *'Japanese girl*, *un'happy* turns into "the 'unhappy boy", *up'stairs* is produced with initial stress in *'upstairs room*, and *twenty'one* turns into "with 'twentyone white swans".

Table 3.11. Some English minimal word pairs that are differentiated by stress only.

Verb	Noun	Adjective
per'fect		*'perfect*
ab'stract	*'abstract*	*'abstract*
sub'ject	*'subject*	*'subject*
up'date	*'update*	
in'sert	*'insert*	
ob'ject	*'object*	
per'mit	*'permit*	
in'crease	*'increase*	
de'crease	*'decrease*	
con'duct	*'conduct*	
ex'port	*'export*	
in'sult	*'insult*	
pro'test	*'protest*	
pro'duce	*'produce*	
sur'vey	*'survey*	
re'cord	*'record*	
con'tract	*'contract*	

For many English words, two stress patterns are possible, even within the same variety of English. This variable stress reflects ongoing language change. In RP, *'exquisite* (used only by conservative speakers any more) is changing into *ex'quisite* (used primarily by younger speakers), *dis'pute* is also heard as *'dispute*, *'comparable* can also be pronounced *com'parable* and *'laboratory* also occurs as *la'boratory*. Equally, *eti'quette* is changing into *'etiquette*, *'integral* into *in'tegral*, *'kilometre* into *ki'lometre*, *'lamentable* into *la'mentable*, *'communal* into *co'mmunal* and *'formidable* into *for'midable*. A look at the history of English shows that stress patterns of words were particularly variable during the Early Modern English period when an enormous amount of loan words entered the language from Latin or French. Very often, the 'original' Latin or French stress pattern existed alongside a version in which stress followed the rule of initial stress for many Germanic words. For example *e'ssay*, *a'spect* and *dis'course* could be heard at that time as well as the stress pattern that has survived in contemporary English.

Differences in stress furthermore exist between present-day varieties of English. Some differences between word stress in RP and GA are shown in Table 3.12. It also includes two examples of differences in compound stress between the two varieties of English.

Table 3.12. Differences in word stress between RP and GA.

RP	GA
ciga'rette	*'cigarette*
'harass(ment)	*ha'rass(ment)*
la'boratory	*'laboratory*
con'troversy	*'controversy*
re'source	*'resource*
re'search	*'research*
'ordinarily	*ordi'narily*
court 'martial	*'court martial*
country 'house	*'country house*

It was stated at the beginning of this section that many multisyllabic words in English have two stressed syllables, one bearing primary stress and the other bearing secondary stress. The word *ˌenter'taining* for example has secondary stress on the first syllable and primary stress on the third syllable. This difference is usually illustrated in a tree indicating the relative stress of the syllables and feet of the word. The phonological theory called **Metrical Phonology** claims that stress is distributed depending on the relative weight of the syllables and feet of a pword. It thus assumes that stress is assigned on each tier (another technical term for level) of the prosodic hierarchy. Figure 3.14 shows the proposed metrical structure of *ˌenter'taining*. The word consists of two feet, and each foot consists of two syllables. It is a central claim of Metrical Phonology that only one of a pair of elements on one tier can receive stress. In our example, of the two feet, the first one is weak and the second strong. Similarly, the two syllables in each foot have a strong-weak pattern.

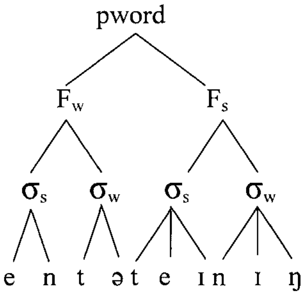


Figure 3.14. The stress pattern of the word ˌenter'taining.

Since the first foot is weaker than the second one, the strong syllable /en/ of the first foot has relatively less stress than the strong syllable /teɪ/ of the second foot. Thus, primary stress lies on /teɪ/ and secondary stress on /en/. An alternative illustration of the stress pattern of words is the **metrical grid**. Figure 3.15 shows how the weight of the syllables and feet of the word ˌ*enter'taining* combine to its final metrical shape. Each syllable receives one metrical weight on the syllable level, both strong syllables receive additional weights on the foot level and the strong foot receives another weight at the pword level.

```
          X          (pword level)
X         X          (foot level)
X    X    X    X     (syllable level)
en   tə   teɪ  nɪŋ
```

Figure 3.15. The stress pattern of the word ˌenter'taining.

3.3.2 Phonological processes occurring at the level of the pword (advanced reading)

Apart from word stress, there are other phonological processes in English that occur at the level of the phonological word (pword). For example, the aspiration of voiceless plosives is restricted to the beginning of a pword. As formulated in rule (1) in section 3.1.2 above, the three voiceless plosives of English are aspirated when they occur at the beginning of a pword, but not in any other position. This is why listeners can make the distinction between *rice pot* and *rye spot.* Both have exactly the same segmental form /raɪspɒt/. In the latter case, however, the /p/ is not aspirated since it occurs in combination with /s/ at the onset of the second syllable. For the same reason, many phonologists assume that "to go" as well as *today* constitute pwords: both have aspirated [t^h]. Likewise, the glottalization of the voiceless plosives and the velarization of /l/, which both occur at the end of syllables, reflect the boundaries of pwords. Since *it'll* has velarization of the /l/, this can be taken as an indication that this phrase constitutes a pword. Nevertheless, in phonological theory, there is still some uncertainty as to which entities in English can be considered pwords. Especially the status of function words and of host+enclitic sequences is still unresolved and seems to be influenced by focus and speaking style (Hall 1999, Raffelsiefen 2004). For example, while in fast and informal speech *it'll* forms one pword, in careful pronunciation the phrase is pronounced "it will" in two separate pwords.

There is some phonetic evidence that the pword is indeed a unit represented in speakers' brains. Experiments involving the exact measurement of articulatory movements of the lips, tongue, velum and vocal folds revealed that there are distinct differences between the articulation of pword-initial consonants and vowels compared to when these sounds occur in the middle of pwords. (Methods with which these articulatory movements can be studied are described in section 2.8). For example, for a /n/ at the beginning of a pword, the tip of the tongue has more and longer contact with the alveolar ridge than for a /n/ at the beginning of a syllable (Fougeron & Keating 1997). The voice onset time (VOT) of a pword-initial voiceless plosive is longer than that of the same consonant in word-medial position, if stress is kept constant (e.g. Turk & Shuttack-Hufnagel 2000). Vowels at pword boundaries furthermore show larger lip opening movements than at syllable boundaries (Byrd & Saltzman 1998, Cho 2002, 2005).

3.4 Theories of the acquisition of English phonology (advanced reading)

How do children and other language learners acquire the phonological representation of phonemes, features, syllables and phonological words for a language? As yet we do not have the means to study directly the structures and processes of the brain relevant for language acquisition. We therefore have to rely on the indirect study of mental representations of phonological structures and rules. For such an indirect study, one can, for example, analyse how speakers pronounce and stress unfamiliar words and draw conclusions about rules they apply automatically. Alternatively, one can present speakers with unfamiliar words and syllables and ask them to judge whether they are 'good' words and syllables in English. Obviously, this method cannot be applied with very young children, with whom one has to rely mainly on their reactions to sounds and words played to them (see section 6.8.1).

3.4.1 English phonology in first language acquisition

There are obvious differences between the pronunciation of words by young children and by adult speakers of a language. It is, however, debated whether productions such as [fɪt] for *fish* can be interpreted to mean that the child's mental representation of the phonological structure of the word *fish* is /fɪt/. The question is: does the child say [fɪt] because he or she 'knows' the word to be pronounced in this way or does the child 'know' that *fish* is pronounced /fɪʃ/ but pronounces it [fɪt] for some other reason? Some researchers argue that the child has an adult-like representation of words but that it is the child's inadequacies in motor control that lead to the deviant production. This view is supported by observations that children understand much more than they are able to produce. When a child understands [fɪʃ], then his or her mental representation is probably also /fɪʃ/. Yet another theoretical approach claims that children have the adult form stored for word recognition but use a different, also mentally represented, 'output' form for their productions. The first phonological unit that seems to be acquired by children is the word (Menn & Stoel-Gammon 1995). Most children have a mental representation of syllables and phonemes only by age six or seven – this seems to be a consequence of learning to read.

The acquisition of word stress proceeds in various stages. The first words children produce have a **trochaic pattern**, i.e. they comprise two syllables of which the first one is strong and the second one weak. It is therefore assumed that the initial mental representation of the metrical structure of words (the minimal word form Wd $_{\text{min}}$) consists of a single foot (F), which comprises two syllables (σ), the first of which is strong, as shown in Figure 3.16.

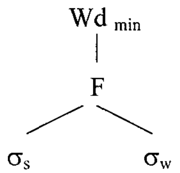


Figure 3.16. The minimal word form as proposed by Fee (1995).

This basic mental representation is reflected in child word forms that constitute mismatches to adult forms such as ['nana] for *ba'nana* (/bə'nɑːnə/). The weak syllable /bə/ preceding the strong syllable /'nɑː/ is not produced because it does not fit into the structure of the minimal word. Children might even **overgeneralise** this trochaic pattern and produce a stress shift in iambic words (which have a weak-strong pattern): *gi'raffe* is often pronounced *'giraffe* or just *'raffe* [waf] by very young children. It is only at around age 2.5 that in trisyllabic words of the SWS (strong weak strong) kind such as *ˌenter'tain* all syllables are produced, but the two strong syllables typically receive the same degree of stress. It is assumed that children only produce iambic word stress and words with an SWS pattern with primary and secondary stress correctly when their mental representation of the metrical structure of words comprises more than one foot.

The theory of **Natural Phonology**, founded by Stampe (1979), is also concerned with the explanation of phonological acquisition. It postulates universal natural phonological preferences that are inborn and form part of human cognition. These include processes that improve the articulation or perception of language and are conceptualized to have a phonetic basis. The process of first language acquisition is understood as a selection of those processes which conform to the language-specific requirements. This is achieved by a suppression, limitation and ordering of the natural processes or preferences, which eventually allows the correct production and perception of the target language's phonological categories. For example, a child is said to prefer the articulation of voiceless plosives to that of voiced plosives (this can be documented in productions such as [dɒk] for /dɒg/). If the language that is being acquired contains voiced plosives, this natural preference has to be suppressed.

3.4.2 English phonology in second language acquisition

As has been shown in section 2.7.2, many differences between the pronunciation of native speakers and second language learners can be explained with the fact that language learners do not have the same mentally stored motor patterns for the activity of the muscles involved in the production of a particular sound sequence, or cannot execute these patterns as fast as native speakers can. This section describes other reasons why some language learners' pronunciation differs from that of native speakers. These reasons lie in the mental representations of phonological units and rules competent speakers of a language have. Language learners, whose first language (L1) requires different phonological representations of phonological units and rules from the second language (L2), might, at least at the beginning of language learning, inappropriately use these units and rules in the L2.

This negative transfer of linguistic knowledge has been noted in many studies. Some learners will make errors that involve altering the syllable structure and syllabification of the L2 when their L1 and L2 have different syllable structures and different syllabification rules (e.g. Ishikawa 2002, Whitworth 2003). For example, some German learners of English insert a glottal stop between a word ending in a vowel and the next one beginning with a vowel such as *the apple*. They probably do this because in German syllables without a consonant in the onset position are rare. If syllables begin with a vowel in spelling (such as *Apfel*), usually a glottal stop /ʔ/ is produced in the onset position (/ʔapfl/). Phonological knowledge of the L1 might also interfere with the perception of a second language. Some learners of English with an L1 that has different rules of syllable structure find it difficult to recognize the number of syllables in an English word or misperceive word boundaries (e.g. Ishikawa 2002, Broselow 1984).

A theory of the acquisition of second language phonology within the framework of Natural Phonology was proposed by Dziubalska-Kołaczyk (1990). In contrast to first language acquisition, which is considered to proceed largely subconsciously, second language acquisition (SLA) of phonology is assumed to be based on learning in a controlled and conscious manner. In the course of SLA, access to universal processes is considered to be more difficult than during first language acquisition, as the phonological system of the L1 is already established and thus limited to selected processes, underlying representations as well as rules. The essential prerequisite for the L2 learning process is that the language learner can access the universal processes. This allows him or her to modify the suppression, limiting and ordering of the universal process types of the L1. Acquiring the L2 phonology might involve the unsuppression of processes which were suppressed in L1 acquisition and the reordering of process types in comparison with the ordering of the L1 phonology. It is assumed that access is facilitated by favourable psycholinguistic conditions such as the amount of formal language instruction, the learner's attitude towards the L2 and his or her general linguistic aptitude.

The Natural Model of phonological acquisition distinguishes between phonemes and surface phonetic segments, and postulates that a speaker processes the sound intention (phoneme) rather than the actual surface realization. A language learner's task is therefore to acquire L2 sound intentions through the means of perceiving L2 surface realizations. As shown in Figure 3.17, in an initial state, a language learner relates L2 surface realizations to L1 phonemes. As a first step, L2 surface realizations need to be perceived without reference to L1 categories. This is supported by a high frequency of L2 realizations, formal language instruction and a favourable attitude towards the L2 and the language learning process by the language learner. In a second step, this perception leads to a representation of L2-specific phonemes (sound intentions), which then triggers the reordering and suppression (or unsupression) of natural processes. The task of a Polish speaker acquiring the aspiration of voiceless plosives in English, for example, can be described as the unsuppression of a natural process. For L1 acquisition of Polish, where aspiration appears only optionally in emphatic styles, it is hypothesised that the natural process of aspiration was suppressed. This suppression now needs to be reversed in order to allow for the pword-initial aspiration of voiceless plosives in English.

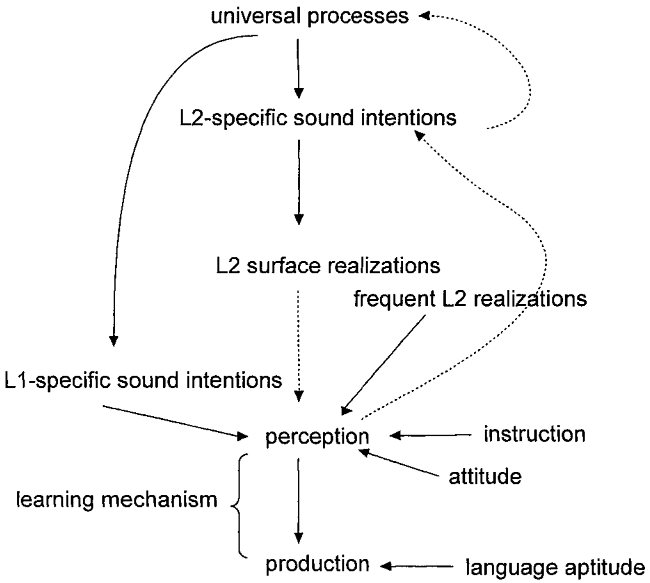


Figure 3.17. The Natural Model of L2 acquisition of phonemes; reprinted with permission from Dziubalska-Kołaczyk (1990).

Archibald (1994) developed a model of L2 word stress learning based on the theory of **Universal Grammar** (UG). This theory claims that all humans are innately endowed with a language faculty that contains grammatical principles and parameters. Principles are grammatical structures shared by all languages of the world, whereas parameters have to be set according to language-specific requirements. It is argued that only inborn knowledge in the form of language-universal principles and parameters can account for the relative uniformity and speed of first language acquisition, which proceeds even with little positive evidence of some language structures in the ambient language ('poverty of the stimulus'). The question whether UG is still available, i.e. whether parameters can be reset, in L2 acquisition, is a matter of dispute. Archibald envisages the L2 word stress learning process as an interaction between UG and the input of the linguistic environment. Learning is conceptualised as the (re-)setting of language parameters. This is influenced by three phenomena: indirect negative evidence in the ambient language, ability to choose appropriate cues and lexical dependency. In an initial state, learner speech will be highly variable or show a preference for the L1 parameter setting, as the parameter for the L2 has not yet been set. After the threshold has been crossed, however, the parameter is set in the L2 and variation should stop. In order to be able to reset parameters, the learner is presumed to possess the ability to choose appropriate cues from the input language. A Czech learner of English, thus, resets his or her parameter – which is set to stress on the initial syllable of each word – to learn the relatively free lexical stress of English.

3.5 Exercises

1. Which units can speech be divided into? Which phonological reasons can be given to support the linguistic plausibility of each of these units?
2. What is the difference between a phoneme, an allophone and a phone? Describe some allophonic rules of English.
3. What is the phoneme inventory of RP and of GA, and how do they differ?
4. What is the difference between phonemic and phonetic transcription?
5. What is the basic structure of an English syllable? Which consonants can occur in which position?
6. What is stress? Which rules for English word stress can be formulated? Why is it difficult to capture English word stress with phonological rules?
7. Transcribe the following words phonemically in both RP and GA (check a good dictionary for the solution):

a. jungle
b. abbreviation
c. confused
d. bridges
e. camera
f. obtain

8. Analyse the following productions by German learners of English. How do they differ from native pronunciations? Can you describe the errors with a phonological rule?

a.	foot	[fu:t]
b.	jeans	[tʃi:ns]
c.	grab	[gɹæp]
d.	houses	[haʊsɪz]
e.	buzzed	[bʌst]

9. Draw the syllable structure of the words *spring*, *faster* and *buzzard*. Remember the rules of syllabification and ambisyllabicity!
10. Divide the following utterances into feet.

    a. It's raining cats and dogs.
    b. This is the house that Jean built.
    c. A bird in the hand is worth two in the bush.

11. In order to illustrate the differences in vowel pronunciation between the world-wide varieties of English, Wells (1982) developed the **lexical sets** – a list of words containing all distinctive vowels in RP and GA. Listen to the recordings of these lexical sets pronounced by an RP speaker (recording 03_exercise11_RP.wav on the CD-ROM), a GA speaker (recording 03_exercise11_GA.wav on the CD-ROM) and a speaker from Nigeria (recording 03_exercise11_Nig.wav on the CD-ROM). Transcribe the vowels phonemically and comment on the differences between the three accents.

3.6 Further reading

The phonemic inventory of RP is described in Roach (2004), that of GA in Giegerich (1992, chapter 3). A description of American English (Far-Western/Mid-Western accent) can be found in Ladefoged (1999). The phoneme inventories of other varieties of English are described in Kortmann & Schneider (2004).

The IPA and its history together with descriptions of the phoneme inventories of a number of different languages are contained in the *Handbook of the International Phonetic Association* (1999). The latest version of the IPA can be downloaded from the homepage of the International Phonetic Association at http://www.arts.gla.ac.uk/ipa/index.html. This website furthermore offers downloads of various phonetic fonts for the computer. An excellent practice CD-ROM containing the sounds of the International Phonetic Alphabet is available from University College London (produced in 1995).

The pronunciation dictionaries by Wells (2000) and Jones (2006) give the transcription of English words and contain a CD-ROM with native pronunciation of these words. Eckert & Barry (2005, chapters III to VI) provide entertaining practice materials on all English phonemes and allophones, stress and weak forms that prove especially challenging for German native speakers. Roach (1991, chapter 12) presents the strong and weak forms of many English words together with training material on audiocassettes.

For advanced readers, Lass (1984) offers an elaborate discussion of feature systems, and feature trees are described in Halle (1992) and Kenstowicz (1994). Giegerich (1992, chapter 6) gives more details on the English syllable structure. See Hogg & McCully (1987) for an overview of Metrical Phonology. Ramus et al. (1999) provide a detailed discussion of traditional and recent measurements of speech rhythm.

For especially interested readers, a concise introduction to the theory of UG can be found in Radford (1997, chapter 1). Vihman (1996, chapter 6) gives an overview of the emergence of phonological structure in first language acquisition. Leather & James (1996) and Hansen & Zampini (2008, chapters 2 to 5) summarize the major issues in second language acquisition of phonology.

4 The Phonology of English: Intonation

Chapter 3 showed how speech can be divided into units of different size and explained that these units are thought to be ordered hierarchically: phonemes, which themselves consist of bundles of phonetic features, form syllables, which in turn combine to make up feet and prosodic words. This can be illustrated with the **prosodic hierarchy** (see Figures 3.1 and 4.1). According to the different levels of the prosodic hierarchy, phonology can be divided into two branches. **Segmental phonology** is concerned with phonological descriptions of the segmental (or phoneme) inventory of languages, whereas **suprasegmental phonology** (also called **prosody**) deals with all units higher than the segment, i.e. syllables, feet, prosodic words and the higher levels intonation phrase, utterance and discourse.

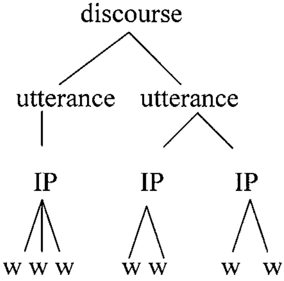


Figure 4.1. Some of the higher levels of the prosodic hierarchy: words (w), intonation phrases (IP), utterances and discourse.

This chapter describes the three highest phonological levels of the prosodic hierarchy illustrated in Figure 4.1. When speakers produce words in a row, we can usually observe that they are structured: individual words are grouped together to form an **intonation phrase.** Section 4.1 explains the phonological unit of intonation phrases and shows how English words combine into intonation phrases. Intonation phrases can coincide with breath groups that were described in section 2.1, but they do not have to. Often, a breath group contains more than one intonation phrase. As with all other phonological units, it is assumed that speakers have a mental representation of intonation phrases, i.e. they know how to produce speech structured into intonation phrases and they rely on this knowledge when listening to the speech of others.

Within an intonation phrase, there is typically one word that is most prominent. Section 4.2 explains the phonological rules describing which of the words in an English intonation phrase is most prominent. Moreover, the pitch of a speaker's voice changes across intonation phrases. The linguistic use of such pitch movements is called **intonation** and is described in section 4.3. When producing speech, speakers can combine intonation phrases into longer **utterances** by means of intonational phrasing and by intonation. Some utterances might contain just one intonation phrase, others might contain several of them. Moreover, speakers can put utterances together to form larger stretches of speech or **discourse**. This also happens with the help of intonational phrasing and intonation, as described in sections 4.1 and 4.3.3. Section 4.4 explains how language learners acquire these phonological units and their rules.

4.1 Intonational phrasing in English

Speakers do not often produce single words when they communicate with each other but usually string several words together. However, speakers do not simply produce one word after the other. In speech, particular groups of words are closer-knit than other groups of words. This is something listeners expect and make use of in interpreting speech. Many people can even hear these word groups in some (related) languages they do not speak – which shows that it is not the meaning alone that makes words belong together but phonological properties of speech (see section 5.5.1 for more details). The groups of words that belong closely together are called phrases. In order to avoid confusion with syntactic phrases, these phrases are referred to as **intonation phrases**. (Note that many other terms are in use for this type of phrase, including intonation group, phonological phrase, tone unit, tone group, breath group and sense group). In general, speakers use **intonational phrasing** in order to structure their discourse into units of information. They thus show a listener which words of an utterance belong together – which words form an intonation phrase. Intonational phrasing in English can have a meaning-distinguishing function. Consider utterances 11a and 11b:

(11a) He washed and fed the dog
(11b) He washed | and fed the dog

If the utterance "He washed and fed the dog" is produced as one intonation phrase, its meaning is that a person both washed and fed a dog. Conversely, if the same utterance is produced as a sequence of two intonation phrases with an **intonation phrase boundary** after *washed* (indicated by the symbol |), the meaning of the utterance changes into 'someone who washed himself and fed a dog'. The end of an intonation phrase – an intonation phrase boundary – is sometimes but not always signalled by a short pause. (Section 5.5.1 explains other ways of indicating the boundaries of intonation phrases, such as final syllable lengthening, anacrusis and change of pitch level).

All examples in this chapter can be listened to on the CD-ROM. Please bear in mind that these examples were produced by a speaker reading printed sentences. As explained below, intonation in spontaneous English speech may show different characteristics.

When considering the function of intonational phrasing it is important to be aware of the fact that there are many different kinds of spoken language, many different kinds of **speaking styles**. Reading a text aloud and reciting a poem or a speech learned by heart require different cognitive processes than telling a story or participating in a discussion. The first types of speaking style are called reading style and prepared speech respectively, whereas the second type of spoken language is referred to as spontaneous or free speech. In spontaneous speech, in contrast to prepared speech, speakers have to speak and plan their speech at the same time. While articulating one portion of the 'message' the speaker will already have to plan and prepare what to say next. Spontaneous speech is therefore delivered in chunks, and 'speaking time' typically consists of a large part (usually up to 30%) of pauses. Clearly, the speaking style influences intonational phrasing. In spontaneous speech, speakers break their message up into smaller chunks than in reading style or prepared speech – both because they need time to plan ahead and because it allows listeners to follow their speech. The words joined into these chunks usually form one piece of information, a meaningful unit called **topic unit**. Often, in spontaneous speech, each intonation phrase corresponds to one meaningful unit. When producing prepared speech or reading a printed text aloud, speakers do not need to plan what to say next. They do not need time to search for the right words and put them into a grammatical order. Consequently, this type of speech is faster and contains fewer pauses. Moreover, intonational phrasing in this type of speech has a very close relationship with punctuation and the grammatical structures of written language.

When producing speech, speakers join several intonation phrases together and produce longer utterances and discourse. With intonation phrase boundaries, speakers can indicate how closely the individual intonation phrases in an utterance belong together. Especially when reading out a text, speakers vary the length of pauses between intonation phrases according to the meaning relationship between these intonation phrases.

(12) He washed | and fed the dog || Then he turned to his own meal ||

In example (12), the intonation phrase boundary between the words *washed* and *and* is shorter (and therefore called a minor boundary) than that between *dog* and *then* (which is referred to as a major boundary). A minor intonation phrase boundary is transcribed with a single |; a major intonation phrase boundary is transcribed with a ||. The major boundary in example (12) indicates the beginning of a new topic – the speaker's now turning to his own meal.

There are many other examples of utterances in which speakers use intonational phrasing in order to indicate the structure of their speech. Coordinated structures, for example, are usually separated by intonation phrase boundaries. This is shown in example (13), which represents an utterance with a coordinate structure consisting of the two components "You could put it over there" and "or leave it where it is". Especially when the two components of such structures are placed in contrast to each other by the speaker, they are separated by an intonation phrase boundary.

(13) You could put it over there | or leave it where it is ||

Similarly, items in enumerations are often produced as separate intonation phrases. This is shown in example (14). It is, however, also possible to produce such lists as a single intonation phrase, as shown in example (15).

(14) I'll buy tomatoes | lettuce | a cucumber | and some carrots ||
(15) Her trousers are red blue green and yellow ||

In an utterance, 'heavy' subjects that consist of noun phrases with many words are typically produced as separate intonation phrases in English. This is illustrated in example (16), in which an intonation phrase boundary is produced after the heavy subject "The inhabitants of our beautiful village".

(16) The inhabitants of our beautiful village | do not care for this bypass ||

Similarly, tags in questions tend to be separated by intonation phrase boundaries. This is shown in examples (17) and (18), which both have so-called reverse-polarity tags, i.e. the verb in the tag is negated when the verb in the main clause is not (example 17) and vice versa (example 18).

(17) We should do it now | shouldn't we ||
(18) You didn't see him | did you ||

Intonational phrasing can furthermore indicate whether the information contained in a relative clause is additional or essential. If it is additional, the relative clause is called non-defining; defining relative clauses contain information about the noun they refer to that is essential for the identification of this noun. In writing, non-defining relative clauses are usually indicated by inserted commas. Non-defining relative clauses seem to be rare in spontaneous language. In reading style and prepared speech, as in examples (19) and (20), they are often separated by intonation phrase boundaries. Thus, the speaker knows that the information "an old woman of about 90" is not essential for the listener in order to know which neighbour he or she is talking about – the speaker probably only has one neighbour. Equally, the speaker of utterance (20) indicates that he or she only has one partner and that the information "who lives in Munich" is merely additional. Defining relative clauses, conversely, tend not to be separated by intonation phrase boundaries in speech. This is shown in example (21). Here, the speaker indicates to the listener that he or she has more than one daughter and that it is the one who lives in Munich who is being referred to.

(19) My neighbour | an old woman of about 90 | has given me this present ||
(20) My partner | who lives in Munich | is 35 years old ||
(21) My daughter who lives in Munich came to see me ||

Speakers use intonational phrasing in order to indicate whether a piece of information is defining or non-defining also in utterances that do not contain relative clauses. In example (22), the dog's name is additional and not necessary for the identification of the dog (the speaker only has one) and thus separated by an intonation phrase boundary. Similarly, in (23) the information "opposite the house" is merely additional. Such utterance elements that contain circumstantial information and which are typically produced as separate intonation phrases are referred to as parenthetical structures.

(22) This is my dog | Tracy ||
(23) Look at the huge tree | opposite the house ||

Furthermore, clause-modifying adverbials in English are commonly separated by an intonation phrase boundary. This applies to adverbials preceding the main clause as in utterances (24) and (25) as well as to the post-modifying adverbial in utterance (26).

(24) Unfortunately | there wasn't enough time for us to see the museum ||
(25) During the last year | not a single bird has been sighted ||
(26) The fight against global warming has begun | apparently ||

Similarly, vocatives and imprecations, when in initial position, tend to be produced as separate intonation phrases. Example (27) illustrates an utterance with a vocative (i.e. an utterance element used to call someone or attract someone's attention); example (28) shows how an imprecation – the calling on a higher power or an expletive – is usually separated by an intonation phrase boundary.

(27) Gillian and Tom | we are leaving ||
(28) For heaven's sake | leave him alone ||

When a subject or an object of a clause is '**topicalised**' in English, it is usually produced as a separate intonation phrase, too. Topicalisation means that the element is in some way emphasised, often by moving it to the beginning of the utterance as in (29) or by recapitulating it at the end of the utterance (30).

(29) Quite interesting | the film | wasn't it ||
(30) Always looks pretty | Jean does ||

Syntactic structures that topicalize sentence elements are cleft sentences and pseudo-cleft sentences. These structures are also typically produced as separate intonation phrases in speech. Example (31) illustrates a cleft sentence and example (32) illustrates a pseudo-cleft sentence.

(31) It was Tom | who was late again ||
(32) What I have also wanted to know | is what you did last Friday ||

When analysing intonational phrasing in spoken English, it becomes obvious that the relationship between the grammatical structures described in examples (13) to (32) and intonational phrasing depends crucially on the speaking style. In reading style, a close relationship between punctuation and type of intonation phrase boundary can be observed – sentences separated by full stops in writing are typically separated by major intonation phrase boundaries; clauses and other sentence elements separated by commas in writing are usually produced as intonation phrases with minor boundaries. Practised readers will produce every printed sentence and every clause as one intonation phrase, and often the length of the pause that is inserted is correlated with the type of syntactic structure. Pauses after complete sentences are longer than those after clauses, which in turn are longer than those after syntactic phrases and subordinate structures.

Similarly, empirical studies have shown that in prepared speech and story-telling the correlation between intonation phrases and syntactic boundaries is very high.

Spontaneous speech, however, is structured by intonation phrases that do not correspond well to syntactic units (e.g. Yang 2004). Speakers do not produce complete grammatical sentences in conversations and spontaneous interactions. Many utterances are elliptical, i.e. do not have a verb or a subject, and are thus much shorter than written sentences. The relationship between type of boundary and pause length is less obvious in spontaneous speech than in reading style (Keseling 1992). Therefore, it is much more difficult to formulate phonological rules that predict where intonation phrase boundaries will occur in spontaneous speech.

The length of intonation phrases seems to be fairly limited. In an analysis of a corpus of spontaneous British English, Crystal (1969) found that, on average, intonation phrases are five words long. Only few intonation phrases (20%) are longer than eight words. This means that, on average, intonation phrases are between one and two seconds long (Tench 1996). This, however, is not due to physiological limitations of speakers. Remember that in section 2.1 it was shown that speakers can produce more than eight seconds of speech before they need to take a breath. The relatively short duration of typical intonation phrases is more likely due to limitations of listeners: it is more difficult to understand speech that is structured into larger units. Think about how difficult it is to follow a presentation that consists of written language read aloud in comparison to a presentation delivered as free speech! On the whole, intonation phrases in spontaneous speech are shorter than when a speaker is producing prepared speech or when reading a text aloud. As explained above, this reflects the fact that in spontaneous speech speakers need to think about what to say next before articulating.

