Tutorial on Phonetics and Speech Analysis

true

Document compiled 30 Jan 2025 23:07

Contents

Preface **7**

Aims . . . 7

How to use this tutorial . . . 7

Recommended software . . . 8

Structure of this tutorial . . . 9

0.1 Recommended textbooks . . . 9

Details . . . 9

Part I: Sounds **13**

1 Sound waves **13**

1.1 Sound . . . 13

1.2 Sound wave . . . 13

1.3 Acoustic media . . . 14

1.4 The speed of sound . . . 15

1.5 Pressure . . . 15

Questions . . . 15

1.6 Oscillogram . . . 16

1.7 Periodic and aperiodic sounds . . . 17

1.8 Key properties of a sound wave . . . 17

1.9 How to work with `Praat` . . . 27

1.10 How to draw an oscillogram . . . 28

2 Converting sound to bytes **31**

2.1 Overview 31
2.2 How to handle a microphone 32
2.3 Key parameters in AD conversion 32
2.4 How to record a sound 34
2.5 How to play back a digital sound 38
2.6 How to manipulate and edit a digital sound 38

3 Complex sounds and spectra **43**

3.1 Introduction 43
3.2 Spectrum 45
3.3 Spectra of aperiodic sounds 47
3.4 Envelope 49

4 Filtering **51**

4.1 Introduction 51
4.2 Types of filters 51
4.3 Properties of filters 53
4.4 Emphasis filters 55

Part II: Speech **59**

5 Speech sounds **59**

5.1 Resonance 59
5.2 Formants 61
5.3 Source-filter theory of speech production 65

6 Spectrograms **69**

6.1 Introduction 69
6.2 Broadband spectrogram 70
6.3 Narrowband spectrogram 72
6.4 How to make a spectrogram 73
6.5 How to read a spectrogram 74

7 Segmenting and labeling speech sounds **77**

7.1 Introduction 77

7.2 TextGrid objects 79

7.3 How to segment and label speech sounds in `Praat` 80

8 Prosody **83**

8.1 Pauses in speech 83

8.2 Durations 83

8.3 Pitch 83

8.4 Intensity 83

8.5 Stresses and accents 83

Preface

Aims

In this tutorial you will learn about **acoustics** (sounds), **phonetics** (speech), and **speech analysis**. You will learn the core concepts in these related fields, as well as the necessary practical skills for speech analysis. The aim of this tutorial is to provide you with the phonetic insights and skills in speech analysis that you need to succesfully conduct phonetic research in your own project (e.g. paper or thesis).

Under construction

This tutorial is a work in progress, resulting from an ongoing revision of an existing tutorial, and meanwhile incorporating other modules and resources.

The existing (outdated) full tutorial is still available at https://resources.lab.hum.uu.nl/resources/phonetics/index.html .

More details on the origins of this tutorial are provided below.

How to use this tutorial

You will learn the most from this tutorial if you

(1) read the explanatory texts in this tutorial,
(2) work through the questions and exercises provided,
(3) practice in applying your new knowledge hands-on, with the `Praat` computer program (detailed below), and
(4) re-read the relevant sections from this tutorial and your textbook, with the help of keywords provided per section.

Questions

Text blocks such as this one will contain questions or exercises inviting you to engage with the tutorial. You will learn most if you attempt to answer these questions (preferably in writing) *before* you proceed and *before* you take a look at the answer provided. (These questions only work in the HTML version of the tutorial; other versions will just show both the question and answer subsequently.)

Question 0.1

What is sound?

Answer 0.1

Sound is a type of energy that travels through a medium (such as air, water, or solid materials) in the form of waves. These sound waves are created by the vibration of objects, which causes the surrounding particles in the medium to move in a back-and-forth motion. This movement, or vibration, transfers energy through the medium, creating waves of high and low pressure.

Recommended software

In this tutorial you will work mostly with `Praat` (Boersma and Weenink, 2024)[1]. This is a popular open-source program for the analysis of speech, developed by Paul Boersma and David Weenink (both at University of Amsterdam). It can be found on its own website (https://www.praat.org), where you will find a wealth of helpful documentation. `Praat` also has extensive `Help` built in, including a full tutorial.
There is an online forum (https://groups.io/g/Praat-Users-List), where users share their knowledge by posting questions and providing answers.

In order to install `Praat` on your computer, go to its webpage at https://www.praat.org/, and then proceed to the download page for the operating system of your computer. Follow the installation instructions on the download page for your operating system.

Instructions for using `Praat`

Text blocks such as this one will contain instructions about how to "do" things in `Praat`.

[1] The Dutch word *praat* / pra t/ means "talk".

Options in software menus, as well as texts in on-screen buttons, will be shown `in this way`. The notation `Main > Sub` means: first choose option `Main` from the main menu, after which a submenu will appear, then choose option `Sub` from the submenu. Commands or formulas that you have to type will be shown `in this way` too. (Commands typically need to be terminated with typing `Enter` or `Return` or or – which however will not be specified in the instructions.)

Structure of this tutorial

TODO add structure description and pointers: Part I sounds (acoustics), Part II speech (phonetics), Part III ??

0.1 Recommended textbooks

This tutorial is intended to be used in addition to one or more textbooks in phonetics, to which this tutorial will provide additional background knowledge. Some excellent textbooks and introductions in phonetics are those by Rietveld and van Heuven (2009) (in Dutch), Johnson (2012), Ladefoged and Johnson (2015), Reetz and Jongman (2020), Beňuš (2021), and Zsiga (2024).

Details

License

Citation

TODO add citation instructions

Technical details

TODO add tech details

History

This work is based on an earlier tutorial (2006-2007) titled Tutorial for self study: basics of phonetics and how to use Praat by Clizia Welker and Hugo Quené. In turn, that 2007 tutorial was partly based on older texts by Hugo Quené, Denise Bruin and Mirjam Wester (1996-2000); these older texts acknowledged valuable comments and suggestions by Paul Boersma, Olga van Herwijnen, Kim Koppen, Eva Sittig, Joyce Vliegen and Mieke van Wijck.

The 2007 version of the tutorial was subsequently revised and adapted to the current version using `R Markdown` (Xie et al., 2018) and `bookdown` (Xie, 2024) in Rstudio by Hugo Quené in 2024–2025.

Part I: Sounds

Chapter 1

Sound waves

Chapter keywords: sound, sound wave, oscillation, propagation, longitudinal wave, transverse wave, medium, speed of sound, force, pressure, Pascal, oscillogram, frequency, Hertz, period, periodic, aperiodic, fundamental frequency, octave, amplitude, intensity, phase, Pascal, `Praat`, object, visualization, picture, figure, harmonic, overtone, timbre, Fourier, spectrum, spectral envelope, noise, impulse.

1.1 Sound

Sound is a type of energy that travels through a medium (such as air, water, or solid materials) in the form of waves. These sound waves are created by the vibration of objects, which causes the surrounding particles in the medium to move in a back-and-forth (oscillatory) motion. This movement, or vibration, or oscillation, transfers energy through the medium, creating waves of high and low pressure.

1.2 Sound wave

A sound wave consists of pressure fluctuations caused by the molecules of the acoustic medium crowding together (compression) and moving apart (rarefaction). A sound wave is spread in all directions from the sound source; we could compare its propagation to that of a circular wave on the surface of a water basin. The molecules themselves move over a very short distance and do not travel along with the wave: instead, after the sound wave (the pressure fluctuation) has passed along, they go back to their equilibrium position.

Sound in air is different from wind. In wind, or in air flow, the air particles move from one position to another (from subtropics to equator, from lungs to mouth, from oceans to continents). In sound, however, there is no net movement of the air particles: the particles only move over a very small distance, and return to their equilibrium after the sound wave has passed. In sound waves, the distance of travel of the air molecules is only about 10^{-11} to 10^{-5} m, depending on the amplitude and frequency of the vibration (more about these key properties in §1.8 below). There are two kinds of waves (also depending on the acoustic medium). In *longitudinal* waves (such as sound waves) the back-and-forth displacement or movement of the medium's particles is in the same direction as the propagation of the wave. In *transverse* waves (such as the waves on the surface of a pond) the back-and-forth displacement of the water particles is perpendicular to the direction of propagation of the wave.
A stadium wave provides a clear example of a transverse wave: a group of persons (the particles) starts the wave by standing up, rising their arms, sitting down, standing up again, and so on. The persons' action is directly followed by that of their neighbours on one side, who do the same and who are again followed by their next neighbours on their side, and so on, until the wave is travelling through the whole stadium. The persons' motion (up-down) is perpendicular to the propagation of the wave (left-right along the bench).

Sound propagates in all dimensions through an acoustic medium, like an expanding sphere, which is indeed the theoretical model used to describe the sound wave propagation pattern. As the sound wave moves away from its source, more particles are involved in the pressure fluctuations. As a consequence, sound waves lose energy while travelling through the medium, as some of the energy is spent in moving increasingly more particles. Finally, sound is perceived as such when the sound wave spread by the sound source and travelling through the acoustic medium finally impinges upon the eardrum of the observer.

1.3 Acoustic media

Air is only one of the media through which sound can propagate. If your head is under water (as in a bath, pool, lake or sea), the water may carry sound waves from the sound source to your eardrums, and you do hear sounds. The propagation of sound waves is faster through liquids than through gases such as air: the closer the molecules of the medium (i.e. the higher its density), the higher the speed of sound in that medium.

You can also put your ear to the ground in order to hear sounds propagated through the soil. The propagation of sound waves in solid soil is even faster than in liquids. Trying this out on dry sand on the beach, one observer noted hearing footsteps until about 25 m distant (Minnaert, 1970, §10).

1.4 The speed of sound

In air, the speed of sound (the speed of propagation of a sound wave, symbol c) is about 332 m/s at 0°C, about 343 m/s at 20°C, and 353 m/s at at 37°C (Shadle, 2010) (all for dry air at sea level). The speed of sound in a gas such as air is affected by only two parameters: - the ambient temperature of the gas (as shown in the numbers above), - the composition of the gas (its mixture and the density and compressibility of its component gases), including its relative humidity: humid air holds more particles (of water), resulting in a slight increase of the speed of sound as relative humidity increases (Harris, 1971)[1].

In sea water, sound travels at about 1435 m/s, in concrete 3400 m/s, in iron (e.g. railroad tracks) about 5100 m/s.

1.5 Pressure

Pressure is the amount of force on a surface. In physics, *force* is defined as an influence causing an object to accelerate. It is expressed in Newton units; a Newton is the amount of force that increases the velocity of a 1-kilogram object by one meter per second (m/s). *Pressure*, in turn, is defined as force per unit of area. It is measured in Pascal units, which correspond to Newton (N) per square meter ($1\ Pa = 1\ N/m^2$). Under normal conditions, atmospheric air pressure is centered at 1013.25 hPa (101325 Pa, an average value[2] on a medium latitude at sea level, at 0°C), with normal meteorological fluctuations of about ± 5000 Pa. Sound wave fluctuations in air pressure are far smaller, ranging from about ±20 µPa (micropascal, or ±0.00002 Pa) at the lower threshold of hearing to about ±20 Pascal at the upper threshold of hearing. Even louder sounds, with variations in air pressure exceeding about ±20 Pascal, are painful and cause hearing damage.

Questions

Question 1.1

Explain why a sound wave loses energy the further it is spread from the oscillation source.

Answer 1.1

The more a sound *wave* moves away from the source, the more particles of the medium (e.g. air) are involved. The amount of initial energy (spread with the

[1] For relative humidity >30% (Harris, 1971).

[2] This is the standard unit of 1 atmosphere. The pressure is due to the Earth's gravition force on the Earth's atmosphere.

source oscillation) is spread over a larger surface, of an expanding imaginary sphere, and consequently the sound wave displaces more particles. The overall amount of energy remains the same. Therefore, the energy on a single medium particle or on a single portion of the sound wave is smaller. Thus, the sound wave fades as the distance to the sound source increases.

Remember that the sound *wave* travels through the medium, but the particles in the medium remain more or less in place.

1.6 Oscillogram

As explained in §1.2 above, a sound wave consists of pressure fluctuations caused by the molecules of the acoustic medium crowding together (compression) and moving apart (rarefaction). These oscillations in air pressure can be measured and visualized, in the form of a so-called **oscillogram** or graphical representation of a sound wave.
(In chapter 2 we'll learn how to measure, record and store a sound wave. Here, we jump ahead and present the oscillogram, in order to explain important properties of sound waves.)

In an oscillogram such as Figure 1.1, the horizontal axis represents the time dimension[3], and the vertical axis represents the air pressure. The pressure fluctuations (compression and rarefaction) are displayed as vertical deviations relative to the horizontal baseline[4]. Thus an oscillogram records the back-and-forth movements of the particles of the medium, indirectly, by recording the fluctuations in relative pressure, at a fixed location.

The oscillogram in Fig.1.1 shows a fragment of speech, taken from the audio file named `1-4-17_City_Council_SLASH_1-4-17_City_Council_DOT_mp3_00029.flac` from the *Peoples Speech* corpus at https://huggingface.co/datasets/MLCommons/peoples_speech. For details about that corpus, see Galvez et al. (2021).

Thus an oscillogram is comparable to a meteorologist's regular measurements of atmospheric air pressure at a fixed location — albeit on far finer scales of time and of air pressure.

The visualisation in an oscillogram may suggest, misleadingly, that the air particles themselves dance "up and down" (transverse) while the sound wave travels "from left to right", like waves on the surface of a body of water. That is not true: sound in air travels in *longitudinal* sound waves, resulting in the air pressure variations that are visualized in the oscillogram.

[3] Often abbreviated to t.

[4] The baseline represents the ambient average air pressure. By convention, higher air pressure is on the top side and lower air pressure at the bottom side of an oscillogram.

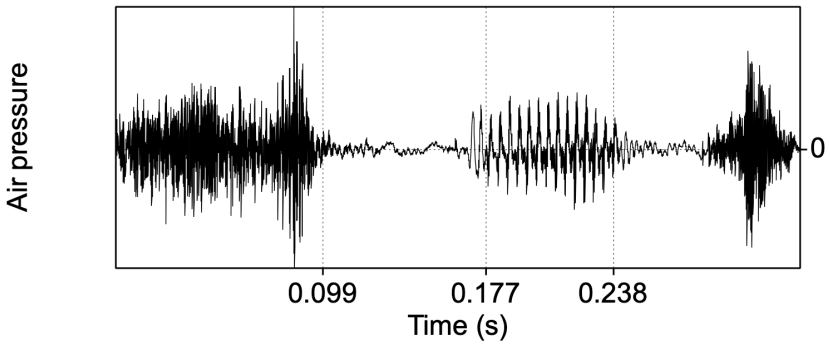

Figure 1.1: Oscillogram of the word *speech*, with boundaries between segments.

1.7 Periodic and aperiodic sounds

There are two classes of sounds which are easily distinguishable in an oscillogram:

- **periodic** sounds, in which there is sound wave pattern that repeats itself after a particular time interval or **period** (symbol T) of a single cycle. Periodic sounds have a perceptible pitch or tone. Vowel sounds such as the /i/ in Figure 1.1 (from 0.177 to 0.238 s) provide clear examples of a periodic sound.