4.2 Nucleus placement in English

Section 3.3.1 showed that every lexical word in English has at least one stressed syllable, i.e. one syllable that is normally accented when the word is pronounced. Consequently, in intonation phrases, which consist of several lexical words, there are several potential places for accents. In every intonation phrase, one of the words (or, to be more precise: one of the syllables of one of the words) will receive the strongest accent; this is called the **nucleus** of the intonation phrase (not to be confused with the nucleus of a syllable!). **Nucleus placement** in English (sometimes also referred to as 'sentence stress'), like intonational phrasing, can distinguish the meaning of utterances. Roach's (1991: 173) examples (33) and (34) illustrate this:

(33) I have plans to LEAVE
(34) I have PLANS to leave

If, as in utterance (33), the nucleus or main accent of the utterance falls on *leave* (the nucleus is printed in capital letters), the utterance means roughly 'I want to leave, I am planning to leave'. If, however, the nucleus is on *plans* as in (34), the speaker expresses that he or she is going to leave some plans (e.g. a drawing of a house by an architect) somewhere.

It is difficult to establish exact rules for nucleus placement in English. Whether an accent is appropriate on one or the other word in an utterance depends a lot on the situational context, the participants of a conversation, the type of speech produced and many other factors. This means that a speaker may produce the same utterance with different accentual patterns for different purposes. For example, a speaker might say

(35) Tomorrow | I am going HOME ||

when telling someone else that she is going home the next day. Alternatively, the speaker might produce this utterance as a reply to the question "Are you going to London tomorrow" and tell her conversation partner that she is not going to London, but rather home (e.g. to Bristol). The same utterance with a different accentuation pattern such as in (36)

(36) Tomorrow | I AM going home ||

might be produced by a speaker wishing to imply that after they have been planning to go home for several weeks it will finally become true tomorrow.

Despite this difficulty of formulating exact rules, some tendencies of nucleus placement in English can be described. As a rule of thumb, it is the last content word of an intonation phrase that receives the nuclear accent. Content words are words that belong to the classes noun (such as *house*), verb (such as *go*), adjective (such as *green*) and adverb (such as *beautifully*). Other types of words such as prepositions (*of*), pronouns (*her*), determiners (*a*) and auxiliaries (*am*, *will*), which are referred to as grammatical or function words, rarely function as the nucleus in an intonation phrase. Thus, in example (37), in which the last word *at* is a function word, the nucleus is placed on the last content word, which is *looking*. Equally, in example (38), the last two words, since they are function words, do not receive the nucleus, which is instead placed on *books*.

(37) Who is she LOOking at ||
(38) She forgot to bring the BOOKS with her ||

In general, nouns are the most likely type of content word to function as a nucleus in an intonation phrase. In cases where a verb follows a noun, it is the noun that receives the nuclear accent and not the verb (examples 39 and 40). This is true despite the fact that, when there is no preceding noun (but e.g. a pronoun instead), the verb is accented – compare examples (40) and (41). This tendency to place the nucleus on nouns even when another content word follows it, can also be seen in many English fixed expressions such as "They got on like a HOUSE on fire", "Let's see which way the WIND'S blowing", "Keep your FINgers crossed" and "He's got a SCREW loose".

(39) The PHONE rang ||
(40) He got his BIKE repaired ||
(41) He got something rePAIRED ||

Some English nouns such as *things*, *people* and *places*, however, have very little meaning and thus tend not to be accented even when they appear in the last position of an intonation phrase (see examples 42 to 44). This is probably because they are used by speakers as vague expressions that do not refer to particular objects or living beings like 'real' nouns.

(42) He's always iMAgining things ||
(43) You shouldn't aNNOY people ||
(44) I've always loved GOing places ||

Only very few function words can receive the nucleus in English. One of them is the preposition of a phrasal verb. Phrasal verbs are verbs that consist of a verb plus preposition (such as *up*, *down*, *off*, *in*) or adverb (such as *back*, *away*). Thus, it is *take IN* (a phrasal verb), but *SIT in* (a prepositional verb). Examples (45) and (46) illustrate that the preposition of a phrasal verb receives the nucleus of an intonation phrase in spoken English. The only exception to this rule is when a noun is inserted between verb and preposition (examples (47) and (48)).

(45) Please take it OFF ||
(46) Just switch it ON then ||
(47) Please take your SHOES off ||
(48) Just switch the LIGHT on then ||

Other function words that attract the nucleus in English are *too* (see example 49), *either* (example 50), *as well* (example 51) and *anyway* (example 52).

(49) I want some spaghetti TOO ||
(50) He didn't see me EIther ||

(51) I am fond of pizza as WELL ||
(52) She was going home Anyway ||

Looking at nucleus placement from a pragmatic point of view, i.e. its function in a conversation, it turns out that speakers can signal a variety of meanings and assumptions by producing the nuclear accent on one particular utterance element. In general, in English utterances, it is **new information** in an utterance that is accented rather than **given information**. The term new information refers to things, events or concepts that have not been mentioned or shown yet in previous conversation and that the speaker assumes the listener does not currently have in his or her consciousness. Conversely, given (or 'old') information has previously been referred to in the conversation or is in some other way present in the listener's mind. Thus, a speaker will typically produce utterance (53) as a reply to the question "Do you like the food".

(53) It is very deLIcious food

The nucleus is on the syllable *li* in *delicious* rather than on *food*. The word *food* had just been mentioned in the question and thus constitutes given information, whereas *delicious* has not previously been referred to and thus constitutes new information.

Given information does not always appear as the exact repetition of the same word in utterance. Sometimes, speakers use synonyms, hyperonyms or hyponyms for words mentioned before, but these do not receive the nucleus. In example (54), the second listener does not place the nucleus on *games* – although it is the last noun of the intonation phrase – because the first speaker has just used the word *football*, a type of game, which caused the concept 'game' to be present in the listener's mind. Consequently, *games* is treated as given information. Equally, in example (55) the mentioning of poodles raises the general topic of dogs in the conversation, so that the word *dogs* later in the intonation phrase can be treated as given information.

(54) Do you like FOOTball ||
No I HATE games ||
(55) Breeding poodles and Other dogs ||

Nucleus placement thus shows whether a speaker considers an element in an utterance as new or given. Alternatively, speakers can use nucleus placement in order to imply givenness of some piece of information. Consider examples (56) and (57). In example (56), the speaker does not accentuate *joke*, treating it as given information and thus implying that Linda's assertion that she hates tomatoes is a joke. If Linda was joking, then Tom's joke is just another joke –

given information – and the word *joke* does not receive the nucleus. Had Linda's assertion that she hated tomatoes been taken seriously, the speaker would have placed the nucleus on *joke* as in example (57). In this utterance the implication is that first Linda said something and then Tom made a joke, which in that case – since no other joke had been made yet – has the status of new information.

(56) Linda said she hated toMAtoes | and then TOM made a joke ||
(57) Linda said she hated toMAtoes | and then Tom made a JOKE ||

Nucleus placement thus serves to highlight certain parts of an intonation phrase, which has also been described as **focus**. A focussed element of an utterance is more prominent than an element that falls outside the focus. New information is generally in focus, whereas given information typically falls outside the scope of focus. Two types of focus can be distinguished: broad and narrow focus. **Broad focus** means that everything – the entire intonation phrase – is brought into focus, as in example (58). With broad focus, the nucleus falls on the last content word. With **narrow focus**, a speaker highlights only a selected part of the intonation phrase. In example (59) the narrow focus is on *sister* so that it is not the last content word of this intonation phrase that receives the nucleus.

(58) What were you doing ||
Everyone gave her a strange LOOK ||
(59) Who came with you ||
It was my SISter who came with me ||

There are also grammatical ways of focussing in English – such as clefts, pseudo-clefts and passive constructions – and normally these fall together with the nucleus. Utterances (59) and (60) illustrate a cleft construction, utterance (61) is an example of a pseudo-cleft and utterance (62) shows a passive construction that coincides with the nucleus.

(60) It was a FLOwer I saw
(61) What he said was RUbbish
(62) The house was hit by a COmet

There are some occasions on which speakers may wish to put focus on utterance elements that constitute given information. This is usually done for purposes of contrast, as in utterance (63). There, the nucleus on *black* indicates which of the two dogs the speaker is referring to. One particular type of contrast is implied in insists, in which a speaker denies something that has been suggested (example 64) or has been implied in a previous utterance (see example 65). Furthermore, given information might be accented in echo questions such as (66).

(63) There was a black and a brown dog || It was the BLACK one that bit her
(64) (Why are you looking at her)
I am not LOOking at her *or*
I am NOT looking at her
(65) (Go to bed now)
But I am not TIRed
(66) (He failed the exam)
He FAILED

As with intonational phrasing, the rules of nucleus placement presented here predict nucleus placement in reading style better than in spontaneous speech. In spontaneous speech some intonation phrases are usually unfinished and do not contain any nucleus at all. This happens, for example, when in conversations speakers are interrupted by others or when they interrupt themselves in order to rephrase or correct what they have said before.

So far, the nucleus has been described as the primary accented syllable of an intonation phrase. Yet, apart from being accented by the mechanisms of increased air pressure and increased tension of the vocal folds described in section 2.1, nuclei are usually also associated with a distinct pitch movement. This is why the nucleus is often described as a **pitch accent**. In fact, most accented syllables in English have a distinct pitch height or pitch movement and most nuclei in English have a characteristic pitch movement. This is described in the next section.

4.3 English tones and their usage

Section 2.2 explained that all voiced sounds of an utterance have a certain height of voice or **pitch**. Normally, the pitch of a speaker's voice moves up and down across an utterance. (It is in fact highly unusual to speak on an unvarying pitch, which creates the impression of a robot talking.) This varying pitch across an utterance is perceived by listeners as a continuous movement of pitch. Only some of these pitch movements are linguistically relevant, whereas others are by-products of physiological mechanisms. Linguistically relevant changes in pitch are

- controlled by the speaker
- perceptible
- associated with a certain meaning or function

What does it mean that linguistically relevant pitch changes are controlled by a speaker? If a speaker's voice goes up and down because he or she is cycling on a bumpy road, these are not linguistically relevant pitch changes. It is only those pitch changes that are intentionally produced by a speaker that can convey linguistic meaning. Equally, there are lots of changes in pitch across utterances that can be measured and seen on a computer (see section 5.5.3), but only those that can actually be perceived by listeners are of linguistic relevance. Section 6.5 explains that only pitch changes that exceed a certain magnitude can be picked up by the human ear; others will simply not be noticed. Lastly and most importantly, pitch patterns that are linguistically relevant are associated with a particular meaning or function. This is true of language in general, where we always have a conventionalised connection between a particular form (a sequence of sounds or a pitch movement) and a particular meaning (the meaning of the sequence of sounds; the meaning of the pitch movement).

Different languages use pitch in different linguistically relevant ways. Section 2.2 explained that in **tone languages** like Chinese or Igbo the pitch height or pitch movement on a syllable or word can determine the meaning of words. This is not the case for **intonation languages** such as English or German. In intonation languages, the meaning of a word does not change with pitch height. Whether you say *two* with a high or a low, a rising or a falling pitch, it will always mean 'two'. Nevertheless, some of the pitch movements occurring in English have distinctive functions. Consider utterances (67) and (68). They differ in meaning although they differ only in the pitch movement on the last word.

(67) You are going \home
(68) You are going /home

In (67) the speaker's voice falls during the production of the word *home* (which is indicated by the symbol \) so that the utterance is perceived as a statement about the addressee going home. In contrast, in utterance (68) the speaker's voice rises on *home* (indicated by the /). This utterance is typically interpreted by listeners as a question with which the speaker asks the listener to verify whether he or she is going home. Thus, some pitch movements (the technical term is **tones)** in English are distinctive just like some speech sounds in English are distinctive. Consequently, English has a tone system or **tone inventory** just as it has a phoneme inventory.

4.3.1 The tones of English and their transcription

Scientific investigation into the tone inventory of English began in Britain in the early 20th century (e.g. Palmer 1922, Kingdon 1958, O'Connor & Arnold 1961). This **British School** of intonational analysis claims that each intonation phrase has one nucleus, which is its most prominent syllable and which is usually associated with a distinct pitch movement (see section 4.2 above). The pitch movement of the nucleus is referred to as the **nuclear tone**. Six basic nuclear tones are proposed for British English (see Table 4.1). These can be divided into the simple tones **fall**, transcribed with the symbol \ preceding the accented syllable, and **rise**, which is transcribed with the symbol /. The **fall-rise** V, the **rise-fall** Λ and the **rise-fall-rise** ΛV are complex tones. Some authors further divide the falls into low and high falls and the rises into high and low rises. A high fall drops from a high point in the speaker's voice to a low point, whereas a low fall starts lower and has less pitch movement. Similarly, a low rise starts low and rises only slightly, whereas a high rise ends very much higher. When the voice sustains its height on a nucleus, this is referred to as **level pitch** and is transcribed with the symbol -. Level pitch can further be divided into high, mid and low level.

Table 4.1. The nuclear tones of English and their transcription.

English nuclear tones	British School transcription	ToBI transcription
fall	\	H* L%
rise	/	L* H%
fall-rise	V	H* L-H%
rise-fall	Λ	L* H-L%
rise-fall-rise	ΛV	L*+H L-H%
level	-	L* L%

The British School proposes other elements of intonation phrases apart form the nucleus. If in an intonation phrase there are any pitch accents preceding the nucleus, they form the **head**. Unstressed syllables preceding the head are referred to as the **pre-head**. Pitch movements in heads can be falling (which is transcribed with the symbol ↘ before the accented syllable), rising (which is transcribed with the symbol ↗ before the accented syllable) or be produced on a level pitch (which would be transcribed as '). In utterance (69), *she* constitutes the prehead, the head includes all syllables from *did* to *to* and the nucleus lies on the last word *go*.

(69) she ↘didn't want to \go
PREHEAD HEAD NUCLEUS

The transcription symbols for the intonation of this intonation phrase indicate that the pitch is falling from *did* to *to* and then falling again on *go*. In order to fall on *go*, the voice will have to go up to a higher level at the beginning of *go*. This can be depicted in an **interlinear transcription** of this utterance given in Figure 4.2. Each syllable of the utterance is represented by a dot: the bigger the dot, the stronger the accent on the syllable. Both the syllable *did* and the syllable *go* in this utterance are strongly accented. In addition, the pitch height of each syllable is transcribed. The two horizontal lines symbolize the upper and lower range of a speaker's voice. The position of the dot between the two lines thus indicates the pitch height of each syllable. The syllable *did* is higher in pitch and more accented than the syllable *she*; the syllable *to* is lower in pitch and less accented then the last syllable *go*. This last syllable of the utterance has a falling pitch, which is symbolised by the falling line attached to the dot.

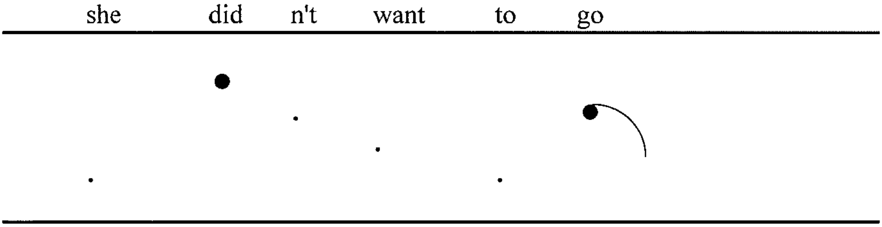


Figure 4.2. Interlinear transcription of the intonation of the utterance "She didn't want to go".

When any further syllables follow the nucleus – which would be called the **tail** in the British School tradition – the nuclear pitch movement begins on the accented syllable and stretches over the following syllables until the end of the utterance. Thus, in utterance (70) the nuclear fall is realized only on the one syllable of the word *go*. In utterance (71), however, the fall begins on *go* and stretches across the words *with* and *him* in the tail. In the transcription, no difference can be seen. In both cases, the transcription symbol for the nuclear fall is placed before the nucleus on *go*. The interlinear transcription of utterance (71) given in Figure 4.3 shows how the pitch of the syllables drops in what is called a step-down. Section 5.5.3 shows that nuclear falls can also be realized as a continuous pitch movement.

(70) She didn't want to \go
(71) She didn't want to \go with him

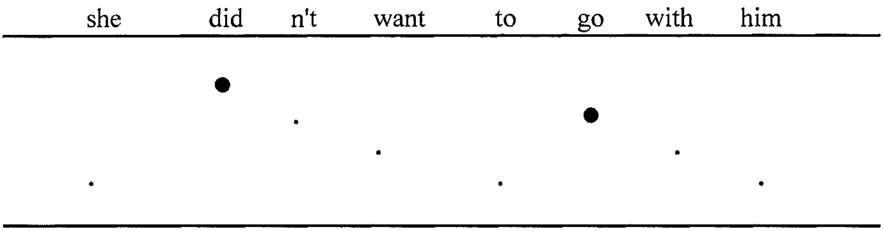


Figure 4.3. Interlinear transcription of the intonation of the utterance "She didn't want to go with him".

In the late 20th century a different system of intonational analysis was developed, the **autosegmental-metrical** (AM) approach, which differs from the system of the British School in some central assumptions. For example, while the British School tradition assumes that speakers have mental representations of pitch *movements*, the AM model claims that speakers have mental representations of the two *tone targets* high (H) and low (L). It is the production of combinations of these tone targets in speech that result in the perception of pitch movements. Within this theoretical framework the transcription system **ToBI** (which stands for Tones and Break Indices) was developed (Silverman et al. 1992). Slightly different ToBI versions exist for the transcription of the tones of American English and British English.

In ToBI, the star (*) signifies that a tone is associated with an accented syllable – an H* thus stands for a high pitch accent. The - refers to a phrase accent and the % to a boundary tone, which both occur at the end of intonation phrases. Thus, by combining a high pitch accent H* and a low boundary tone L% the speaker's voice is falling – which would be transcribed with a \ preceding the accented syllable in the British School system. Most language teachers use the transcription system of the British School when teaching English intonation, whereas most research articles on English intonation use the ToBI system.

Despite these fundamental theoretical differences, the nuclear tone inventory of English proposed by both models is very similar. Table 4.1 above lists the six basic nuclear tones of English together with their transcription in the British School and in ToBI. The two transcription systems are compared in utterance (72), whose intonation was illustrated with an interlinear transcription in Figure 4.2. The British School transcription given in (72a) marks a falling head and a falling nucleus. The equivalent ToBI transcription in (72b) indicates that a high pitch accent H* occurs on the syllable *did*, followed by a low tone L. The syllable *go* receives another high pitch accent H* and the nucleus falls towards the low boundary tone L%.

(72a) She ↘didn't want to \go
(72b) She didn't want to go
H*+L H* L%

4.3.2 The function of English tones and tunes

The meaning or function of particular tones in English has been analysed on at least three different levels:

- the **attitudinal** function
- the **pragmatic** function
- the **discoursal** function

Many tones in English can convey certain **attitudes** or emotions of speakers such as surprise, politeness, deference or judiciousness. The reply "Yes" to the question "Do you like the food" can convey very different meanings depending on the tone the speaker chooses. A falling tone usually indicates real approval whereas a fall-rising tone conveys doubtfulness. Moreover, tones may be used with certain **pragmatic** aims and indicate different **speech acts**. Thus, the utterance "You've finished" can function as a request for information when uttered with a rising nucleus. When produced with a falling tone, the utterance usually has the pragmatic function of a statement. In interactive conversations or **discourse**, tones can signal when a speaker has ended his or her turn and gives the floor to someone else. Furthermore, as shown in section 4.2 above, nucleus placement signals to the listeners what is new information and what is given information.

In practice, it is difficult to separate the three functions of tones clearly – often the attitudinal function overlaps with the discoursal function and the pragmatic function. They all have in common that they indicate the relationship between certain linguistic elements of the utterance and the context in which they are uttered. This in turn means that it is impossible to give detailed and exact rules for the use of tones by speakers. The possibilities of different utterances occurring in different contexts are far too many and the rules would be far too complex to be of any use for language learners. In the following, a few general tendencies of co-occurrence of some English tones will be given that may serve as guidelines for students of linguistics and English language learners. It has been suggested that in some respects the function of tones differs between American English and British English. Research has furthermore shown that many regional accents spoken on the British Isles use tones, especially rises, differently from Standard British English. In the following, only the function of tones in Standard British and Standard American English, RP and GA, will be described.

Falling tones in English are in general associated with finality, completeness and definiteness. A nuclear fall typically indicates to the listener that the content of the utterance a speaker produces is complete and does not require any additions. Thus, they are most often used in declarative utterances (see example 73). Commands are also usually uttered with a falling nucleus (example 74). Similarly, participants of a conversation indicate the end of their **turn** with a falling nucleus. This signals to the other participants that the speaker has nothing further to add and someone else can contribute to the conversation now.

(73) They burst out \laughing
(74) Give me the \book

Furthermore, wh-questions in English, i.e. questions that contain a wh-word such as "Where do you live" and "What's your name", are also typically produced with a falling nucleus (see examples 75 and 76. The ' indicates a high level head). When question (77) is spoken with a rising nucleus it sounds very friendly and might be appropriate when talking to children or a frightened person. This rise has also been labelled 'encouraging rise'.

(75) 'Where do you \live
(76) 'What's your \name
(77) 'What's your /name

Rising tones mostly indicate non-finality and that the speaker is seeking or anticipating information. Moreover, rising nuclei are commonly used in English conversation to indicate that the listener is expected to draw a conclusion from what was said and to respond. Typical utterances produced with rising nuclei in English are yes/no questions – questions that can be answered by "Yes" or "No" – as in (78). Here, British English speakers typically produce low rises, whereas American English speakers tend to produce high rises. A yes/no question produced with a fall as in example (79) sounds insisting and perhaps threatening in English. High rises in British English can imply surprise or disbelief. Thus, when the pitch of utterance (78) rises very high, the speaker might express that it was impossible to have seen her because (the speaker knows) she could not have been there at that time.

(78) Did you /see her
(79) Did you \see her

Equally, rising tones are used in English communication to indicate that the speaker has not finished yet and wishes to continue. In utterance (80), it is used to link two intonation phrases. In British English, these rises tend to be low rises (transcribed with ⁄, starting low and with a finishing point at mid-level). In American English, by contrast, high rises are preferred in this context.

(80) I was cycling home from ⁄work | when I saw this big \bird

Rising nuclear tones can also be used to imply particular meanings in English. For example, a low rise on an imperative as in utterance (81) implies that something nice is going to happen and thus appears more reassuring and encouraging than a fall.

(81) Come ⁄here | and I'll ↗give you a \present

Increasingly, declarative utterances (i.e. statements) in English can be heard with nuclear rises. As in example (82), speakers use a rising nuclear tone although they are not producing the utterance in order to ask a question or because they want to check whether something is the case or not. This phenomenon, which is called **upspeak**, **uptalk** or **high rising terminal** (HRT) and which occurs in about 3 to 4% of all declarative utterances (e.g. Britain 1992), was first documented for young female speakers in Australia but can now be found in many other accents of English, such as New Zealand English, Canadian English, different varieties of American English and some varieties of British English. In general, this intonation pattern tends to be used by younger speakers and to be frowned upon by older generations. In Belfast English and some British regional accents, by contrast, a rising nucleus on declarative utterances is the typical intonation pattern and therefore constitutes a different phenomenon than the HRT.

(82) (Where do you come from)
/London ||

In British English, low rises are associated with utterances that express consolation, encouragement and confirmation. Nuclear **fall-rises**, by contrast, often imply something that is thought to be unpleasant or unwelcome. They often occur with warnings, bad news, threats or disagreement. Thus, an imperative with a nuclear fall-rise conveys a threat or warning as in (83):

(83) Go a\/way | or I'll \hit you

Fall-rises are often produced on utterances that express doubt or reservation and constitute an appeal to the listener to reconsider. A fall-rise on an utterance like "Yes" or "No" typically expresses a response with reserve – when replying "Yes" with a falling-rising tone the speaker gives only limited agreement. Thus, the response "Can't she" in (84) indicates doubt or polite disagreement:

(84) (She can't be serious)
\/Can't she

Furthermore, a fall-rise can be used to signal tentativeness as in (85) or to politely correct someone else as in (86). British English speakers also use fall-rises on yes/no questions to indicate politeness.

(85) (Is this the way to the station)
I \/guess so
(86) (She got an A)
A \/B

In conversations, a fall-rise can indicate that the speaker has not finished yet and that there is more speech still to come, as illustrated in examples (87) and (88). Utterances can often be divided into material of primary importance, the **topic**, and material of secondary importance, the **comment**. Typically, the topic is associated with a falling tone, whereas the comment is produced with a non-falling one. This also applies to constructions in which the comment follows the topic, as in (89).

(87) If you're \/ready | we can \start ||
(88) \/First | I'll show you the \bathroom ||
(89) He was a \doctor | my \/father ||

A **rise-fall,** when used with a declarative utterance, typically involves a sense of completeness and finality. It can, however, also convey surprise or strong feelings of approval or disapproval. Utterances with rising-falling nuclei can express arrogance, confidence, self-satisfaction, challenge, putting-down and mocking. A speaker saying "That was nicely done" (see example 90) with a rising-falling nucleus is highly sarcastic and utters a mockery rather than praise.

(90) That was /\nicely done

Surprise is also expressed by a speaker when using a **rise-fall-rise** as in (91). As a nuclear tone, however, this tone is not very common in speech.

(91) She really went to /\/France again

Tag questions in English have different pragmatic functions when produced with a falling or a rising tone. In (92), the speaker invites an expression of confirmation and does not envisage disagreement. This utterance is appropriate when the speaker knows that the listener shares his or her belief. Conversely, the rising tone in (93) invites contradiction and is used by a speaker who expects the listener to know better. Clearly, the speaker did not see the play himself or herself and indicates that he or she is only reporting someone else's opinion.

(92) The play was \nice | \wasn't it
(93) The play was \nice | /wasn't it

In British English, furthermore, a low rise on a tag question might be interpreted as a menace as in (94):

(94) You are \leaving | ˏare you

Although the nuclear pitch movements are considered the most important aspect of intonation in English, they are rarely associated with a distinct meaning or function by themselves. In general, it is a combination of nucleus plus preceding pitch accents (the head in the British School tradition) that are associated with different meanings or functions. This combination of nuclear tone and preceding pitch accents in the head is referred to as a **tune**.

O'Connor & Arnold (1973) give a complete list of the attitudinal meaning of ten different tunes in colloquial British English. For example, the tune consisting of a rising head followed by a falling nucleus is often used to express surprise in exclamations such as (95):

(95) ↗Where do you \live

If a high level head is combined with a low fall (transcribed with the symbol ˎ) as in example (96), an utterance appears categorical, weighty or judicial. It can be used in answers to questions and statements where it gives extra weight to approvals or disapprovals.

(96) He is the 'ugliest man I ˎknow

Utterances that consist of a high level head and a low rise tend to sound soothing and reassuring when produced on a statement (see example 97). It is the typical tune for asking yes/no questions in which genuine information is sought and for softening commands and wh-questions which might appear too inquisitive.

(97) (Where are you going)
'Just to post a ↗letter

English speakers sometimes make use of **stylised patterns** of particular tones, which appear sung rather than spoken. This is because distinct pitch levels are produced instead of a sequence of pitch movements. The calling contour in (98), which consists of a step down by an interval of about a minor third from one pitch level to another, is an example in point.

(98) –Pe_ter

Such stylised patterns are stereotypical, i.e. highly conventionalised and almost ritualistic. They occur very frequently in child's play and in formulaic utterances such as greetings and conversation openers and closings.

4.3.3 Pitch range, key and register

The previous sections have shown that, in English intonation phrases and utterances, pitch moves up and down in a systematic, linguistically meaningful way. These movements of pitch can differ in their width. A speaker might produce a high pitch accent with an extremely high pitch, the highest that is physiologically possible. Alternatively, a high pitch can be just a little higher than a low pitch accent. Speakers normally produce speech in the bottom third of their potential pitch range. For particular linguistic reasons, however, the **pitch range** of an utterance can be varied. The term pitch range refers to the difference in pitch height between the highest and the lowest pitch produced in an intonation phrase or utterance. Figure 4.4 shows the pitch range of the reading of the sentence "A tiger and a mouse were walking in a field." by a female British English speaker. The curving line on top illustrates the movement of pitch across the utterance (see section 5.5.3 for details on automatic pitch-tracking) in relation to the words produced. The boundaries of the words are marked by the vertical lines. The pitch curve is interrupted where voiceless sounds such as the [t] in *tiger*, the [s] of *mouse*, the [k] in *walking* and the [f] in *field* occur. The two pitch accents of this utterance can easily be seen in the pitch peaks on the syllable *ti* of *tiger* and the syllable *walk* of *walking*. The pitch range of this utterance comprises the area from the highest point reached on *tiger* to the lowest pitch reached on *field*, as indicated by the arrows (see section 5.5.3 for more information on how to measure the pitch range of an utterance).

By changing their pitch range, speakers also change the **key** of an utterance. The term key describes the pitch level of an utterance in relation to the speaker's normal pitch level. A wide pitch range is associated with high key, whereas a narrow pitch range goes hand in hand with low key.

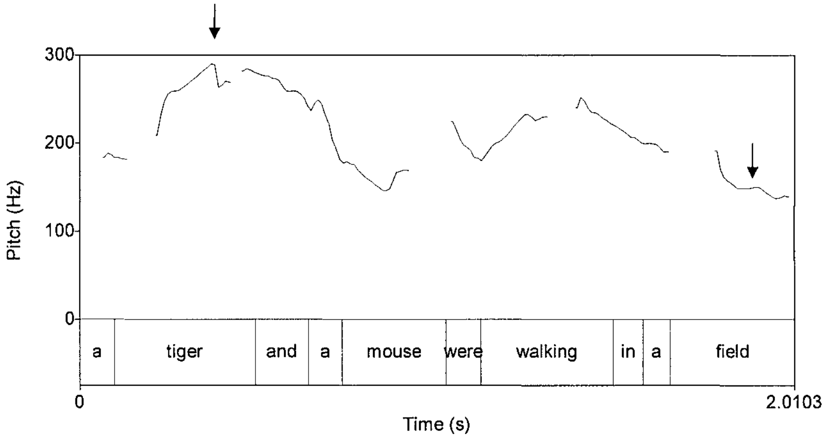


Figure 4.4. Pitch range of a reading of "A tiger and a mouse were walking in a field".

Both pitch range and key are used to link together successive utterances into larger discourse. When speakers produce speech, they do not simply string utterances together, but organize their spoken discourse (sometimes also called text) in a way so that the individual utterances are grouped together into topic units. Pitch range and key are used to indicate the beginning and end of the topic units. Typically, high key is produced in order to indicate the beginning of a new topic or the change of subject, and low key indicates the end of a topic. The use of tones in relation to topic units has also been called **paratone**, a term used in analogy to the term paragraph for written texts. Just like written texts are structured into orthographic paragraphs according to their topics and content (indicated by the beginning of a new paragraph in print), spoken texts are organized into paratones. At the beginning of a paratone, speakers usually use an introductory expression (for example "Guess what happened"). The first intonation phrase of a paratone is uttered with high key, whereas the end of a paratone is usually marked by low key and a lengthy pause. Figure 4.5 shows an example of the intonational structure of the first two paratones of a fairy tale. H stands for high key, M indicates mid high key and L refers to low key. (There is no agreed transcription system for key, unfortunately.)

The topic of the first paratone – the king – is typically introduced with high key; the second intonation phrase "he lived in a beautiful castle" is generally produced with mid high key, whereas the last intonation phrase "together with his wife and daughter" is usually produced with low key. When a new topic is introduced – the prince – a new paratone typically begins with high key. Like in the first paratone, the following two intonation phrases in the second paratone are normally produced with mid and low key.