- **aperiodic** sounds, in which there is not a repetitive but instead a random pattern in the air pressure variations. Aperiodic sounds do not have a perceptible pitch but instead we hear them as noise. Some consonant sounds, e.g. the /s/ in Figure 1.1 (from 0 to 0.099 s), are clear examples of such noisy, aperiodic sounds[5].

1.8 Key properties of a sound wave

A periodic sound wave can be characterized by three key properties, which are illustrated in the oscillogram in Figure 1.2 and which are further discussed in the following sub-sections.

[5]The clearest examples are provided by unvoiced fricative consonants, such as /f, s/.

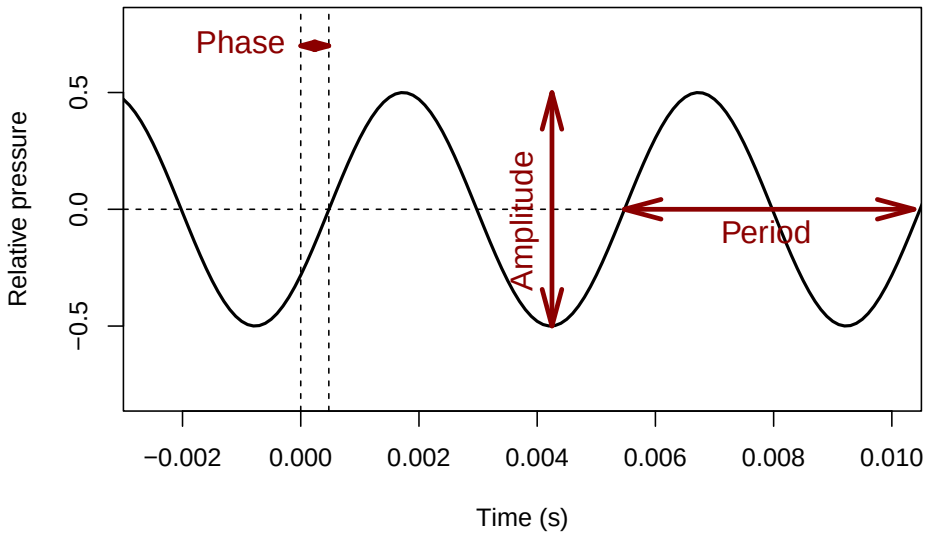

Figure 1.2: Oscillogram of a periodic sound (in black), with indications of the key properties frequency (1/period, see below), amplitude, and phase. The oscillogram is recorded over time (along the horizontal axis), at a fixed position in space.

1.8.1 Frequency

The frequency (symbol f) of a sound wave is the number of repeated cycles or periods (of air pressure variations) within a time interval. Only periodic sounds do have such repetitions, and thus a frequency. The frequency of a sound is perceived as its *pitch* or tone. Frequency is expressed in periods per second, or Hertz units, named after Heinrich Rudolf Hertz (1857-–1894)[6]. Each of these periods or cycles corresponds to the repeating fluctuation between two consecutive maxima, or between corresponding 'zero crossings' of adjacent periods. In Figure 1.1, in the vowel /i/, we count 14 periods in 0.065 seconds, so $f \approx 14/0.065 \approx 215$ Hz. These periods have a duration of about $T \approx 0.065/14 \approx 0.0046$ s[7]. Period T and frequency f are each other's inverse, so $f = 1/T$ and $T = 1/f$.

In the transverse stadium wave, the period T is the time interval between two consecutive actions of standing up by the same person (e.g. $T = 5$ s), and frequency f is the number of actions that occur within a given time unit (e.g. $f = 1/T = 1/5$ Hz).

[6] In older texts you may find 'cycles per second', abbreviated 'cps'.
[7] More exactly, the period *is* 0.0046 s.

1.8.1.1 Octave

An octave refers to a frequency ratio of $1 : 2$ or $2 : 1$, that is, doubling or halving of the frequency. An octave is the distance between 12 semitones (guitar frets or piano keys, counting black and white keys). If you sing or play a musical note (e.g. note `A2` with $f = 110$ Hz), and then jump to the next higher octave, then the new note `A3` has a frequency of $2 \times 110 = 220$ Hz. The next higher octave note `A4` has a frequency of $2 \times 220 = 440$ Hz.
(As an aside: doubling the frequency means halving the wavelength, and vice versa, see §1.8.5 below).

1.8.2 Amplitude

The amplitude (symbol A) of a sound wave is the extent of the variations in air pressure (due to compression and rarefaction), measured in Pascal units of pressure. With some simplification, the amplitude of a sound wave is perceived as its *loudness* or 'volume'. In an oscillogram, the amplitude corresponds directly to the maximum vertical displacement, that is, to the peak deviation in air pressure relative to the ambient reference pressure).

In the transverse stadium wave, the amplitude could be thought of as the extent to which persons raise their hands: the heighth of the wave crests.

In practice, the amount of energy in a (longitudinal) sound wave is better assessed in the form of *intensity*, which is discussed in §1.8.3 below.

Study tip: The sections on amplitude and on intensity (including decibels) are rather difficult. In first pass, just read these sections globally. After having finished the entire chapter (or more), return for a deeper second pass through these sections. Compare the explanations in these sections with those in your textbook(s), and then return for a third pass through these sections.

1.8.2.1 RMS amplitude

How can we calculate the mean or average amplitude of a sound wave, over a particular stretch of time? We do this by squaring the amplitude values, then averaging the squared values, and then taking the Root of the Mean of the Squares. The resulting value is termed the **RMS** amplitude, and it is equivalent to the *standard deviation* of the amplitude, computed over a particular stretch of time.

In a simple sinewave tone, as in Figure 1.2, the RMS of the amplitude corresponds to $0.707\times$ the maximum or peak amplitude.

1.8.3 Energy, power, intensity

Amplitude is related to the pressure variations of a sound wave (with pressure defined as force per unit area; §1.5). The *energy* of the sound wave is its capacity to do work, that is, to exert a force on (and thus causing displacement of) particles of the medium, that is, to propagate. Thus we want to consider the amount of acoustic energy, imparted by the vibrating sound source, and transmitted throughout the medium by causing particles to bump into each another. Energy is expressed in Joule (J) units; one Joule "is equal to the amount of work done when a force of one newton displaces a mass through a distance of one metre in the direction of that force" (< https://en.wikipedia.org/wiki/Joule>).

We saw earlier that we may regard the sound propagating through the medium as an imaginary expanding sphere (see §1.2). Consider the sound of a single impulse, say a hammer hitting a nail, with air as the medium of propagation. The energy of the sound remains constant, but as the sound propagates, the imaginary sphere expands, and the original energy is distributed over a wider area of that imaginary sphere. The variations in air pressure decrease as the sphere expands, that is, as the distance to the sound source increases.

Energy is independent from the time dimension: a continuous sound having an energy of 10 J for 1 second involves as much total energy as a continuous sound having an energy of 5 J for 2 seconds. When studying sounds, however, we typically do take time into account: we are interested in the amount of energy of a sound *per second*. This is termed the *power* of the sound. Power is expressed in Watt (W) units; one Watt equals 1 Joule per second. By analogy, you may consider a lamp of 10 W burning for 1 hour, and another lamp of 5 W burning for 2 hours. Both lamps have consumed equal amounts of energy (in principle), but the lamp with higher power converts more energy *per hour* into light: it shines brighter.

For sound waves, we are interested in the amount of power, but now expressed per unit area on the imaginary sphere of propagation of the sound wave in the medium (see §1.2). How much acoustic energy is transferred in the wave, by particles colliding with their neigboring particles, per time and per area? This property is termed the **intensity** of the sound. Intensity is expressed in units of energy per square meter perpendicular to the direction of propagation (W/m^2).

The intensity of a sound drops off as the distance to the sound source increases[8]. The sound of a hammer hitting a nail has an intensity $I/1 = I$ at a distance of $1 \cdot r$, but only an intensity of $I/3^2 = I/9$ at a triple distance of $3 \cdot r$. If you double the distance from r to $2r$, the intensity is significantly reduced by -14 dB (§1.8.3.2)[9]. **Protect your ears** against loud sounds by moving away from

[8] The intensity on the surface of the sphere undergoes a dropoff which can be calculated as $I = I_s/4\pi r^2$, with I_s being the intensity at the source, and r being the radius of the sphere, that is, the distance from the source.

[9] Even increasing the distance by only 25% reduces the intensity significantly by -4 dB.

the sound source!

The intensity of a sound is proportional to the maximum pressure variations of that sound, with

$$I = \frac{1}{2} \cdot \frac{p^2}{\rho \cdot c}$$

where p is the RMS amplitude of pressure variations, ρ is density of the medium (for air, $\rho = 1.29$ kg/m^3) and c is speed of sound (in air, $c = 332$ m/s at 0°C, see §1.4). A simple sine wave (as in Fig. 1.2) with frequency $f = 1000$ Hz and with RMS amplitude of pressure of $p = 2 \cdot 10^{-5}$ N/m^2 has an intensity $I = 4.7 \cdot 10^{-13}$ W/m^2. Because 1/2 and ρ and c are approximately constant, the relationship between intensity I and air pressure p is often simplified as $I \propto p^2$.

Experiments have shown that the sensitivity of the human ear corresponds approximately with the *logarithm* of the intensity (at the eardrum). This makes it attractive to use logarithmic scales for sound pressure and for the intensity of a sound. Such a scale allows us to represent very large and very small values with equal perceptual accuracy across the scale.

1.8.3.1 Logarithm

The *logarithm* of a number x, or $\log(x)$, is the exponent or power to which you must raise a given base number in order to obtain x. For the base number, we often use 10 or 2. So-called 'natural logarithms' have $e \approx 2.7$ as the base number, and these are often indicated as $\ln(x)$. For example,

$$^{10}\log(1000) = 3$$

since

$$10^3 = 1000$$

One advantage of using logarithms is that multiplication of two numbers is simplified to addition of their logarithms (using a common base):

$$100 \cdot 1000 = 100000$$

$$10^2 \cdot 10^3 = 10^{2+3} = 10^5$$

$$^{10}\log(100) + ^{10}\log(1000) = 2 + 3 = ^{10}\log(100000) = 5$$

Another advantage is that large *ratios* (e.g. of sound intensities or frequencies) are easier to express as logarithms:

$$10 : 1000 = 10^1 : 10^3$$

$$^{10}\log(10) : ^{10}\log(1000) = 1 : 3$$

By definition, $x^0 = 1$, and $^x\log(1) = 0$, for any base x.

Intensity is perceived in an approximately logarithmic fashion: each tenfold increase of intensity (logarithm using base 10) is perceived as one equal step in intensity, as further explained below. In a similar fashion, frequency too is perceived in logarithmic fashion: each twofold increase of frequency (doubling, logarithm using base 2) is perceived as one equal step in frequency, viz. by one octave (§1.8.1.1).

1.8.3.2 Decibel

Sound intensity or sound pressure is not expressed in absolute units, but in relative decibel units. The *decibel* unit of intensity (or of sound pressure) is based on three steps (Fry, 1979, 91ff):

1. Take the **ratio** of the intensities (pressures) of a target sound, and that of some reference sound – more about the choice of reference sound in the next paragraphs;
2. Take the **logarithm** (§1.8.3.1) of that intensity ratio (pressure ratio); this results in a log ratio, in *Bel* units, named after Alexander Graham Bell (1847-1922, inventor of the telephone);
3. **Multiply** the logarithm **by 10** to obtain the intensity ratio in *decibel* units.

Several different references are commonly used:

- a reference sound that has a standard RMS pressure of $p_0 = 2 \cdot 10^{-5}$ Pa, and/or a standard intensity of $I_0 = 10^{-12}$ W/m^2. This is the weakest sound perceived by humans[10]. As this reference p_0 is expressed in RMS units of sound pressure (in Pascal units), ratios using this p_0 as absolute reference are termed `dB Sound Pressure Level` or `dB(SPL)`, as in Table 1.1 below.
  (Aside: In the reference sound, the air particles move back and forth over a tiny distance, roughly the same as the size of the molecules in air! If our hearing would be just slightly more sensitive, we would just hear the constant (brown) noise of the Brownian movement of the particles in the atmosphere (see §3.3.1).)
- the smallest sound pressure perceived by a particular *individual* listener (test participant, patient) at a particular frequency in a particular ear, this is called the 'sensation level' (of that ear, listener, frequency). Comparisons using this relative threshold are termed `dB (SL)`; this allows us to present auditory stimuli with an equal subjective loudness across listeners and across ears.

[10] This reference is the *minimal* hearing threshold, achieved only by the 1% most sensitive listeners. The *average* hearing threshold, achieved by 50% of listeners, is at about +16 dB relative to this reference value (Fletcher, 1953, Fig.96), see below.

- the *loudest* sound (highest pressure, or voltage, or intensity) that a device can process without distorting the input signal; incoming sounds are expressed in negative dB units below this reference.

Putting the three steps together, we see that for sound intensity levels:

$$L_I = 10 \cdot^{10} \log\left(\frac{I}{I_0}\right)$$

where L_I is the level of intensity, in dB, I is the intensity of the target sound, and I_0 is the intensity of the reference sound (see above), e.g. $I_0 = 10^{-12}$ W/m^2. The three steps may be followed by working through this formula from the inside out: (1) we take the target-to-reference ratio, (2) we take the logarithm of that ratio (using 10 as base number), and (3) we multiply by 10 to obtain *deci*bel units.

If the target sound has an intensity I which is 10000× as large as I_0, then

$$L_I = 10 \cdot^{10} \log\left(\frac{I}{I_0}\right) = 10 \cdot^{10} \log\left(\frac{10000}{1}\right) = 10 \cdot 4 = 40 \text{ dB}$$

For sound pressure levels, using RMS amplitude of sound pressure:

$$L_p = 10 \cdot^{10} \log\left(\frac{I}{I_0}\right) = 10 \cdot^{10} \log\left(\frac{p}{p_0}\right)^2 = 20 \cdot^{10} \log\left(\frac{p}{p_0}\right)$$

where L_p is the level of sound pressure, in dB, p is the RMS amplitude of the target sound, and p_0 is the RMS amplitude of the reference sound (see above), $p_0 = 2 \cdot 10^{-5}$ Pa.

A doubling of the sound pressure, that is, of the RMS amplitude (factor 2, ratio 2) means an increase of the sound pressure level by +6 dB:

$$20 \cdot^{10} \log\left(\frac{2}{1}\right) = 20 \cdot^{10} \log(2) = 20(0.30103) = 6.0206 \approx 6$$

A doubling of the intensity means an increase of the sound intensity by +3 dB:

$$10 \cdot^{10} \log\left(\frac{2}{1}\right) = 10 \cdot^{10} \log(2) = 10(0.30103) = 3.0103 \approx 3$$

By definition, a value of 0 dB means that the target sound has the same sound pressure and/or same intensity as that of the chosen reference sound.

Question 1.2

Check that the last sentence above holds true.

Table 1.1 shows some examples of sound intensities under various circumstances (Fry, 1979, 94).

Table 1.1: Examples of sound intensities in dB (SPL), in various situations.

dB (SPL)	situation
192	Loudest sound possible in air, rarefactions are vacuum (https://en.wikipedia.org/wiki/Sound_pressure#Examples_of_sound_pressure)
172	Krakatoa eruption, 1893 (at ~160 km distance) (Henderson, 2023, 43)
130	Jet engine (at 30 m distance)
120	Threshold of pain
103	Recommended maximum level in clubs and venues in the Netherlands (average over 15 min, peaks may be higher)
100	Symphony orchestra playing 'fortissimo'
92	Outside on emergency lane of busy highway (200 cars/min, 60 mph) (Dahl et al., 2007)
85	Hearing protection is obligatory for employees in the Netherlands
80	Shouting, singing (at 1.5 m distance); street noise
65	Pedestrian city area
60	Average conversation (at 1.5 m distance)
40	Quiet room
35-40	Outside in residential area, at night
30	Whispering
27	Outside in Hermit Basin, at the bottom of the Grand Canyon, USA (Dahl et al., 2007)
20	Background inside audio studio
16	Average threshold of hearing for 50% of listeners (Fletcher, 1953, Ch.8)
0	Reference value; absolute threshold of sine wave sound with $f = 1000$ Hz

1.8.4 Phase

The phase of a sound wave (symbol φ) is the starting time of a sound wave period, relative to the duration of that period. It's easiest to explain by comparing two sounds. When listening, a single sound will arrive at our two ears at slightly different arrival times. (Unless the sound source is directly behind or in front, the sound will have a slightly longer path to travel to the further ear than to the nearer ear.) Thus the two sounds heard by the two ears will differ in phase: the starting time of a period in one ear will differ slightly from the starting time of a period in the other ear, and the difference can be expressed as the proportion of a period by which they differ.
The brain of the listener uses this phase difference between the two ears to estimate the direction of the sound source relative to the head. You may appreciate the effect by listening to a music record in mono or in stereo.

In addition, we use phase unconsciously to assess atmospheric and acoustic conditions. For example, when listening to a sound in a room, we hear not only the direct sound but also the indirect reflections from the floor, walls, ceiling, furniture, people, etc. The brain uses the phase relations among multiple reflections to assess the dimensions and conditions of the room.

Phase is expressed relative to the period T, but it is not expressed in time (seconds) but in proportions, often expressed as degrees in the period cycle (which runs from 0° to 360°). So, a phase difference of $\varphi = 180^\circ$ and $\varphi = 0.5$ mean the same: the time difference between the two signals is half a period, whatever the duration of that period is.

In the transverse stadium wave, phase corresponds to the difference in time between the sit-down-moment of one group of persons, and the comparable sit-down-moment of another group of persons in a different section of the stadium. Imagine two waves rolling along the stadium benches: one wave on the lower benches, and a different wave on the upper benches. The two waves may be out of phase (lower and upper persons sit down at different times) or in phase (lower and upper person sit down at the same time) – irrespective of whether the two waves have the same or different frequencies.

Question 1.3

Continue this thought experiment, with two stadium waves having the same frequency on the lower and upper benches, and with phase $\varphi = 0.5$ between the lower and upper sections. What would the resulting wave pattern look like?

1.8.5 Wavelength

The wavelength (symbol λ) is the length of a single cycle or period, as a distance in the medium in which the sound wave propagates, between repeated patterns

in the wave. It is expressed as a distance in meters. The wavelength depends on the propagation speed c of the sound wave in m/s (see §1.4), and on its frequency f in Hz (see §1.8.1).

$$\lambda = c/f$$
$$\lambda = c\,T$$

Sounds with higher frequency have shorter wavelength, and vice versa. A periodic sound with $f = 440$ Hz has a wavelength in air of $\lambda \approx 343/440 = 0.7795$ meter.

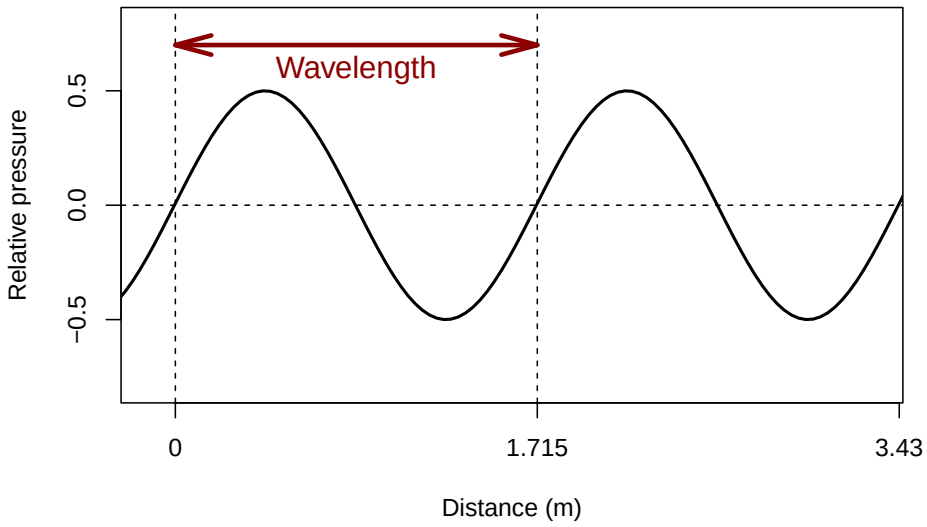

Figure 1.3: Snapshot of the pressure of a sound wave in air, varying with distance from the sound source (along the horizontal axis). The snapshot was taken at a fixed moment in time, with $c = 343$ m/s in air.

Question 1.4

What is the frequency of the sound depicted in Figure 1.3?

Answer 1.4

$f = c/\lambda = 343/1.715 = 200$ Hz [(m/s)/(m) = 1/s]

In the transverse stadium wave, the wavelength λ is the distance in meters (on the same bench) between two persons in two consecutive periods who reach the highest point of their sit-stand-arms-up cycle. This is the distance the wave has traveled between the two moments at which a person repeats the same action. If we assume that $c = 10$ m/s and $f = 0.2$ Hz (as very rough estimates), we find that the wavelength $\lambda = c/f = 10/0.2 = 50$ meter.

1.9 How to work with Praat

Praat is a computer program designed to process, analyze and visualize speech sounds.

After starting the program, Praat opens two windows: an *Objects* window (typically on the left) and a *Picture* window (typically on the right).

Objects

In Praat, signals and derived representations are all seen as objects. Objects may have different types, e.g. Sound, Spectrum, Pitch, etc[11].
Each type of object comes with pre-defined operations that are possible. If you select an object of a different type, then the buttons (operations) change with the object type. As an analogy, consider the various types of objects in your room: clothing items and human bodies may be washed, food may be cooked but human bodies may not be cooked, furniture and food items may be opened but human bodies may not be opened, clothing can be inside furniture, etc. Moreover, relations between object types are also specified: for example, a plant can become a food item (by means of cooking), but a human body may not.

Before working with an object, you need to select that object, by clicking on it in the list of objects displayed in the Praat Objects window.

Objects of any type may be saved and opened using the Save and Open options in the top menu of the Objects window. This is a great way to save the objects resulting from your phonetic analyses, i.e., to save your results.

The buttons at the bottom of the Objects window are *always* available for objects of any type: Rename..., Copy..., Inspect (to take a deeper look), Info, and Remove.

We will often work with objects of the *Sound* type. Such a Sound object is a digital sound (sampled audio), which you can Play or Scale or Convert or Combine, etc. You may analyze a Sound, which will typically result in an object of a different type (e.g. Pitch). Sound objects can be opened from disk, and they can be saved as audio files in a wide variety of audio formats.

Picture window

Praat will draw its visualisations (figures) in its Praat Picture window. The figure will be scaled to the area with the pink boundary, the so-called viewport.

[11] By convention, object types are written with a capital; this helps to distinguish physical properties (e.g. the intensity of a sound) from the Praat representations of those properties (e.g. the Intensity object computed from a Sound object, both within Praat).

By changing the viewport after drawing a part of a figure, you may obtain multiple visualizations in a single figure, as will be illustrated in this tutorial.

The combined figure in the viewport may be saved (`File > Save`) or printed (`File > Print`) in the top menu of the Praat Picture window. You may also save the figure in a different way, as a "recipe" set of instructions to re-create the figure (`File > Save as Praat picture file`), for later reuse.

1.10 How to draw an oscillogram

1.10.1 Open an audio file

- In the `Praat` Objects window, go to the top menu, choose `Open`, then `Read from file...` and pick the desired audio file. It will be loaded as a Sound object at the bottom of the list of objects in the `Praat` Objects window.

For the time being, selecting a short audio will make it easier for you to explore how `Praat` works; choose an audio file containing only a few seconds of sound.

- **The last-added object in the list is selected automatically.**
- In the `Praat` objects window, go to the bottom menu, choose `Info`. If it does not exist, `Praat` will open an additional window, titled `Praat Info`. In this window you'll see all kinds of meta-data (data about data) of the selected Sound object, with numbers in scientific notation (see below).

1.10.1.1 Scientific notation

The scientific notation of a number such as `9.876e-2` is to be read as $9.876 \times 10^{-2} = 9.876 \times 0.01 = 0.09876$; for background see https://en.wikipedia.org/wiki/Scientific_notation.

1.10.2 Draw an oscillogram

- In the `Praat` Objects window, select a Sound object.
- Next, in the `Praat` Object window, choose `Draw > Draw...` and use default values for all arguments or options (button `Standards`).
- After pressing `OK`, the dialogue menu disappears, and the oscillogram of the Sound is drawn in the selected viewport area in the Picture window, with some basic labeling of axes (if option `Garnish` is on).

- If you are not happy with the resulting figure: in the `Praat` Picture window, choose `Edit > Erase all`, or `Edit > Undo`, and start again at the top of this subsection.

1.10.3 Save a picture

- Check the page setup: in the `Praat` Picture window, choose `File > Page setup...`, and check the settings.
- You can save the selected Viewport area in several formats: `PDF` is best for publication-quality figures. `PNG` format is best for on-screen images (blogs, web pages, documentation). A graphical resolution of 300 pixels per inch-on-screen (dots per inch, or dpi) if often adequate, since that graphical resolution is typically better than most computer screens (but you can try out 600 dpi for your setup). `EPS` format is nowadays less popular, but may offer flexibility for downstream image processing (check documentation about this format).
- Store the graphical (image) file in an appropriate folder for your project on your computer, under an appropriate file name, that will still make sense to you in a year's time. Make a note in your journal about where you've saved this file, and about the contents of the file.

Chapter 2

Converting sound to bytes

Chapter keywords: analog-to-digital conversion, digital-to-analog conversion, AD, DA, ADC, DAC, microphone, sound insulation, directional, filtering, sampling, sampling frequency, nyquist frequency, amplitude resolution, quantization, rounding, noise, recording level, gain, clipping, audio formats, codec, lossy, lossless, scale, fade, concatenate, chain.

2.1 Overview

In order to process sounds by means of a computer program, or telephone, we first need to convert that sound, the variations in air pressure, to numbers that are then further processed by a computer or by a telephone device. This is a two-step process, involving at least two key components in order:

(1) the **microphone**: this device transforms variations in air pressure into matching variations in an electrical signal. The microphone has a thin membrane, and displacements of the membrane (caused by the sound pressure wave hitting the membrane) are transformed into proportional fluctuations in electric current (Ampere), electric voltage (Volt) or electric resistance (Ohm), depending on the design of the microphone. For instructions about how to handle a microphone, see the text box in §2.2) below. The analog electrical signal is then passed on from the microphone to…

(2) the **analog-to-digital-converter** (ADC): this device converts a continuous, analog electrical signal into a stream of discrete, digital numbers. The ADC repeatedly measures the input signal, and reports the digital output value of that input signal. This process is also called 'sampling'. Sampling a signal is done with a certain 'sampling frequency' (number of repeated

measurements per second) and with a certain precision of measurement (known as 'amplitude resolution'), both explained below. The result is an output stream of digital numbers (in bytes), to be handled further by computer software (e.g. to be displayed, compressed, transmitted, stored, played back, etc.)[1] [2]

Very soon, whenever you want to hear sound from a computer or from a telephone connection, you will also need

(3) a **digital-to-analog-converter** (DAC): this device converts a stream of discrete, digital numbers into a continuous analog electrical signal, with a pre-specified conversion frequency and amplitude precision. The result is an output analog electrical signal, to be handled further by audio hardware (e.g. to be amplified, sent to a loudspeaker, etc.)

2.2 How to handle a microphone

- A good microphone is a very sensitive and very expensive device. Treat it with great care. Never blow into a microphone (it's far better to just say `test` or `check` or anything with plosive and fricative consonants). Do not tap on its surface.
- Do not plug or unplug the microphone into/from a "hot" port. First set the port's input/output volume to zero, then plug/unplug.
- Do not speak *into* the microphone, but just over it or alongside. The microphone should measure sounds, but *not* the flow of air coming out of a speaker's mouth and nose. If the microphone comes with a foam cap to dampen airflow, then use it.
- Do not touch the microphone while it is picking up sound; this will result in undesired (and often loud) contact sounds in the output signal.

2.3 Key parameters in AD conversion

The digital signal obtained by analog-to-digital conversion is an approximation of the original (analog) sound. Two key parameters determine the accuracy of the digital approximation, and thus the quality of the digital sound recording.

[1] The input signal to be sampled often comes from a microphone, but other signals may also be sampled, e.g. the signal coming from an electro-encephalogram (EEG) electrode.

[2] In a speaker's telephone, the stream of numbers (output from the ADC) constitutes the input for subsequent processing and data compression, even before speech data are transmitted to the receiving phone.

The first parameter is the number of samples taken per second: the *sampling frequency*, and the second parameter is the number of bits used to describe the amplitude value of the sample: the *amplitude resolution*. These two parameters are further explained below.

2.3.1 Sampling frequency

The sampling frequency (symbol f_s) is the frequency with which digital samples are taken and stored from the original analog sound. With a higher sampling frequency, the digital signal better (more closely) approximates the analog source in the *time* dimension, resulting in a better digital recording. The sampling frequency is expressed in samples per second, in Hertz units (cf. §1.8.1). A sampling frequency of 2 kHz (2000 Hz) means that the sound is sampled 2000× per second.

The sampling frequency f_s must be at least 2× the highest frequency f in the analog source sound. Thus, the source sound may not contain any components with frequencies above $f_s/2$, the so-called 'nyquist frequency' (this requirement follows from the so-called 'Nyquist theorem'). In practice this requirement is guaranteed by *low-pass* filtering the source sound (see §4.2 about filtering), with the nyquist frequency as cutoff, thus removing any components with frequencies higher than the nyquist frequency. This filtering is routinely done before AD conversion, by the AD conversion hardware.

For speech, most acoustic information is contained in the frequency range up to 8 kHz. Given the previous paragraph, this means that we need a sampling frequency of at least 16 kHz[3] or higher. In most phonetic projects, the most relevant phonetic information is contained in the frequency range up to 16 kHz, for which a sampling frequency of $f_s = 32$ kHz[4] is adequate. For music, relevant information may be contained in the full audible range up to 22 kHz, and the standard sampling frequency is 44.1 kHz[5].