H	Once upon a time there was a rich king|	
M		he lived in a beautiful castle|
L	together with his wife and daughter||	

H	One day a prince knocked at the door|	
M		he had come on a black horse|
L	that was stamping its feet||	

Figure 4.5. Key of the first two paratones in a fairy tale. The symbols H, M and L refer to H (high), M (mid) and L (low) key.

Empirical studies have shown that paratones are marked in this way in different speaking styles. In general, the pitch range of speakers is wider when reading a text than when retelling a story or producing spontaneous speech in English. This can be seen when you compare Figures 4.4 and 4.6. When a speaker reads out a sentence, as illustrated in Figure 4.4, the pitch range is much wider than when a speaker spontaneously produces the utterance "So the tiger and the mouse walking along together" in a retelling of the story.

Several empirical investigations have shown that British newsreaders structure their speech into intonationally marked paratones. They produce the first paratone with a wide pitch range and on a high key, whereas the last paratone is read with a narrow pitch range and a low key. The intonation phrases in between the first and last one have an increasingly narrower pitch range. Similarly, when you present readers with a written text split into two paragraphs they will produce a pause before the second paragraph and begin it with high key. When reading a story, English speakers typically introduce a new topic with high key, whereas elaborations of a topic are associated with intonation phrases beginning with mid high key and background information is marked by low key.

Not only is the reading of texts but also free speech is structured into topic groups by the means of pitch range and key. In lectures, speakers raise their pitch by about 50% between the end of one topic and the beginning of a new topic (Wennerstrom 1998). Intonation groups that constitute new topics have the widest pitch range, followed by intonation groups that are reformulations or extensions of what has been said before. The narrowest pitch range is found in intonation groups that form subordinations to the topic. This is illustrated in example (99), where the first intonation phrase constitutes a new topic, the third is an extension of this topic, and the second intonation phrase can be considered additional material or subordination.

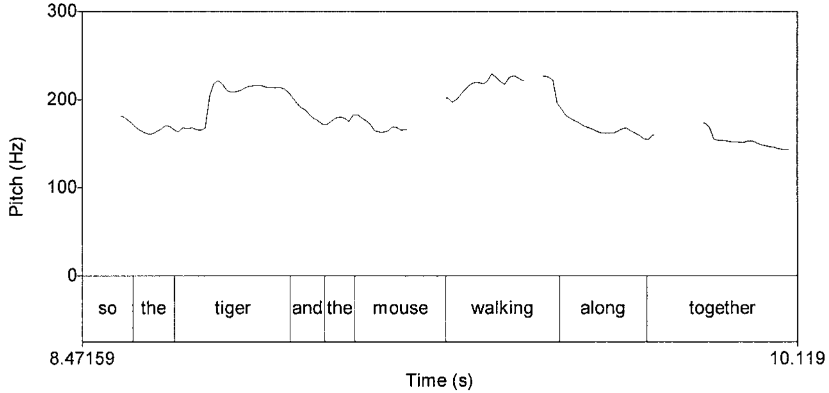
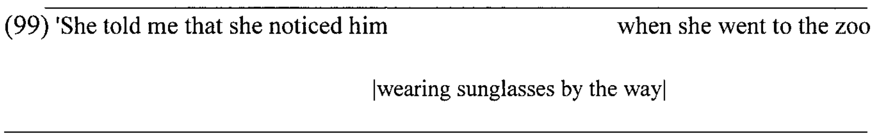


Figure 4.6. Pitch range of the utterance "So the tiger and the mouse walking along together" produced spontaneously by the same speaker as in Figure 4.4.

Low key with a compressed pitch range is often used to indicate additional information. In addition, words that present additional information are often spoken with a much faster speaking rate and a lower overall loudness than the adjacent intonation groups. This can be observed predominantly in parenthetical remarks such as asides, certain tag-questions, background information and meta-discoursal remarks (see example 99 in Figure 4.7). While the topic is produced with high key, loud voice and in a 'normal' speech tempo, the aside "wearing sunglasses by the way" is spoken on a low key, low voice and in a faster tempo before the speaker picks up the key, loudness and tempo of the first intonation phrase again.



(99) 'She told me that she noticed him when she went to the zoo

|wearing sunglasses by the way|

Figure 4.7.Pitch range varying with type of information.

Speakers sometimes also use different **registers** in their utterances. In contrast to differences in pitch range, register involves raising the average pitch level. In English and many other languages, a high register is associated with certain social or emotional roles. Putting on a 'high voice' in English does not signal any linguistically relevant information but is often used by speakers to indicate a subservient or deferential role. Since an overall higher pitch is caused by tensioning the vocal folds (see also section 2.2), a high register often goes hand in hand with speakers' emotional states of stress and tension.

4.4 The acquisition and teaching of English intonation (advanced reading)

In order to become competent speakers, both children and adult learners of a language have to acquire the phonological rules of intonational phrasing and nucleus placement as well as the association between certain nuclear tones, entire tunes and the different types of pitch range with their respective linguistic functions. It seems that the acquisition of these prosodic features of speech proceeds independently of the acquisition of the speech sounds of a language.

4.4.1 The acquisition of English intonation in first language acquisition

Producing English speech that is structured appropriately into intonation phrases is a difficult task for children: they do not master this before age six. It appears that the acquisition of intonational phrasing is closely intertwined with children's general mental and physical development. The first utterances by children consist of a single syllable, later a single word (one-word stage). At around two years of age, most children produce two words in a row. Yet, these two words, although they belong together semantically, are initially not produced as a single intonation phrase but rather as two separate intonation phrases separated by a pause. Unfortunately, it is impossible to decide whether this is due to the limited mental abilities of children (can they only think of one thing at a time?), due to restrictions in their control of the muscles necessary for speech production (can they not keep up the airstream necessary for the production of longer intonation phrases?) or due to a lack of a mental representation of the phonological unit of the intonation phrase. Only at around age 3;5 (three years and five months) do children begin to use intonation phrase boundaries systematically (see e.g. Gut 2000a) in utterances such as (100)

(100) one piece is /missing | and one piece \missing

The phonological use of nucleus placement with the function of marking new versus given information and focussing is also acquired relatively late by children. Utterances produced by most two-year-olds do not have a perceptible nucleus: all words are accented equally. The marking of new information begins at around age three. 9-year-olds place the nucleus on new information in statements in 68% of all cases and in questions in 90% of all cases.

Conversely, the production and meaningful usage of nuclear tones in English are acquired fairly early. Some children systematically produce different speech acts such as questions and statements with distinct nuclear tones – rises with questions and falls with statements – from two years of age on. Similarly, rising pitch is used for enumerations and for attracting attention from two years on. When children begin to produce utterances with subordinate structures from about 2;6 years on, the subordinate part is often linked to the main clause with a rising tone. Some authors (e.g. Crystal 1981) even claim a distinct order of acquisition for the different nuclear tones. Falls and rises are distinguished first; then the difference between high and low falls and rises is acquired. Only after that do children produce nuclear rise-falls and fall-rises.

Moreover, very young children seem to be able to mark utterances with different communicative function by a combination of nuclear pitch movement, loudness and pitch range. For example, utterances during which eye contact with another person is held are typically marked with higher pitch, higher loudness and a wider pitch range than utterances during which children appear to be talking to themselves as early as at age two.

4.4.2 The acquisition of English intonation in second language acquisition

Learners of a second language already have lots of knowledge about the rules of the use of pitch in their first language(s). If this language is a tone language, the phonological rules concerning the form and function of pitch patterns will be very different from the rules of nucleus placement, intonation and usage of pitch range in English. In tone languages, the pitch level or pitch movement on syllables determines the meaning of words. This means that nuclear tones, the tunes of utterances and pitch range are not used for linguistic purposes as they are in English: nucleus placement does not indicate new and given information or focus; tones, tunes and pitch range are not used for the expression of attitudinal meanings or certain discoursal functions in the same way in tone languages. Consequently, native speakers of tone languages have to acquire an entirely new phonological system with a whole range of new phonological rules.

When acquiring English, native speakers of an intonation language like German or French have already learned the fundamental concepts of using intonation phrases, nucleus placement, pitch patterns and pitch range. Nevertheless, individual phonological rules of the meaning and usage of particular nuclear tones may differ between the language learner's first and second language. For example, in British English a low rising or falling-rising nucleus is often produced on statements or commands, which carries certain attitudinal implications such as leaving open the possibility of agreement or disagreement. Utterance (101) thus invites the conversational partners to voice their own opinion.

(101) That's ↗hot

In German, this usage of a low rise or fall-rise does not exist and falls are produced instead. German learners of English who produce a falling nucleus in these cases may thus inadvertently create the impression that they want to impose a belief on the conversation partner. Furthermore, learners of English with an intonation language as their native language may find it especially difficult to acquire the phonetic aspects of English intonation. For example, pitch range in British English is much wider than in German although it is used for the same linguistic purposes (Mennen 2007).

Analyses of the intonation of learners of English show that differences compared to native English intonation exist in all areas. German learners of English, for example, produce more and shorter intonation phrases when reading a text than British English speakers do. Moreover, they insert intonation phrase boundaries at places where native speakers do not. By the same token, learners of English produce more pitch accents than English native speakers and do not use nucleus placement for the marking of new and given information in the same way as native speakers do. Often, no distinction is made between the two types of information status, and given information receives the same accent as new information (e.g. Grosser 1997). For example, in the utterance "You don't even like cheese" where *cheese* is given information, many learners of English produce a nuclear pitch accent on the word *cheese*. In general, there seems to be a tendency for learners of English to put the nucleus on the last word of an utterance. Similarly, while British English speakers always read sentences 102a and 103a with nucleus placement on *phone* and *hi*, German learners of English tend to produce 102b and 103b with the nucleus on *rang* and *Nick*.

(102a) The PHONE rang
(102b) The phone RANG

(103a) Oh HI Nick
(103b) Oh hi NICK

Furthermore, learners of English use tones and tunes in a different way from native speakers. On the whole, they produce fewer nuclear fall-rises when speaking English, and while native English speakers often produce the tune rising head + falling nucleus (↗+\), learners of English have a tendency to produce falling heads. When measuring the phonetic extent of nuclear falls, it can be shown that learners of English produce shorter falls that do not reach the same bottom baseline. Finally, the overall pitch range in all speaking styles is narrower for learners of English than for native English speakers. Equally, new topics are marked by a wider pitch range by native speakers than by language learners.

4.4.3 Teaching English intonation

As with any other aspect of English phonetics and phonology, the acquisition of English intonation is not impossible – even for late learners who begin learning English after early childhood. Some learners acquire the rules of English intonation perfectly and are indistinguishable from native speakers. It is still not well established in what way teaching can contribute to successful acquisition. It appears that creating **language awareness** is beneficial for learning a new intonational system. Exercises that raise language awareness in the area of intonation can for example show learners in what way their intonation of English differs from the intonation of native speakers. This can be achieved by visualising the pitch movement or pitch range of utterances with the help of specialized software (see sections 5.5.3 and 5.7). Figure 4.8 shows the pitch movement and pitch range of a reading of the sentence "A tiger and a mouse were walking in a field." by a German learner of English. Comparing this to Figure 4.4, which illustrates the pitch range of a native English (RP) speaker, distinct differences can be seen: the pitch range of the learner of English is much reduced and far more pitch accents (visible as the pitch peaks indicated by the arrows) are produced.

By being able to see differences between their own intonation and the intonation of a native speaker, language learners gain language awareness and might be able to change their intonational patterns in the direction of native speaker values.

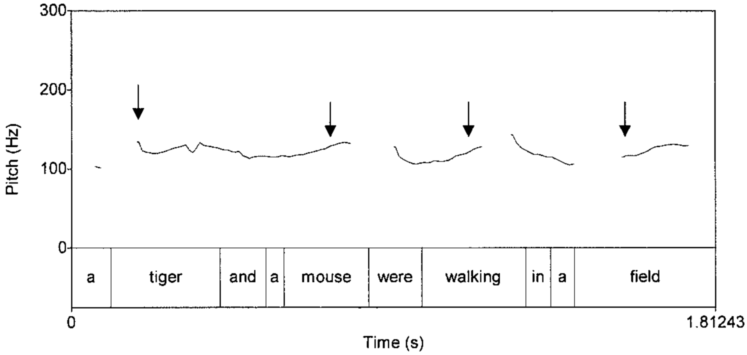


Figure 4.8. Pitch range of a reading of "A tiger and a mouse were walking in a field." produced by a German learner of English.

4.5 Exercises

1. What is an intonation language?
2. What is a pitch accent?
3. How can one show that pitch movement is linguistically relevant?
4. What is intonational phrasing used for by English speakers?
5. How do speakers signal the end of their turn?
6. Mark where an intonation phrase boundary is appropriate in the following utterances:

a. I'm not absolutely sure to be honest

b. She seemed to spend most of the day in bed with a crime novel

c. He usually leaves at nine doesn't he

d. No that's not acceptable

e. I went to buy a jacket a skirt and a pair of shoes

7. Mark the nucleus placement in the following utterances:

   a. Did you have a lot to do
   b. Yesterday I went to the doctor twice
   c. Has your mother left already
   d. Do you know when the first train leaves
   e. She is trying to fix the washing machine

8. Mark the appropriate tone in the following utterances:

   a. Is this a joke
   b. That's wonderful (enthusiastic)
   c. All right (doubtful)
   d. Would you like some tea

9. Listen to recordings 04_exercise9a, b and c on the CD-ROM and transcribe the intonation of the utterances using the British School transcription system.
10. Record yourself saying the utterances in recordings 04_ exercise9a, b and c (instructions are included in the Praat manual on the CD-ROM) and compare your nucleus placement, intonational phrasing and intonation with that of the native speaker. Which differences do you notice?
11. What are the major difficulties for learners of English regarding intonation? Why?

4.6 Further Reading

Wells (2006) presents a very detailed and accessible description of (British) English intonation, including intonational phrasing, nucleus placement and the form and meaning of pitch movements. It is written for students of English and contains many examples and practical exercises as well as a CD-ROM. Cruttenden (1997) also offers an excellent overview of all aspects of (mostly British) English intonation. Roach (1991, chapters 15 to 19) gives a detailed introduction to British English intonation in the British School tradition together with practical exercises on audio cassettes. For general introductory reading on the attitudinal function of tunes in British English see O'Connor & Arnold (1973), who also offer many practical exercises.

For advanced students, short overviews of the British English and the American English intonation systems are presented in Hirst (1998) and Bolinger (1998) respectively, but they are based on yet another system of intonation transcription. The development of the ToBI transcription system and its history are described in Beckman et al. (2005) and can be found on the website <http://www.ling.ohio-state.edu/~tobi/>.

Tench (1996) gives many examples of the grammatical function of intonational phrasing in British English and compares it to the grammatical function of intonational phrasing in German. Many aspects of the discourse intonation of English are described in Wennerstrom (2001). For advanced students, Pierrehumbert & Hirschberg (1990) present a description of the meaning of American English intonation in the ToBI framework. Brazil (1975) describes the usage of key in English discourse. More details on the intonational marking of paratone are given in Brown & Yule (1983).

Vihman (1996, chapter 8) gives an introduction to the prosodic development in first language acquisition. Trouvain & Gut (2007) contains a collection of articles on second language acquisition and teaching of various prosodic features, and Chun (2002) describes discourse intonation by language learners.

5 Acoustic properties of English

Chapter 2 explained how speech is produced by humans; chapter 6 explains how speech is perceived by humans. However, after having been produced by a speaker and before being perceived by a listener, speech exists as a physical property – it 'floats through the air'. This chapter is concerned with the properties of speech as it is transmitted by air. The study of the physical properties of speech and of how they are perceived by listeners is called **acoustics**. Acoustic studies of speech have only become possible relatively recently: new technologies have enabled researchers to make speech 'floating through the air' visible and to measure its acoustic features. With the help of phonetic analysis software, we can see many of the acoustic properties of speech, for example the pitch movement across the utterance "Please help me" produced by a British English speaker shown in Figure 5.1. The acoustic study of speech has also enabled researchers to synthesize speech, i.e. to build machines that speak (nowadays many children's toys 'speak', and artificial voices talk to us from nearly every device ranging from our washing machine to the navigation system in our car). Similarly, with our increasing knowledge of the acoustic properties of speech, machines have been developed that 'understand' speech. You will have encountered them in telephone bookings or telephone enquiries of any kind.

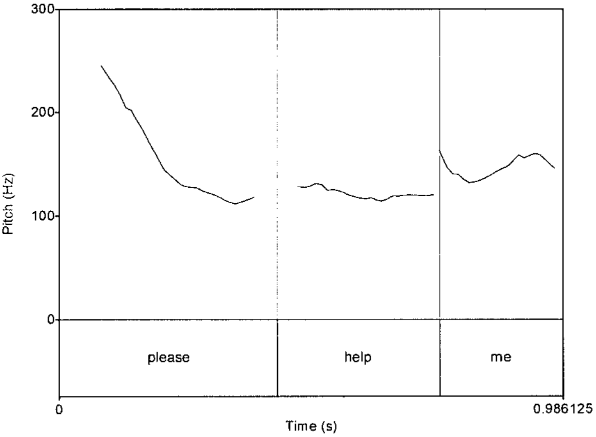


Figure 5.1. Pitch movement across the utterance "Please help me".

The possibility to analyse the acoustic structure of speech has enabled linguists to improve their descriptions of some sounds. For example, it is easier to characterize vowels by describing their acoustic nature than their underlying articulatory movements. Furthermore, an acoustic study of speech recordings reveals many properties of speech that cannot be heard even by trained listeners. It appears, for example, that when English speakers produce an accented syllable, they do this mainly by increasing the height of pitch on the vowel, whereas speakers of German typically produce greater loudness on an accented syllable compared to an unaccented one. This finding is of course of great importance for a native German speaker aiming to acquire the pronunciation of English.

The study of the acoustic properties of speech can thus be applied in many areas such as language teaching, clinical linguistics and speech therapy, and forensic linguistics. Teachers of English, for example, can show their students differences between their speech and that of native speakers. This knowledge can further be used to develop new teaching materials and techniques. By the same token, differences between the speech of people with speech disorders and healthy speakers can be described and used to develop therapeutic interventions. Finally, forensic phoneticians can give acoustic evidence at court in trials where it is important to ascertain whether the voice recorded on an answerphone is the one of a person accused of blackmail or not. The next section provides a brief introduction to the basic principles of the acoustic structure of sound. Sections 5.2 and 5.3 are concerned with the acoustic properties of English vowels and consonants. In sections 5.4 and 5.5 the acoustic properties of connected speech and English intonation are described. Section 5.6 shows how acoustic phonetics can be applied in English language learning and teaching. Section 5.7 explains what needs to be considered in order to make a good speech recording.

5.1 Acoustic properties of sound

The sensation of sound is caused by movement. Whether you drop a saucepan lid, whether you slam a door or whether you make an airstream pass from your lungs through your mouth – you have caused air to move. To be more precise, you have caused small variations in air pressure that occur very rapidly. You can feel them by putting your hand in front of your mouth while talking. These variations in air pressure travel (the technical term is 'propagate') through the air, and when they reach the eardrum of a listener, they set in motion the process which results in the perception of a sound (see chapter 6). When recording speech, it is these variations that are picked up by the microphone. The **propagation of sound** does not depend on the medium air of course. It is perfectly possible to transmit speech under water – yet, the properties of the medium water create a very different sensation of sound.

Sounds can travel across long distances and they do so in the form of **sound waves**. Since the type of air pressure variations caused by the production of different speech sounds can differ considerably, their corresponding sound waves differ as well. For a production of a vowel, for example, the vibrating vocal folds chop up the airstream into regular pulses that alternate between high pressure and low pressure. For plosives, the airstream is stopped completely and then rushes out with great force. Accordingly, the sound waves of vowels and plosives that are transmitted through the air have very different properties. The properties of the various sounds of English will be described in detail in sections 5.2 and 5.3.

5.1.1 Acoustic properties of sound waves

Sound waves can be made visible in various forms. When you record some speech and transfer this recording to a computer, you can employ speech analysis software to create a **waveform**, the visible representation of the sound waves of the sounds in the recording. Figure 5.2 shows the waveform of the utterance "Please help me". Listen to it on the CD-ROM (05_please.wav).

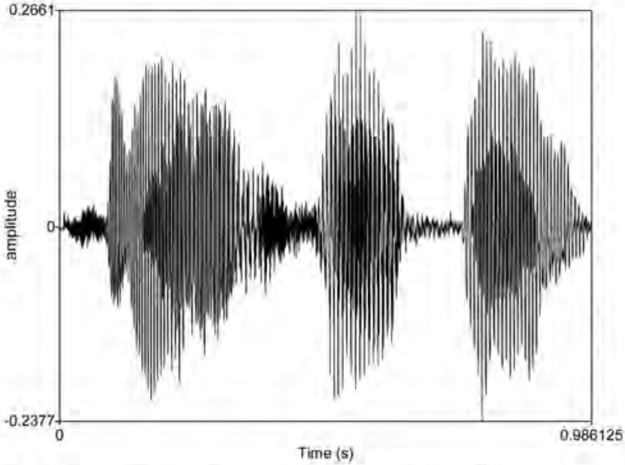


Figure 5.2. Waveform of the utterance "Please help me".

Two acoustic properties of this utterance are represented by the two axes of the waveform: duration and intensity. The horizontal axis shows the **duration** of the speech signal (the utterance is approximately 0.98 seconds long). Moreover, the duration of each part (e.g. word, syllable and phoneme) of the utterance can be measured. Since speech events are typically very short, the unit of measurement is either seconds (s) or milliseconds (ms). The vertical axis in Figure 5.2 indicates the variation in air pressure across the various parts of the utterance. This variation of pressure – which is perceived as loudness – corresponds to the physical property of **intensity**. Intensity is proportional to the **amplitude** of the waveform, which is visible in the relative distance of each point from the zero line. In other words, the height of each portion of a waveform above the zero line indicates the height of air pressure. Portions of the waveform below the zero line correspond to low air pressure. In Figure 5.2 you can see that there are three places of very high amplitude and some intervening places with lower amplitude in the utterance. Intensity is a relative value and has to be calculated with reference to some specific intensity level. In speech analysis software packages, it is usually expressed relative to the smallest amplitude value in the waveform. The most widely used unit of measurement for the intensity of a speech signal is **decibel** (dB). Because sound intensity is proportional to the square of the amplitude, decibel is a logarithmic measurement. This means that a small increase in dB values corresponds to a much larger increase in intensity and perceived loudness. As a rule of thumb, when a sound is perceived as about twice as loud as another, the increase in intensity is 5 dB (see section 6.5 for details on the relationship between acoustic properties and human perception). Figure 5.3 illustrates the intensity changes across the utterance "Please help me" measured in dB. You can see that the intensity peaks occur exactly where the vowels in the utterance are. The two plosives [p] in *please* and *help* have the lowest intensity.

One other important phonetic fact can be seen when observing the waveform in Figure 5.2. It immediately becomes clear that there are no boundaries between individual speech sounds and individual words in an utterance (this was explained with reference to articulation in section 2.5). Rather, each utterance in speech consists of a continuous sequence of sounds which have no visible (and indeed audible) boundaries between them. In Figure 5.3, the vertical dotted lines indicate the beginning and end of each sound in the utterance "Please help me". Due to the continuous nature of speech, these boundaries are not easy to draw. When you listen to each of the sounds between the given boundaries, you will easily be able to tell the influence of the neighbouring sounds. For acoustic analyses of individual speech sounds, especially vowels, it is thus the linguistic convention to indicate the 'stable parts' of the sounds. Stable parts are those parts of a speech sound that are not influenced by the neighbouring sounds.

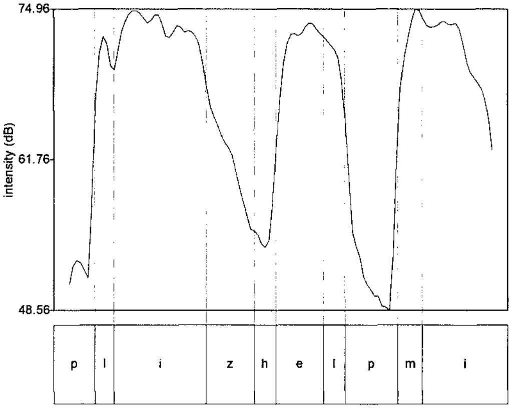


Figure 5.3. Intensity curve of the utterance "Please help me".

When you select from Figure 5.2 just the portion of the waveform that represents the vowel [i] and zoom it in on the computer, the waveform illustrated in Figure 5.4 will be displayed to you. This waveform appears very regularly patterned with regularly occurring peaks and valleys in air pressure.

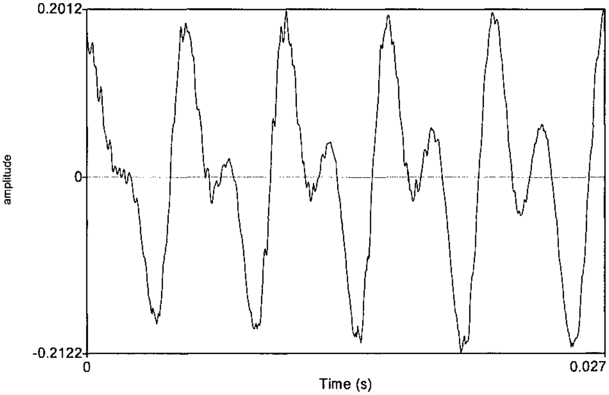


Figure 5.4. Waveform of the vowel [i] in please.

Each repetition of the pattern is called a cycle. When you measure the period of time each cycle takes, you can determine the **frequency** of a sound. The frequency of a sound is defined as cycles per second. In the [i] shown in Figure 5.4, one cycle is 0.0054 seconds long (the entire recording is 0.027 seconds long and contains 5 cycles, so you divide 0.027 by 5 and arrive at 0.0054). To calculate the number of cycles per second of the [i], you now divide 1 (one second) by the length of one cycle (0.0054) and arrive at approximately 185. This means that the cycles of this vowel occur about 185 times per second, which in turn tells you that the speaker's vocal folds were vibrating approximately 185 times per second. The vibration of his vocal folds chopped up the airstream into approximately 185 puffs per second. Measurements of the frequency of a sound are expressed with the unit **Hertz** (abbreviated Hz). The frequency of the [i] in Figure 5.4 is thus 185 Hz.

5.1.2 Simple and complex waveforms (advanced reading)

The waveform shown in Figure 5.4 is a **periodic** waveform, which means that its cycles occur at regular periods of time. Periodic or quasi-periodic waveforms are typically produced when speakers articulate vowels and voiced consonants (because they involve vocal fold vibration and thus a regular chopping up of the airstream). The periodic waves of voiced speech sounds are called **complex** periodic waves. This term reflects that complex periodic waves are composed of a combination of several **simple** periodic waves or **sine** (short for **sinusoidal**) **waves**. A sine wave is a wave that has only one frequency. The waveform of a sine wave with the frequency of 200 Hz is displayed in Figure 5.5. Simple sine waves do not occur in real speech but are often used in the assessment of hearing. There they are referred to as **pure tones** (see section 6.6 for details).

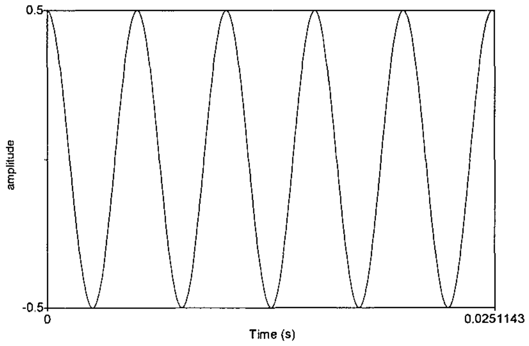


Figure 5.5. Waveform of a pure tone of 200 Hz.

A complex wave consists of several sine waves that differ in frequency. Thus, acoustically speaking, the voiced sounds in speech consist of a combination of various sine waves. Each of them has a different frequency and they differ in **phase**. The term phase refers to the relative beginning of each cycle of the wave. Two sine waves with the same frequency but which are out of phase are displayed in Figure 5.6. You can see that they reach the amplitude maximum at different points in time.

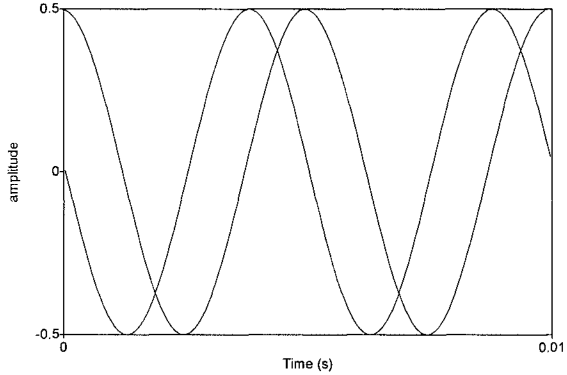


Figure 5.6. Waveform of two pure tones of 200 Hz, out of phase.

When you combine several sine waves with different frequencies and different phase you arrive at a complex period waveform as the one illustrated in Figure 5.4 for the vowel [i] in *please*. You can see that within each cycle several peaks of amplitude occur – these reflect the peaks of each of the underlying simple waves and show you how they are out of phase with each other. Figure 5.7 illustrates the components of a complex sound wave consisting of three sine waves with the frequencies 150 Hz, 300 Hz and 450 Hz, respectively.

Listen to each of the individual sine waves shown in Figure 5.7 (05_sine150.wav; 05_sine300 .wav; 05_sine450.wav) on the CD-ROM. The sound that results when you play all three of the sine waves at the same time – the complex waveform displayed in Figure 5.8 – sounds like 05_complexwave.wav on the CD-ROM.

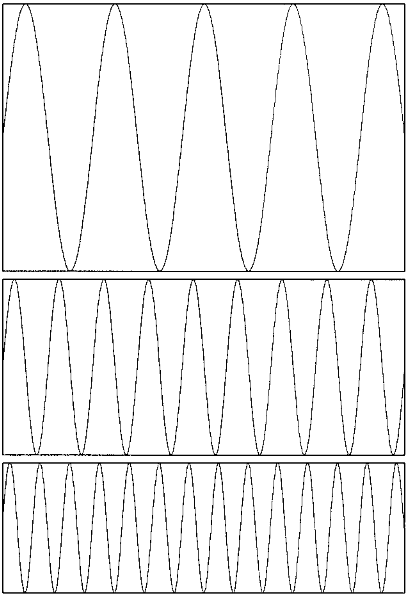


Figure 5.7. Waveforms of three sine waves with 150 Hz (top), 300 Hz (middle) and 450 Hz (bottom) frequency and different amplitudes.

The sine wave with a frequency of 150 Hz has a higher amplitude than the sine wave with a frequency of 300 Hz, which in turn has a higher amplitude than the sine wave with a frequency of 450 Hz. In Figure 5.8, you can see the complex waveform that results when you combine these three simple waves. When you listen to the three sine waves and the resulting complex wave on the CD-ROM, you will notice that the complex wave has the same **pitch** (height of tone, see section 2.2 for a description of the articulatory properties of pitch) as the sine wave with 150 Hz frequency. The frequency of the lowest sine wave in a complex periodic wave is called the **fundamental frequency,** abbreviated as **F0**. This is the frequency that determines the **pitch** that is perceived. (Since the pitch of the vowel [i] in Figure 5.4 is at 185 Hz, you now know that the lowest sine wave of this sound lies at 185 Hz). The sine waves with frequencies above this fundamental frequency are referred to as the **harmonics** (they are called overtones in music). Their frequency stands in a mathematical relationship to the fundamental frequency. To be precise, they are integer multiples of the fundamental frequency. Thus, when the fundamental frequency is 150 Hz, the first harmonic will have a frequency of 300 Hz and the next a frequency of 450 Hz. The tenth harmonic has a frequency of 1500 Hz and so on.

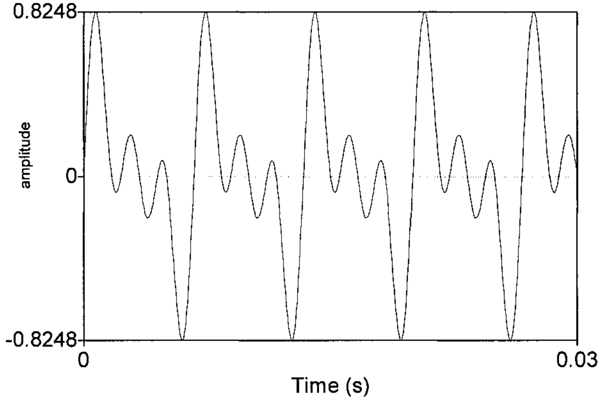


Figure 5.8. Complex waveform consisting of three sine waves with 150 Hz, 300 Hz and 450 Hz frequency, respectively.

It is the vocal fold vibration in the larynx that produces the complex periodic wave of voiced speech sounds. The fundamental frequency of this wave corresponds nearly exactly to the rate of vocal fold vibration. Thus, a sound with a fundamental frequency of 150 Hz is produced by the vocal folds completing 150 opening and closing cycles per second. The amplitude of each of the harmonics decreases so that the amplitude of the third harmonic is lower than that of the second but higher than that of the fourth and so on.

If you heard a recording of the complex wave produced by vocal fold vibration it would not sound like speech to you. This is because when passing the vocal tract, the complex wave is modified significantly. These alterations that result in what we perceive as speech sounds are caused by the physical properties of the pharynx and the oral and nasal cavity. In acoustic terms, the vocal tract acts as a **resonator** and **acoustic filter**. In a simplified manner this can be explained as follows: when the vocal folds chop up the airstream into little puffs of vibrating air and this vibrating air reaches the pharynx and oral and nasal cavities, specific frequencies in the complex wave cause specific tissues and membranes in the vocal tract to vibrate also. It also causes the bones in your head to vibrate, which you can feel especially clearly when you cover your ears while speaking or singing. When one vibrating body causes another body to vibrate with maximum amplitude at its natural frequency, which also happens to be the natural frequency of the first vibrating body, this is called **resonance**.

When one body vibrates with maximum amplitude at one particular frequency, this in turn means that other frequencies are dampened. This selective enhancing and damping of frequencies is called filtering. An acoustic filter thus passes or blocks specific frequencies of a complex wave. **Low-pass filters** block or reduce in intensity some higher frequencies, **high-pass filters** block or reduce in intensity some lower frequencies. By altering the position of your tongue, velum and the other articulators you alter the resonating function of your vocal tract. For example, when you change the position of your tongue and lips from the articulation of [a] to articulating [u], the shape of your vocal tract changes and thereby modifies its resonance, i.e. the particular frequencies of the complex wave produced by the vocal fold vibration that are enhanced and dampened. When you close your lips and lower your velum, you again cause different patterns of resonance in your vocal tract, particularly in the nasal cavity, which result in the sound [m].

As an upshot of all this, it can be summarized that the complex wave leaving a speaker's mouth consists of several sine waves of different frequencies, some of which have a higher intensity than others. The intensity of each frequency in a voiced speech sound can be made visible in a **power spectrum**. Figure 5.9 illustrates the spectrum of the vowel [i] in *please*. The vertical axis expresses the intensity (sound pressure level) and the horizontal axis the different frequencies contained in the vowel. You can see that only some of the frequencies have a high intensity (in particular the fundamental frequency at 185 Hz, which has the highest intensity). This is caused by the particular position of the speaker's tongue during the articulation of this vowel – a high tongue position with the front of the tongue closest to the roof of the mouth.

Another way of visualizing the intensity of the different frequencies of voiced speech sounds is the **spectrogram**. Speech analysis software usually creates spectrograms from recordings by performing a **Fourier analysis** (named after Jean Baptiste Fourier, who showed in the early 19th century that complex waves can be analysed as the sum of their sine wave components). With this analysis, the individual frequencies of each speech sound and their corresponding intensities can be made visible. Another technique for examining the spectral properties of sounds, which is offered by many speech analysis packages, is the **linear prediction coefficient** analysis (**LPC**).

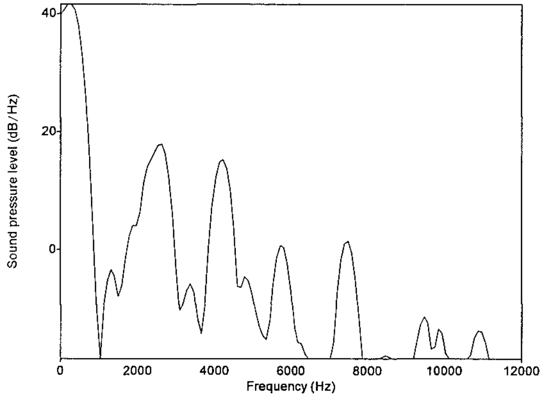


Figure 5.9. Spectrum of the vowel [i] in please.

Figure 5.10 shows a spectrogram of the utterance "Please help me". The boundaries of each speech sound are indicated by the dotted lines. You can see that in a spectrogram, time or duration is displayed on the horizontal axis, the different frequencies are shown on the vertical axis and the intensity of particular frequencies is expressed by the colour. Black areas thus indicate frequencies at high intensity, whereas grey areas indicate frequencies at lower intensity. You can see that all the voiced sounds, the vowels as well as [l] and [m], have bands of frequencies with high intensity – indicated by the black bars in the spectrogram. These frequency areas with high intensity are called **formants** and will be discussed in detail in the next section.

There are two ways of displaying the different intensities of different frequencies of sounds in a spectrogram. In a **broadband spectrogram** such as shown in Figure 5.10, the individual harmonics are not clearly visible but are rather lumped together as broad areas of high energy. It is thus relatively easy to detect changes in these formants over time. In the creation of **narrowband spectrograms**, by contrast, the frequency dimension of speech signals is displayed very accurately. However, this is at the expense of exact resolution in the time dimension, which is higher in broadband spectrograms. Figure 5.11 shows a narrowband spectrogram of the utterance "Please help me". Especially in the lower frequency areas, the individual harmonics are clearly visible as thin lines.

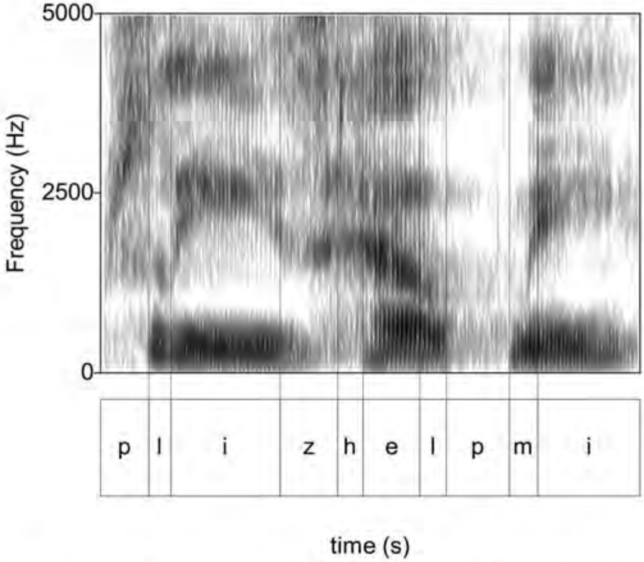


Figure 5.10. Broadband spectrogram of the utterance "Please help me".

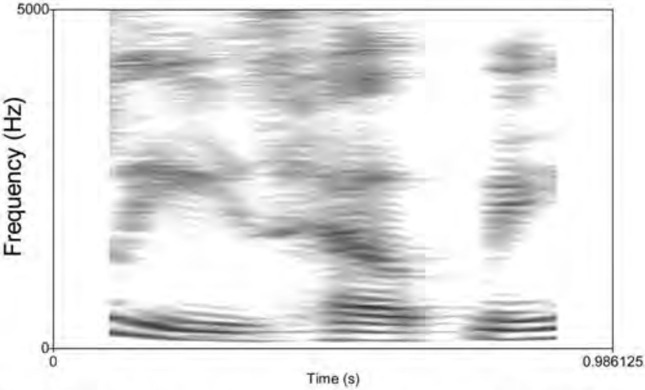


Figure 5.11. Narrowband spectrogram of the utterance "Please help me".

In Figure 5.10, you can also see that the voiceless speech sound [p] looks different from the voiced sounds in the spectrogram. It has one large area of high intensity that covers many different frequencies. This is because [p], like all other voiceless speech sounds, does not consist of periodic but of **aperiodic** waves. This means that its waves do not have a regular pattern like voiced speech sounds but instead consist of waves of many frequencies with random amplitude and phase. This is illustrated in Figure 5.12, which displays the waveform of a [s]: no regular pattern is visible. Sounds that have underlying aperiodic waves are called noise, so that all voiceless speech sounds are strictly speaking noise rather than sounds. Section 5.3 gives more details about the acoustic properties of voiceless sounds.

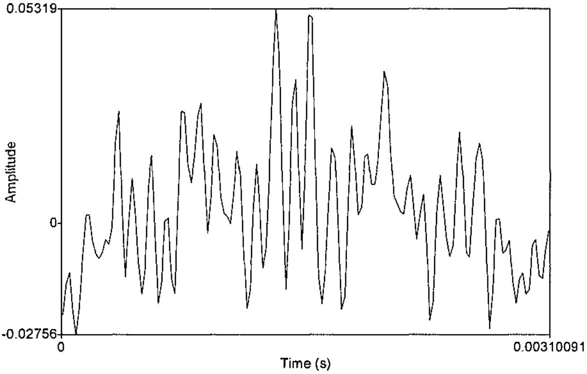


Figure 5.12. The aperiodic waveform of a [s].

The complex wave that is produced by vocal fold vibration in the larynx is referred to as the **sound source**. As described above, this is changed significantly when it passes the vocal tract (i.e. through the throat and oral and nasal cavities). You have also seen that, in acoustic terms, the vocal tract can act as a frequency-sensitive **filter**. Thus, the acoustic model of speech production is called the **source-filter model**.

Table 5.1 summarizes how the various acoustic measurements are related to articulatory features and perceptual features. (For speech perception see chapter 6.) What is measured in the speech signal as the fundamental frequency corresponds to the quasi-periodical vibrations of the vocal folds and is perceived as pitch. The measurement of intensity captures articulatory effort and subglottal air pressure and is perceived as loudness. The duration of speech elements corresponds to the duration of the speech gestures and is perceived as length. The formant values measured for vowels reflect the vocal tract configuration and are perceived as vowel quality.

Table 5.1. Relationship between articulation, acoustic features and perception.

Articulation	Acoustic features	Perception
quasi-periodic vibration of vocal folds	fundamental frequency (F0) measured in Hz	pitch: low – high
articulatory effort, subglottal air pressure	intensity measured in dB	loudness: soft – loud
duration of speech gestures	duration measured in ms	length: short – long
vocal tract configuration	formant values measured in Hz	vowel quality: reduced – full types of vowels

5.2 The acoustic properties of English vowels

Section 5.1.2 above explained that vowels (and other voiced speech sounds) consist of periodic sound waves with different frequencies at different intensities. The lowest frequency is the one listeners perceive as pitch (the height of voice), the higher frequencies, which are called harmonics or overtones, provide the particular quality of the vowel. Of the frequencies produced by vocal fold vibration in the larynx, some are enhanced and some are blocked by the resonating properties of the vocal tract, depending on the size of the vocal tract and the shape and position of the speaker's articulators such as the tongue, lips and the velum. This means that the shape and position of the articulators and the size of the vocal tract determine the quality of vowels, i.e. the type of vowel a speaker produces and a listener perceives. You can test this easily by pronouncing the vowels [a], [i] and [u] in a row and on the same pitch level. You can feel the change in the position and shape of your tongue as well as in the position of your lower jaw when going from [a] to [i]. The movement of your tongue and jaw changes the size and shape of your oral cavity, and this in turn influences which of the frequencies are enhanced in intensity and which are blocked. When going from [i] to [u], it is the position of your tongue and lips that changes and thereby causes different frequencies in the waves to be enhanced or blocked. Those frequencies that have very high intensity (= energy) are called **formants**. Thus, formants are the most important acoustic cues for vowels – they distinguish the different vowels of a language.

With the help of a speech analysis program it is possible to see and measure the formants of English vowels. Figure 5.13 presents a spectrogram of the vowels [i], [ɪ], [e], [æ], [ɒ], [ɔ], [ʊ] and [u]. The dark horizontal bands indicate the greater relative intensity of particular frequencies, i.e. the formants, in each vowel. Most of the vowels in Figure 5.13 have between three and five distinguishable formants. The formant with the lowest frequency is called **F1**, the second **F2**, the third **F3** and so on. For the differentiation of vowels in a language it is sufficient to refer to the first two or three formants. The higher formants are especially important in music and singing, where they differentiate between the timbres of different musical instruments and voices and, in the case of the singer's formant at around 3000 Hz, enable singers to be heard over an orchestra.

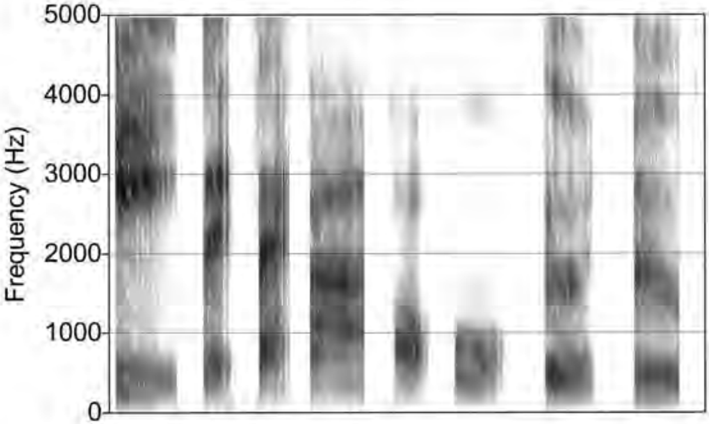


Figure 5.13. Spectrogram of the vowels [i], [ɪ], [e], [æ], [ɒ], [ɔ], [ʊ] *and* [u] *produced by an RP speaker.*

Figure 5.13 illustrates that the first formant in the vowel [ɪ] is higher than in the vowel [i]. Equally, it is higher in [e] than in [ɪ] and higher in [æ] than in [e]. When you compare F1 in the vowels [ɒ], [ɔ], [ʊ] and [u] you can easily see that its height decreases steadily. Thus, it becomes clear that the position of F1 is correlated with tongue height. The lower the tongue during the production of the vowel (as in [æ] and [ɒ]), the higher is F1. Vowels produced with a high tongue position such as [i] and [u], correspondingly, have a low F1. There is thus an inverse relationship between tongue height and height of F1 in vowels. The relationship between vowel type and the second formant is more complicated. Figure 5.13 illustrates that there is some correlation between back vowels such as [ɒ], [ɔ], [ʊ] and [u] and a relatively low F2 – the second formant in the front vowels [i], [ɪ], [e] and [æ] is clearly higher. Yet, F2 is also considerably affected by whether the vowel is produced with rounded or spread lips.

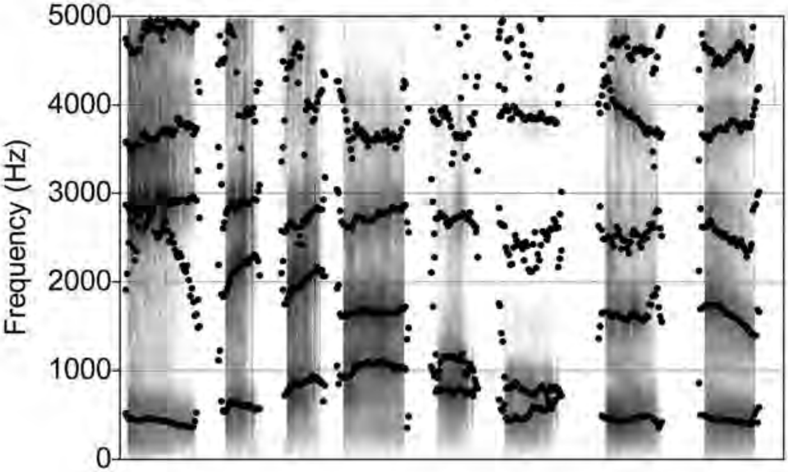


Figure 5.14. The formants of the vowels [i], [ɪ], [e], [æ], [ɒ], [ɔ], [ʊ] *and* [u] *produced by an RP speaker.*

Speech software packages usually offer an automatic analysis of formants that allows a measurement of the actual frequencies of the formants. Figure 5.14 displays the formants of the vowels in *beat*, *bit*, *bet*, *bat*, *pot*, *bought*, *put* and *boot* as given by the speech analysis software *Praat* (which can be downloaded for free at http://www.praat.org). The exact value of each formant frequency is displayed and can be measured and expressed in the unit Hertz (Hz). When measuring formant height in individual vowels, it is important to choose an appropriate place for this. The sounds preceding and following a vowel influence its formant shape. It is thus customary to select the 'stable part' at the mid-point of the vowel for formant measurements. This is of course far from straightforward when the vowel in question is very short.

Several researchers have measured the average frequencies of F1 and F2 in the different vowels of English. Table 5.2 gives these values for both Standard British English and Standard American English.

Table 5.2. Typical formant frequencies (in Hz) of British English and American English vowels (adapted from Kent & Read 2002, p. 111f. and 122).

	F1		F2	
	British English	American English	British English	American English
i	300	290	2300	2250
ɪ	360	400	2100	1930
e	570		1970	
ɛ		550		1800
ɑ	680	710	1100	1100
ʊ	380	440	950	1220
u	300	330	940	1200
ʌ	720	600	1240	1300
ɜ	580	470	1380	1350

When interpreting the values in Table 5.2, you have to bear in mind that these formant values are average values that were measured across several speakers. However, considerable variation in formant frequencies can occur between different speakers. As explained in section 5.1.2, the harmonics of every fundamental frequency have a mathematical relationship; they are multiples of each other. This means that the formant frequencies depend on the **fundamental frequency** of a sound. The fundamental frequency itself corresponds to the rate of vibration of the vocal folds, which, in turn, depends on the length, tension and thickness of the vocal folds. Thus, speech produced by speakers with relatively short vocal folds such as children has a much higher fundamental frequency and consequently different formant frequencies than speech produced by speakers with longer vocal folds (see section 5.3 for details). Equally, the length of a speaker's vocal tract influences the position of the formants, so that the same vowel produced by speakers with shorter pharynxes has different formant frequencies than when produced by speakers with longer pharynxes. However, even when speakers vary in the acoustic realization of vowels, they always maintain the phonological contrast between different vowels. For example, even if the [e] produced by one speaker has similar acoustic properties to the [ɪ] produced by another speaker, the first speaker's [e] and [ɪ] will be sufficiently distinct from each other. Listeners adjust to the particular acoustic properties of individual speakers and normalize them in perception (see also section 6.7).

Formant frequencies further vary between varieties of English. Table 5.2 shows that F2 in the vowels /u/ and /ʊ/ is much higher in American English than in British English. This means that these vowels tend to be more fronted in American English and/or have less lip rounding than in British English. Furthermore, you can see that the IPA symbols used for the different vowels are only approximations and that they cannot represent the actual acoustic properties of vowels. Tradition may be the reason why a vowel with similar F1 and F2 values is transcribed as /e/ in British English and as /ɛ/ in American English. This variation of formant frequencies between speakers and within speakers means that listeners cannot rely on absolute frequency values for the identification of vowels. It is the relative position of F1 and F2 that make listeners perceive one vowel or another (see section 6.7 for more details on vowel perception).

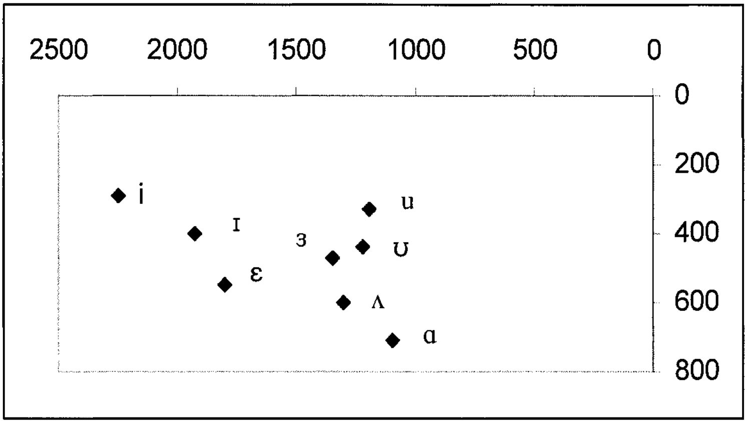


Figure 5.15. Acoustic map for some American English vowels. F2 values on top, F1 values on the right.

The average formant frequencies of vowels can be taken to plot an acoustic map, in which the vowels are positioned according to the frequency of their first two formants. Figure 5.15 shows the formant chart for American English vowels based on the measurements of the formant frequencies in Table 5.2. The values of F1 are displayed in inverse order on the vertical axis, whereas the values of

F2 are displayed on the horizontal axis, again in inverse order. Thus, in a slightly unusual arrangement, the zero values of both axes appear in the top right-hand corner. This arrangement shows immediately that the vowels are positioned in such an acoustic map similar to how they are usually arranged in a vowel chart like the one drawn in section 3.1.3, where vowels were positioned based on their articulatory properties. Thus, the formants of vowels demonstrate a clear relationship to the traditional articulatory description of vowels. Tongue height is inversely correlated with height of F1. F2 varies with tongue retraction (back and front position) and lip rounding. Nevertheless, the better we are able to examine real articulatory movements with techniques such as X-ray and EMMA (described in section 2.8), the more obvious it becomes that vowels and the differences between vowels can be expressed much more accurately in acoustic terms than in articulatory terms. When analysing real speech, a speaker's tongue position when articulating an [i] is far less reliably high and front than F1 being low and F2 high. Listeners perceive an [i] always with this particular constellation of the first two formants but may also perceive [i] with a tongue position other than high and front. It is thus much easier to establish relationships between acoustic and perceptual parameters of vowels than between articulatory and perceptual parameters.

Some researchers have proposed further ways of measuring the acoustic properties of vowels apart from measuring formant frequencies. These include intrinsic vowel duration, intrinsic vowel fundamental frequency as well as formant movement due to adjacent consonants. It was found that some vowels are generally (i.e. intrinsically) longer than others. For example, low vowels such as [a] are longer than high vowels such as [u]. Moreover, high vowels such as [i] have a slightly higher fundamental frequency than low vowels such as [a] (Lehiste 1970). In addition, the speech sounds surrounding vowels, especially plosives and fricatives, modify the position of the formants in vowels. Bilabial sounds such as [b] and [p] cause a general lowering of the formants. Velar sounds such as [k] and [g] cause F2 and F3 to come close together at the beginning of the sound and to separate as the velar plosive is released. These changes in vowel formant frequencies are referred to as **formant transitions** and will be described in detail in the next section.

The acoustic properties of diphthongs are similar to those of monophthongs. Diphthongs are vowels that consist of two vocalic parts. Unlike monophthongs, they do not have a single vocal tract shape – for the articulation of a diphthong the tongue changes its position. This means that the acoustic properties of diphthongs are saliently different from the acoustic properties of monophthongs. Figure 5.16 shows the spectrogram of the diphthongs [eɪ], [aɪ], [ɔɪ], [aʊ] and [əʊ] produced by an RP speaker. You can see that the formants change during the production of the diphthongs. In [eɪ], the starting position (or **onglide**) has a higher F1 and a lower F2 than the final position or **offglide**. For [aɪ], this is also the case. By the same token, the formant frequencies change considerably during the production of the diphthongs [ɔɪ], [aʊ] and [əʊ].

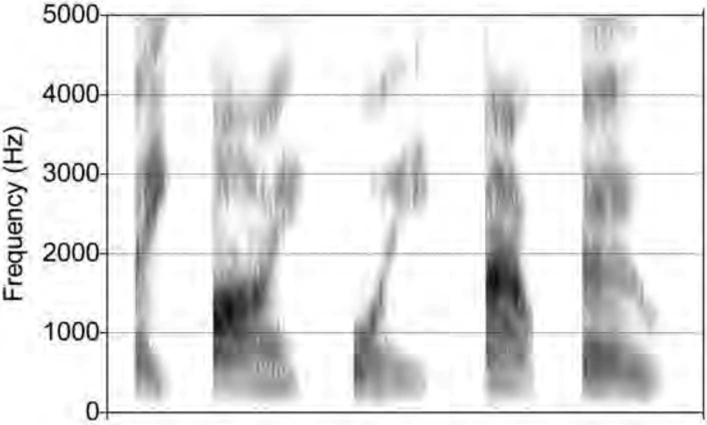


Figure 5.16. Spectrogram of the diphthongs [eɪ], [aɪ], [ɔɪ], [aʊ] *and* [əʊ].

5.3 The acoustic properties of English consonants

The acoustic properties of English consonants, like English vowels, can be visualized and measured with the help of a spectrogram. Yet, there is no single acoustic measurement that can be applied to distinguish all consonants in English. Due to the varying manners of articulation that underlie the different consonants of English, it is useful to group them accordingly for an acoustic analysis. Generally speaking, voiced consonants, like vowels, are characterized by a formant structure. This is especially distinctive for the group of nasal consonants, as will be described below. Voiceless consonants, by contrast, are often characterized by stretches of high noise energy. This applies especially to fricatives as shown below. Plosives are produced with a complete obstruction of the airstream in the vocal tract, which gives them their specific acoustic pattern.

The acoustic pattern of plosives can be described as a sequence of at least three separate acoustic events: the **closure** part during the blockage of the airstream, the **burst** when the pent-up airstream is released suddenly, occasionally **aspiration**, and the formant transition part. Figure 5.17 shows the spectrogram of a speaker uttering the voiced plosives [b], [d] and [g] in the nonsense words *aba, ada, aga.* You will immediately notice the gap in the spectrogram after the first [a], which of course reflects the time when no air passes through the mouth during the closure of the lips.

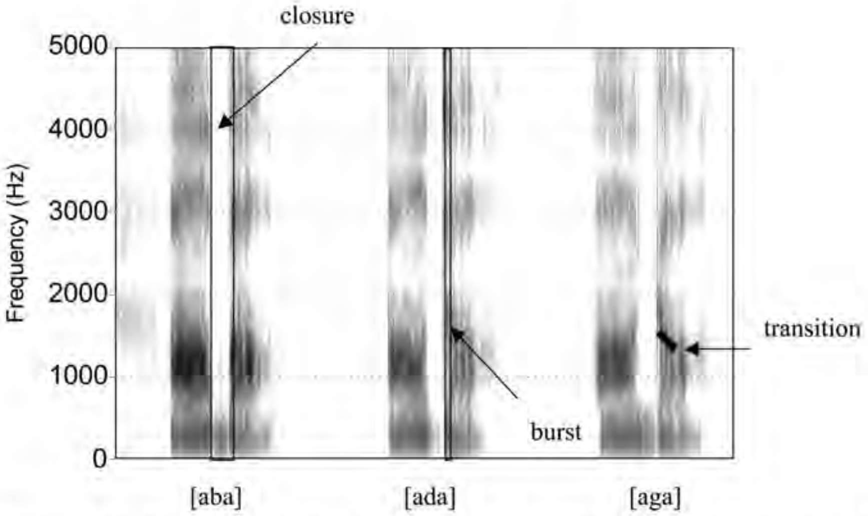


Figure 5.17. Spectrogram of aba, ada, aga *illustrating the closure, burst and transition of plosives.*

Typically, the closure part of plosives is between 50 and 150 ms long. Longer closures are perceived as speech pauses by listeners. The burst of plosives is one of the shortest acoustic events in speech lasting for between 5 and 40 ms only. It is visible in spectrograms in the random high energy along the entire frequency spectrum. Voiceless plosives at the beginning of a stressed syllable are aspirated in English when they do not form part of a consonant cluster. The term aspiration refers to the breathy noise generated as the airstream passes through the partially closed vocal folds. The acoustic characteristics of aspiration therefore are essentially the same as the acoustic characteristics of the fricative [h]. Aspirated [p^h], [t^h] and [k^h] can be seen in Figure 5.18, which shows the spectrogram of [ap^{h}a], [at^{h}a] and [ak^{h}a]. Aspiration in these examples occurs, on average, for 100 ms.

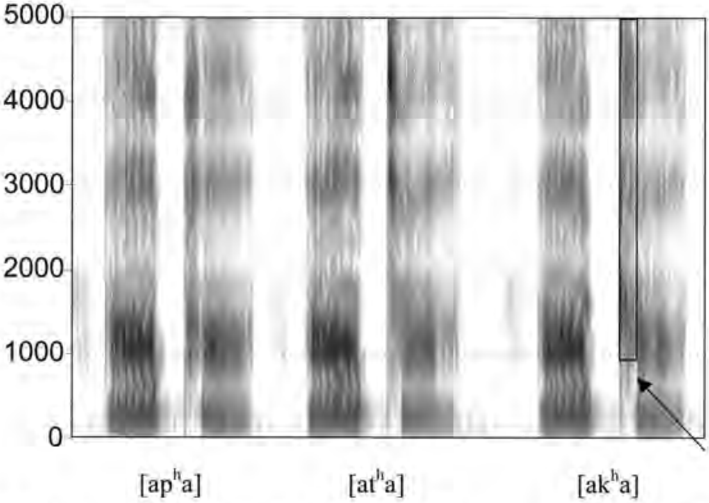


Figure 5.18. Spectrogram of ap^ha, at^ha, ak^ha *illustrating aspiration (the arrow points towards the aspiration phase of [k^h], which is marked by the box).*

As far as the voiceless consonants are concerned, in many cases it is difficult to distinguish them clearly in a spectrogram. The difference between the voiced plosives [b], [d] and [g], however, can be seen in the beginning of the formant structure of the following vowel. The difference between the three consonants lies in the formant transitions, the movements of the second and third formant at the beginning of the vowel in syllables like [ba] and [da], and the movements of the formants at the end of the vowel in syllables like [ag] and [ad]. These formant transitions reflect the changes in resonance connected with the movement of the articulators for the closure phase of the plosive. They therefore carry information about the place of articulation of the plosive. Figure 5.19 illustrates schematically the formant transitions of [b], [d] and [g] when these plosives precede the vowel [a]. For [b], F2 and F3 begin low and rise. For [d] and [g], F3 rises into the vowel and F2 drops into it. It is the specific characteristic of [g] that F2 and F3 are close together at the beginning of the vowel. The formant transitions of the voiced plosives following a vowel in words like *rob*, *odd* and *dog* are practically mirror images of the formant transitions of these sounds preceding a vowel.

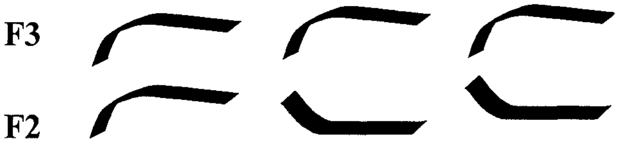


Figure 5.19. Schematic representation of formant transitions in ba, da, ga.

Differences between voiced and voiceless plosives can be measured in **voice onset time** (VOT). This term refers to the time period between the burst and the beginning of voicing. For English voiced plosives, the burst typically occurs between 20 ms before and after voicing begins. In other words, when producing [ba], [da] or [ga], speakers usually release the blockage of the airstream for the [b], [d] or [g] between 20 ms before and 20 ms after their vocal folds start vibrating for [a]. The corresponding VOT values are thus –20 ms to +20 ms. If voicing actually starts before the burst, the consonant is called **prevoiced**. For voiceless plosives in English, the typical VOT ranges between +40 and +80 ms. Aspirated plosives can have a VOT of up to +120 ms, which means that there can be a 120 ms interval filled with friction between the release of the airstream obstruction and the vowel in words like *pat*, *tack* and *cap*.