- Check the sampling frequency before making a digital recording, set it to an appropriate value, write down the sampling frequency in your lab journal, and mention it in your report.
- Using a higher sampling frequencies will result in proportionally larger digital sound files, which require longer processing times and more computer storage.

Aside: In the past, analogue telephones used a limited bandwidth of 300 to 3400 Hz; this allowed low-cost transmission of speech that was still intelligible. Such 'telephone speech' may be sampled at $f_s = 8$ kHz, capturing acoustic

[3] This is the typical sampling frequency in VoIP, "wideband speech".
[4] This is the standard sampling frequency for FM radio.
[5] This is the standard sampling frequency for audio CDs.

information in the frequency range up to 4 kHz. You may encounter legacy speech recordings sampled at 8 kHz, and/or research reports mentioning this f_s value.

2.3.2 Amplitude resolution

The amplitude resolution, or quantization, refers to the number of separate steps in amplitude (voltage) that are discerned during sampling. Again, with a higher amplitude resolution, the digital signal better (more closely) approximates the analog source in the *amplitude* dimension, resulting in a better digital recording. The amplitude resolution is expressed in bits[6] or bytes[7].

The recorded amplitude values are discrete, and because of the "jump" from one discrete amplitude step to the next-higher or next-lower value, the amplitude values are "rounded" to some extent. This rounding or quantization results in audible noise in the digital signal. This rounding noise amounts to half a step of possible amplitude values. If we have more amplitude values (higher amplitude resolution) then the rounding off becomes less noticeable[8].

In phonetics, the most common amplitude resolution is 16 bits, or $2^{16} = 65536$ different amplitude steps[9]. The quantization noise has an amplitude of 1/65636 of the maximum amplitude; this corresponds to a signal-to-quantization-noise ratio of about 98 dB. This small amount of rounding noise is negligible.

2.4 How to record a sound

For any audio recording, there are a few **essential precautions** that you'll have to attend to, in order to obtain high-quality recordings suitable for subsequent analysis and re-distribution.

2.4.1 Remove non-target sounds

In order to obtain a high-quality recording, it helps to attenuate all non-target sounds, in various ways:

[6] 1 bit or binary digit is a single digit in the binary system. A binary digit can only have 2 possible values, 0 or 1 (just as a decimal digit can have 10 possible values, 0 to 9).

[7] 1 byte is 8 bits, or $2^8 = 256$ possible values.

[8] Notice that one bit is required to record the sign of the value (positive or negative), so with 1 byte of resolution we can in principle record 256 possible amplitude values, running from -128 to +128, with rounding noise having an amplitude of 0.5/128 or 1/256 of the maximum amplitude. This corresponds to a signal-to-quantization-noise ratio of 50 dB. In practice, however, amplitude values are not stored as integer numbers but as floating numbers.

[9] This is also the standard amplitude resolution for audio CDs.

- If available, use a sound-attenuating cabin or booth. The booth will help to insulate your target signal from unwanted other sounds. Close the door of the booth properly. Leave non-essential equipment (watches, phones) outside the booth.
  If a booth is not available, then find the quietest space available. Try using thick curtains, carpets, and cushions, and other sound-dampening materials, to improve your recording. Make lots of test recordings, listen critically, and attempt phonetic analyses before you proceed with your recordings.

Background: Phonetic analyses typically aim at finding acoustic properties in the speech signal that may be related and relatable to the speaker's articulations and prosody. However, similar spectral properties (e.g. resonances, formants, see §5.2) may also arise from acoustic reflections in the recording room, and it may be difficult or impossible to disentangle similar spectral properties coming from different origins. Hence it helps to minimize any acoustic reflections in the recording room[10].

- If there is a lot of noise or non-target speech, try using a *directional* microphone, which only or mostly picks up sounds coming from one direction, and which attenuates sounds from other directions. Vary the position of the microphone and make test recordings.

- Switch off any non-essential equipment, and try to attenuate non-target sounds from elsewhere. Even if you cannot hear a difference, the equipment sounds and outside sounds may interfere with the target sound signal, resulting in unwanted artefacts.

- Despite all these precautions, outside sounds may still interfere. This happens in particular with low-frequency signals, e.g. due to traffic outside, elevators elsewhere in the building, and so forth. These interfering sounds are typically outside the frequency range of speech. Therefore we can easily remove them, by high-pass filtering the target speech signal, *before* DA conversion.

  - Use a high-pass filter that will discard frequencies below a certain cut-off frequency (see §4.2).
  - Set the cut-off frequency well below the lowest possible frequency (component) in your target speech, say, a cutoff frequency of about 50 to 60 Hz.

[10] *Frequency* and *formant* measurements may be suspect if the corresponding *wavelength* (for frequencies) or wavelength ×4 or wavelength ×2 (for formants) corresponds to one of the dimensions of the room (see §1.8.5 and §5.2.3); or if the reported formant frequency is below 200 Hz; or if the reliability of the frequency measurement is low.

2.4.2 Check recording level

During your recording, check the level of the recording.
If the recording level is too low, then the target signal is too weak, and the background noise (including quantization noise) is relatively strong. It's very difficult or impossible to fix the signal-to-noise ratio later, so you need to **fix this now**, during the recording.
If the recording level is too high, then the loudest portions of the target signal will be too strong, leading to "clipping" or distortion. It is impossible to fix this later, so you need to **fix this now**, during the recording.
There are several ways to adjust the level of the recording: - by adjusting the level of the input channel (maybe called `Gain`), in your computer settings, - by varying the distance from the sound source to the microphone (in general: the closer the better, but also depending on the type of microphone), - in speech: by instructing the speaker to speak more loudly or more softly.

2.4.3 Avoid lossy audio formats

It it tempting to record sounds digitally on smart devices which store lots of audio in compressed (lossy) formats such as MP3 or MP4. However, the lossy compression of audio data may lead to difficulties in subsequent phonetic analyses. The results may sound quite right to your ears, but details in timing or in spectral details may nevertheless have been lost in the compression. It depends on your interests and your research questions whether or not this constitutes a problem. For phonetic research, it is generally better to record in 'lossless' audio formats that do not compress the audio data, rather than in 'lossy' formats. Check whether and how your smart device can make lossless recordings: if possible at all, this will probably require a deep dive into the settings on the recording device.

2.4.4 On your computer, using `Praat`

We assume that you use a computer equipped with a **microphone**, or with an analog **input port** for an analog signal coming from an external microphone. If using an external microphone, plug it into your computer (see §2.2). Check that audio input is arriving in your computer, using your computer system settings for sound input (if using an external microphone, select the appropriate channel).

There is a helpful option in `Praat`, from the *main* menu (not from the Objects window menu): `Praat > Settings > Sound recording settings...` where you may adjust certain settings as needed.

2.4.4.1 Record

- In the `Praat` Objects window, select `New > Record mono sound`. In most situations it is not necessary to record stereo sounds.
- Choose the appropriate sampling frequency, e.g. 22050 Hz (see §2.3.1). You may receive an error message if the chosen sampling frequency is incompatible with your computer.
- Click `Record`, and speak a test sentence into the microphone, e.g. *The source of the huge river is the clear spring*[11], or make a test recording of your sound source. After testing, click `Stop`.
- **While recording, check the level of the recording** (§2.4.2).
  In `Praat`, make sure that your recording level is in the yellow zone, with occasional peaks in the red zone, but without clipping.
- Click `Play` to listen to your recording. **Repeat the recording** until the recording level is good.
- Enter a name for the recording, e.g. `river`, in the lower right corner (this name will be used within `Praat`).
- Save the recording: `Save to list` (i.e., to the list of objects in the `Praat` Objects window).

Your speech recording is now an object within `Praat`.

2.4.4.2 Save

For storage, you should save this object as an audio file on your computer's hard disk.

- To do so, in the `Praat` object window, select the Sound object that you just recorded. (Normally this new object has been added at the BOTTOM of the list of objects).
- From the top menu, choose `Save > Write to WAV file...` or choose any of the other audio formats. Save the Sound object in a folder and under an unambiguous name that you will remember and understand a year from now – not just `river.wav`. Note in your journal the folder and filename of your sound recording. Keep projects in separate folders on your computer.

[11] One of the so-called Harvard test sentences, from list 3, https://en.wikipedia.org/wiki/Harvard_sentences. In Dutch, a popular test sentence is *Het leven is mooi als de zon schijnt.*

2.4.4.3 Open

- In order to open an audio file from your computer hard disk, from the top menu in the Objects window, choose `Open > Read from file...`, and pick the target audio file.

2.5 How to play back a digital sound

Once again, we assume that you use a computer equipped with an analog **output port** for playing back sounds. Check that audio output is audible, using your computer system settings for sound output to your loudspeakers or headphones.

There is a helpful option in `Praat`, from the *main* menu (not from the Objects window menu): `Praat > Settings > Sound playing settings...` where you may adjust certain settings as needed.

In the Objects window, select the Sound object(s) that you want to play back. Then press the `Play` button in the Objects window. You might interrupt the playback with the `Esc` key, but this depends on your playback settings (see the paragraphs above). (If you have selected multiple Sound objects, then they will be played consecutively, *without* a pause in between; see §2.6.4.)

2.6 How to manipulate and edit a digital sound

2.6.1 Scale

Even if you have heeded all the warnings in this chapter, it may be necessary to adjust the amplitude or intensity of your digital recording. Sample values in the digital sound will be multiplied with a certain scale factor. `Praat` has different recipes to choose this scale factor.

2.6.1.1 Scale amplitude

`Modify > Scale peak...`
The amplitude scale factor is chosen so that the maximum sample value is a particular proportion $\times$ the maximum *possible* amplitude value (which in `Praat` is 1 by definition). Sensible values for this proportion are 0.99 (the default) or 0.995, resulting in peak values which will be 0.2 dB or 0.1 dB below the maximum possible amplitude value, respectively. (If you choose a proportion > 1, the result will contain clipped samples, with amplitude values exceeding the maximum possible value. This will lead to distorted sound when played back. Do not choose a proportion > 1 here.)

2.6.1.2 Scale intensity

`Modify > Scale intensity...`
The amplitude scale factor is chosen so that the average *intensity* of the resulting sound is the specified target intensity in dB SPL (see §1.8.3). `Praat` warns you that the result may contain clipped samples, with amplitude values exceeding the maximum possible value. Such clipped samples will lead to distorted sound when played back.

2.6.2 Excise part of a sound recording

In many projects, we wish to excise (cut out) a part of a recording, and save that part as a separate audio file. `Praat` contains a so-called SoundEditor to perform these tasks.

Reliably segmenting or delimiting fragments of a speech utterance requires the background knowledge provided in the following chapters — and practice. Segmenting speech will also be discussed in Chapter 7. Below are some brief starter instructions.

- Select an input Sound object in the `Praat` Objects window.
- Choose `View & Edit`. This will open a SoundEditor window. Figure 2.1 shows an example of a SoundEditor, displaying the fragment *…a small window to get this…* spoken by a male speaker of American English; the single word *window* is highlighted.

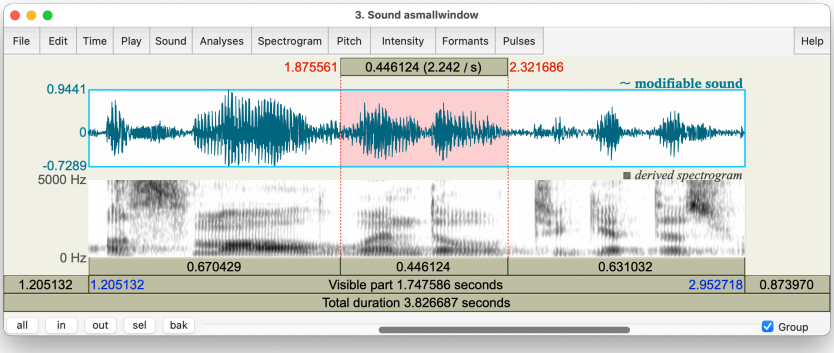

Figure 2.1: SoundEditor window, showing oscillogram and spectrogram of the fragment *...a small window to get this...*, with the word *window* highlighted.

The SoundEditor in Fig.2.1 shows a fragment of speech, taken from the audio file named `10-18-17_Council_SLASH_10-18-17_Council_DOT_HD_DOT_mp3_00285.flac` from the *Peoples Speech* corpus at https://huggingface.co/datasets/MLCommons/peoples_speech. For details about that corpus, see Galvez et al. (2021).

The SoundEditor contains a lot of information! In the top panel you always see an oscillogram (§1.6). Below, there may be additional and overlapping visualisations of the Spectrogram (Ch.6), Pitch, Intensity, Formants (§5.2), and/or voice Pulses. These additional visualisations will be discussed later in this tutorial guide; for now, you may switch them `off` by choosing the appropriate buttons in the top of the SoundEditor window, and then unselecting the `Show` option.

- In the SoundEditor window, you can **select** a part of the recording with your mouse. The selected area is marked in pink in the oscillogram panel of the SoundEditor window, as shown in Fig.2.1.
- You can **play back** the selected part, or the parts before/after it, by clicking the appropriate buttons (labeled with the durations of those parts). For more information about audio playback from `Praat`, consult §2.5 above.
- In the SoundEditor window, you can use the buttons in the lower left corner to **zoom** `in`, `out`, to zoom out to the entire sound (`all`), or to zoom to the `sel`ected part.
- You can also select a part of the recording by specifying the start and end times explicitly: choose `Time > Select...` and enter the start and end times.
- By convention, the selected part should begin and end at **positive zero crossings** (where the signal changes from negative to positive; see §2.6.3 below). You can find these precise places as follows (disregarding shortcuts in `Praat`):
  - put the cursor in the approximate area where the boundary should be located;
  - choose `Sound > Move cursor to nearest zero crossing`;
  - ask `Praat` to report the cursor location in its Info window: `Time : Get cursor`, and note this time;
  - do this for both start and end of the selection part;
  - use these noted times to set the selection part as described above.
- In this way, the word *window* was selected in the SoundEditor in Fig.2.1.
- To **save** the selected part as a separate audio file: in the SoundEditor, choose `File > Save selected sound as WAV file...`. Provide the appropriate folder and file name, and click `OK`.

- You may also **extract** the selected part into a new Sound object in `Praat`. In the SoundEditor window, choose `Sound > Extract selected sound (preserve times)` if you wish to maintain the same times as in your source recording: the new Sound object will have a starting time equal to the starting time of the selection in the source recording. If you want your new Sound to start at zero, then choose `Sound > Extract selected sound (time from 0)`.
- The new Sound object may need a fade-in and fade-out (see §2.6.3 below).
- Remember to save this new Sound object (see §2.4.4.2) if you wish to keep it.

2.6.3 Fades

An abrupt change in amplitude may occur at the beginning and/or at the end of a sound file. Such an abrupt change will be heard as an impulse sound (see §3.3.2), that is, as a clicking sound. This may happen with speech cut out of a longer speech utterance. Due to the workings of the ear, such an impulse may briefly mask subsequent sounds, i.e., affect the perception of subsequent sounds. In order to avoid these artefacts, it is common practice to enforce a gradual fade-in (at the onset, from zero) and fade-out (to zero). This is especially important if you intend to play back the sound to listeners.