The group of nasal consonants is characterized by their specific formant structure. Figure 5.20 shows the spectrogram of [ama], [ana] and [aŋa]. Due to the resonating characteristics of the nasal cavity, the formant structure of nasals is clearly distinguishable from that of the vowel. The first formant of nasals is low at about 250 Hz to 300 Hz, and there is very little energy at frequencies above this up to about 2500 Hz where another **nasal formant** lies. Both formants can only be seen weakly in Figure 5.20. The region in between the two nasal formants (which appears almost white for [m] and [n] in Figure 5.20) is called the region of the **antiformants**. In general, nasals have less overall energy than vowels. The difference between [m], [n] and [ŋ] lies in the formant transition into and from the vowel. Usually, a vowel preceding a /m/ shows a downward movement of F2 and F3, whereas in a vowel preceding a /ŋ/ F2 and F3 come together. Very little formant movement can typically be seen in a vowel before a /n/.

The voiced approximants [l], [ɹ], [w] and [j] also have formants not unlike those of vowels. Figure 5.21 shows a spectrogram of a speaker saying [ala], [aɹa], [awa] and [aja]. You can see that the [l] has a low F2 (at around 1700 Hz) and a high F3 (at around 3000 Hz) which, however, generally have little energy. Postvocalic 'dark l' has a higher F1 and lower F2 than prevocalic 'clear l' in British English (Lehman & Swartz 2000). [ɹ] has a low frequency in the third formant – typically well below 2000 Hz – which rises sharply at the beginning of the vowel. [w] has a low position of all formants with a sharp rise in F2. The approximant [j] has a low F1 and a high F2 which move towards each other. In fact, the formant structure of [w] resembles that of [ʊ] very much just like the formant structure of [j] resembles that of an [ɪ]. This explains why these two approximants are called semi-vowels and why some linguists transcribe the diphthong /aɪ/ as /aj/ and /aʊ/ as /aw/.

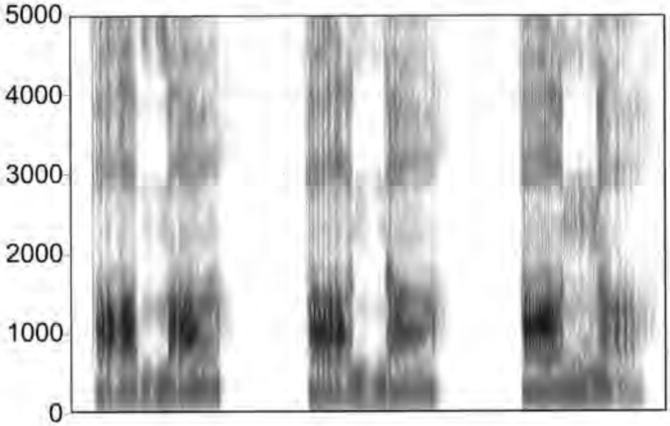


Figure 5.20. Spectrogram of ama, ana, aŋa *illustrating the nasal formants at about 250 Hz and 2500 Hz.*

The last group of consonants in English with a distinct acoustic pattern are the fricatives. During their articulation a narrow constriction in the vocal tract causes turbulence in the airstream, which is associated with noise in the acoustic signal. This noise is much more intense in the stridents (sometimes also referred to as sibilants) [s, z, ʃ, ʒ] than in the non-stridents [f, v, θ, ð, h]. The voiced fricatives only have a very faint formant structure where each formant has only little energy. This is illustrated in the spectrogram of [ava], [aða], [aza] and [aʒa] in Figure 5.22. [z] and [ʒ] are clearly characterized by their noise, the random energy in the higher frequency regions.

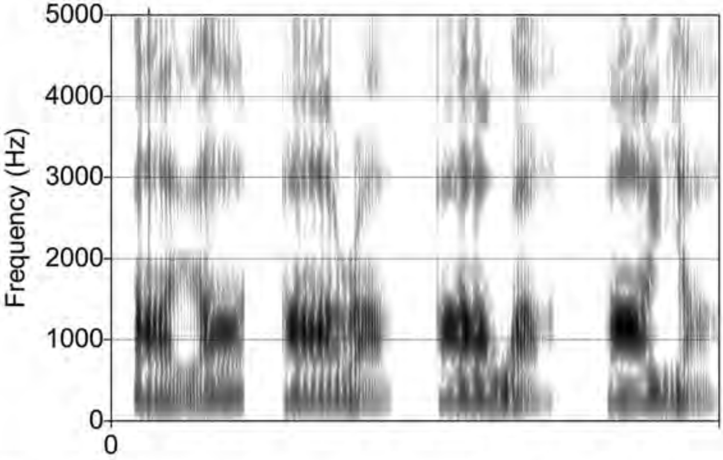


Figure 5.21. Spectrogram of ala, aɹa, awa, aja *illustrating the formant transitions.*

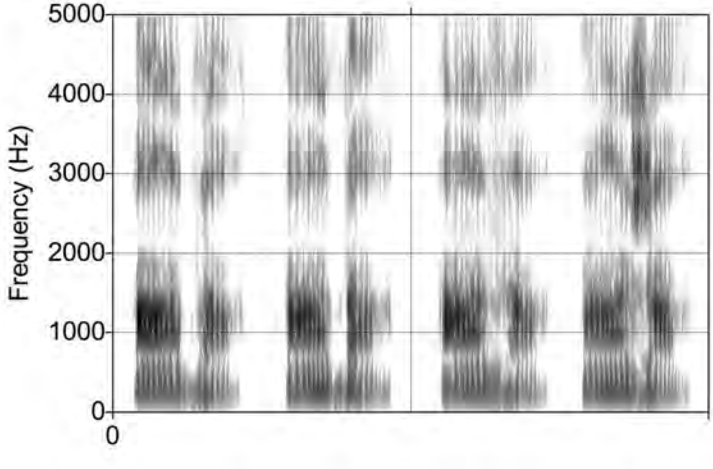


Figure 5.22. Spectrogram of ava, aða, aza, aʒa *illustrating the faint formant structure and friction noise in [z] and [ʒ].*

The friction of voiceless fricatives is, like the burst of the voiceless plosives, visible as random noise in a spectrogram. In contrast to the burst of plosives, however, the noise is restricted to the upper frequency regions and varies with the place of articulation. Figure 5.23 shows a spectrogram of the syllables [fa], [θa], [sa], [ʃa] and [ha]. You can see that the noise of the labiodental [f] and the interdental [θ] is randomly distributed from about 1000 Hz onwards. The difference between the two speech sounds lies in the formant transitions: for [θ] F2 moves downwards at the beginning of the following vowel, whereas it stays relatively stable for [f]. The energy peaks in the noise of a [s] lie between 5000 and 8000 Hz. Conversely, in [ʃ] the random noise has a peak at about 2500 Hz with high energy between about 4000 to 8000 Hz. Both of these consonants are also associated with distinct downward movements of F2 at the beginning of [a]. The [h] does not show any formant transitions because its articulation does not involve any articulators in the vocal tract. During the production of [h] a speaker can already form the vocal tract appropriate for the articulation of the following vowel.

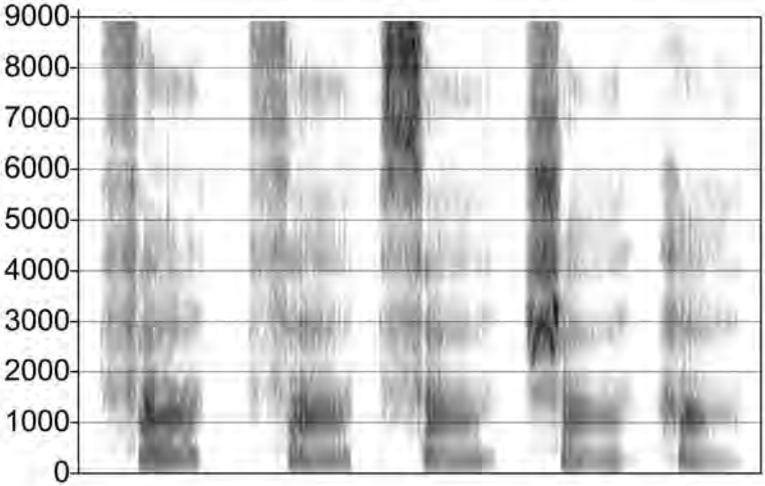


Figure 5.23. Spectrogram of fa, θa, sa, ʃa, ha.

Table 5.3 gives an overview of some acoustic properties of consonants. These descriptions, however, can only function as rough guidelines since both the adjacent sounds and the prosodic function of the consonant (whether it occurs in a stressed or an unstressed syllable, at the beginning, in the middle or at the end of a word or an utterance and so forth) influence the articulation and hence also their acoustic features.

Table 5.3. Acoustic properties of consonants (rough guidelines).

Place or manner of articulation	Acoustic property
approximant	formant structure like vowel
nasal	formant structure with nasal formant at about 250 Hz, 2500 Hz and 3250 Hz
lateral	formant structure with F1 at about 250, F2 at about 1200 and F3 at about 2400
fricative	random noise in higher frequency depending on place of articulation
plosive	gap followed by noise at all frequencies for voiceless plosives and formant structure for voiced plosives
bilabial	low F2 and F3
alveolar	F2 at about 1700 Hz
velar	F2 high; F2 and F3 joined at beginning of formant transition
retroflex	F3 and F4 low

5.4 Acoustic aspects of connected speech in English (advanced reading)

Section 2.5 showed that speech does not consist of a sequence of speech sounds but that the articulatory gestures underlying one speech sound are always affected by the articulatory gestures underlying the surrounding speech sounds. This fact can be easily made visible and measured with the help of speech analysis software. In general, a comparison between a speech sound produced in isolation (= on its own) and the same speech sound produced in connected speech shows distinct differences between the two. The overall duration of a speech sound is shorter when produced in a sequence of sounds than when produced in isolation. Moreover, consonants in consonant clusters are shorter than single consonants in the onset position of a syllable. Thus, the [w] in *wing* is longer than the [w] in *swing*. By the same token, the syllable [stɪk] is longer in *stick* than in *sticky* and shortest in *stickiness* (Lehiste 1972). Furthermore, the duration of speech sounds varies with their phonotactic position. For example, a vowel followed by a voiced sound is longer than when followed by a voiceless sound. This means that the [i] in *heed* is longer than the [i] in *heat*. The same consonant in syllable coda position is longer than in syllable onset position. Thus, the [d] in *head* is longer than the [d] in *do*. The length of speech sounds is furthermore determined by their prosodic position. For example, the VOT of a word-initial voiceless plosive such as the [t] in *tack* is longer than when the [t] appears in the middle of a word as in *attack* (e.g. Turk & Shuttack-Hufnagel 2000). Equally, a vowel in a word at the beginning of an utterance is longer than the same vowel appearing in the middle of an utterance (Cho 2005).

The duration of individual speech sounds depends crucially on the **speaking rate**. In careful speech – which speakers produce in an effort to be highly intelligible, for example when addressing young children, a hard-of-hearing person or foreigners – speech sounds are lengthened and the duration of pauses is increased. On average, between two and three syllables are produced per second in careful speech. In conversational speech, by contrast, between 5 and 11 syllables are produced per second. This means that vowels and consonants often lose some of their acoustic distinctiveness as they are modified or reduced. In very rapid speech, some speech sounds are even 'deleted' altogether, i.e. their articulatory gestures are not carried out to an audible extent. Typically, only 5% of three-consonant coda clusters are realized. In other words, a word containing three consonants in the coda position such as *facts* /fækts/ is nearly always produced as [fæks]. This reduction is much more frequent in function words (e.g. prepositions, conjunctions and auxiliaries) than in content words (e.g. nouns, verbs, adjectives).

Figure 5.24 shows the waveform and spectrogram of the sentence "Give her the post." read at a relatively slow speaking rate by a British English speaker. The duration of each speech sound is indicated by the boundaries given in the annotation below the spectrogram. When you measure the length of the vowel in *give*, you can see that it is 102 ms long. The [ð] in *the* is 40 ms long. Compare this with Figure 5.25, which shows the spectrogram and waveform of the same sentence produced at a higher speaking rate. First of all you can see that one speech sound is deleted: there is no acoustic trace of the /h/ in *her*. Furthermore, you can see and measure that many of the speech sounds have a much shorter duration when produced with higher tempo. The vowel in *give*, for example, is only 43.9 ms long in the utterance produced at the higher speaking rate. The /ð/ in *the* is so short (13 ms) that it is perceived as a [d] rather than a [ð]. When you carry out a segmentation of the individual speech sounds in an utterance – that is, when you indicate the beginning and end of each speech sound in the annotation – you will notice that this is increasingly difficult with increasing speaking rate. For some speech sounds it will be impossible to determine their exact boundaries.

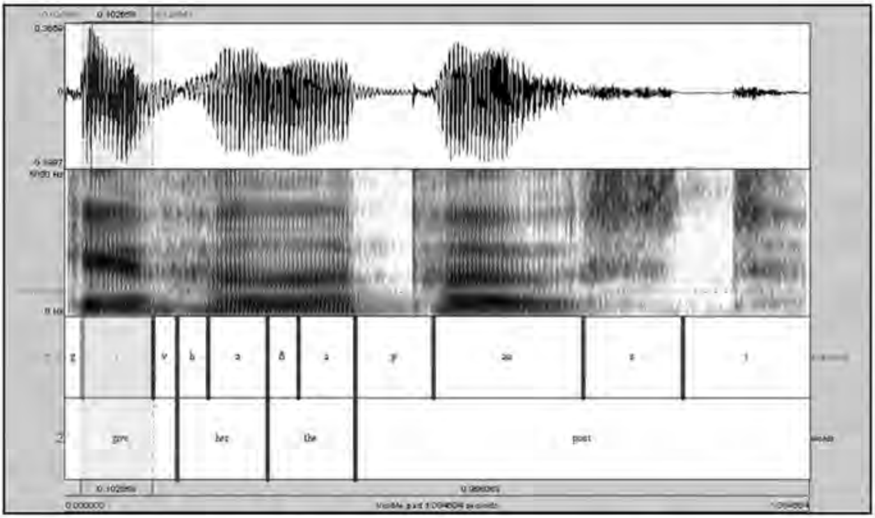


Figure 5.24. Duration of the individual speech sounds in a slow reading of the sentence "Give her the post." by a British English speaker (listen to it on the CD-ROM: 05_slow.wav).

This difficulty to determine the boundaries between speech sounds in connected speech is due to coarticulatory processes. Section 2.5 described that during speech production the different articulators move in an overlapping way. While producing the [s] in words like *stew*, the lips are already rounded in anticipation of the following vowel [u]. This overlap of articulatory gestures for speech sounds in connected speech is referred to as **coarticulation** and can occur both as anticipatory coarticulation as with the [s] in *stew* and as perseverative coarticulation, in which case the articulatory gestures of one speech sound continue during the production of the following sound. For example, perseverative coarticulation can be seen in the word *climb*, where the [l] is produced as voiceless due to the voicelessness of the preceding [k]. Coarticulatory effects can be viewed in a spectrogram. The vowel [æ] in *am*, for example, will typically be nasalized, i.e. the spectrogram will show the nasal resonance due to the lowering of the velum for the following [m] already during the production of the vowel. This is visible in the presence of low-frequency nasal formants and the appearance of antiformants with severely reduced energy in particular frequency regions. By the same token, spectrograms can show that in rapid speech many vowels between voiceless consonants – as in the word *pit* – are voiceless themselves. This is especially frequent in unaccented syllables in English.

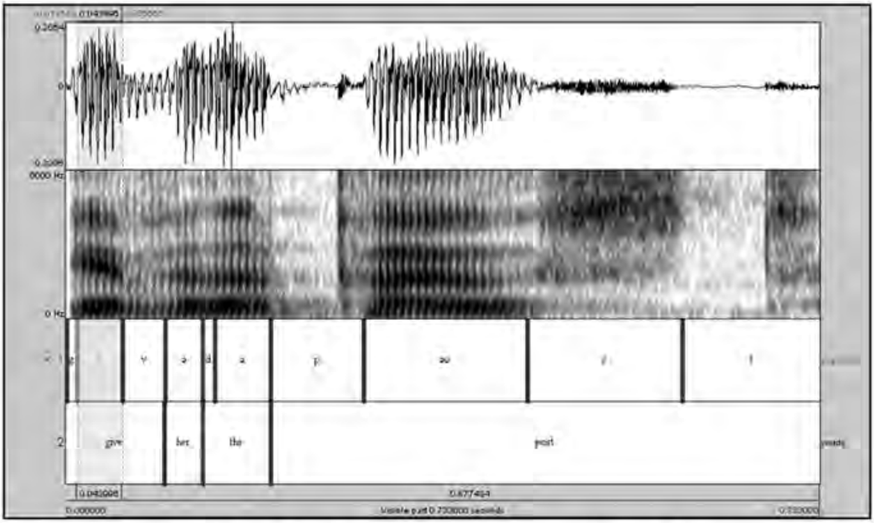


Figure 5.25. Duration of the individual speech sounds in a fast reading of the sentence "Give her the post." by a British English speaker (listen to it on the CD-ROM: 05_fast.wav).

5.5 The acoustic properties of English intonation

Chapter 4 has shown that intonation in English comprises both the placement of accents in an utterance and the employment of pitch and pitch movements across and between utterances. Nucleus placement has the main function of signalling which elements in the utterance are in focus (i.e. which are new and which are especially emphasized). Pitch is employed by speakers for various purposes including the marking of the function of an utterance (for example as a request for information or as an aside) and the expression of specific attitudes. The acoustic properties of both – accents and pitch movements – can be visualised and measured with the help of speech analysis software packages.

5.5.1 The acoustic properties of intonation phrases in English

Section 4.1 explained that one of the major functions of intonation in English is the structuring of utterances and discourse. These basic units of speech are referred to as **intonation phrases**. Many acoustic cues conspire to give the perceptual impression of an intonational phrase boundary. Acoustic properties correlating with intonation phrases in English include

- length of following pause
- final syllable lengthening
- change in pitch level after a phrase boundary
- tempo of unstressed syllables after a phrase boundary

Typically, pauses occurring after intonation phrase boundaries are between 100 and 600 ms long (Campione & Veronis 2002). There seems to be a tendency for speakers to produce longer pauses after more significant phrase boundaries (such as between the last utterance in a topic unit and the first utterance of the following topic unit) than after minor intonation phrase boundaries (e.g. between an adverbial and the main clause within an utterance), at least in reading style. In spontaneous speech, the relationship between pause length and status of intonation boundary is less clear (Keseling 1992).

The length of the final syllable before an intonation phrase boundary is the most important cue indicating the end of an intonational phrase. The term **final syllable lengthening** describes the phenomenon that the last syllable of a word is distinctly longer when this word occurs at the end of an intonation phrase than when it occurs in the middle of an intonation phrase. Thus, the syllable *ing* in *going* is longer in the utterance "Where are you going" than in the utterance "I am going home". The length of the final syllable increases with the rank of the intonation phrase boundary: there is less final syllable lengthening before less important boundaries (e.g. Yang 2004).

Another acoustic correlate of intonation phrase boundaries is the **resetting of pitch** after the boundary. As will be explained in section 5.5.3 below, pitch usually drops throughout an utterance and reaches its lowest point at the end of the last intonation phrase. At the beginning of the next intonation phrase, pitch is picked up at a higher starting position, which is clearly visible in automatic pitch contour displays. Finally, the first unstressed syllables of a new intonation phrase are produced at a very high speaking rate. Thus, when a speaker produces utterance (104), the unstressed syllables *but* and *un* after the intonation phrase boundary are likely to be articulated very quickly. This sudden acceleration in speech tempo is called **anacrusis**.

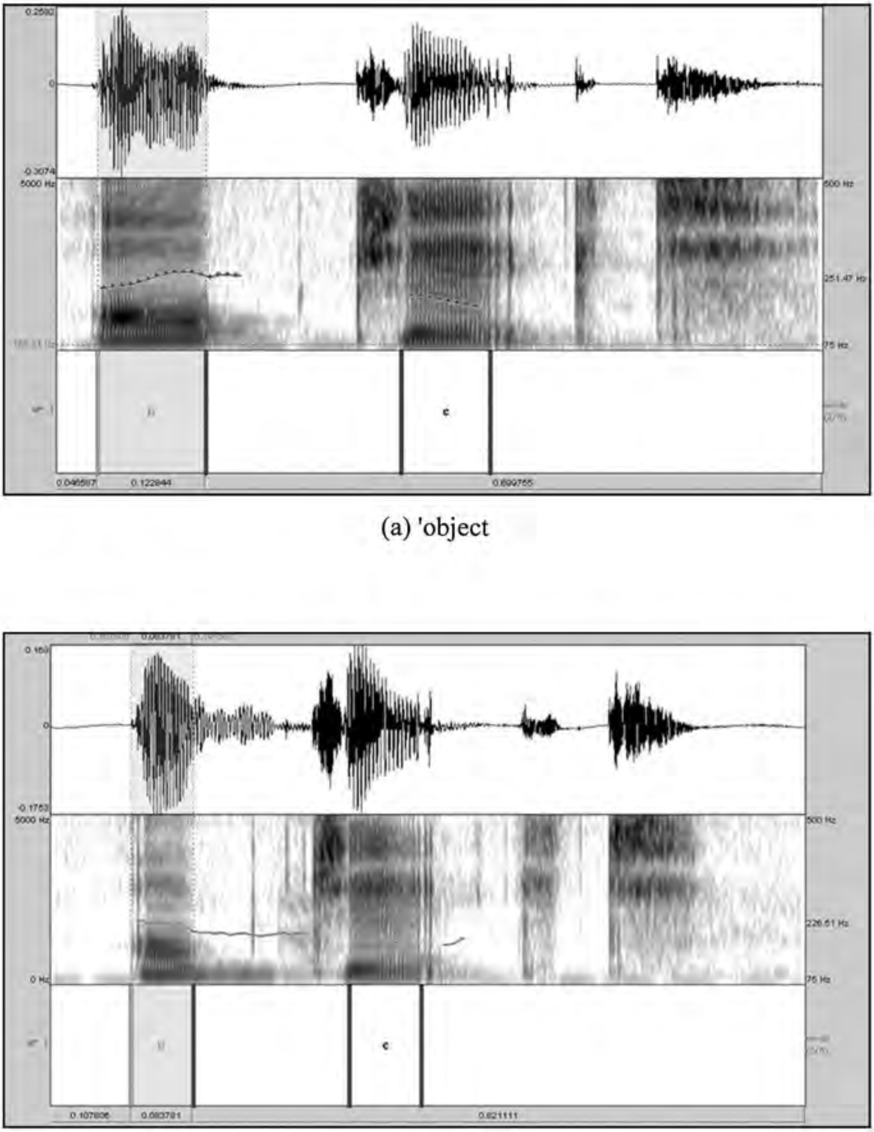
(104) I was going home | but unfortunately I had forgotten my keys

5.5.2 The acoustic properties of accents in English

In English, syllables can be accented or unaccented. The first syllable of *English*, for example, is usually accented when a speaker produces this word, and the second one is unaccented. The acoustic differences between the two types of syllable can be measured in terms of their

- duration
- intensity
- pitch
- vowel quality

Various studies have shown that accented syllables are longer than unaccented syllables in English (e.g. Fant et al. 1991, Williams & Hiller 1994). On average, accented syllables are about 300 ms long, whereas unaccented syllables are about 150 ms long. This difference in length between the two types of syllable is mainly due to the processes of **vowel reduction** and **vowel deletion**, which occur in unaccented syllables in English. Vowels in unaccented syllables are much shorter than vowels in accented syllables, and often they are not realized at all (see e.g. Delattre 1981). Vowels in accented syllables furthermore have a higher intensity and a higher pitch than vowels in unaccented syllables. Since the intensity of vowels in accented syllables is typically increased, English has been described to have **dynamic accentuation**. Yet, intensity seems to be the least consistent and least salient property of accentuation in English. Perceptual experiments have shown that a rise in pitch on a syllable is the most important cue for accentuation in English. Vowels in accented syllables are associated with a prominent pitch peak, whereas vowels in unaccented syllables generally have lower pitch. Vowels in accented and unaccented syllables further have a different quality, which is reflected in a different formant structure in a spectrogram. Vowels in unaccented syllables are produced with a more central position of the tongue, with narrow jaw-opening and without lip rounding, which results in nearly evenly spaced formants. Figure 5.26 shows the spectrograms of the words *'object* and *ob'ject.* When you analyse the acoustic parameters of the vowels [ɒ] and [e] in both words, you can see that [ɒ], when it is accented, is 122 ms long, has a pitch peak of 262 Hz and a peak intensity of 72 dB. When it is produced in an unaccented syllable in *ob'ject*, however, it lasts 83 ms, has a pitch peak at 226 Hz and a peak intensity of 72 dB. This pattern is also true for the second vowel [e]. In accented position it is 96 ms long, has a pitch peak of 179 Hz and a peak intensity of 72 dB. In unaccented position, by contrast, it has a length of 90 ms, a pitch peak at 170 Hz and a peak intensity of 73 dB.



(a) 'object

(b) ob'ject

Figure 5.26. Waveform and spectrogram of the words 'object *(a) and* ob'ject *(b).*

5.5.3 Measuring pitch and pitch movement in English

Section 2.2 and section 5.1.1 above explained that the pitch of a speaker's voice is determined by the rate of vocal fold vibration, which in turn depends on the length and thickness of the vocal folds. By altering the tension (and thereby the length and thickness) of the vocal folds, speakers can vary the rate of vocal fold vibration and can produce different levels of pitch and pitch movements. What is perceived as pitch is in acoustic terms the **fundamental frequency** of the complex period wave underlying voiced speech sounds. The fundamental frequency (F0) of speech is measured acoustically in Hertz (Hz). It is possible to do this in a waveform (by counting the peaks of amplitude per second) or in a spectrogram (by reading off the frequency of F0, especially a narrowband spectrogram which has good frequency resolution, see section 5.1.2 above), but most speech analysis software packages also offer a more convenient **pitch tracking** tool. This tool extracts the fundamental frequency from the speech signal and displays it as a pitch contour. By clicking on any point of this contour, a measurement of F0 can be taken. Figure 5.27 shows the pitch track of a British English speaker reading "A tiger and a mouse were walking in a field." at the beginning of a story.

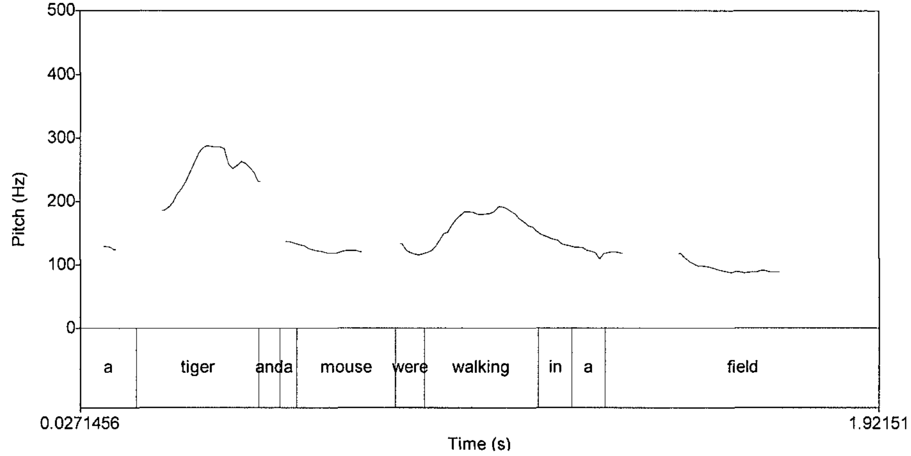


Figure 5.27. Pitch track of the sentence "A tiger and a mouse were walking in a field." read by a British English speaker.

The first noticeable feature of the displayed pitch movement is that there are gaps in the pitch contour. These indicate the presence of voiceless speech sounds such as the [t] in *tiger*, the [s] in *mouse* and the [f] in *field*, which do not have pitch. It is furthermore clearly visible which syllables are accented in this utterance: both the syllable *ti* in *tiger* and the syllable *wal* in *walking* are produced with a distinct pitch peak; the syllable *field* has a smaller pitch peak. When comparing the pitch peaks, it becomes obvious that the first peak on the syllable *ti* is higher than the second on the syllable *wal,* which in turn is higher than that on *field.* This phenomenon is referred to as **declination**, a term that describes the overall drop of pitch height across utterances. It is assumed that declination is caused by the decreasing air pressure during speech production. The longer a speaker produces speech on an airstream, the less air there is left in the lungs and the lower the air pressure is. Listeners adjust for declination, which means that usually the overall drop of pitch is not noticeable. In fact, synthetic speech without declination sounds very unnatural.

Several acoustic measurements can be taken that describe English intonation. First of all, the acoustic properties of **tones** (see section 4.3), for example the extent of nuclear falls and rises in utterances can be determined. The nucleus of the utterance displayed in Figure 5.27 lies on the last word *field.* A transcription of the intonation of this utterance is given both in the British School tradition (105a) and in ToBI (105b):

(105a) A ↘tiger and a mouse were ↘walking in a \field
(105b) A tiger and a mouse were walking in a field
H*+L H*+L H* L%

By measuring the exact pitch height of the highest pitch at the beginning of the word *field* (117.64 Hz) and the pitch height at the end of the word *field* (88.69 Hz), one can calculate that the extent of the nuclear fall is 28.95 Hz (117.64 minus 88.69). By the same token, the **pitch range** of this utterance can be measured. This is calculated by subtracting the lowest pitch at the end of the utterance (88.69 Hz at the end of the word *field*) from the highest pitch peak at the beginning of the utterance (287.23 Hz on the syllable *ti*). The pitch range of this utterance is thus 198.54 Hz. Section 6.5 explains that the distance between two frequencies expressed in Hz is not perceived equally for all frequency ranges. In short, the same difference of 198.54 Hz between two points in the pitch range sounds greater when produced in a low frequency region (i.e. by a low voice) than in a high frequency region (i.e. produced by a higher voice). In order to be able to compare the pitch range of different speakers, it is therefore common to employ the unit **semitones** as a measurement. A pitch range of the same number of semitones thus has exactly the same perceptual range for every speaker.

For every utterance, the average pitch height can be measured and speakers can be compared according to their average pitch. Acoustic measurements of this kind have shown that, on average, women speak with higher pitch than men do and that children's average pitch is highest (dropping with increasing age). This is of course related to the relative length of a speaker's vocal folds, which co-determines the rate of vocal fold vibration. Vocal folds of adult males have an average length of 17-25 mm while adult women's vocal folds are, on average, 12 to 17 mm long. The vocal folds of newborn babies are about 2 mm long. Table 5.4 gives some average values of pitch height for adult males, females and children.

Table 5.4. Average pitch height of male speech, female speech, and children's speech.

	Average pitch	Pitch range
male speech	100-150 Hz	70-250 Hz
female speech	150-250 Hz	80-400 Hz
children aged 2	500 Hz	300-1000 Hz
children aged 6	400 Hz	300-700 Hz
children aged 10	300 Hz	200-500 Hz

Empirical studies suggest that pitch range is language-specific (e.g. van Bezooijen 1995). This means that speakers of different languages habitually use different pitch height and pitch range. For example, English speakers typically have a higher average pitch or a wider pitch range than speakers of German (Mennen 2007). On average, for female speakers the pitch range in English is 7.16 semitones, while it is only 5.42 semitones in German.

Pitch range and pitch height are used by English speakers to signal new information and the beginning of new topics (see also section 4.3.3). Acoustic measurements have shown that utterances containing a new topic have a wider pitch range and begin at a higher pitch than utterances that merely give additional information or constitute reformulations of a previous utterance (e.g. Wennerstrom 1998, Wichman 2000). When a speaker produces parenthetical material such as an aside or a meta-discoursal remark, this is often associated with a narrow pitch range. For example, a speaker would produce the aside "but you know this already" with a narrower pitch range than "Grandma told me that her neighbour has moved to Thailand" in example (106).