These smooth fades can be made in `Praat` by choosing `Modify > Fade in...` and `Modify > Fade out...`, with a fade duration of about 0.005 to 0.010 second. Samples affected by the fade are multiplied with a factor increasing from 0 to 1 during fade-in, or decreasing from 1 to 0 during fade-out.

2.6.4 Concatenate

In phonetic research, we often want to construct a "chain" of sounds in a particular order e.g. (1) a warning beep, (2) a silent portion of fixed duration, and (3) a speech stimulus. This can be achieved in several ways; the most basic operation is to "concatenate", that is, to create a "chain" of sound objects.

In the `Praat` Objects window, select the Sound objects *in the order in which they are to be concatenated*, e.g. (1) beep, (2) silence, (3) stimulus. In order to select multiple Sound objects from the list, in a specified order, press `Command` while selecting objects. Then choose `Combine > Concatenate` in the Objects window. The resulting "chain" of Sounds will be added to the BOTTOM of the list of objects. Remember to save this new Sound object (see §2.4.4.2).

Chapter 3

Complex sounds and spectra

Chapter keywords: sinewave sound, complex sound, spectrum, fourier transform, fourier analysis, fast fourier transform, FFT, harmonics, fundamental, fundamental frequency, f0, overtone, component, timbre, octave, noise, white noise, brown noise, impulse.

3.1 Introduction

The *sine wave*, depicted in Fig. 1.2, is the simplest sound possible. It is composed of the simplest back-and-forth variation or oscillation in air pressure, similar to the regular swing pattern or oscillation of a pendulum[1]. We only encounter sine wave sounds if they are artificially generated, and hardly ever in nature – although the sound of a tuning fork comes quite close to a sinewave pattern.

By contrast, *complex sounds* have more complex wave patterns. All natural periodic sounds are complex sounds. **A complex sound can be regarded as the sum of multiple sine wave sounds.** This relation has been described by the French mathematician, baron J.B.J. Fourier (1768–1830). The sine waves are termed 'frequency components' of the complex sound. Each of these components has its own frequency, amplitude, and phase. In a so-called 'Fourier analysis' or 'Fourier Transform' of a complex sound, these frequency components are being estimated from the waveform.

If the complex sound has a repeating waveform, then we have a periodic complex sound, of which Figure 3.1 provides an example. The resulting sound has

[1] Drawing the position of a swinging pendulum over time will result in the same figure.

been obtained by adding three frequency components, drawn in dotted lines, of 100 Hz ($T = .01$) and 200 Hz ($T = .005$) and 400 Hz ($T = .0025$), respectively. Note that the frequency with which the complex sound repeats itself, 100 Hz, is the same as that of the lowest component. This lowest component is called the *fundamental*, and its frequency is called the *fundamental frequency* (symbol f_0) of the complex periodic sound; we hear this f_0 as its pitch. The higher components are called *overtones*. The fundamental and overtones are collectively called *harmonics*: the fundamental is the first harmonic, the first overtone is the second harmonic, etc. **In a periodic complex sound, the frequencies of the overtones are integer multiples of the fundamental.**

TODO crossref pitch, missing fundamental

Typical examples of periodic complex sounds are the vowel sounds in normal speech. The properties of a periodic complex sound depend on the amplitudes, frequencies and phases of its component harmonics.

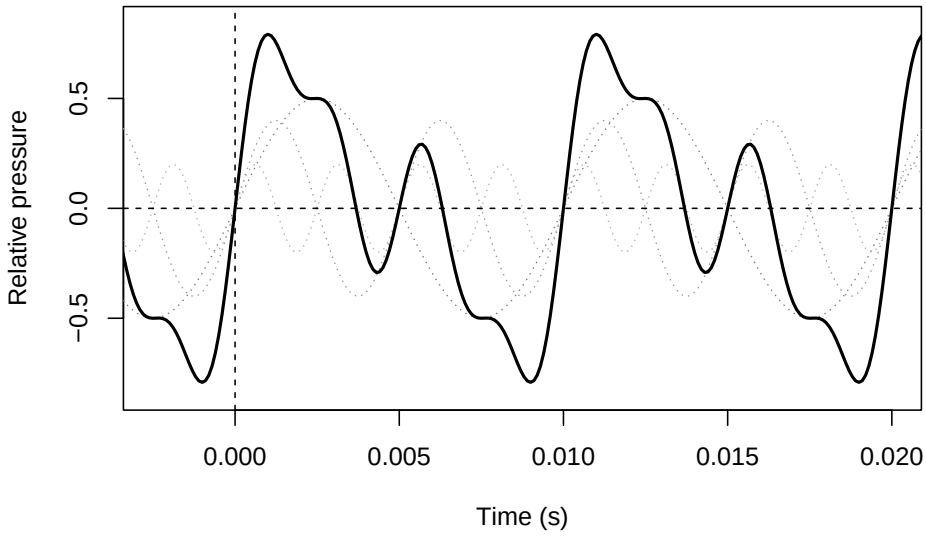

Figure 3.1: Oscillograms of three sinewave sounds and their resulting complex periodic sound.

3.1.1 Timbre

Two periodic complex sounds, having the same overall amplitude and fundamental frequency, may differ strongly in their character. The general name for this property is *timbre*. Timbre depends on the relative amplitudes of the harmonics, and hence very many different timbres are possible. For aperiodic

sounds, timbre also depends on the relative amplitudes of the (infinitely many) frequency components. A sound may have a dull or sharp timbre, or rich or thin, warm or metallic. The difference between distinct vowels, such as /a/ vs. /i/, spoken by the same person at the same pitch and amplitude, is also a matter of timbre, as is the difference between similar but distinct consonant sounds, such as /s/ vs. / /.

Timbre is not a one-dimensional property of a sound (as frequency and amplitude are), but a multi-dimensional property.

3.2 Spectrum

A sound can be represented in two equivalent ways: as a function of time (in an oscillogram), or as a function of frequency. The latter representation is called a *spectrum*. A spectrum is useful to assess the frequency components of a signal, which are far easier to determine in a spectrum than in an oscillogram. A rainbow reveals the spectum of sunlight: water droplets refract the incoming light into its frequency components (colors). Similarly, complex sounds may be broken down into their (high and/or low) frequency components.

Figure 3.2 shows an oscillogram on the left (of the same complex sound as shown in Fig. 3.1), and its matching spectrum on the right. The spectrum shows the amplitude along the vertical axis, of each frequency component along the horizontal axis. (The phases of the frequency components are ignored.) Thus, a spectrum shows the frequency and amplitude of each component.

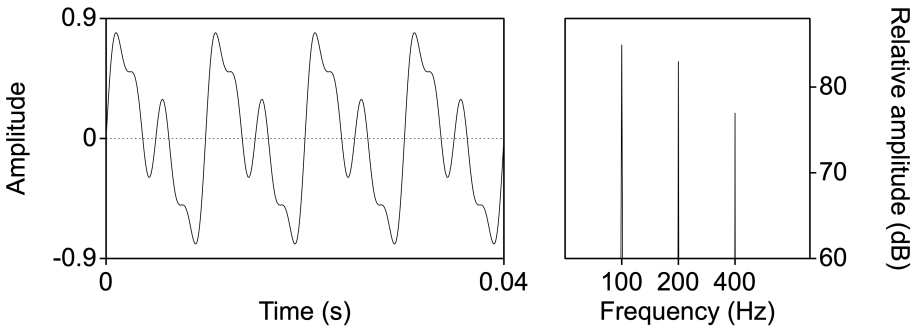

Figure 3.2: Oscillogram (left) and spectrum (right) of a complex periodic sound.

Questions

Question 3.1

Draw the spectrum of a sinewave sound with a frequency of 450 Hz and an amplitude of 40 dB SPL.

3.2.1 How to obtain a spectrum

- In the `Praat` object window, select a Sound object.
- Next, in the `Praat` object window, choose `Analyse spectrum >` and then `To Spectrum...`.
- `Praat` offers two versions, regular and fast, of the Fourier analysis to estimate a spectrum. The faster version of the Fourier transform is termed 'Fast Fourier Transform' or FFT, and it requires significantly fewer computations. FFT requires that the number of samples to be analysed is a power of 2. If you choose the `Fast` version by ticking the box, then `Praat` adds zeroes to your sound in order to meet this requirement.
- The result is a Spectrum object calculated *over the entire Sound* (plus additional zeroes, in FFT), and this Spectrum object is added at the bottom of the list of objects.
- Remember to `Save` this Spectrum object if you wish.
- As the spectral representation (spectrum, horizontal axis is frequency) is equivalent with the temporal representation (oscillogram, horizontal axis is time), the Spectrum object may be reconverted again into Sound, by means of "inverse fourier transform". To do so, select the Spectrum object, select button `Sound >` and then `To Sound`.

3.2.2 How to obtain a spectral 'slice'

Only rarely are we interested in the spectrum of an entire Sound object. Typically, we want to inspect the spectrum of the sound only over a small time window within that sound. For example, we might want to inspect the spectrum of the /i/ vowel sound in Figure 1.1, at $t \approx 0.210$ seconds.

This can be done by means of two features in `Praat` which we will explore in depth only later, viz. the SoundEditor and the Spectrogram (Ch.6.

- In the `Praat` object window, select a Sound object.

- Next, in the `Praat` object window, choose `View & Edit`. This will open a so-called SoundEditor window, with the oscillogram as its main feature.

  TODO crossref SoundEditor

- In the SoundEditor window, go to `Spectrogram...` and then `Spectrogram settings`. Here we may need to adjust the setting for `Window length`. `Praat` will average the resulting spectrum over the time window of 2× this value, centered around the position of the cursor in the oscillogram (see §6.4).
  A *short* window (e.g. 0.005 s) will show more detail in the time domain (showing individual periods and transient sounds), but the resulting spectrum will be smeared in the frequency domain (so that individual harmonics are invisible). A *long* window (e.g. 0.015 s) will show less detail in the time domain (so that individual periods and clicks will be smeared in the time domain), but more detail in the frequency domain (so that individual harmonics may be visible). Read §6.4, read the `Praat` Help information available in the menu window, then try various window lengths, and notice the differences in the subsequent spectral slices.

- In the SoundEditor window, go to `Spectrogram...` and then choose `View spectral slice`. As a first result, a new Spectrum object is added in the Objects window. The spectrum in this object is estimated over the window length on either side of the cursor position. Secondly, this new Spectrum object is opened in a SpectrumEditor window, see Fig.3.3 for an example.

- In the SpectrumEditor window, if you click inside the spectrum, the frequency and amplitude coordinates are shown. Placing the cursor on a spectral peak in the SpectrumEditor can be done by means of the button `Spectrum` in the top row, then choose `Move cursor to nearest peak`.

- In Fig.3.3, the individual harmonics are clearly visible in the spectrum. The 10th harmonic is at 2195 Hz, which suggests that $f_0 \approx 219$ Hz (see §3.1). Also notice the overall downward slope of the spectrum (see §4.4 and §5.3.1).

- Remember to `Save` the Spectrum object if you wish.

3.3 Spectra of aperiodic sounds

Stable noise and brief impulses are two types of aperiodic signals (§1.7): the variations in air pressure do not follow a regular periodic pattern[2]. Aperiodic

[2]Or, you might say that the period is infinitely long.

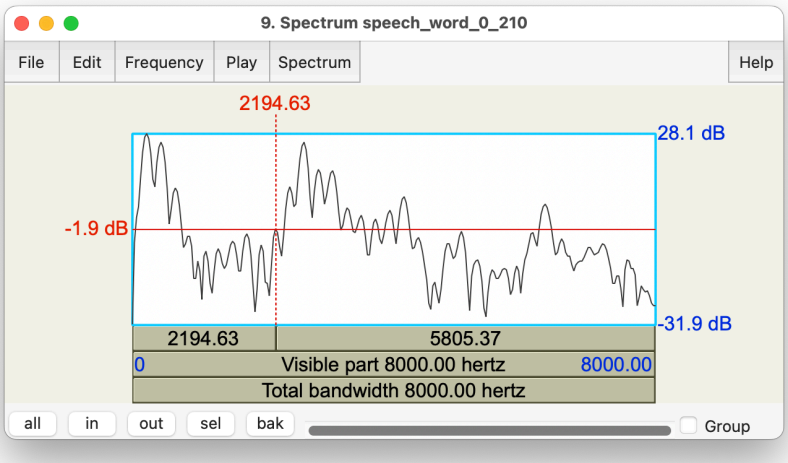

Figure 3.3: Spectral slice of the /i/ vowel in the word *speech*, estimated over 0.015 s on either side of the cursor at *t*=0.210 s.

sounds do not have a fundamental frequency (because there is no regular period), and their phase is undefined, but aperiodic sounds do have an amplitude and a spectral composition.

3.3.1 Noise

First we discuss *stable* aperiodic sounds: **noise**. You might say that a noisy sound has an infinite number of frequency components. That is, the components are not only harmonics of the fundamental frequency (as with periodic complex sounds), but may be found at *every* frequency. The relative amplitudes of the many frequency components determines the timbre of the noise.

In *white noise*, all frequency components are equally strong[3], and thus the spectral envelope is flat. In so-called *brown noise*, the spectral envelope decreases by −6 dB per octave, so that lower frequencies are more dominant than higher frequencies. Because this spectral envelope resembles that of speech (cf. §5.3.1), brown noise is often used in phonetic research whenever we need to mask speech.

The random deviations from the ideal, smooth spectral envelope are due to (a) the random variability inherent in noise, and (b) the fact that the spectrum

[3]This is called 'white' noise by analogy with white light, in which all frequency components in the visible part of the electromagnetic spectrum are equally strong.

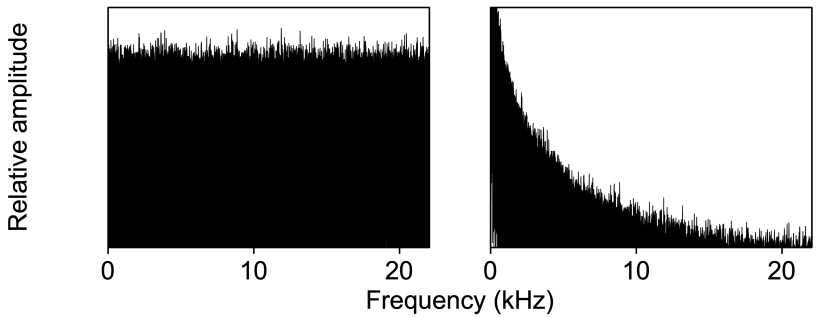

Figure 3.4: Spectra of white noise (left) and of brown noise (right), with a linear frequency axis (in kHz).

was calculated over a finite amount of time[4], with (c) a particular sampling frequency of the noise.

3.3.2 Impulses

An **impulse** is a very brief and *transient* sound, such as a hand clap or tick. Acoustically, a very brief impulse sound is like a brief burst of white noise, with a flat spectral envelope. The shorter the impulse, the flatter the spectral envelope becomes.

An impulse may occur unintentionally if the amplitude suddenly increases from zero to a high value, e.g. at the onset of a sound recording starting at a nonzero value. The resulting noise burst should be effectively removed by *fading in* the sound, see §2.6.3 for more.

3.4 Envelope

The *envelope* of a sound describes how the properties of that sound change over time. This concept is best described by regarding the amplitude of a sound: the 'amplitude envelope'. However, a sound may at the same time have multiple and different envelopes for its amplitude, for its (fundamental) frequency, and for (a singular parameter of) its timbre. Even the properties of a filter may follow an envelope, that is, they may change over time (see Ch. 4).

The concept of the *envelope* of a sound property stems from electronic music (synthesizers); critical time points are the onset and offset of a key being pressed

[4] If you would listen to white noise for an infinitely long time, then all frequency components would indeed be equally strong.

on the keyboard of the synthesizer. However, the envelope is also a helpful concept for describing analog musical sounds (e.g. picking a guitar string) and speech sounds (e.g. plosive vs fricative consonants).

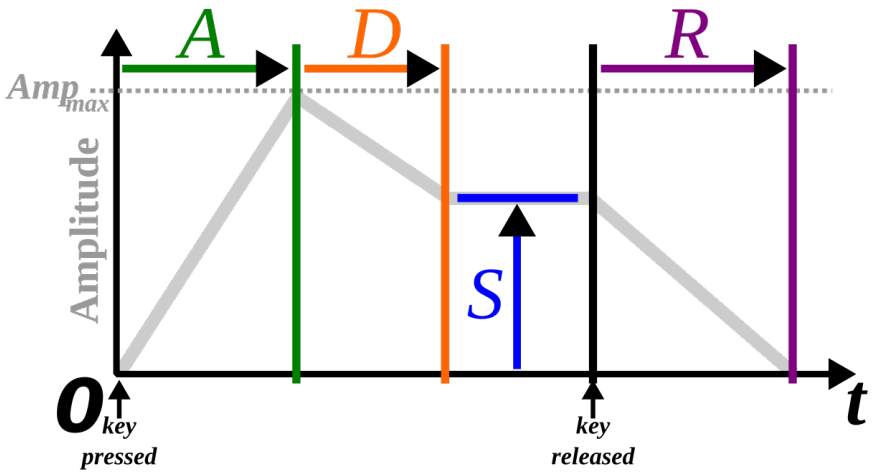

Figure 3.5: A typical amplitude envelope (in gray), with four key parameters Attack, Decay, Sustain, Release describing the changes of amplitude over time, relative to the onset and offset of a synthesizer keyboard key press. Image taken from *https : //commons.wikimedia.org/wiki/File : ADSR*$_p$*arameter.svg*, used under CC-BY-SA license.

TODO: add text

A D S R

In terms of its amplitude, we may regard a brief impulse sound (click or pulse, see §3.3.2 above) as having very short attack and decay times, a zero sustain level, and zero release time.

Chapter 4

Filtering

Chapter keywords: filter, resonance, resonator, Helmholtz, frequency, frequency component, cutoff, slope, bandwidth, emphasis, pre-emphasis, de-emphasis, spectral envelope.

4.1 Introduction

A filter is a device that changes the spectrum of the input signal, by enhancing certain frequency components and/or by attenuating others. Filters play an important role in speech analysis. Filters can be of an acoustic nature (e.g. an organ pipe, or the human vocal tract) or they can work by electronic means. We already encountered filters as a routine component in (or just before) analog-to-digital conversion of a sound wave (§2.3.1) to prevent aliasing.

4.2 Types of filters

There are four different basic types of filters, differing in their frequency characteristics (specifying which frequency components are attenuated and which are enhanced).

- **Low-pass** filters (Fig.4.1 allow lower frequencies to pass through, but higher frequencies are attenuated.

- **high-pass filters** (Fig.4.2) do the reverse: they allow higher frequencies to pass through, but lower frequencies are attenuated.

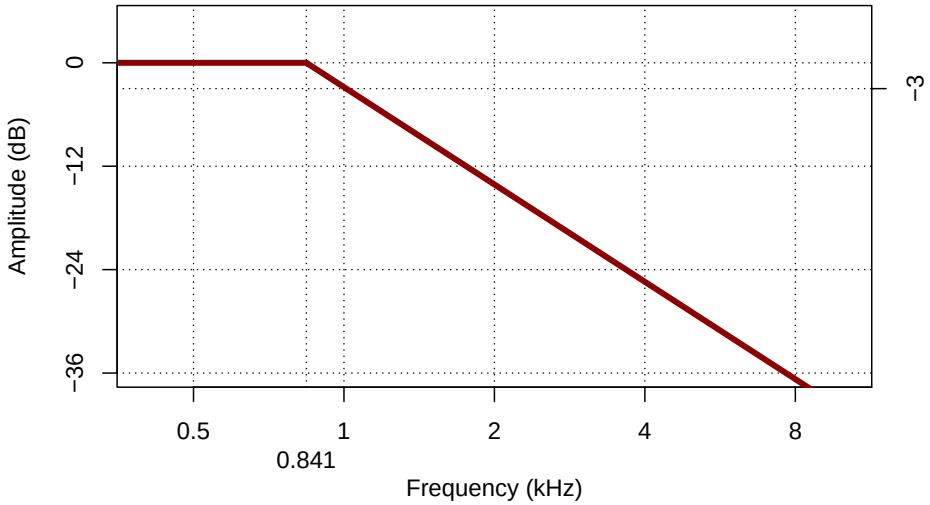

Figure 4.1: Frequency characteristic of a low-pass filter, with cutoff frequency 1000 Hz, and slope of -12 dB per octave.

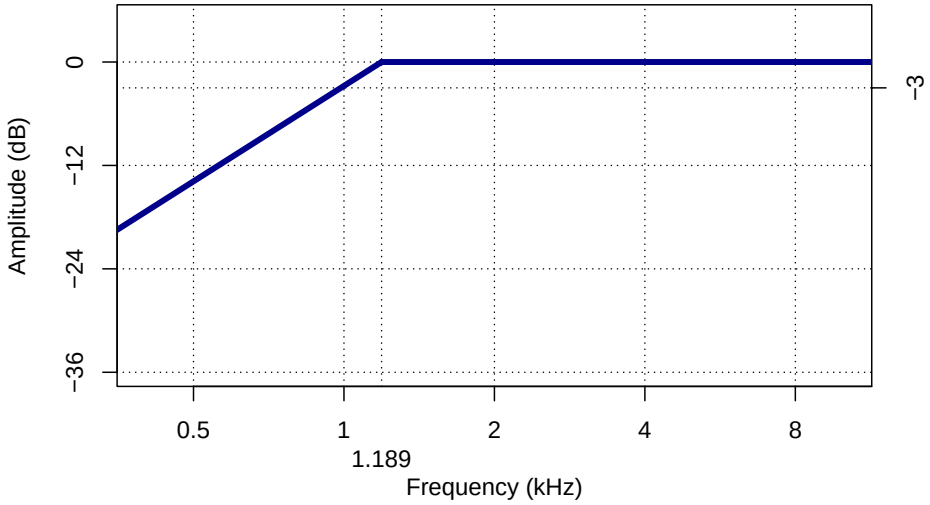

Figure 4.2: Frequency characteristic of a high-pass filter, with cutoff frequency 1000 Hz, and slope of -12 dB per octave.

- **band-pass filters** (Fig.4.3) allow frequencies within a certain frequency band to pass through, and they attenuate frequencies outside this pass band.

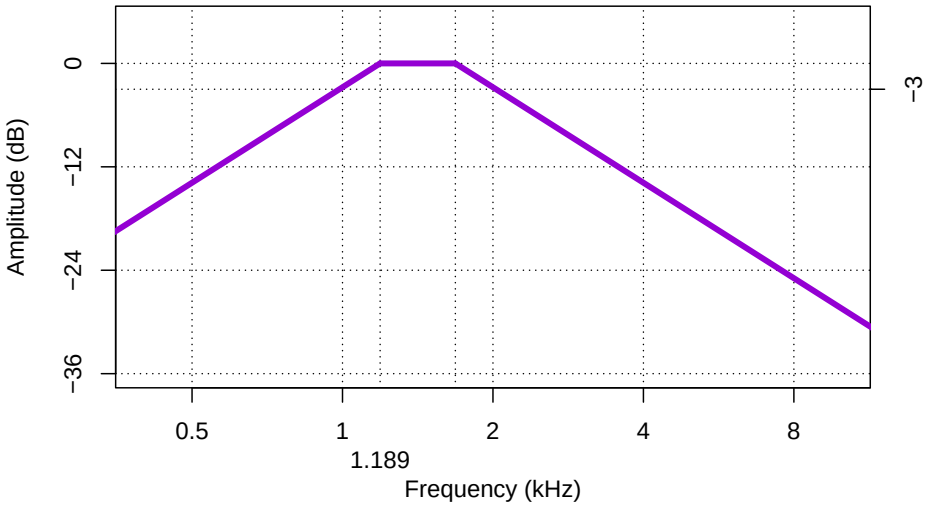

Figure 4.3: Frequency characteristic of a band-pass filter, with a pass band from 1 to 2 kHz (one octave), and with slopes of -12 dB per octave on both sides.

– A tuneable band-pass filter is essential for producing a spectrogram.

TODO crossref spectrogram

– A telephone works as a bandpass filter with a fixed pass band of 300 to 3400 Hz.

- **band-reject** or **notch filters** (Fig.4.4) again do the reverse: they attenuate frequencies within a certain frequency band, and allow frequencies outside this band to pass through.

4.3 Properties of filters

As shown in the figures above, a filter is characterised by two properties, viz. the *cutoff frequency* and the *slope*. Band-pass and band-reject filters are also characterised by their *bandwidth*.

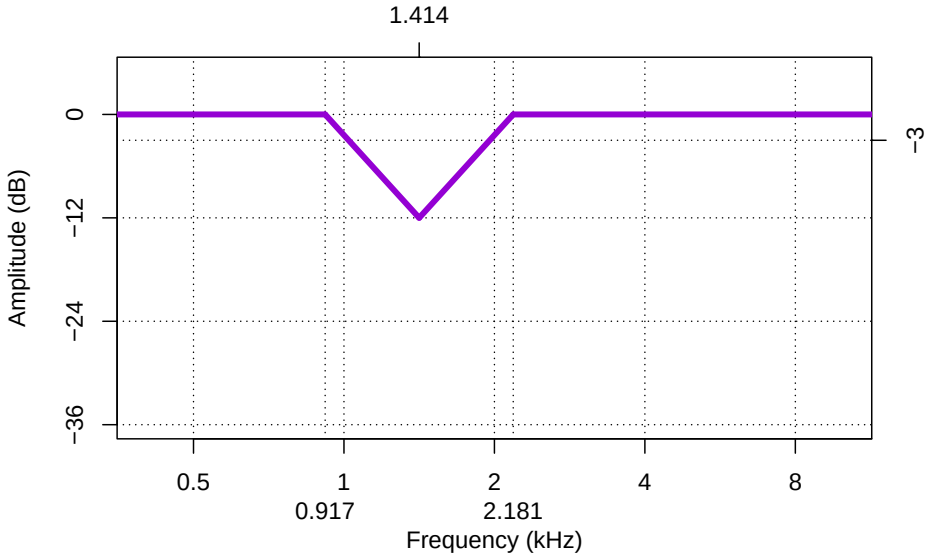

Figure 4.4: Frequency characteristic of a band-reject filter, with a reject band from 1 to 2 kHz (one octave), and with slopes of -24 dB per octave on both sides.

4.3.1 Cutoff frequency

The cutoff frequency(ies) separates the pass band(s) and reject band(s) of the filter, that is, the frequency components that are attenuated and those that are passed through unattenuated. It is defined at the frequency where the attenuation is −3 dB, as illustrated in the filter characteristics above.

4.3.2 Slope

The slope of the filter indicates the steepness of the attenuation between the pass band(s) and reject band(s). It is commonly expressed in dB attenuation per octave change in frequency (§1.8.1.1), that is, in dB per octave[1].

4.3.3 Bandwidth

Band pass filters and band reject filters are also characterised by their bandwidth: the width of the frequency span affected by the filter, equal to the distance between the *two* cutoff frequencies of such filters.

[1]Note that the band reject filter in Fig.4.4 has steeper slopes than the band pass filter in Fig.4.3.

This distance may be expressed as the musical interval between the lower and higher cutoff frequencies. Both filters shown in Figures 4.3 and 4.4 are so-called "octave band" filters, because the cutoff frequencies are one octave apart (§1.8.1.1) with a frequency ratio of 1:2.

In a musical third interval, the frequencies have a ratio of 4:5; filters with these cutoff frequency ratios are so-called "third band" filters. Both octave-band and third-band filters are widely used in phonetic research.

4.4 Emphasis filters

For reasons that we will see later (in §5.3.1), the spectrum of speech has a typical overall spectral slope of −6 dB per octave (in its voiced parts). This downward overall slope was also visible in the spectral slice in Fig.3.3. Thus, the amplitude of higher frequency components decreases by about −6 dB on average for each doubling of the frequency (§1.8.1.1). Consequently, spectral details of higher-frequency components tend to be poorly visible in the analysis. As a remedy, we can apply a so-called *pre-emphasis* filter, which modifies the overall spectral slope by +6 dB per octave. This boosting of higher-frequency components ideally results in a flat spectral envelope of speech[2], with equal amplitudes for all frequency components in speech. This makes the spectral details of speech equally discernible across the full spectral range.

The reverse operation is called *de-emphasis* filtering: this changes the overall spectral slope by −6 dB/octave. This de-emphasis filtering was applied, for example, to obtain the brown noise (−6 dB/oct) from the white noise (0 dB/oct) in Fig.3.4.

4.4.1 How to apply a (pre|de)emphasis filter

- Select an input Sound object in the `Praat` Objects window. Then choose `Filter > Filter (pre-emphasis)...` for pre-emphasis filtering, or choose `Filter > Filter (de-emphasis)...` for de-emphasis filtering. The filter only applies above a certain cutoff frequency, to be specified. Choose a cutoff frequency below the lowest speech frequency, e.g. 60 Hz.
- The resulting filtered output Sound object is again added at the bottom of the list of objects.

[2]The overall average slope of the spectral envelope of speech changes from −6 to −6+6 = 0 dB per octave.

Part II: Speech

Chapter 5

Speech sounds

Chapter keywords: resonance, formant, formants, tube, tube model, spectral slope, source, filter, radiation, source-filter model, speech production, vowels, phonation, consonants, ingressive, egressive, noise, closure, constriction, phonation, voicing.

5.1 Resonance

Imagine that you tap a glass object with a metal object – briefly and not too hard. The glass resonates at a particular frequency, which depends on the shape and mass of the glass, and you hear a resounding 'ting' sound. What is going on?

Your tapping force initiates many different vibrations within the glass. Some of these vibrations have a wavelength (§1.8.5) that equals the dimensions of the glass. The vibrations at these particular wavelengths are reinforced, while other non-matching vibrations are attenuated. This is termed **resonance**. The amplified frequencies are termed 'resonant frequencies' or 'natural frequencies'. The primary resonant frequency, which is amplified the most, is also termed the 'eigen' frequency of the resonating object. Very quickly, the non-resonant frequencies are dampened out, and the glass continues to vibrate at its resonant frequency (or frequencies). The vibrating glass constitutes a sound source, the sound of which propagates in the surrounding air (see Ch.1), and we hear a 'ping' sound. The envelope (§3.4) of this 'ping' sound depends on the material (composition), mass, and shape (and contents) of the vibrating object.