(106) Grandma told me that her neighbour has moved to Thailand | but you know this already

Acoustic measurements of speech can be employed to describe differences in voices between speaker groups, for example between young and old speakers, men and women, healthy speakers and speakers with handicaps such as cleft palate (see section 2.3) or dysarthria, a neurological speech disorder associated with Parkinson's disease. In general, voices of older speakers have less regular vocal fold vibration, which is visible in pitch trackings. Moreover, it has been found that women's voices are often breathier than men's voices and that speakers born with a cleft palate and/or cleft lip show a high degree of nasalization in their speech. Some acoustic measurements of speech can be used for speaker identification, for example in **forensic phonetics**. At court, phoneticians are sometimes asked to make positive identifications of criminals whose speech was recorded. Speaker identification is based on the fact that some aspects of speech are relatively characteristic for individual speakers due to the particular physiological characteristics of each speaker. For example, F4 and higher formants in vowels are associated with a particular voice quality, and the higher nasal formants depend on individual vocal tract characteristics. Equally, long-term average spectra (LTAS) reflect individual voice characteristics, and the rate of formant transitions after voiced plosives as well as the length of aspiration of voiceless plosives vary systematically among speakers.

When measuring and interpreting an automatically generated pitch contour, it is important to know about human perception of pitch. Chapter 6 describes in detail how the frequencies in a sound wave are related to the sensation of pitch height. It shows that humans, on average, can only pick up frequencies between 20 and 20,000 Hz and that not every change in frequency is directly related to a change in the perception of pitch. Furthermore, changes in frequencies are not perceived as equal in the lower and the higher frequency ranges, which is why differences between speakers are often described using units other than Hz (for example semitones). Similarly, the automatic pitch tracking algorithm displays many small changes in pitch that are linguistically unimportant because listeners do not notice them. For example, prevocalic stops and fricatives produce short-term perturbations (= a jump up or down) of the pitch contour, as you can see for the [g] in *tiger* in Figure 5.27. Some tools have therefore been developed that display only those pitch changes across utterances that are perceptually relevant (e.g. Mertens 2004) and that display 'smoothed' pitch curves which exclude short-term perturbations.

5.6 Acoustic properties of L2 learner English and the use of acoustic phonetics in pronunciation teaching (advanced reading)

Many empirical studies have shown that the acoustic properties of speech produced by second language learners of English differ from those of speech produced by native speakers of English. Barry (1989) analysed how ten German learners of English, aged 17-19, who had been learning English at school for six years produce the vowels in words like *bat*, *bet* and *but*. The acoustic measurements showed that these learners of English confuse the vowels /æ/ and /e/ in words such as *bat* and *bet*. Especially the F2 values are roughly the same for both vowels instead of being distinctly lower for the vowel /æ/ than for /e/. It was further examined whether these learners of English produce the difference in duration of a vowel before the voiced plosives /b, d, g/ and the voiceless plosives /p, t, k/. Native speakers of English produce a longer vowel (approximately 200 ms) in *dog* than in *dock* (approximately 135 ms). At the same time, the duration of the closure of /b, d, g/ is shorter (approximately 130 ms) than the closure of /p, t, k/ (approximately 150 ms) in native English. It was found that many German learners of English do not produce such a clear distinction between closure durations and vowel durations.

Many studies have been concerned with the acoustic measurement of voice onset time (VOT, see section 5.3 above) of plosives in English. For example, it was found that some Spanish native speakers produce a shorter VOT for the plosives /t/ and /p/ in English (Schmidt & Flege 1996) than native English speakers do. This means that the Spanish speakers' productions of /p/ are closer to what English speakers identify as /b/ and their productions of /t/ are closer to an English native speaker /d/.

Section 5.5.2 above showed that several acoustic changes can be measured between accented and unaccented syllables. These include a longer vowel with different formants as well as higher pitch and higher intensity of the vowel in accented syllables, with high pitch being the most important perceptual cue. When German learners of English produce an emphasized syllable, they sometimes make different use of these acoustic parameters than English native speakers (Gut 2000b). For some speakers, the greatest difference between an emphasized and an unemphasized syllable lies in a difference in intensity rather than in a difference in pitch height. This might be due to the fact that accentuation in German also relies more on intensity changes than on pitch changes. Most learners of English furthermore do not reduce vowels in unaccented syllables to a native-like extent. An analysis of the LeaP corpus (Gut 2007a) of non-native English has shown that the mean length of an unaccented syllable is 101.8 ms in native English but 155.07 ms in learner English. When an unaccented syllable follows an accented one, as in the word *measure*, the accented syllable is on average 2.45 times longer than the unaccented one in native English. Many learners of English do not succeed in producing the same durational difference: their accented syllable is, on average, only 1.98 times longer than the following unaccented one. Acoustic studies of the intonation that learners of English produce also point out differences between native and learner speech. Dutch learners of English have a smaller declination rate across utterances than native English speakers (Willems 1982). An analysis of the LeaP corpus of non-native English has shown that the nuclear falls produced by learners of English at the end of an utterance are smaller (they extend only across 3.64 semitones) than the falls produced by native English speakers (7.81 semitones on average) (Gut 2007a). Moreover, some learners of English use pitch height for the marking of new and given information to a different degree than native English speakers do. Wennerstrom (1998) found in her study that some learners of English do not produce higher pitch on new information. By the same token, these learners produce little difference in pitch height between the end of a topic and the beginning of a new topic. There is less resetting of pitch level after an intonation phrase boundary in learner English than in native English (Willems 1982). Another study showed that Japanese and Thai learners of English do not mark the first utterance in a new paragraph with a wider pitch range like native speakers do (Wennerstrom 1994).

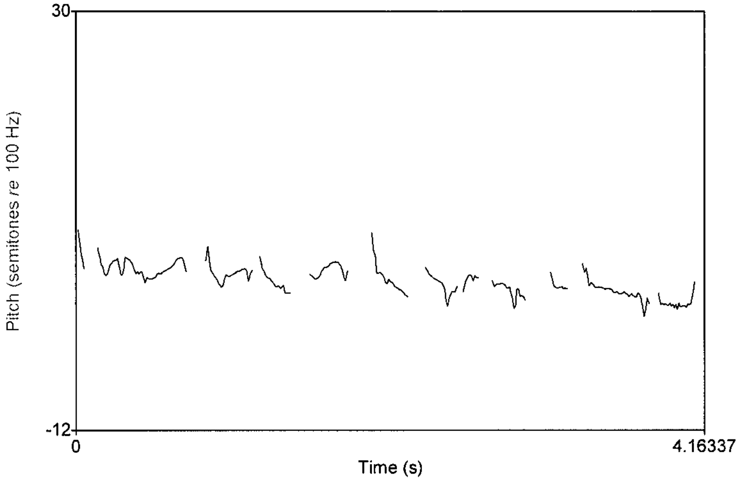


Figure 5.28. Pitch movement across the sentence "A tiger and a mouse were walking in a field when they saw a big lump of cheese lying on the ground." read by a German learner of English.

Acoustic measurements of pitch range have shown that learners of English, on average, produce a narrower pitch range than native speakers do. Conversely, when English native speakers speak German, they have a wider pitch range. Figure 5.28 illustrates this with the pitch contour of the sentence "A tiger and a mouse were walking in a field when they saw a big lump of cheese lying on the ground." read by a German learner of English who was aged 24 at the time of recording and had just spent six months studying at a British university. The pitch range of his utterance is displayed in semitones. Measured in Hz, the highest point of the pitch curve at the beginning of the utterance is 153 Hz, the lowest is 102.6 Hz.

Compare this to Figure 5.29, where a British English speaker read the same sentence. His pitch ranges from 286.6 Hz at the highest point at the beginning of the utterance to 85 Hz at the end of the utterance.

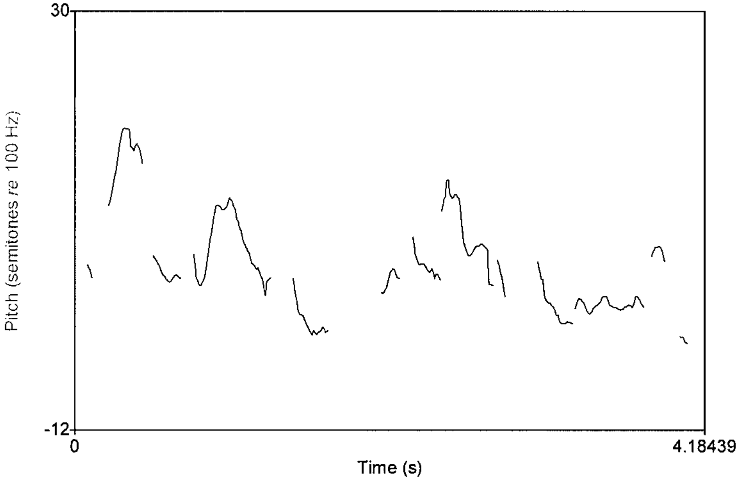


Figure 5.29. Pitch movement across the sentence "A tiger and a mouse were walking in a field when they saw a big lump of cheese lying on the ground." read by a British English speaker.

The methods and findings of acoustic phonetics have been used variously in pronunciation teaching and self-study courses (e.g. James 1977, Herry & Hirst 2002, Gut 2006, 2007a, b). In an early study, James (1977) found the visualization of the intonation curve of utterances to be effective in the acquisition of French. Similarly, Herry & Hirst (2002) devised a computer-assisted prosody learning tool based on simultaneous visualisation and audition for students of English, which was used successfully at a French university. Acoustic measurements can moreover be used in order to explore typical learner errors in the pronunciation of individual speech sounds. For example, students can compare the vowel qualities of /æ/ and /e/ in words such as *have* and *said* in both their own and native speech. This activity can be based on an auditory analysis combined with an acoustic measurement of the formant structure of these vowels. As shown in section 5.2, the first two formants, F1 and F2, are related to tongue position during articulation. By identifying the relevant vowels in a recording and measuring the formants as shown in Figure 5.30, students can explore differences between their own and a native speaker's speech.

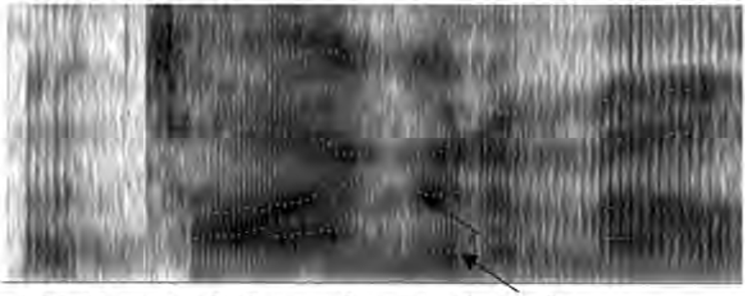


Figure 5.30. Measuring the first two formants of a vowel.

By the same token, students can measure the pitch range of different speakers, as well as the length of pauses and the acoustic properties of pitch, loudness and length as they are employed by individual speakers in order to achieve accentuation. This type of occupation with the properties of sounds is assumed to increase language awareness of pronunciation problems in general.

5.7 How to make a good speech recording

In order to carry out reliable measurements of the acoustic properties of speech, it is necessary to obtain high-quality speech recordings. The basic prerequisite for a good speech recording is a suitable recording environment. If you do not have access to a sound-treated room that excludes all external noises, you need to pick a location with minimal background noise. Common sources of background noise contaminating a recording include electronic equipment such as computers, fluorescent lighting, nearby corridors, bathrooms, lifts, roads, parks and playgrounds. In general, it is important to remember that a microphone will pick up many sounds that human perception filters out or reduces automatically, for example the speaker's shoe hitting the table leg or some background noise from outside. It is therefore always advisable to listen to the recording while making the recording and to make test runs before the real recording.

Both the recording device and the microphone you use need to be sensitive enough to capture all acoustic cues that are perceptually important. The choice of a suitable microphone – there are full-size microphones mounted on a table, clip-on microphones and head-mounted microphones – depends on whether the speakers are required to move around the room or can be placed on a chair behind a table. Since sound waves travelling through air lose energy (it takes energy to move the molecules in air), the microphone should always be placed close to the speakers mouth. However, in order to avoid disturbances of the recording by the blowing noises produced by the speaker's breathing or the aspiration of plosives, the microphone should be either attached to the speaker's blouse or be kept to the side of the speaker's mouth as in a headset.

When recording speech you have two basic options: either using **analogue** devices such as cassette tapes or recording **digitally** onto minidisks and the like or directly onto a computer. Analogue recording devices convert the continuous air pressure variations into equally continuous electrical signals. In order to be analysed on a computer, they have to be converted into digital signals, which entails some loss of information. In digital recordings, the continuous speech signal is converted into discrete points in time and stored as numbers. This means that the computer or digital recording devices represent the speech signal as a sequence of **samples**. They check the signal at regular intervals of time and store the properties of the signal at these points in time. When recording with a digital device you therefore have to decide on the **sampling rate** – how many points in time will be represented per second in your recording. Since speech events can be extremely small in duration (compare section 5.5.3), you need to choose a sampling rate that does not risk losing important information of the speech signal. All perceptible properties of speech are definitely recorded with a sampling rate of 48 kHz, but since recordings at this sampling rate require a lot of storage on your computer, many linguists record their data with a sampling rate of 22 kHz. This sampling rate still guarantees that all important acoustic aspects of speech are captured.

Another decision that has to be made when recording speech is that of **quantization**. This term refers to the accuracy of representation of the different amplitudes in the speech signal, i.e. the number of separate amplitude levels that are represented on a computer. They are stored in binary digits (bits) so that quantization is expressed in **bits**. Most researchers encode speech signals with 12 or 16 bits. When the volume of the speech that is being recorded goes beyond the range that can be represented, clipping occurs, which makes the recording useless for any phonetic analysis. You can see **clipping** in the waveform displayed in Figure 5.31: it is in one part cut off at the upper and lower ranges of amplitude. Most recording devices indicate whether the volume is within the critical range with a bar in the colour green, which switches to yellow and finally to red when the range is exceeded.

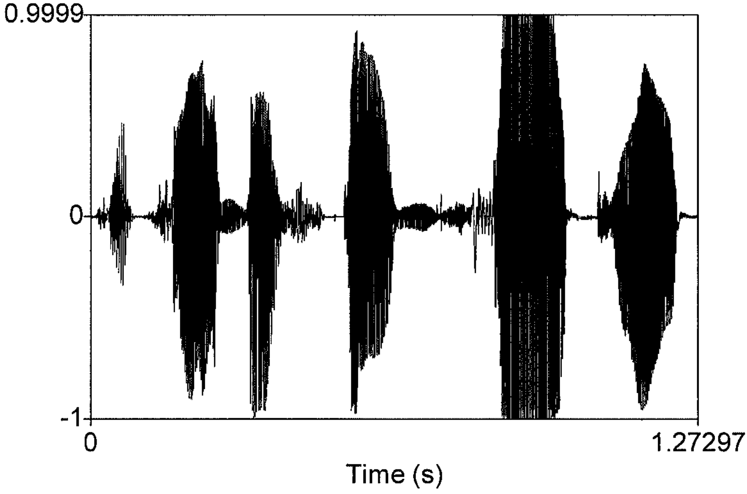
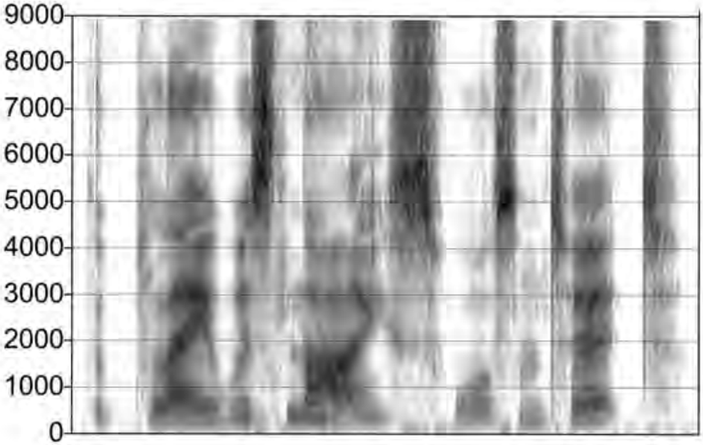


Figure 5.31. Speech recording with clipping due to too high recording volume.

5.8 Exercises

1. What are the acoustic properties of sound?
2. Which features are displayed in a waveform, in a spectrogram and in a pitch track?
3. What are the differences between vowels and sonorants on the one hand and voiceless consonants on the other, in acoustic terms?
4. What are the characteristic properties of vowels that are displayed in a spectrogram?
5. What does aspiration look like in a spectrogram?
6. Why does the pitch track of an utterance have gaps?
7. How is pitch range measured?
8. What are the acoustic correlates of accentuation in English?
9. Which aspects of English pronunciation can be taught with the help of acoustic phonetics?

10. Why is it so difficult to segment speech into individual speech sounds in a recording?
11. Record yourself saying "A tiger and a mouse were walking in a field when they saw a big lump of cheese lying on the ground" and compare your pitch range to that displayed in Figure 5.29.
12. Identify the vowels and fricatives in the following spectrogram. Can you guess what the utterance is?



5.9 Further Reading

Ladefoged (2001b, chapter 8), Yavaş (2006, chapter 5) and Davenport & Hannahs (2005, chapter 5) offer good, basic introductions to acoustic measurements of vowels and consonants. Clark, Yallop & Fletcher (2007, section 7.19), Kent & Read (2002, chapter 7) and Ashby & Maidment (2005, chapter 10) give short introductions to the acoustic properties of intonation and accentuation.

Students interested in the finer details of the acoustic analysis of the different speech sounds are recommended to read Kent & Read (2002, chapters 3 to 6) and Clark, Yallop & Fletcher (2007, chapter 7). For advanced readers, acoustic theories of speech production are explained in Johnson (1997, chapters 5 to 9). Details on the source-filter theory of speech production can be found in the classic texts by Fant (1960) and Stevens & House (1961).

Several software tools for sound visualisation, acoustic measurements and speech manipulation are available for free on the Web, e.g. the software *Wavesurfer* at <http://www.speech.kth.se/wavesurfer/> and the software *Praat* at <http://uvafon.hum.uva.nl/praat/>. A *Praat* manual describing how to record yourself and some basic steps in measuring acoustic aspects of speech is available on the CD-ROM (Praat_Manual.pdf).

6 Speech perception

Our ability to speak depends crucially on our ability to hear. Babies who are born deaf do not acquire speech (they use sign language for communication instead), although they have fully functioning speech organs. Not being able to hear, they cannot learn about the relevant units (such as phonemes, syllables and words) and rules of speech and cannot form mental representations (i.e. knowledge stored in their memory) of them. Yet, as chapters 3 and 4 explained, without mental representations of the units and rules of speech, the production of speech is not possible. This chapter describes the anatomy (structure) and physiology (function) of the organs that are involved in the perception of speech, although it has to be stressed that even today our knowledge about some of the fundamental processes of hearing is still incomplete. Like the speech organs used for articulation, the organs of speech perception are divided into 'systems', namely the **peripheral auditory system** – the ear – and the **internal auditory system** – the relevant parts of the brain. The three components of the peripheral auditory system, namely the outer, the middle and the inner ear, are described in sections 6.1 to 6.3. Section 6.4 illustrates the function of the internal auditory system. The perception of the acoustic cues of the speech signal and the methods of measuring these are described in sections 6.5 and 6.6. Section 6.7 presents theories of the perception of sounds and words, and section 6.8 describes what we know about the acquisition of perceptual abilities in first and second language acquisition.

6.1 The outer ear

Figure 6.1 illustrates the three parts of the human **peripheral auditory system**, which is divided into the outer, the middle and the inner ear. The **outer ear** consists of the **pinna**, the only visible part of the human auditory system, and the ear canal, or meatus. The pinna's main function is to protect the entrance of the ear canal from dirt and potentially harmful objects. The fact that humans have two ears, one on either side of the head, enables them to localize sounds, i.e. to tell where a sound is coming from. It appears that the pinnae contribute to this ability, especially for the higher frequencies of sounds.

The **ear canal** is a tube with a length of between 25 and 50 mm, which serves as the pathway for acoustic signals to reach the middle ear. The ear canal has two functions. On the one hand, it protects the complex musculature of the middle ear; on the other, it acts as a resonator for the incoming sound waves. The ear canal provides resonance, which means that it enhances the properties of a great range of sounds, especially those between 500 and 4000 Hz. This range includes all the major cues to linguistically relevant sounds, as shown in sections 5.2 and 5.3. The ear canal further appears to be particularly sensitive to frequencies between 2000 and 4000 Hz, which means that it plays an important role in the perception of fricatives.

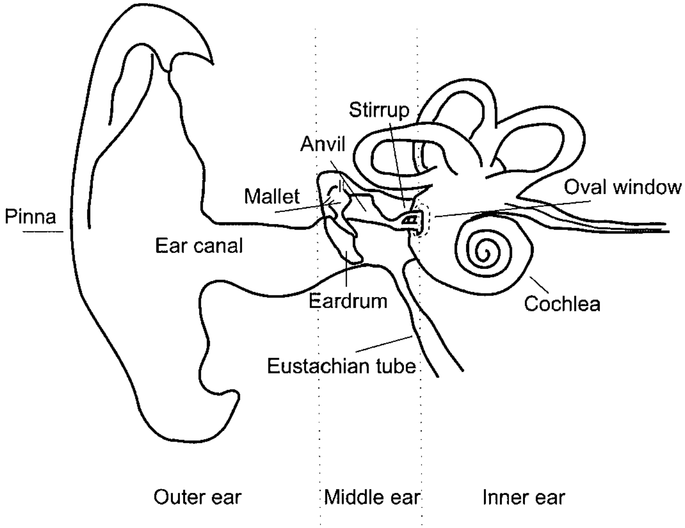


Figure 6.1. The human peripheral auditory system. Adapted from Clark, Yallop & Fletcher (2007: 299).

6.2 The middle ear

The **middle ear** is a small air-filled cavity within the skull that comprises the eardrum, a set of three interconnected bones – called the auditory ossicles – and their associated muscles (see Figure 6.1). The ear canal ends in the **eardrum**, which consists of a membrane that provides a seal between the outer and the middle ear. It is connected with the first of the **auditory ossicles**, the mallet, which in turn is attached to the second ossicle, the anvil. The anvil itself is attached to the stirrup (incidentally the smallest bone we have in our body with about 3.3 mm of length), which in turn is connected to the **oval window**, a small gap in the skull's bone structure that constitutes the interface to the inner ear.

One of the functions of the middle ear is that of passing on the incoming sounds to the inner ear. It does so by transforming the sound pressure variations in the air into equivalent mechanical movements. When sound pressure variations reach the eardrum via the ear canal, the eardrum is deflected and sets the mallet into vibration. The mallet sets the anvil into motion, which in turn causes the stirrup to vibrate.

Apart from the transmission of the speech signal, the middle ear also has the functions of amplification and regulation of the incoming sound level. Amplification of the incoming sound is based on two facts. First, the eardrum has a much larger surface area than the oval window, and the same pressure applied to a smaller area is greater than when applied to a bigger area. Second, the three auditory ossicles act like a mechanical lever system because the mallet is longer than the anvil. Due to this, the movement of the ossicles produces a greater force than that which hits the eardrum. Taken together, these factors cause the acoustic energy at the oval window to be about 35 times greater than that at the eardrum. Especially the acoustic energy between 500 and 4000 Hz is thus maximized during the transmission of the speech signal to the inner ear. The musculature of the middle ear further regulates the sound level, for example by protecting the inner ear against damage with a reflex called the **acoustic reflex** mechanism. When excessively loud sounds reach the middle ear, the musculature contracts and adjusts the ossicles so that the efficiency of sound transmission to the inner ear is reduced. Similarly, the ossicles contract just before speaking. It is assumed that this has the function of reducing for speakers the perceived sound volume of their own voice.

Figure 6.1 illustrates that the **Eustachian tube** links the middle ear and the oral cavity and thus provides an air pathway that can be opened in order to equalize pressure differences between the outer and the middle ear. Sometimes, after going up or down rapidly, for example in an airplane or lift, or when entering or leaving a tunnel, the changes in air pressure are felt painfully in the tension of the eardrum. By swallowing you open the Eustachian tube and thus allow the excess pressure to be released from the middle ear.

6.3 The inner ear

The **inner ear** is located within the skull and has a complex structure of which the cochlea is the most important organ for speech perception. Figure 6.1 illustrates that the **cochlea** forms part of the labyrinth, which also contains the vestibular system that is essential for our sense of balance. The cochlea is a snail-like structure, a conical chamber of bone that is rolled up about 2.5 times. On the one end, the base, it connects with the stirrup at the oval window; the other end is called the apex. Looking inside the cochlea, one can see that it consists of three tube-like canals filled with various fluids and lying on top of each other. Figure 6.2 is an illustration of a cross-section through the cochlea. It shows that the three canals are divided by two membranes, the vestibular membrane and the basilar membrane. The **basilar membrane** is of central importance for the perception of speech. It contains the **organ of Corti**, which has little hair cells attached to it that are connected to the auditory nerve. It is important to note that the basilar membrane does not have the same thickness in all places. It is stiffer and thinner at the end with the oval window and less stiff and wider at the inner end, the apex.

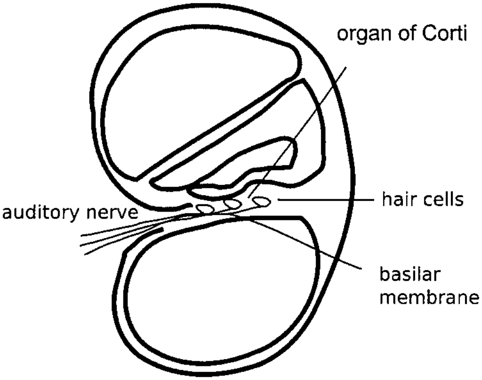


Figure 6.2. Cross-section through the cochlea.

The function of the inner ear is to transform the mechanical movements transmitted by the auditory ossicles into neural signals that can be passed on to the brain. This happens in the following way: first, the movements occurring at the oval window (the flexible membrane connected to the stirrup) are transmitted through the cochlear fluid and cause movement along the basilar membrane. This motion has been described as a travelling wave. The amplitude of this wave is greatest at the point where the frequency of the incoming sound matches the frequency of the movement of the basilar membrane. This in turn depends on the stiffness and width of the basilar membrane, which varies as described above. Thus, different frequencies of the incoming sound affect different areas of the basilar membrane – this is called tonographic organization. High frequencies result in movement near the oval window, whereas low frequencies produce movement at the curled-up inner end. Figure 6.3 shows the approximate locations of sensitivity to specific frequencies along the basilar membrane. Basically, the basilar membrane performs a frequency analysis of incoming sounds like the one performed by the Fourier analysis for the creation of a spectrogram (see section 5.1.2).

The movement of the basilar membrane causes movement in the hair cells of the organ of Corti, which transforms them into neural signals. As different frequencies affect the basilar membrane in different places, the different frequencies of an incoming speech signal can be transmitted individually to the brain via the auditory nerve, as will be explained below in section 6.4. A very strong movement of the cochlear fluid due to very loud noise may cause some hair cells to die. This is a common reason for partial **hearing loss**, which is why users of heavy machinery or producers and listeners of loud music should wear earplugs or earmuffs.

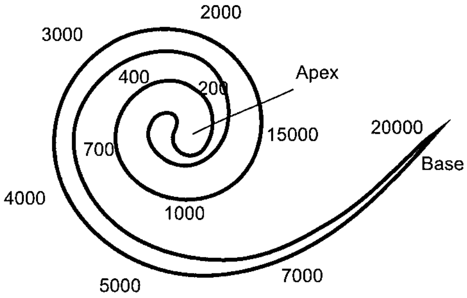


Figure 6.3. Frequency resolution on the basilar membrane.

Like the outer and the middle ear, the inner ear does not simply transmit properties of sounds, but contributes to shaping the incoming sound. Active vibrations of the hair cells of the organ of Corti pre-amplify the sound and thus contribute to the frequency resolution, the detection of the different frequencies of an incoming sound. In addition, neural impulses from the brain can make the hair cells move the basilar membrane and thus create sound, a phenomenon called otoacoustic emission, which can be recorded with a microphone in the ear canal. It is thought that this process supports the ability of the basilar membrane to detect differences in the acoustic properties of incoming sounds.

6.4 The internal auditory system

The internal auditory system consists of the auditory nerve and the auditory cortex. The latter forms the essential organ of hearing, since perception does not happen in the ear but only when the auditory sensations have reached and been processed by the relevant area in the brain. Information from the inner ear is passed on to the brain in the following way: the organ of Corti on the basilar membrane converts the motion of the basilar membrane and the cochlear fluids into electrical signals that are communicated via neurotransmitters to thousands of nerve cells (approximately 28.000 in each organ of Corti). The nerve cells themselves create action potentials that travel along the **auditory nerve** to the relevant areas in the brain for further processing. Each fibre of the auditory nerve is connected to a small area of the basilar membrane or even to a single hair cell alone. Since specific areas on the basilar membrane react to specific frequencies only (see section 6.3 above), each fibre of the auditory nerve is optimally stimulated by a specific frequency. The ability to hear certain sounds thus depends on the frequency resolution of the fibres. Many sounds are too high (have a too high frequency) for human ears. Age-related hearing loss, which apparently begins as early as in the twenties, is due to the loss of transmission of high frequencies by the relevant fibres. Damage of the inner ear (for example through excessive noise) leads to an increased insensitivity of the fibres, and their 'tuning' to certain frequencies is lost. This means that acoustically complex signals like speech (see sections 5.2 and 5.3) cannot be analysed precisely anymore, and speech perception is impaired.

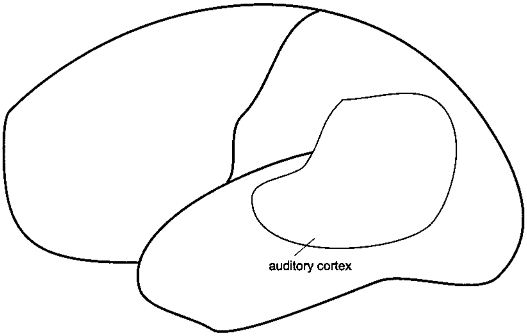


Figure 6.4. The auditory cortex.

The auditory nerve does not only transmit the frequency of a sound to the brain but also passes on information about the duration of the speech signal and its intensity. The area of the brain that is active in speech perception is the **auditory cortex**, which can be further divided into the primary, secondary and tertiary auditory cortex. It is located on the temporal lobe, as illustrated in Figure 6.4.

This area is activated during the perception of speech and its decoding into words and utterances. Like for speech production (see section 2.6), the left hemisphere of the brain is crucially involved in the perception of speech for right-handed people (for some left-handed people this does not apply): more activity can be measured in the left hemisphere than in the right hemisphere when listeners hear speech. The left hemisphere appears to be responsible for decoding the meaning of a linguistic message, whereas the right hemisphere seems to be more involved in the perception of music and the prosodic characteristics (especially pitch) of language. Investigations of patients who suffered damage to the left hemisphere show that they are likely to have difficulties with understanding simple words and with selecting an object that was named. Patients with damage in the right hemisphere, conversely, often lose their ability to understand metaphoric speech or idioms. Instead of understanding the idiomatic meaning of an expression such as "He kicked the bucket", they will interpret this utterance literally. A study by Shaywitz et al. (1995) suggests gender-specific language processing: women tend to process phonologically relevant information with both hemispheres, whereas men use only one.

At present, little is known about the exact location or function of the parts of the brain relevant for sound discrimination and sound identification, and most of our current knowledge is based on animal experiments. It is assumed that the various neurons of the auditory cortex are organised according to the frequency of sound to which they respond best. Furthermore, the area of the brain called the primary auditory cortex seems to be responsible for the differentiation of voiced and voiceless sounds, whereas changes in pitch are perceived in the secondary auditory cortex. Tone, the linguistic use of pitch for the differentiation of meaning (see section 4.3), appears to be perceived also in places in the cortex other than the auditory cortex.