Almost all physical bodies have a tendency to resonate at their natural frequencies: not only a glass but also a pendulum, a swing, a bridge, all musical instruments, a rope on a flagpole (Minnaert, 1970, §61), the column of air in

an organ pipe, the volume of air in a hall or venue – and, importantly, also the body of air in the human vocal tract.

As a first approximation, let us consider a simple straight tube, uniform in diameter over its entire length, filled with air of 37°C, open at one end and having a sound source at the other closed end, as illustrated in Figure 5.1. This is an *idealization* of e.g. an organ pipe, a flute, or of a human speaker producing a continuous schwa vowel /ə/. (It differs somewhat from a concert hall, in that the tube is open at one end.).

Figure 5.1: Uniform straight tube of length L, open at one end, and having a sound source at the closed end. After Stevens (1998), Figure 3.8.

The principle of resonance is explained by Christian Huygens as follows (in a letter in November 1693):

> "Because every sound, or rather noise, which is repeated in equal and very small intervals, produces a [periodic] musical tone; similarly the length of an organ pipe determines its [resonant] tone, because the vibrations of air arrive at small intervals of time which are equal to the time which [the vibrations] need to move back and forth through the length of the organ pipe, as it is closed at one end." (Huygens, 1905, 571, transl. HQ)

The natural frequencies resonate (are amplified) because their wavelength matches the length L of the tube: the compressions and rarefactions of a wave period going "one way" coincide with those of the reflected wave going the "other way", so that the amplitudes of the pressure variations add up, and thus increase, for some frequencies, viz. the resonant or natural frequencies. Sound waves with frequencies close to the resonant frequencies are thus amplified and those with other frequencies are attenuated[1].

[1] In other words, the tube acts as a bandpass filter around its natural frequencies, albeit with a filter characteristic different from the one illustrated in Fig.4.3.

5.2 Formants

5.2.1 one uniform tube

Above, we discussed the resonance in an ideal, uniform tube (straight, and closed at one end, see Fig.5.2.1). This tube may be regarded as a single tube with the same (uniform) diameter across its length, and it is similar to the vocal tract of a human speaker producing a continuous schwa vowel /ə/. This vocal tract acts as a single resonator: it amplifies its natural frequencies and attenuates other frequencies.

The natural frequencies of such a uniform tube are

$$F_n = \frac{2n-1}{4} \cdot \frac{c}{L}$$

where n is the ranking number of the natural frequency, c is the speed of sound (353 m/s, see §1.4), and L is the length of the tube (Stevens, 1998, 139) [2]. If we assume that the vocal tract of an adult male speaker has length $L = 0.17$ m, then the first resonant frequency of this speaker's vocal tract would be at $F_1 = \frac{1}{4} \cdot \frac{353}{0.17} = 519$ Hz. In reality, the natural frequency is slightly lower[3] at $F_1 \approx 500$ Hz. Similarly, his $F_2 \approx 1500$ Hz and $F_3 \approx 2500$ Hz, etc, as shown in Figure 5.2. A female vocal tract is slightly shorter than that of a male speaker (we assume $L = 0.15$ m for a female vocal tract), thus we find that for a female vocal tract her $F_1 \approx 550$ Hz, $F_2 \approx 1650$ Hz and $F_3 \approx 2750$ Hz.

These natural frequencies of the vocal tract are termed its **formants**, a term coined by Ludimar Hermann (1838–1914)[4].

Figure 5.2 shows the transfer function and formants of the uniform vocal tract (open at one end) of a male speaker. This figure is to be interpreted like a spectrum (§3.2) or bandpass filter response (Fig.4.3).

A single formant is characterised by two properties:

- its peak *frequency*, or formant frequency,
- its *bandwidth* or 'peakness', that is the width of the spectral envelope 3 dB below the peak amplitude of that formant; a larger bandwidth value corresponds with a wider, flatter, and less peaked formant[5].

[2] And the corresponding *wavelengths* of these natural frequencies are $\frac{4}{2n-1} \cdot L$ (Nooteboom and Cohen, 1984, 60).

[3] This is due to absorption and radiation and to the vocal tract being not straight but bended, all of which factors are outside the scope of this introductory text.

[4] Formants are a property of the tube or of the vocal tract (and of the medium in it), and not of the sounds coming out of that tube or tract, but this distinction is seldom relevant and seldom made.

[5] The term *formant quality* may refer to the inverse of the formant bandwidth, i.e. 1/bandwidth; a larger quality value corresponds with a narrower, higher, and more peaked formant.

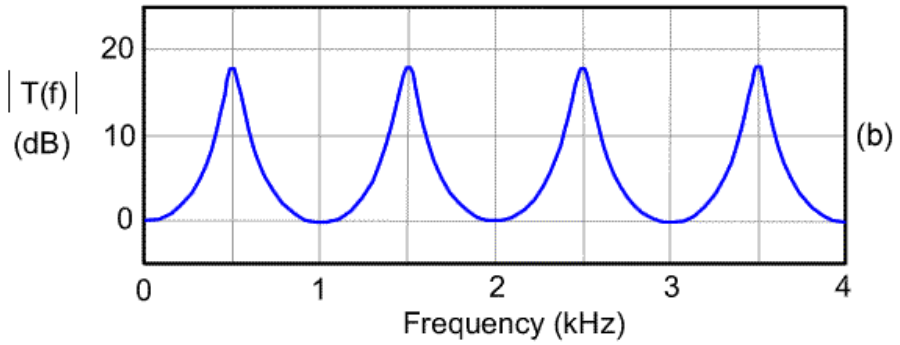

Figure 5.2: Overall transfer function of the (uniform and open) vocal tract of a male speaker. Adapted from Stevens (1998), Fig.3.3(b); from <https://www.mq.edu.au/faculty-of-medicine-health-and-human-sciences/departments-and-schools/department-of-linguistics/our-research/phonetics-and-phonology/speech/acoustics/acoustic-theory-of-speech-production/vocal-tract-resonance>

5.2.2 multiple tubes

A special property of the *human* vocal tract is that is not a fixed tube, such as the one-tube idealization above, but that it can be *articulated*[6] into multiple tubes, with varying shapes and varying cross-sections. This *articulation* of the vocal tract allows humans to produce many different vowel and consonant sounds, with different phonetic properties. Thus **the articulated and articulating vocal tract allows humans to convey linguistic meaning by means of speech.**

First we look at the two-tube configuration in Figure 5.3 (after Stevens, 1998, Fig.3.13, p.143); this is similar to the human vocal tract producing a low vowel / /. Due to the low position of the jaw and the tongue, the pharynx is relatively narrow while the oral cavity is relatively wide. This system of two tubes acts as a single system of two *coupled* resonators. The natural frequencies depend on the respective lengths L_1 and L_2 and the respective cross-sections A_1 and A_2, that is, the natural frequencies of the narrower tube affect those of the wider tube, and vice versa. This coupling results in a higher F_1 and lower F_2 for this two-tube configuration (Stevens, 1998, Fig.3.14), relative to those for the uniform tube (§5.2.1). Indeed, for / /, average formant values are (for males) $F_1 = 730$ and $F_2 = 1090$ Hz, and (for females) $F_1 = 850$ and $F_2 = 1220$ Hz.

Even more complex is the configuration in Figure 5.4 (after Stevens, 1998, Fig.3.15, p.144); it is similar to the human vocal tract producing a high vowel

[6] in the classical sense of "having two or more sections connected by a flexible joint" (*New Oxford Dictionary of English* 1998).

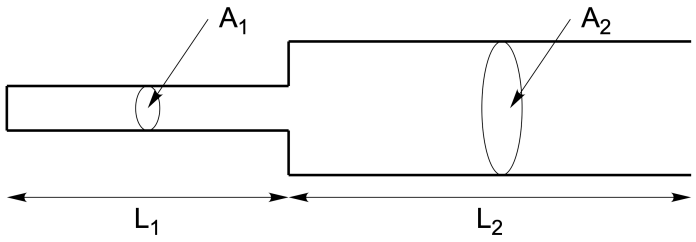

Figure 5.3: Two-tube configuration, with two coupled resonators. After Stevens (1998), Figure 3.13.

/i/. Due to the high and front position of the tongue, the vocal tract is divided (articulated) into a back and a front cavity, with a relatively narrow constriction separating the back and front cavities. The overall length L of the vocal tract remains constant. Each of the three tube sections has its own natural frequencies, which also affect one another, the interaction depending on the lengths and areas of the tube sections. For the configuration shown in Fig.5.4, the natural frequencies come out as approximately $F_1 = 300$ Hz (mainly due to constriction), $F_2 = 1900$ Hz (back cavity), and $F_3 = 2200$ Hz (front cavity) (Johnson, 2012, 137) for a male vocal tract.

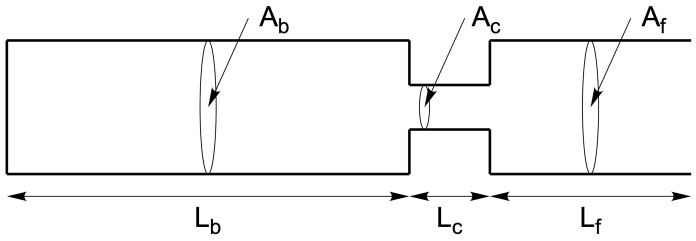

Figure 5.4: Three-tube configuration, with a back cavity (left), constriction, and front cavity (right). After Johnson (2012), Figure 6.3.

In the spectrum of the recorded /i/ vowel produced by a *female* speaker, in Fig.5.5, we see that harmonics are amplified around formants $F_1 \approx 400$ Hz and $F_2 \approx 2600$ Hz. The $F_3 \approx 3000$ Hz is poorly visible in this spectrum. This formant pattern characterizes an /i/ vowel spoken by an adult female speaker (Peterson and Barney, 1952, Table II).

Formants are numbered by their peak frequency, and are named 'F1', 'F2', 'F3',

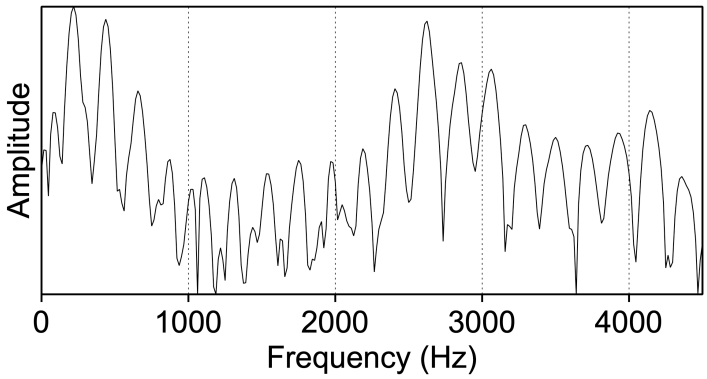

Figure 5.5: Spectral slice of the /i/ vowel in the word *speech*.

etc. (The fundamental frequency of a sound is called f_0, see §3.1.) The different vowels in any language have different formant patterns, and the first two or three formants are most important for distinguishing vowels. Because various sections of the articulated vocal tract have *interacting* resonances (as explained above), formant frequencies cannot be mapped easily on articulatory features. Nevertheless, with considerable simplification, there is a negative correlation between vowel F1 and the height of the tongue (lower tongue, higher F1, see Fig.5.3), and between vowel F2 and the front–back position of the tongue (more fronted tongue, higher F2, see Fig.5.4). These patterns are typically illustrated in a diagram depicting vowel formants in the F1-by-F2 plane[7], in your phonetics textbook. F3 is involved in the contrast between rounded and unrounded vowels (such as /i/ vs. /y/).

For further background on *formants*, see Aalto et al. (2018).

5.2.3 How to measure formants

The safest method for measuring formants is to *bypass* the spectrum, and measure formants directly from a Sound object.

- Select an input Sound object in the `Praat` Objects window. Then choose `Analyse spectum >` and then `To formant (burg)...`.
- As we saw above, for a male speaker we may expect 5 formants in the range from 0 to 5000 Hz. For a female speaker, having a slightly shorter vocal

[7] In that diagram, the scales may be displayed in Bark units, and the F1 axis may have been reversed, in order to clarify the relation between articulatory features and formant frequencies.

tract than a male speaker, we may expect 5 formants in the range from 0 to 5500 Hz. Choose appropriate `Maximum number of formants` and choose the matching `Formant ceiling` value for your speaker; these values depend on the length of the speaker's vocal tract and on your sampling frequency (§2.3.1)[8]. Leave other arguments at their default (standard) values. Click `OK`.

- The resulting Formant object is again added at the bottom of the list of objects.
- You may `Draw` or `Tabulate` or interactively `Query` the formant values (frequencies and bandwidths per formant per time frame).
- Remember to `Save` the Formant object if you wish.

Remember that the results from your formant analysis may be **suspect**:

- if the reported formant frequency is low, say below 200 Hz,
- if the reported formant bandwidth is large (formant quality is low),
- if the corresponding *wavelength* of the reported formant frequency (§1.8.5) matches one of the dimensions of the recording room: for example, an $F1 \approx 86$ Hz may be due to the length of $L = 2.00$ m of the recording room [for a tube closed at *both* ends, like the room, $F_n = (n/2) \cdot (c/2.00)) \approx 86$ Hz].

For background, see §2.4.1 above.

5.3 Source-filter theory of speech production

According to the source-filter theory, speech is produced by first generating a source sound, which is subsequently modified by an acoustic filter (Fant, 1970). The theory postulates that source and filter are two independent and subsequent components in speech production, in particular, that the filter does not affect the source (although source-filter interactions do occur in actual speech production; Tokuda (2021), §4).

Several types of **source** sounds may be generated, separately or in combination, of which the most important are:

- phonation (periodic sound produced by vibration of the vocal cords),
- whisper (aperiodic sound produced by air turbulence at the glottis),

[8] See the `Help` information in the `Praat` menu, in particular its last paragraph.

- explosion or frication (aperiodic sound produced by air turbulence at a constriction in the vocal tract),
- ingressive sounds (e.g. clicks; aperiodic and transient sound produced at a constriction in the vocal tract, with ingressive airstream).

These source sounds are then modified by the acoustic **filter**, formed by the (part of the) vocal tract that is between the sound source and the outside air. The shape of that (part of the) vocal tract may be varied (articulated), with noticeable and meaningful phonetic effects. The (part of the) vocal tract works as a filter, by means of resonance (§5.1). Figure 5.6 summarizes the model for the production of vowels with phonation as source sound.