Both the peripheral and the internal auditory system are able to **adapt** to characteristics of the incoming speech signal and thus contribute to the differentiation of perception. For example, the hearing threshold can be altered so that the subjective difference between the loudness of two sounds is increased. This means that even in noisy conditions humans can focus their attention on a single sound source. This ability was first described by Cherry (1953), who termed it the Cocktail party effect. It comprises two different phenomena: first, in a mixture of conversations, listeners are able to concentrate on a single speaker. Second, even when concentrating on one conversation, when someone mentions your name in another conversation you will notice this. Both abilities are based on active noise suppression by the auditory system: the sound source one is concentrating on seems up to three times louder than the other ambient noises.

6.5 The perception of loudness, pitch and voice quality

It is one of the characteristics of human speech perception that changes in the acoustic properties of speech sounds do not correspond linearly to changes in the sensations experienced by a listener. This holds true for the perception of pitch and its relation to the frequency of tones. It further applies to the perception of loudness, where a mismatch between objectively measured sound pressure and subjectively perceived loudness exists, as well as for the perception of voice quality, or timbre. Studies concerned with the relationship between acoustic properties and human speech perception are called **psychoacoustic studies**.

Humans can perceive a wide range of different frequencies (see section 5.1.1 for a description of the physical property frequency) ranging from 20 to about 20,000 Hz (with age, the ability to perceive the higher frequencies drops). Frequencies lower than 20 Hz do not activate the hair cells of the organ of Corti and are thus below the **threshold of hearing**. Psychoacoustic experiments have shown that changes in absolute frequency are not perceived equally well for all frequency regions. Humans are by far most sensitive to changes in frequency below 1000 Hz. Above 1000 Hz the ability to notice small changes in absolute frequency decreases progressively. The experiments show that a pitch change that corresponds to a change from 100 to 500 Hz is not perceptually equivalent to a pitch change experienced when the frequency of a tone is changed from 6100 to 6500 Hz. This characteristic of human hearing is caused by the structure of the basilar membrane. Figure 6.3 above illustrates that a large part of the thicker end of the basilar membrane responds to frequencies below 1000 Hz, whereas only a small part responds to frequencies of above 12,000 Hz. This means that changes in the lower frequency range are more easily detected than changes in the higher frequency range.

In order to depict the relationship between subjective perception of pitch and absolute frequencies a scale was developed, the **Mel scale**. Figure 6.5 illustrates how the perceptual difference between two frequencies depends on the absolute magnitude of the frequencies. A sound wave with a frequency of 1000 Hz has the Mel value of 1000. It is perceived as twice as high as a sound wave with 500 Mel. This sound wave, however, has a frequency of 400 Hz. It can be seen that below 1000 Hz there is a fairly direct relationship between frequency and perceived pitch, but above this point the relationship becomes logarithmic. The units of a logarithmic scale represent powers of ten, i.e. the difference between 2 and 3 corresponds to the difference between 10^2 and 10^3.

The findings of psychoacoustic studies on the perception of frequency are not of direct relevance to the perception of sounds since they use simple sinusoidal waves with one frequency only. Many speech sounds, however, are complex and consist of a combination of different sinusoidal waves with different frequencies (see section 5.1.2). The lowest of them, the fundamental frequency, is perceived as **pitch**, whereas the others contribute to the timbre of a sound. The perception of the **timbre** of a sound is independent of its loudness and pitch. It is the energy distribution of the different frequencies of a sound that determines the perception of its timbre. Thus, a sound of the same loudness and same pitch produced by a flute or a human voice yields a very different sensation for listeners. Experiments have shown that even when the fundamental frequency is removed from a sound it can still be perceived by listeners. It seems that listeners decode the composition of the sound (remember that there is a linear relationship between the fundamental frequency and the harmonics, see section 5.1.2) and derive the fundamental frequency from it.

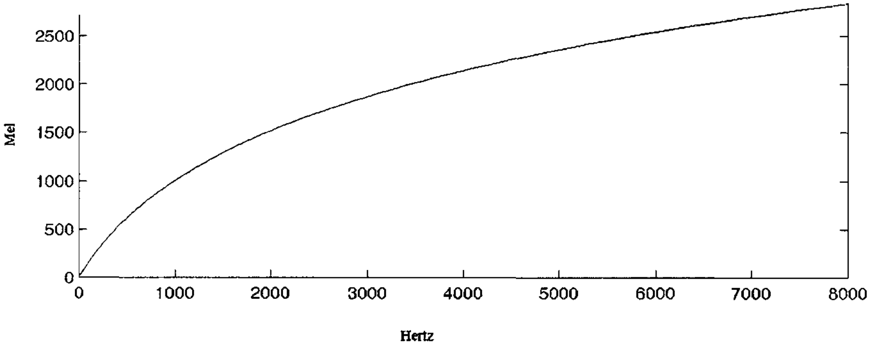


Figure 6.5. The Mel scale.

Psychoacoustic experiments concerned with the sensation of loudness show that humans can perceive a remarkable range of sound intensities. The hearing threshold lies at about 20 microPascals (μPa), whereas the highest perceptible intensity is about a million times greater. Like with the perception of pitch, there is no linear relationship between changes in sound pressure and changes in the sensation of loudness. In a typical experiment on the relationship between sound pressure and the perception of loudness, listeners are asked to adjust the loudness of one sound so that they perceive it as twice as loud or as half as loud as another. It has been found that the relationship between sound intensity and perceived loudness is far from linear. A doubling of the intensity of a sound is not perceived as a doubling of the sensation of loudness. In fact, the correspondence between intensity and perceived loudness is nearly logarithmic.

With the help of these psychoacoustic experiments, a scale of subjective loudness can be drawn, similar to the Mel scale for the perception of pitch. Figure 6.6 illustrates the relationship between the subjectively perceived loudness, measured in the unit **sone**, and objective sound pressure level, measured in microPascal (μPa). It can be seen that for quiet sounds relatively small changes in sound pressure lead to large changes in perceived loudness – the left end of the curve rises steeply. Conversely, for loud sounds, changes in loudness are only perceived when relatively large changes in sound pressure occur.

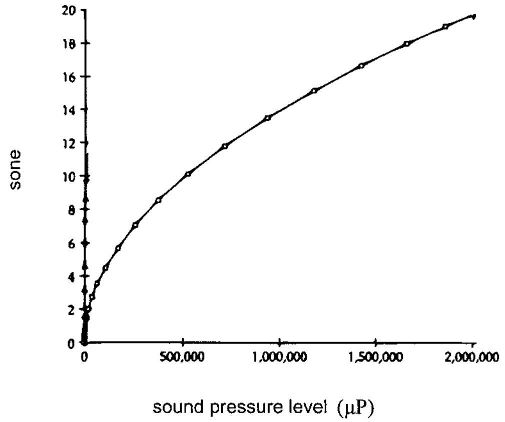


Figure 6.6. Sound pressure level in relation to sone.

Table 6.1. Sound pressure levels and corresponding dB values for different sensations.

Sensation	Sound pressure (μPa)	dB
threshold of hearing	20	0
whisper, quiet garden	200	20
very quiet speech, tearing of paper	2,000	40
normal conversation	20,000	60
car passing, vacuum cleaner	100,000	70
noise inside a jet, loud music	200,000	80
brass band, train	2,000,000	100
thunder, threshold of pain	20,000,000	120
ear damage	200,000,000	140

The other measurement of relative loudness used to describe human perception of loudness is **decibel** (dB). Its relationship to sound pressure is logarithmic and thus represents human perception very well. Table 6.1 gives the dB levels and corresponding sound pressure levels of some sensations. It is important to remember that sound intensity expressed in dB, since it is a logarithmic measurement, is always relative to some reference point. This reference point is usually the threshold of hearing.

Listening experiments have further shown that perceived loudness varies as a function of frequency. The human auditory system is not equally sensitive to all frequencies. In psychoacoustic studies, listeners are asked to adjust tones of different frequencies so that they are perceived as equally loud. Very low frequencies and very high frequencies need to be amplified so that a human perceives them as equally loud as frequencies in the middle range. Figure 6.7 illustrates the hearing threshold, i.e. the lowest sound pressure level necessary to perceive a sound for different frequencies. Clearly, the human auditory system is most sensitive to frequencies between 2000 and 5000 Hz. In this frequency region, only a very low sound pressure level is necessary to cross the hearing threshold. Very low frequencies require the highest sound pressure levels in order to be perceptible to humans.

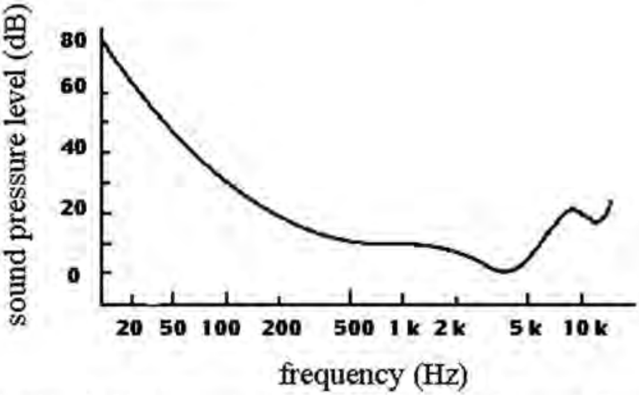


Figure 6.7. Threshold of hearing – the relationship between frequency and intensity in human perception (1 k = 1000 Hz).

6.6 Measuring hearing sensitivity

The method concerned with measuring hearing sensitivity is called **audiometry**. It is used both for the detection of hearing impairment and hearing loss as well as the examination of normal hearing. Hearing impairment or a complete hearing loss can have different causes. An infection of the middle ear can lead to so-called conductive hearing loss, in which case the sound pressure waves are not passed on through the ossicles in the middle ear. Other illnesses such as measles or meningitis can cause damage of the cochlea or the auditory nerve. A sudden loud noise or continued exposure to high noise levels (such as machines or music) can cause permanent damage to the hair cells on the basilar membrane, which then cannot transmit sounds effectively anymore. For the testing of hearing sensitivity, simple sinusoidal waves, so-called **pure tones**, with different frequencies (e.g. 250 Hz, 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz) are played successively to each ear, first at a very low intensity and then increasingly louder until the listener indicates that she or he can perceive it. As shown in Figure 6.7, the threshold of normal hearing for these frequencies is around –10 dB to +20 dB. The extent of hearing loss is usually indicated using the unit dB. A hearing loss of 50 dB, for example, means that a particular frequency must have an intensity level of 50 dB before it can be perceived. When the hearing threshold for each frequency is established, an audiogram can be made for each ear. Figure 6.8 illustrates a pure tone **audiogram** for a normal hearing person.

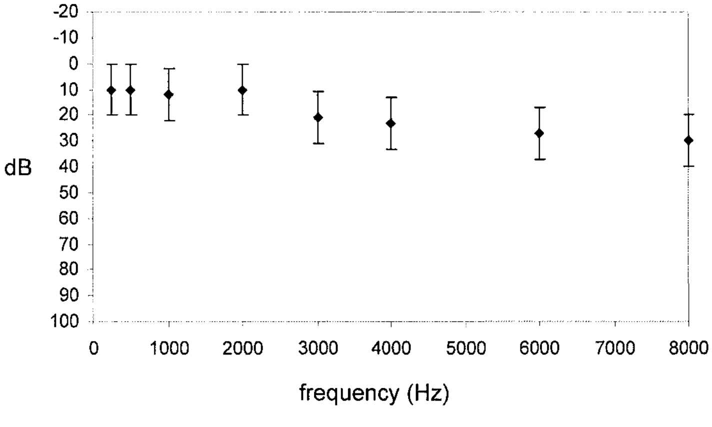


Figure 6.8. Average audiogram for a normal hearing person. (The bars indicate the error interval of the measurement.)

Several methods exist to treat hearing impairment and hearing loss. **Hearing aids** worn behind the pinna contain small microphones that amplify the incoming sounds. The amplification of specific frequencies is regulated for each individual patient depending on which frequencies are affected by hearing loss.

The weakness of hearing aids compared to normal hearing lies in the fact that they cannot adapt like the auditory system. As described above in section 6.4, these adaptive abilities of the internal auditory system enable listeners to tune into a particular sound source and to fade out others. A hearing aid, conversely, amplifies every sound equally much and does not perform this kind of focussing. Babies born without hearing often receive a cochlear implant that stimulates the auditory nerve with electrical impulses. This implant is placed under the skin behind the pinna. An additional small device containing a microphone, a speech processor and a transmitter is worn behind the pinna, filters sounds and transmits them to the cochlear implant. An implant does not create normal hearing but, together with speech therapy, enables children to learn to speak. Yet, several studies have shown that the acoustic properties of speech children with cochlear implants produce vary systematically from those of speech produced by normal-hearing children (e.g. Mildner & Liker 2003).

6.7 Theories of speech perception (advanced reading)

Section 6.5 described how the various parameters of speech, in particular pitch, loudness and timbre, are perceived by humans. In fact, psychoacoustic experiments have shown that the auditory system enables speakers to identify differences in frequency and intensity as well as in time (duration) to a much higher degree than is necessary for the discrimination and identification of speech sounds. They have also shown that these perceptual abilities – which are referred to as **psychoacoustic** abilities – of humans are a necessary prerequisite for the understanding of speech but do not suffice to explain how speech perception works. **Speech perception** involves more than the perception of frequencies, intensities and duration: it requires the linguistic interpretation of the neural signal transmitted to the auditory cortex. In other words, speech perception describes how a listener decodes the acoustic signal as a meaningful message instead of as a sequence of noises with different frequencies and loudness. Speech perception is by no means a trivial task. Section 2.5 showed that the same sound and the same word can have very different acoustic realizations when produced by speakers with different anatomies of the laryngeal and the vocal tract, when produced at a different speaking rate or in different phonetic contexts. How are these variable productions of a sound, word or utterance identified as the same by listeners? Section 2.5 further showed that there are no boundaries between successive speech sounds or words in a breath group. How do listeners extract meaningful units from this continuous acoustic stream?

These are the questions all of the models and theories of speech perception are concerned with. It is still not possible to examine the activity of the brain in such detail that it would allow the identification of areas or processes underlying speech perception. We still cannot measure exactly how and where in the brain an auditory signal is converted into a linguistic message that can be interpreted according to its grammatical and semantic content. Therefore, we have to rely on theories and models to explain the various questions concerning speech perception: How are individual speech sounds identified? How can listeners cope with the variability of the speech signal? How is an auditory signal decoded and interpreted as a speech sound, a word or a sequence of words? Which mental representations of speech do speakers have? Which linguistic units are stored? How are they stored?

Several experiments have been concerned with the question of how individual speech sounds are discriminated, i.e. how listeners can tell that they are the same or different. It has been found that humans are very good at identifying small differences between two speech sounds, an ability which has been termed **categorical perception**. Categorical perception means that gradual differences between acoustic properties of sounds are not perceived in a linear fashion, but that listeners rather group sounds into categories. This can be exemplified with a typical experimental set-up: the two plosives [p] and [b], for example, differ in voicing. This means that during the production of the voiced plosive [b], the vocal folds usually vibrate, whereas during the production of a [p] they do not. Yet, for all [b]s produced by different speakers the exact timing of the vocal fold vibration can vary – it might already have started before the lips close to obstruct the airstream, it might coincide exactly with the closure of the lips, start a short while after the closure, start with the opening of the lips or any time after it. In the production of a [p] vocal fold vibration might start immediately after the opening of the lips or several milliseconds later when the following sound is already being articulated. Many variations of this relative voice onset time (VOT, see section 5.3 for a detailed description) occur in plosives and determine whether they are perceived as voiced or voiceless. In English, a typical [b] has a VOT of about -20 to +20 ms, i.e. on average voicing begins between 20 ms before and 20 ms after the opening of the lips. Typical realisations of [p], by contrast, have a VOT of between +40 and +80 ms, which means that 40 to 80 ms pass after the lip opening and before the beginning of voicing.

In an experiment on categorical perception, a set of synthesised, artificially created sound examples consisting of a plosive and a following vowel (e.g. [ba, pa]) are prepared that vary in VOT in a systematic way. Participants listen to these plosive-plus-vowel stimuli in a random order. For example, in one of them, the plosive has a VOT of -20 ms, the next might have a VOT of +15 ms, the next one a VOT of -10 ms and so forth. When listeners hear these stimuli and are asked to decide whether they perceive [pa] or [ba], they do not hesitate over those stimuli that form borderline cases between voiced and voiceless plosives (e.g. with a VOT of +20 ms). Rather, listeners shift abruptly from the perception of [b] to the perception of [p] despite the fact that the listening examples are evenly spaced on a continuum in terms of their VOT. *Within* the two phonetic categories of [b] and [p], listeners do not differentiate between sounds with a slightly longer or shorter VOT; discrimination only occurs *between* the categories.

Categorical perception enables listeners to differentiate between different speech sounds. But how do they identify these speech sounds? Listeners cannot only tell that a [b] is different from a [p], but are also able to recognize a [b] as a /b/ and a [p] as a /p/. For this recognition and identification of speech sounds, speakers must have mental representations of speech sounds that they can compare to the incoming speech signal. Chapters 3 and 4 explained that phonological theories assume that speakers have abstract mental representations of many different units of speech. The smallest unit that is claimed to be represented mentally is that of speech sounds, the phonemes. What does such a representation of a speech sound look like? This representation must be able to deal with the variability of the sounds speakers produce: depending on the particular shape of a speaker's speech organs, on the neighbouring sounds and on the speaking rate, a simple sound like [i] can have very different acoustic properties (see sections 2.5 and 5.2). Still, listeners can identify the different physical signals as the same sound.

Some authors argue that this is due to fact that there are invariant acoustic correlates of speech sounds, which enable listeners to recognize different sounds (e.g. Stevens & Blumstein 1981). These invariant cues lie on the level of the phonetic features and are thus smaller units than phonemes. Sections 5.2 and 5.3 describe in detail which acoustic properties the individual speech sounds of English and classes of speech sounds in general have. For example, one acoustic cue that supports the identification of fricatives is the spectral structure of the 'noise' or friction created by the narrow constriction of the vocal tract during articulation. Depending on the place of articulation, this friction has different acoustic properties: in a typical [ʃ], for example, there is a peak at about 2500 Hz with additional energy between about 4000 and 8000 Hz, whereas a /s/ has most of the noise in the frequency range of 5000 to 8000 Hz (see section 5.3 for further details). The identification of voiced plosives such as /b/, /d/ and /g/ relies on the so-called formant transitions, i.e. the change in formants from a preceding consonant to a vowel or from a vowel to a following consonant (see section 5.2 for an explanation of what a formant is). The particular place of articulation can be inferred from the movement of the second formant, F2 (see section 5.3 for details). The different manners of articulation for consonants also have invariant acoustic cues. Nasals, for example, have a band of frequency with very high energy at about 250 to 300 Hz and little energy above that and below 2000 Hz (see section 5.3 for details). The perception of vowels is more complicated. Perception experiments have shown that vowels are distinguished from each other with the help of the relative position of the first three formants. Speakers are able to 'normalize' acoustic differences between vowels that occur in the speech of different speakers (for example, female, male and children's voices); they identify vowels not on the basis of absolute acoustic values but decode how the formants occur relative to each other (see section 5.2).

However, most theories of speech perception assume that the identification of individual speech sounds is not necessary for the understanding of words and utterances. There is no doubt that listeners can identify individual speech sounds and have mental representations of them, but it is very unlikely that they use this ability when they listen to real speech. There is some evidence that listeners use the unit of the syllable for speech perception. The motor theory of speech perception (Liberman et al. 1967, Liberman & Mattingly 1985), for example, claims that listeners decode a sequence of sounds in a syllable with reference to their articulatory knowledge. They are assumed to compare internally the incoming sound sequence to their mental representations of the muscle activities underlying the production of this sound sequence. Listeners have 'perceived' a syllable when a match to a stored motor sequence is made. Furthermore, studies on speech errors also claim that the syllable is a basic unit of representation for speech perception.

In summary, there is still little agreement concerning the size or type of basic perceptual units in speech perception at present. Is it mental representations of phonetic features, phonemes or rather syllables that enable speakers to interpret acoustic signals as meaningful language? Most probably, a combination of all units that have been suggested so far contribute to speech perception. Possibly depending on the task – whether we are listening to speech in noisy conditions, whether we are familiar with the language and whether the speaker is articulating carefully or speaking at very high speed – the different mental representations are employed for speech perception to a different degree.

Listeners are not only able to recognize and identify sounds and syllables, but they also understand words. Experiments have shown that words are recognized very quickly, in between 200 and 500 ms after the first sound is produced by the speaker (Marslen-Wilson 1987). It is assumed that spoken **word recognition** involves matching an acoustic signal to a representation in memory. All theories about the recognition of words share the assumption that this mental representation of words has the form of a mental lexicon. Apart from phonological information, the mental lexicon is considered to contain information about the spelling of each word and its syntactic, semantic and pragmatic properties. The mental lexicon is probably represented in a special area of memory located in the brain's left hemisphere and comprises approximately 75,000 words for an adult speaker of English with average education. Moreover, there is evidence that the mental lexicon does not only contain words, but also groups of word forms for each word (the various inflected forms) as well as formulaic sequences of words, such as phrases used very frequently in conversation (e.g. "how are you", "you know what I mean") and idioms.

Several competing theoretical models exist that describe the stages necessary for the recognition and identification of words. Generally, it is assumed that the acoustic information reaching the brain activates several hypotheses in the mental lexicon (e.g. McClelland & Elman 1986, Norris 1994). At this stage, acoustic variation due to coarticulation and rate of speech is 'normalized'. This, however, does not exclude that information about a speaker's voice is used in speech perception (cf. Pisoni & Lively 1995). In the second stage, these hypotheses 'compete' with each other so that finally all hypotheses except one are discarded. The remaining hypothesis is what is perceived by the listener. There is empirical evidence that this **lexical access** is faster for familiar words than for unfamiliar words. That is, words that occur very often in a language and for a particular speaker/listener are recognized faster and more accurately than words that speakers/listeners use less often.

Most theories on speech perception agree that understanding speech is not a passive process. It is not true that listeners simply process the incoming signal in a linear way, match the units one after another to a stored template and finally interpret them accordingly. This process, which has been called bottom-up (from ear to brain), cannot describe why listeners understand words even if they are mispronounced or when some sounds are distorted by noise. Much rather, listeners must also employ top-down processes (i.e. an active participation of mental representations) in the interpretation of speech. This can be demonstrated with many different examples. Experiments show that listeners are able to perceive a word even when one or two sounds in the sound sequence are replaced by noise (Warren 1970, Samuel 1981) or when speech is produced in very noisy conditions. Even when part of the signal is effectively covered, as when the [n] in *sand* is masked by a loud cough, the listener will perceive [sænd] and will not be able to say which of the sounds was obscured by the noise. Furthermore, listeners often perceive what was not said. A speaker might say "The women has come" but the listeners will still claim to have heard "The woman has come". These examples are taken to confirm the top-down nature of speech perception, i.e. the active role of the representations in the brain in speech perception. Probably, both bottom-up and top-down processes combine to ensure fast and robust speech perception.

Speech perception does not rely on acoustic cues only. McGurk & McDonald (1976) showed in an experiment how much speakers rely on visual cues in speech perception. They played a video of a face saying one sound sequence, e.g. [gaga], but combined it with the audio recording of another sound sequence, e.g. [baba]. Listeners not looking at the face heard [baba] and watchers of the video with the sound turned off identified the motion of the articulators as [gaga]. However, when simultaneously watching and hearing the non-matching sound and video, listeners perceived something like [dada], even those listeners that had heard or seen the video or the recording on their own before. This finding, which was named the **McGurk effect**, was interpreted in the following way: when the brain receives conflicting information from the auditory and the visual channel, it will settle on a middle ground, in this case [dada]. Many later studies have confirmed the effect; it is stable even with babies (Rosenblum, Schmuckler & Johnson 1997) or when the video screen is turned upside down (Campbell 1994) or with a combination of male face and female voice (Green, Kuhl, Meltzoff & Stevens 1991).

6.8 Speech perception and language acquisition (advanced reading)

6.8.1 Speech perception in first language acquisition

Newborn babies have a wide range of perceptual abilities: they can recognize their mother's voice, they can recognize their parents' language and they are able to differentiate a number of different speech sounds. These abilities result from the fact that a foetus in the womb can already hear during the last weeks of pregnancy. The middle ear physiology starts working at around 3 months before birth, and the foetus can perceive sounds – although these sounds have different acoustic characteristics, since they are not transmitted by air but by the fluids in the womb. Yet, these pre-natal sensations seem to be sufficient for babies to be born with first perceptual representations of speech.

Several methods have been developed to study infant speech perception. The **High-Amplitude Sucking technique** (HAS) is based on the observation that humans, on the one hand, become bored by repeated sensations and, on the other, react to changes they perceive. In HAS studies, infants suck on a dummy that records their sucking rate. When the infant's 'normal' sucking rate – the baseline – is established, a repeating sound sequence such as [ba ba ba] is played. An immediate increase in sucking rate shows that the infant has perceived this sound. After a while, the sucking rate will decrease again, which is interpreted as waning interest. Then a new sound sequence is introduced, for example [ga ga ga] spoken at the same rate and by the same voice. If the sucking rate increases, it is concluded that the infant is interested, i.e. has noticed the difference and therefore is able to differentiate between the two sounds [b] and [g]. Another method of testing infant speech perception is the head-turn method, for which sounds are presented from different sources. If infants turn their head towards the source of a new sound, it is concluded that they can perceive the difference.

With the help of these methods, it has been found that right after birth babies prefer speech recordings from their native language to those of other languages (Mehler et al. 1988). Infants furthermore can discriminate between nearly every acoustic contrast, including those that do not occur as functional contrasts in their language (see Werker & Pegg 1992 for a summary). Babies growing up in an English-speaking environment are able to discriminate voicing contrasts as for example between [ta] and [da], differences in place of articulation as in [ta] and [ka], and differences in manner of articulation such as between [ta] and [sa]. Similarly, they can discriminate vowels such as [i] and [a], but vowel discrimination seems to be – as for adults – continuous rather than categorical. Moreover, they can discriminate voiceless aspirated [t^{h}a] and breathy voiced dental stops [d^{h}a], a contrast that is not phonologically relevant in English but in Hindi.

While infants up to about 8 months of age are able to discriminate sounds that play no contrastive role in their native language/s, this ability is lost at around 9 months. From then on, babies can only discriminate sounds that function as phonemes (see section 3.1 for the definition of phoneme) in their native language/s. This is interpreted as the result of the influence of the ambient language/s – it probably stems from a reorganization of the initial auditory capabilities. It is assumed that a phonological system of the native language/s is constructed gradually over the first years of life. Yet, some phonemic contrasts remain difficult to discriminate for English-speaking children even at age three. These difficulties concern mainly the pairs /ɹ/ and /w/ and /θ/ and /f/. The construction of a mental lexicon with phonological representations of words proceeds slowly and continues well into the teens. At around 2.5 years of age, mental representations of some words seem to begin to be established. Babies, for example, look at a picture of a dog significantly longer when they hear [dɒg] than when [bɒg] is played to them. In general, perception seems to precede production. There are many anecdotes and studies that show that children have stable mental representations of the phonological shapes of words before they can pronounce them correctly themselves. Yet, even school-age children are still much slower to recognize words than adults.

6.8.2 Speech perception in second language acquisition and teaching

When comparing the acquisition of speech perception in first language (L1) and second language (L2) acquisition, it becomes clear that the two differ distinctly. One of the major differences lies in the fact that learners of a second language have already acquired perceptual categories for the sounds, words and other linguistic units of their first language. The course of development of perceptual representations in an L2 can be roughly sketched as follows: initially, when listening to the foreign language, second language learners do not understand anything. This is clearly due to the fact that they have no mental representations of speech sounds, syllables and words and their meaning to which the incoming acoustic signal could be matched. It is the task of language learners to acquire all phonological representations necessary for speech perception. This probably involves representations of phonetic features, phonemes, syllables and words as well as intonational units (section 6.7 showed that our knowledge in this area is still sketchy). Until these representations have reached a native-like form, language learners will always find it more difficult and take more time to understand L2 speech than native speakers do, especially in unfavourable conditions such as in noisy places and without visual cues (Werker et al. 1992 found that L2 learners also use visual information in speech perception). This can be explained by the fact that language learners have fewer top-down processes available (compare section 6.7).

Unfortunately, we still have very little empirical data on the acquisition of perception in an L2. As is the case with speech production, only few longitudinal studies exist that document the development of perceptual abilities across longer stretches of time. Usually, learners or learner groups are studied at one point in time only. These studies have shown that learners of a second language do not perceive speech in the L2 in the same way as native speakers of this language do. For example, Barry (1989) found that German learners of English confuse the vowels /e/ and /æ/. They rely on durational differences in the identification of these vowels, whereas American English listeners use formant differences for identification (Bohn & Flege 1990). L2 learners also perceive intonation and word stress differently than native speakers (e.g. Grabe et al. 2003, Archibald 1997). Various factors seem to influence how well L2 learners perceive sounds of the L2. For Italian immigrants to Canada, age at arrival and amount of continued use of the L1 influence perceptual abilities (Flege & MacKay 2004). Those who arrive at a later age and continue to speak more Italian have more difficulties in perceiving English speech sounds than speakers emigrating at a younger age and not using their L1 very often anymore. It seems, however, possible to acquire native-like perception – the ability to perceive new linguistic contrasts is not lost at any point in life. Training studies have shown that adult learners can acquire the differentiated perception of foreign sounds (see overview in Rvachew & Jamieson 1995). In Bohn & Flege's (1990) study, those German learners with a large amount of experience with English (more than five years of residence in the U.S.) perceived the vowels /ɛ/ and /æ/ similar to native listeners, whereas learners with shorter periods of residence did not. Differences between children's and adults' perceptual abilities appear not to be based on physiological differences, but rather on their mental representations of L1 sounds that make them pay attention to the particular acoustic cues relevant in the L1.

On the basis of these empirical findings, several theories of L2 perceptual acquisition have been formulated. Most of these theories focus on the unit of speech sounds (however, one of them – Archibald [1994] – is concerned with the acquisition of the perception and production of word stress). The theories disagree about whether the mental representations underlying the identification of speech sounds consist of phonemes (e.g. Flege 1995) or articulatory gestures (e.g. Best 1995, Dziubalska-Kołaczyk 1990), but they agree on the assumption that a language learner perceives the L2 sounds through categories of the phonological structure of the L1. These categories constrain which non-native sounds can be perceived correctly and, in turn, can be learned to produce correctly.

Not all L2 contrasts and categories are equally easy or difficult to perceive. The acquisition of L2 sounds that do not exist in the speaker's L1 and that are perceptually very different from L1 sounds is relatively easy: new perceptual categories are created quickly for these sounds. The greater the perceived phonetic dissimilarity between an L2 sound and L1 categories, the higher is the chance that a new mental category will be established. A German learner of Zulu, for example, will have little trouble noticing that the click sounds (see section 2.1) are very different from German speech sounds and will easily acquire a mental representation of them. Yet, this category for the L2 sound may still be different from the category of a native speaker.

Flege (1995) predicts that when no phonetic differences between two sounds are perceived by the learner – a process which is labelled equivalence classification – category formation for an L2 sound will be blocked and the learner will end up with the representation of a single phonetic category for both sounds. On the one hand, this happens when two L2 sounds are assimilated to the same L1 category (for example, some German learners perceive both /w/ and /v/ in English as /v/). On the other hand, two L2 sounds might be assimilated into two different L1 categories (/θ/ and /ð/ in English might be categorized as /s/ and /d/, respectively). Not until learners notice these wrong categorizations will they be able to establish new perceptual categories.

A theoretical model of L2 phonological acquisition within the general framework of **Optimality Theory** was proposed by Boersma (1998): Functional Phonology. It distinguishes between articulatory and perceptual representations and features, as illustrated in Figure 6.9. In speech perception, a hearer is confronted with the acoustic input of speech. This acoustic input consists of physical properties such as frequency, loudness and noise and is put into square brackets in Figure 6.9 because it is language-independent. The hearer's perception grammar, which consists of a perceptual categorization system, converts the raw acoustic input into a more perceptual, language-specific representation. This perceptual input is put between slashes in Figure 6.9. It is interpreted by the recognition system, which converts it into an underlying form (written between pipes). A speaker takes the perceptual specification of a word or syllable as stored in his or her lexicon as the input to the production grammar, which determines the surface form of the word or utterance. This results in an articulatory output in terms of articulatory gestures. The speaker's perceptual output is his or her acoustic output as perceived by himself or herself. This can be compared to the perceptual input of an acoustic input by another speaker and thus forms the essential part of the learning system.

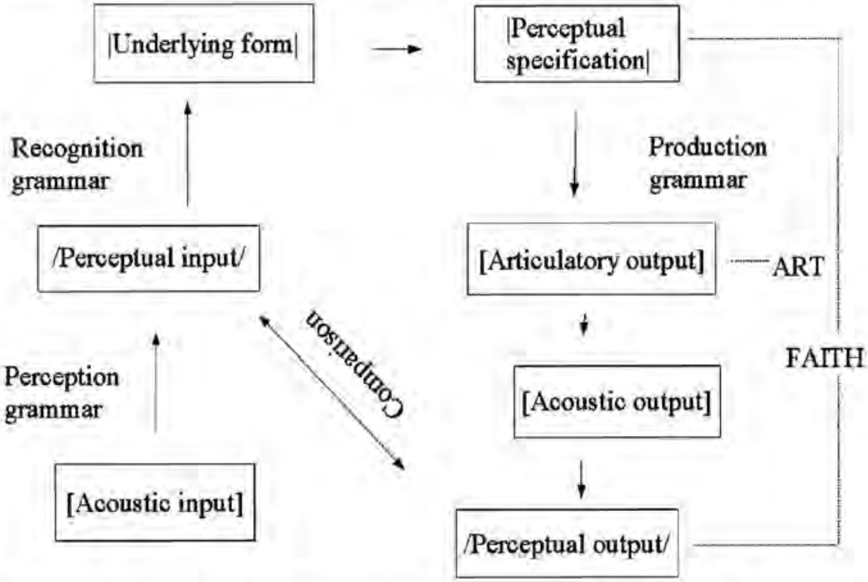


Figure 6.9. Perceptual representations proposed by Boersma (1998). FAITH stands for faithfulness constraints which monitor the correctness of pronunciation.

Mental representations are imagined, as in Optimality Theory, to consist of ranked constraints, either articulatory or perceptual in nature. Since constraints are assumed to be violable and can dominate each other, the articulatory output depends on the ranking of a set of competing constraints. Functional Phonology describes language acquisition in terms of learning how to rank constraints. Second language learners of Hausa, for example, have a fully specified perception and production grammar and perceptual specifications for all phonological features of their native language/s, but have no specifications for the perception and production of ejectives, if their native language does not have them as phonemes. In this initial stage, the language learner will hear the syllable [k'a] containing the ejective / k'/ as the acoustic input, but will perceive it as /ka/. The resulting underlying form will also be |ka|, and the speaker will generate the articulatory candidate [ka] and perceive this as /ka/. The first learning step consists of the acquisition of the correct perceptual categorization. This happens by listening to the language input and noticing that the phonetic difference is associated with differences in meaning. The learner will now categorize [k'a] as /k'a/ and have an underlying form of |k'a|. In production, however, he or she will still pronounce [ka]. Only when the difference between the learners' perception of their own output as /ka/ and the perception of the native speaker's acoustic output as /k'a/ is noticed, is the next learning step triggered. This second stage in the language acquisition process consists of the acquisition of the sensorimotor skills for the production of the particular phonological feature. In stage 3, in many cases, a faithful rendering of the perceptual target becomes possible (especially in imitation and in the production of isolated words, Boersma [1998] claims), but in normal speech production the articulatory constraint still dominates. For an accurate production in continuous speech, a fourth step is necessary, in which the learners must break the identity between the perceptual input and the underlying form. In the fourth stage, the learners will produce overly faithful speech because they have not noticed yet that in certain cases, for example, /k'/ is realized as the allophone [k]. The change of the underlying form to |k'a| allows the learners to categorize both forms in the perceptual input. Stage 5 involves learning the rules for these alternating constraints. The learners must see patterns in the words of the language based on morphological and other alternations and must construct abstract underlying forms. The input representation of the learners' production grammar shifts from the native speaker word to this new underlying representation.

Which implications for second language teaching can be drawn from the findings and theories of perceptual acquisition? Since our knowledge in this area is still very sketchy, only tentative recommendations can be made. It seems that successful perception training should involve multiple voices, the same sounds in different syllabic positions and in different phonetic contexts (Pisoni & Lively 1995, Rvachew & Jamieson 1995) because perception of L2 speech sounds does not occur on the level of phonemes but rather on that of allophones or phonetic features. In a perception experiment by Pisoni & Lively 1995, language learners learned an L2 contrast in a context-dependent manner and did not seem to encode phonetic contrasts in an abstract unit such as the phoneme. They rather seemed to store every instance of a sound in every phonetic context in which they heard it. Thus, listeners must learn what kind of variation is important and what specific attention should be paid – at least at the beginning of language learning – to allophonic and coarticulatory variation. This means that the perception of speech sounds is supported more when listeners hear real language than when they are exposed to synthetic, artificially created speech. Before attempting to produce sounds, language learners should be exposed to the language and simply listen. It has also been suggested that exposure to orthography should be delayed because this may activate the wrong responses and create patterns of perception that interfere with appropriate sound perception. Yet, the establishment of perceptual categories does not seem to be a necessary requirement for correct production in all cases. Some speakers can produce a phonetic distinction without being able to perceive it (e.g. Sheldon & Strange 1982).

6.9 Exercises

1. Which organs are involved in the perception of speech, and how do they work?
2. What is the role of the brain in speech perception?
3. How can perceptual abilities be measured?
4. Which relationship is there between acoustic features of sound such as frequency and intensity on the one hand and human perception of pitch and loudness on the other?
5. Which units of speech are stored in the brain as mental representations, and how are they used for understanding speech?
6. Why can humans perceive and understand speech even in very noisy conditions?
7. What is categorical perception, and how does this support the identification of speech sounds?
8. What seem to be the major difficulties for second language learners in perceiving the second language?

6.10 Further reading

The anatomy and physiology of the peripheral auditory system are described in Clark, Yallop & Fletcher (2007, chapter 8). Details on the neural processing of speech can be gained from Delgutte (1997) and Moore (1997). Johnson (1997, chapter 4) describes psychoacoustic experiments, and Lively et al. (1994) give an overview of research on the different models of word recognition. Speech perception in first language acquisition is described in detail in Vihman (1996). An overview of speech perception by second language learners can be found in Strange (1995) and Strange & Shafer (2008).

7 List of References

Abercrombie, D. (1967): *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Allen, G. & Hawkins, S. (1978): "The development of phonological rhythm". *Syllables and Segments*, ed. A. Bell & J. Hooper. Amsterdam: North Holland Publishing Company. 173-185.

Archibald, J. (1994): "A formal model of learning L2 prosodic phonology", *Second Language Research* 10, 215-240.

Archibald, J. (1997): "The acquisition of English stress by speakers of tone languages: lexical storage versus computation", *Linguistics* 35, 167-181.

Ashby, M. & Maidment, J. (2005): *Introducing Phonetic Science*. Cambridge: Cambridge University Press.

Barry, W. (1989): "Perception and production of English vowels by German learners: instrumental-phonetic support in language teaching", *Phonetica* 46, 155-168.

Beckman, M., Hirschberg, J. & Shattuck-Hufnagel, S. (2005): "The original ToBI system and the evolution of the ToBI framework", *Prosodic Typology*, ed. S.-A. Jun. Oxford: Oxford University Press. 9-54.

Behrens, H. & Gut, U. (2005): "The relationship between prosodic and syntactic organization in early multiword speech", *Journal of Child Language* 32, 1-34.

Best, C. (1995): "A direct realist view of cross-language speech perception", *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed. W. Strange. Timonium: York Press. 171-204.

Blevins, J. (1995): "The syllable in phonological theory", *The Handbook of Phonological Theory,* ed. J. Goldsmith. Oxford: Blackwell. 206-244.

Boersma, P. (1998): *Functional Phonology*. The Hague: Holland Academic Graphics.

Bohn, O.-S. & Flege, J. (1990): "Interlingual identification and the role of foreign language experience in L2 vowel perception", *Applied Psycholinguistics* 11, 303-328.

Bolinger, D. (1981): *Two Kinds of Vowels, Two Kinds of Rhythm*. Bloomington: Indiana.

Bolinger, D. (1998): "Intonation in American English", *Intonation Systems*, ed. D. Hirst & A. Di Cristo. Cambridge: Cambridge University Press. 45-55.

Bongaerts, T., van Summeren, C., Planken, B. & Schils, E. (1997): "Age and ultimate attainment in the pronunciation of a foreign language", *Studies in Second Language Acquisition* 19, 447-465.

Boysson-Bardies, B. de, Sagart, L. & Durand, C. (1984): "Discernible differences in the babbling of infants according to target language", *Journal of Child Language* 11, 1-15.

Boysson-Bardies, B. de, Vihman, M., Roug-Hellichius, L., Durand, C., Landberg, I. & Arao, F. (1992): "Material evidence of infant selection from the target language: A cross-linguistic phonetic study", *Phonological development: Models, research, implications,* ed. C. Ferguson, L. Menn & C. Stoel-Gammon. Timonium, Maryland: York Press. 369-391.

Brazil, D. (1975): *Discourse Intonation.* Birmingham: Birmingham University.

Britain, D. (1992): "Linguistic change in intonation: The use of high rising terminals in New Zealand English", *Language Variation and Change* 4, 77-104.

Broselow, E. (1984): "An investigation of transfer in second language phonology", *International Review of Applied Linguistics* 22, 253-269.

Browman, C. & Goldstein, L. (1992): "Articulatory phonology: an overview". *Phonetica* 49, 155-180.

Brown, C. (2000): "The interrelation between speech perception and phonological acquisition from infant to adult", *Second Language Acquisition and Linguistic Theory,* ed. J. Archibald. Oxford: Blackwell. 4-63.

Brown, G. & Yule, G. (1983): *Discourse Analysis.* Cambridge: Cambridge University Press.

Bybee, J. (2001): *Phonology and Language Use.* Cambridge: Cambridge University Press.

Bybee, J. (2002): "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change", *Language Variation and Change* 14, 261-290.

Byrd, D. & Saltzman, E. (1998): "Intragestural dynamics of multiple prosodic boundaries", *Journal of Phonetics* 26, 173-199.

Campbell, R. (1994): "Audiovisual speech: Where, what, when, how?", *Current Psychology of Cognition* 13, 76-80.

Campione, E. & Véronis, J. (2002): "A large-scale multilingual study of silent pause duration", *Proceedings of Speech Prosody 2002,* ed. B. Bel & I. Marlin. Aix-en-Provence : Laboratoire Parole et Langages. 199-202.

Cherry, E. (1953): "Some experiments on the recognition of speech, with one and with two ears", *Journal of the Acoustical Society of America* 25(5), 975-979.

Cho, T. (2002): *The effects of prosody on articulation in English.* New York: Routledge.

Cho, T. (2005): "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English", *Journal of the Acoustical Society of America* 117, 3867-3878.

Chomsky, N. & Halle, M. (1968): *The sound pattern of English.* New York: Harper & Row.

Chun, D. (2002): *Discourse Intonation in L2.* Amsterdam: Benjamins.

Clark, J., Yallop, C. & Fletcher, J. (2007): *An Introduction to Phonetics and Phonology.* Oxford: Blackwell. 3rd edition.

Clements, G. (1985): "The geometry of phonological features", *Phonology Yearbook* 2, 223-252.

Couper-Kuhlen, E. (1993): *English Speech Rhythm. Form and Function in Everyday Verbal Interaction.* Amsterdam: John Benjamins.

Cruttenden, A. (1997): *Intonation.* Cambridge: Cambridge University Press. 2nd edition.

Crystal, D. (1969): *Prosodic Systems and Intonation in English.* Cambridge: Cambridge University Press.

Crystal, D. (1981): *Clinical Linguistics.* Wien: Springer Verlag.

Cucchiarini, C., Strik, H. & Boves, L. (2002): "Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech", *Journal of the Acoustical Society of America* 111, 2862-2873.

D'Odorico, L. & Carubbi, S. (2003): "Prosodic characteristics of early multi-word utterances in Italian children". *First Language* 23, 97-116.

Dauer, R. (1983): "Stress-timing and syllable-timing reanalysed", *Journal of Phonetics* 11, 51-62.

Davenport, M. & Hannahs, S. (2005): *Introducing Phonetics and Phonology.* London: Hodder Arnold. 2nd edition.

Delattre, P. (1981): "An acoustic and articulatory study of vowel reduction in four languages", *Studies in Comparative Phonetics*, ed. P. Delattre. Heidelberg: Groos. 63-93.

Delgutte, B. (1997): "Auditory neural processing of speech", *The Handbook of Phonetic Sciences*, ed. W. Hardcastle & J. Laver. Oxford: Blackwell. 507-538.

Dziubalska-Kołaczyk, K. (1990): *A Theory of Second Language Acquisition within the Framework of Natural Phonology.* Poznan: AMU Press.

Eckert, H. & Barry, W. (2005): *The Phonetics and Phonology of English Pronunciation.* Trier: Wissenschaftlicher Verlag.

Fabricius, A. (2002): "Ongoing change in modern RP", *English World-Wide* 23, 115-136.

Fant, G. (1960): *Acoustic theory of speech production.* The Hague: Mouton.

Fant, G., Kruckenberg, A. & Nord, L. (1991): "Durational correlates of stress in Swedish, French and English", *Journal of Phonetics* 19, 351-365.

Fee, E. (1995): "Segments and syllables in early language acquisition", *Phonological Acquisition and Phonological Theory*, ed. J. Archibald. Hillsdale, N. J.: Erlbaum. 43-61.

Flege, J. & Eefting, W. (1987): "Production and perception of English stops by native Spanish speakers", *Journal of Phonetics* 15, 67-83.

Flege, J. & MacKay, I. (2004): "Perceiving vowels in a second language". *Studies in Second Language Acquisition* 26, 1-34.

Flege, J. (1995): "Second language speech learning theory, findings and problems", *Speech perception and linguistic experience: Issues in cross-linguistic research,* ed. W. Strange. Timonium: York Press. 233-277.

Fougeron, C. & Keating, P. (1997): "Articulatory strengthening at edges of prosodic domains", *Journal of the Acoustical Society of America* 101, 3728-3740.

Furrow, D. (1984): "Young children's use of prosody". *Journal of Child Language* 11, 203-213.

Giegerich, H. (1992): *English phonology*. Cambridge: Cambridge University Press.

Gimson, A. (1962): *An Introduction to the Pronunciation of English*. London: Edward Arnold.

Gordon, M. (2001): *Small-town Values and Big-city Vowels: A Study of the Northern Cities Shift in Michigan*. Durham, N.C.: Duke University Press.

Grabe, E., Rosner, B., García-Albea, J. & Zhou, X. (2003): "Perception of English Intonation by English, Spanish, and Chinese Listeners", *Language and Speech* 4, 375-401.

Green, K., Kuhl, P. Meltzoff, A. & Stevens, E. (1991): "Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect", *Perception & Psychophysics* 50, 524-536.

Grosser, W. (1997): "On the acquisition of tonal and accentual features of English by Austrian learners", *Second Language Speech – Structure and Process,* ed. A. James & J. Leather. Berlin: Mouton de Gruyter. 211-228.

Gut, U. (2000a): *Bilingual acquisition of intonation*. Tübingen: Niemeyer.

Gut, U. (2000b): "The phonetic production of emphasis by German learners of English". *Proceedings of New Sounds 2000*, Amsterdam, 155-157.

Gut, U. (2003): "Non-native speech rhythm in German". *Proceedings of the ICPhS Conference,* Barcelona, Spain, 2437-2440.

Gut, U. (2005): "Nigerian English prosody", *English World-Wide* 26, 153-177.

Gut, U. (2006): "Learner speech corpora in language teaching", *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, ed. S. Braun, K. Kohn & J. Mukherjee. Frankfurt: Lang. 69-86.

Gut, U. (2007a): "Learner corpora in second language acquisition research and teaching", *Non-native Prosody. Phonetic Description and Teaching Practice*, ed. J. Trouvain & U. Gut. Berlin: Mouton de Gruyter. 145-167.

Gut, U. (2007b): "Sprachkorpora im Phonetikunterricht", *Zeitschrift für interkulturellen Fremdsprachenunterricht* 12(2).

Hall, A. (1999): "Phonotactics and the prosodic structure of German function words", *Studies on the phonological word*, ed. A. Hall & U. Kleinhenz. Amsterdam: John Benjamins. 99-131.

Halle, M. (1992): "Phonological features", *International Encyclopaedia of Linguistics*, ed. W. Bright. Oxford: Oxford University Press. 207-212.

Handbook of the International Phonetic Association (1999). Cambridge: Cambridge University Press.

Hansen Edwards, J. & Zampini, M. (eds.) (2008): *Phonology and Second Language Acquisition*. Amsterdam: John Benjamins.

Hawkins, S. & Midgley, J. (2005): "Formant frequencies of RP monophthongs in four age groups of speakers", *Journal of the International Phonetic Association* 35, 183-199.

Herry, N. & Hirst, D. (2002): "Subjective and objective evaluation of the prosody of English spoken by French speakers: the contribution of computer assisted learning", *Proceedings of Speech Prosody 2002*, ed. B. Bel & I. Marlin. Aix-en-Provence : Laboratoire Parole et Langages. 383-386.

Hirst, D. (1998): "Intonation in British English", *Intonation Systems*, ed. D. Hirst & A. Di Cristo. Cambridge: Cambridge University Press. 56-77.

Hogg, R. & McCully, C. (1987): *Metrical Phonology: A Coursebook*. Cambridge: Cambridge University Press.

Ishikawa, K. (2002): "Syllabification of intervocalic consonants by English and Japanese speakers", *Language and Speech* 45, 355-385.

Jakobson, R. & Halle, M. (1956): *Fundamentals of Language*. The Hague: Mouton.

James, E. (1977): "The Acquisition of Second-Language Intonation Using a Visualizer". *Canadian Modern Language Review* 33, 503-506.

Jassem, W. & Gibbon, D. (1980): "Re-defining English accent and stress". *Journal of the International Phonetic Association* 10, 2-16.

Johnson, K. (1997): *Acoustic and Auditory Phonetics*. Oxford: Blackwell. 2nd edition.

Jones, D. (2006): *Cambridge English Pronunciation Dictionary*. Cambridge: Cambridge University Press.

Jonson, B. (1640): *The English Grammar*. Published by Menston: Scolar Press, 1972.

Kenstowicz, M. (1994): *Phonology in generative grammar*. Oxford: Blackwell.

Kent, R. & Read, C. (2002): *The acoustic analysis of speech*. Albany: Delmar, Thompson Learning. 2nd edition.

Kent, R. (1983): "The segmental organization of speech", *The Production of Speech*, ed. P. MacNeilage. New York: Springer. 57-90.

Kent, R. (1997): *The Speech Sciences*. San Diego: Singular Publishing Group.

Keseling, G. (1992): "Pause and intonation contours in written and oral discourse", *Cooperating with Written Texts. The Pragmatics and Comprehension of Written Texts*, ed. D. Stein. Berlin: Mouton de Gruyter. 31-66.

Kingdon, R. (1958): *The Groundwork of English Intonation*. London: Longman.

Kortmann, B. & Schneider. E. (2004): *A Handbook of Varieties of English. Volume 1: Phonology*. Amsterdam: Mouton de Gruyter.

Labov, W. (1989): "The child as the linguistic historian", *Language Variation and Change* 1, 85-97.

Labov, W., Yaeger, M. & Steiner, R. (1972): *A Quantitative Study of Sound Change in Progress*. Philadelphia: US Regional Survey.

Ladefoged, P. & Maddieson, I. (1996): *The Sounds of the World's Languages*. Oxford: Blackwell.

Ladefoged, P. (1999): "American English", *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press. 41-44.

Ladefoged, P. (2001a): *Vowels and Consonants*. Oxford: Blackwell.

Ladefoged, P. (2001b): *A Course in Phonetics*. Boston: Heinle & Heinle. 4th edition.

Laeufer, C. (1996): "The acquisition of a complex phonological contrast: voice timing patterns of English final stops by native French speakers". *Phonetica* 53, 117-142.

Lass, R. (1984): *Phonology: an introduction to basic concepts*. Cambridge: Cambridge University Press.

Laver, J. (1994): *Principles of Phonetics*. Cambridge: Cambridge University Press.

Leather, J. & James, A. (1996): "Second language speech", *Handbook of Second Language Acquisition,* ed. W. Ritchie & T. Bhatia. San Diego: Academic Press. 269-316.

Lehiste, I. (1970): *Suprasegmentals*. Cambridge, Mass.: MIT Press.

Lehiste, I. (1972): "The timing of utterances and linguistic boundaries", *Journal of the Acoustical Society of America* 51, 2018-2024.

Lehman, M. & Swartz, B. (2000): "Electropalatographic and spectrographic descriptions of allophonic variants of /l/", *Perceptual & Motor Skills* 90, 47-61.

Levelt, W. (1989): *Speaking: From Intention to Articulation*. Cambridge, Mass: MIT Press.

Liberman, A. & Mattingly, I. (1985): "The motor theory of speech revised", *Cognition* 21, 1-36.

Liberman, A., Cooper, F., Shankweiler, F. & Studdert-Kennedy, M. (1967): "Perception of the speech code", *Psychological Review* 74, 431-461.

Lieberman, & Blumstein, S. (1988): *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press.

Lieberman, P. (1984): *The biology and evolution of language*. Cambridge, Mass.: Harvard University Press.

Lively, S., Pisoni, D. & Goldinger, S. (1994): "Spoken word recognition: research and theory", *Handbook of Psycholinguistics*, ed. M. Gernsbacher. San Diego: Academic Press. 265-301.

Marslen-Wilson, W. (1987): "Functional parallelism in spoken word-recognition", *Cognition* 25, 71-102.

McCarthy, J. (1988): "Feature geometry and dependency: a review", *Phonetica* 45, 84-108.

McClelland, J. & Elman, J. (1986): "The TRACE model of speech perception", *Cognitive Psychology* 18, 1-86.

McGurk, H. & J. McDonald (1976): "Hearing lips and seeing voices", *Nature* 264, 746-748.

Mehler, J., Jusczyk, P. Lambertz, G. Halsted, N. Bertoncini, J. & Amiel-Tison, C. (1988): "A precursor of language acquisition in young infants", *Cognition* 29, 143-178.

Menn, L. & Stoel-Gammon, C. (1995): "Phonological development", *The Handbook of Child Language*, ed. P. Fletcher & B. MacWhinney. Oxford: Blackwell. 335-359.

Mennen, I. (2007): "Phonological and phonetic influences in non-native intonation", *Non-native Prosody. Phonetic Description and Teaching Practice*, ed. J. Trouvain & U. Gut. Berlin: Mouton de Gruyter. 53-76.

Mertens, P. (2004): "The prosogram: semi-automatic transcription of prosody based on a tonal perception model". *Proceedings of Speech Prosody 2004*, Nara, Japan, 549-552.

Mildner, V. & Liker, M. (2003): "Acoustic analysis of the speech of children with cochlear implants and comparison with hearing controls", *Proceedings of the International Congress of Phonetic Sciences*, Barcelona, 2377- 2380.

Missaglia, F. (2007): "Prosodic training for adult Italian learners of German: the Contrastive Prosody Method", *Non-native Prosody. Phonetic Description and Teaching Practice*, ed. J. Trouvain & U. Gut. Berlin: Mouton de Gruyter. 237-258.

Moore, B. (1997): "Auditory processing related to speech perception", *The Handbook of Phonetic Sciences*, ed. W. Hardcastle & J. Laver. Oxford: Blackwell. 538-565.

Moyer, A. (2004): *Age, Accent and Experience in Second Language Acquisition: An Integrated Approach to Critical Period Inquiry*. Clevedon: Multilingual Matters.

Nespor, M. & Vogel, I. (1986): *Prosodic Phonology*. Dordrecht: Foris.

Neu, H. (1980): "Ranking of constraints on /t,d/ deletion in American English", *Locating Language in Time and Space*, ed. W. Labov. New York: Academic Press. 37-54.

Norris, D. (1994): "Shortlist: A connectionist model of continuous speech recognition", *Cognition* 52, 189-234.

O'Connor, J.D. & Arnold, G. (1961): *Intonation of Colloquial English*. London: Longman. 2nd edition 1973.

Palmer, H. (1922): *English Intonation with Systematic Exercises*. Cambridge: Heffer.

Pierrehumbert, J. & Hirschberg, J. (1990): "The meaning of intonational contours in discourse", *Intentions in Communication,* ed. P. Cohen, J. Morgan & M. Pollack. Cambridge, Mass.: MIT Press. 271-311.

Pisoni, D. & Lively, S. (1995): "Variability and invariance in speech perception: A new look at some old problems in perceptual learning", *Speech perception and linguistic experience. Theoretical and methodological issues in cross-language speech research*, ed. W. Strange. Timonium: York Press. 433-459.

Radford, A. (1997): *Syntax: a minimalist introduction*. Cambridge: Cambridge University Press.

Raffelsiefen, R. (2004): "Paradigm uniformity effects versus boundary effects", *Paradigms in Phonological Theory*, ed. L. Downing, T.A. Hall & R. Raffelsiefen. Oxford: Oxford University Press. 211-262.

Ramus, F., Nespor, M. & Mehler, J. (1999): "Correlates of linguistic rhythm in the speech signal", *Cognition* 73, 265-292.

Roach, P. (1982): "On the distinction between 'stress-timed' and 'syllable-timed' languages", *Linguistic Controversies, Essays in Linguistic Theory and Practice,* ed. D. Crystal. London: Edward Arnold. 73-79.

Roach, P. (1991): *English Phonetics and Phonology*. Cambridge: Cambridge University Press. 2nd ed.

Roach, P. (2004): "British English: Received Pronunciation.", *Journal of the International Phonetic Association* 34, 239-245.

Rosenblum, L., M. Schmuckler & J. Johnson (1997): "The McGurk effect in infants", *Perception & Psychophysics* 59, 347-357.

Roy, N., Merrill, R., Thibeault, S., Parsa, R., Gray, S. & Smith, E. (2004): "Prevalence of voice disorders in teachers and the general population", *Journal of Speech and Hearing Research* 47 (2), 281-293.

Rvachew, S. & Jamieson, D. (1995): "Learning new speech contrasts: Evidence from adults learning a second language and children with speech disorders", *Speech perception and linguistic experience. Theoretical and methodological issues in cross-language speech research*, ed. W. Strange. Timonium: York Press. 379-410.

Sampson, G. (1980): *Schools of linguistics: competition and evolution*. London: Hutchinson.

Samuel, A. (1981): "Phonemic restoration: insights from a new methodology", *Journal of Experimental Psychology: General*, 110, 474-494.

Schmidt, A. & Flege, J. (1996): "Speaking rate effects on stops produced by Spanish and English monolinguals and Spanish/English bilinguals". *Phonetica* 53, 162-179.

Selkirk, E. (1986): "On derived domains in sentence phonology", *Phonology Yearbook* 3, 371-405.

Shaywitz, B., Shaywitz, S., Pugh, K., Constable, R., Skudlarski, P., Fulbright, R., Bronen, R., Fletcher, J., Shankweiler, D., Katz, L. & Gore, J. (1995): "Sex differences in the functional organization of the brain for language", *Nature* 373, 607-609.

Sheldon, A. & Strange, W. (1982): "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception", *Applied Psycholinguistics* 3, 243-261.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Pierrehumbert, J. & Hirschberg, J. (1992): "ToBI: a standard for labeling English prosody", *Proceedings, Second International Conference on Spoken Language Processing 2*, Banff, Canada, 867-70.

Stampe, D. (1979): *A Dissertation on Natural Phonology*. New York: Garland.

Stevens, K. & Blumstein, S. (1981): "The search for invariant acoustic correlates of phonetic features". *Perspectives on the Study of Speech*, ed. P. Eimas & J. Miller. Hillsdale, NJ: Erlbaum. 1-38.

Stevens, K. & House, A. (1961): "An acoustical theory of speech production and some of its implications", *Journal of Speech and Hearing Research* 4, 303-320.

Strange, W. (ed.) (1995): *Speech perception and linguistic experience. Theoretical and methodological issues in cross-language speech research*. Timonium: York Press.

Strange, W. & Shafer, V. (2008): "Speech perception in second language learners: The re-education of selective perception", *Phonology and Second Language Acquisition*, ed. J. Hansen Edwards & M. Zampini. Amsterdam: John Benjamins. 153-191.

Tench, P. (1996): "Intonation and the differentiation of syntactic patterns in English and German", *International Journal of Applied Linguistics* 6, 223-256.

Trouvain, J. & Gut, U. (eds.) (2007): *Non-native Prosody. Phonetic Description and Teaching Practice*. Berlin: Mouton de Gruyter.

Turk, A. & Shattuck-Hufnagel, S. (2000): "Word-boundary-related duration patterns in English", *Journal of Phonetics* 28, 397-440.

Upton, C. (2004): "Received Pronunciation", *A Handbook of Varieties of English*, ed. B. Kortmann & E. Schneider. Berlin: Mouton. 217-230.

Van Bezooijen, R. (1995): "Sociocultural aspects of pitch differences between Japanese and Dutch women", *Language and Speech* 38, 253-265.

Vihman, M. (1996): *Phonological development*. Oxford: Blackwell.

Warren, R. (1970): "Perceptual restoration of missing speech sounds", *Science* 176, 392-393.

Wells, J. (2000): *Longman Pronunciation dictionary*. London: Longman.

Wells, J. (2006): *English Intonation*. Cambridge: Cambridge University Press.

Wells, J., Barry, W., Grice, M., Fourcin, A. & Gibbon, D. (1992): "Standard computer-compatible transcription". *Technical Report*, SAM Stage Report Sen.3 SAM UCL-037.

Wells, J.C. (1982): *Accents of English*. Cambridge: Cambridge University Press.

Wennerstrom, A. (1994): "Intonational meaning in English discourse: a study of non-native speakers", *Applied Linguistics* 15, 399-420.

Wennerstrom, A. (1998): "Intonation as cohesion in academic discourse", *Studies of Second Language Acquisition* 20: 1-25.

Wennerstrom, A. (2001): *The Music of Everyday Speech*. Oxford: Oxford University Press.

Werker, J. & Pegg, J. (1992): "Infant speech perception and phonological acquisition", *Phonological development*, ed. C. Ferguson, L. Menn & C. Stoel-Gammon. Timonium: York Press. 285-311.

Werker, J., Frost, P. & McGurk, H. (1992): "La langue et les lèvres: Cross-language influences on bimodal speech perception", *Canadian Journal of Psychology* 46, 551-568.

Whitworth, N. (2003): "Prevocalic boundaries in the speech of German-English bilinguals", *Proceedings of the 15th International Conference of Phonetic Sciences*, Barcelona, 1093-1096.

Wichmann, A. (2000): *Intonation in Text and Discourse*. London: Longman.

Willems, N. (1982): *English Intonation from a Dutch Point of View*. Ambach: Intercontinental Graphics.

Williams, B. & Hiller, S. (1994): "The question of randomness in English foot timing: a control experiment", *Journal of Phonetics* 22, 423-439.

Yang, L.-C. (2004): "Duration and pauses as cues to discourse boundaries in speech", *Proceedings of Speech Prosody 2004*, Nara, Japan. 267-270.

Yavaş, M. (2006): *Applied English Phonology*. Oxford: Blackwell.

Zsiga, E. (2003): "Articulatory timing in a second language", *Studies in Second Language Acquisition* 25, 399-432.

8 Index

Textbooks in English Language and Linguistics (TELL)

Edited by Joybrato Mukherjee and Magnus Huber

Band 1 Ulrike Gut: Introduction to English Phonetics and Phonology. 2009.

www.peterlang.de