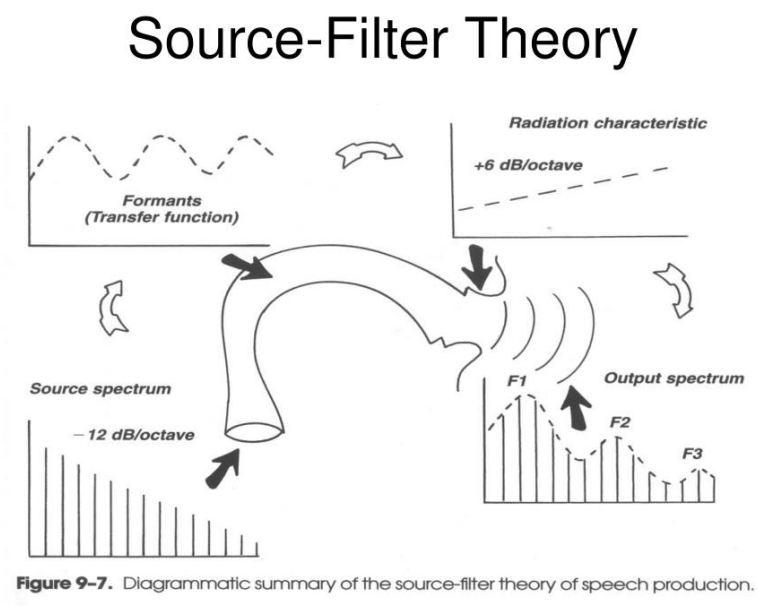

Figure 5.6: Diagrammatic summary of the source-filter theory of speech production; figure by Perry C. Hanavan, from <https://image.slideserve.com/325813/source-filter-theory34-l.jpg>.

Figure 5.6 also illustrates a third component of the source-filter theory of speech production, viz. the effect of sound **radiation** at the lips. The radiation may be regarded as a high-pass filter with an overall slope of +6 dB per octave (§4.2)[9]

[9] Components with lower frequency suffer more from acoustic impedance at the border between the vocal tract and the outside air, and these lower-frequency components are therefore attenuated in the resulting speech sound.

[10].

5.3.1 Production of vowels

Figure 5.6 illustrates how **vowels** are typically produced. The source sound is the voicing sound (phonation) generated by the vibration of the vocal cords in the larynx. (The *glottis* is the opening between the two vocal folds.) This phonation sound has a spectral slope of about −12 dB per octave. This phonation sound is subsequently modified or filtered by the vocal tract, which amplifies certain frequencies (*formants*, §5.2). Finally the filtered sound is radiated from the lips, which modifies the spectral slope by +6 dB per octave. The resulting vowel sound has an overall spectral slope of −6 dB per octave, as mentioned in §4.4 above.

Remember that *harmonics* including f_0 depend on the source signal (phonation), while *formants* depend on the filter (vocal tract). Consequently, a formant frequency does not have to coincide with a harmonic: in the output spectrum in Fig.5.6, for example, F1 falls between the second and third harmonic.

5.3.2 Production of consonants

For **voiceless** (and egressive) consonants, e.g. /s, k/, the source sound is the turbulent noise generated at a constriction or closure of the vocal tract; this source sound may be regarded as white noise (§3.3.1). This noise typically has a flat spectral slope (0 dB per octave). This noise sound is subsequently modified or filtered by the part of the vocal tract that is between the place of closure or constriction, and the outside air; certain frequencies are amplified and other frequencies may be attenuated.

Ingressive consonants are produced by the same acoustic-phonetic processes, albeit with an ingressive (inward) air stream.

For **voiced** (and egressive) consonants, e.g. /z, b/, the situation is more complex: typically two source sounds are generated *simultaneously*: a turbulent noise sound generated at a constriction or closure in the vocal tract, as well as a voicing sound (phonation) generated by the vibration of the vocal cords in the larynx. The periodic voicing sound is filtered by the entire vocal tract, as vowels are, while the aperiodic (noise) sound is filtered only by the 'downstream' part of the vocal tract. Because the egressive airstream from the lungs is used to drive the phonation at the larynx, the noise generated inside the vocal tract is considerably weaker for voiced plosive and fricative consonants than for their unvoiced counterparts. In fact, it is somewhat difficult to generate two sounds

[10]This spectral effect of radiation is often regarded as if it were a spectral property of the source sound, after which this radiation is further ignored.

simultaneously, as the two sound generation processes tend to counteract each other.

Chapter 6

Spectrograms

Chapter keywords: spectrogram, center frequency, bandwidth, broadband, narrowband, detail, smearing, formant, amplitude, frequency, time, component, harmonic, glottal pulse.

6.1 Introduction

As we saw, a spectrum (§3.2) shows the various frequency components of a complex signal, such as the formants of a vowel (§5.2). However, a spectrum either shows the frequency composition of an entire digital sound, or it shows the frequency composition of a brief 'slice' or frame or window of that sound (box 3.2).

In order to track spectral changes over time, we need a series of spectral slices, measured at fixed intervals. Thus we wish to depict three dimensions: time, amplitude, and frequency. This most important visualisation in speech analysis is called a **spectrogram**. As in an oscillogram, *time* is shown along the horizontal axis. Like in a spectrum, but rotated by 90°, *frequency* is shown along the vertical axis. *Amplitude* is shown in the degree of blackness (or in coloring). The example spectrogram in Figure 6.1 visualizes a blackbird song, showing the amplitudes (blackness) of the frequency components (vertically), varying over time (horizontally).

In this spectrogram you see an initial steeply falling 'chirp' sound, followed by three repetitions of a two-tone 'syllable', against a background of other birdsong.

Spectrograms used to be made by means of a special device, a sonagraph or spectrograph, which worked by means of a bandpass filter (Fig.4.3). This bandpass filter was initially set to a low *center frequency*, then applied to the sound recording, and the resulting output amplitude of the filter was printed as varying degrees of blackness, resulting in a single horizontal 'line' of the figure. Then

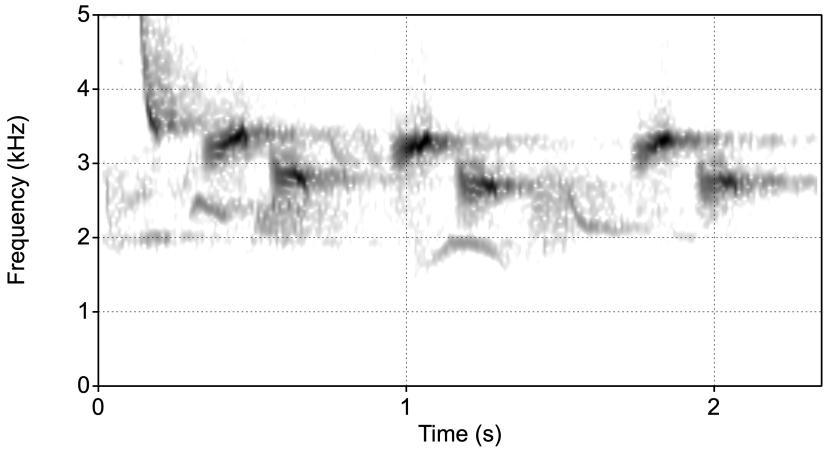

Figure 6.1: Spectrogram of a brief part of song by a blackbird.

the frequency band of the filter was increased slightly, to produce the next 'line' of the figure. This was repeated across the entire range of frequencies[1]. Moreover, a spectrograph used to have two fixed settings for the *bandwidth* of the band-pass filter (§4.3.3): 260 Hz for so-called broadband spectrograms, and 43 Hz for narrowband spectrograms, respectively. (Nowadays, in digital speech analysis, bandwidth is controlled indirectly, via the length of the window for spectral analysis, see §3.2.2 and see below).

6.2 Broadband spectrogram

In a broad-band spectrogram, as in Fig.6.2, the bandwidth used in the spectral analysis is relatively wide (by convention, 260 Hz).

The wide bandwidth results in less detail (more smearing) in the frequency dimension, while allowing more detail in the time dimension. We can see temporal events in great detail, such as individual periods of voiced parts, and noise bursts in plosive or affricate consonants. On the other hand, frequencies are blurred or smeared. This makes it easier, however, to see formants (§5.2) in the vowels, and to see other "broad" spectral properties in the consonants.

[1] Thus a spectrograph required several minutes to produce a single spectrogram.

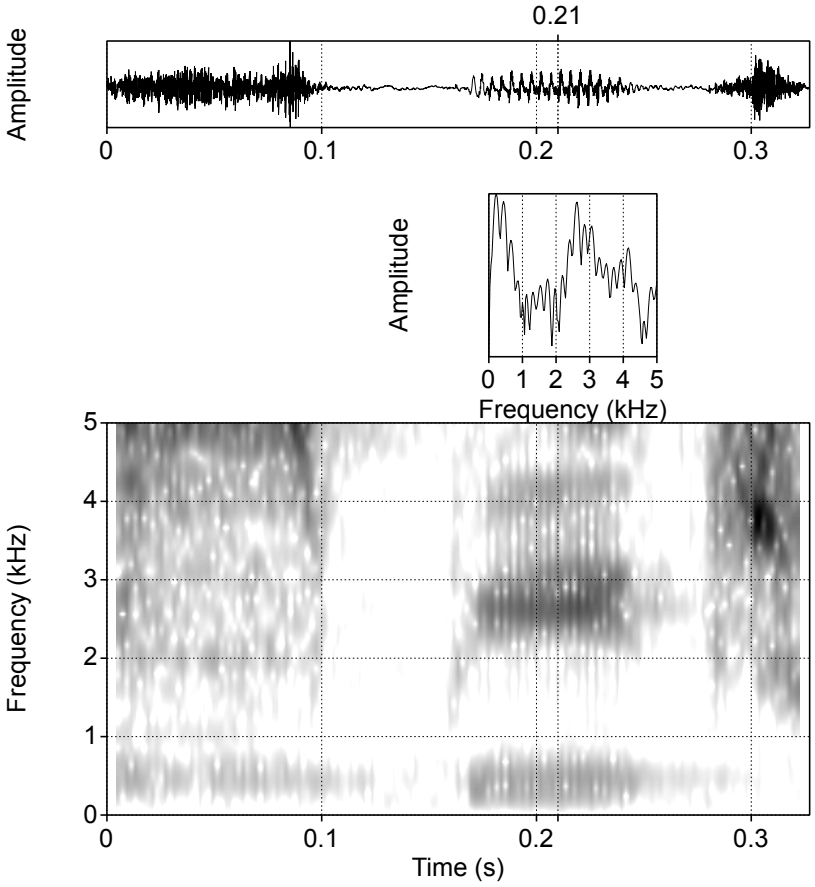

Figure 6.2: Oscillogram, broad-band spectrum at 0.210 s, and broad-band spectrogram of the word *speech*.

6.3 Narrowband spectrogram

In a narrow-band spectrogram, as in Fig.6.3, the bandwidth used in the spectral analysis is relatively narrow (by convention, 43 Hz).

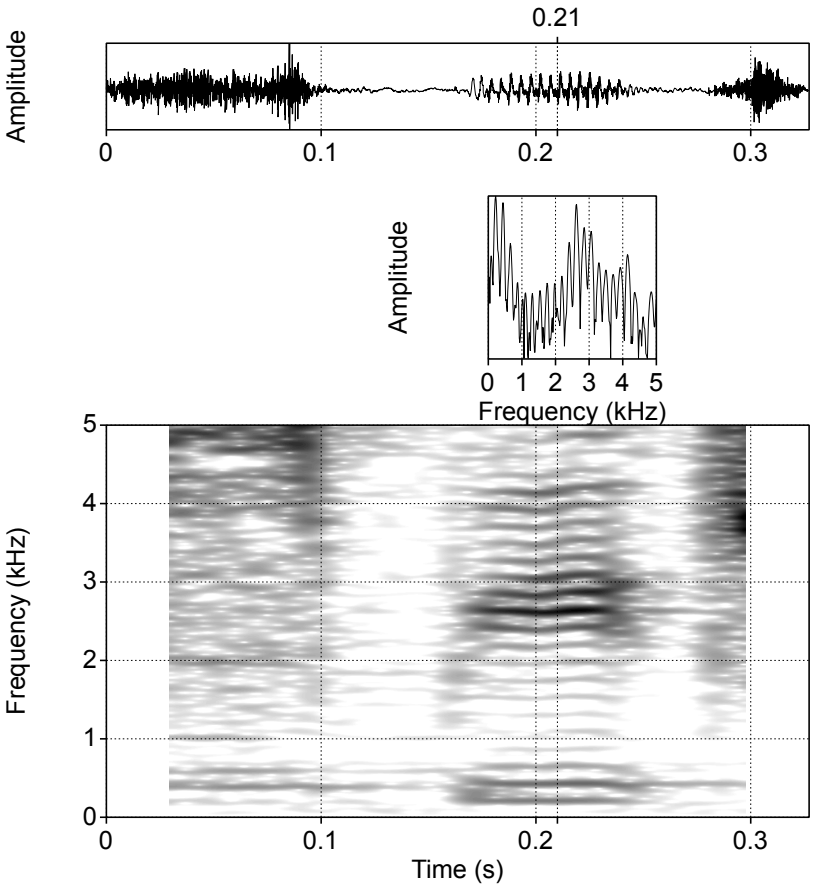

Figure 6.3: Oscillogram, narrow-band spectrum at 0.210 s, and narrow-band spectrogram of the word *speech*.

The narrow bandwidth results in more detail in the frequency dimension, while allowing less detail (more smearing) in the time dimension. We can see frequencies in great detail, such as individual harmonics in voiced parts. On the other hand, brief temporal events such as noise bursts are blurred or smeared.

6.4 How to make a spectrogram

There are two ways to create a spectrogram of a sound recording: either directly, or by means of the SoundEditor.

6.4.1 to Spectrogram directly

- Select an input Sound object in the `Praat` Objects window. Then choose `Analyse spectum >`, then `To Spectrogram...`.
- The parameter `Window length` determines the filter bandwidth used to create the spectrogram. For a broadband spectrogram, choose 0.005 s; for a narrowband spectrogram, choose 0.030 s. For the spectrogram of birdsong in Fig.6.1, an intermediate value of 0.015 s was used.
  Set `Window shape` to `Gaussian`.
  Leave other arguments at their default (standard) values. Click `OK`.
- The resulting Spectrogram object is again added at the bottom of the list of objects.
- You may `Draw > Paint...` the spectrogram. Specify the time and frequency range, as well as the dynamic range (of amplitude in dB, between darkest and lightest colour).
  The `Pre-emphasis` slope helps in discerning visual details in the higher frequencies, as explained in §4.4 and §5.3.1; the default value of +6 dB per octave works well in most situations.
- Remember to `Save` the Spectrogram object if you wish.

6.4.2 from SoundEditor

This method is similar to the spectral analysis described in §3.2.2. Using this method, your spectrogram in the SoundEditor is immediately updated if you change (`Apply`) values for the analysis parameters (which may be convenient), but extracting and saving the Spectrogram requires some extra work.

- In the `Praat` object window, select a Sound object.
- Next, in the `Praat` object window, choose `View & Edit`. This will open a so-called SoundEditor window, with the oscillogram as its main feature.

  TODO crossref SoundEditor

- In the SoundEditor window, go to `Spectrogram...` and then `Spectrogram settings`.

- The parameter `Window length` determines the filter bandwidth used to create the spectrogram. For a broadband spectrogram, choose 0.005 s; for a narrowband spectrogram, choose 0.030 s. For the spectrogram of birdsong in Fig.6.1, an intermediate value of 0.015 s was used.
  Set the `Dynamic range` (of amplitude in dB) between darkest and lightest colour to some value between 40 and 55 dB (try different values by using `Apply`, and notice the differences in the resulting spectrogram). Leave other arguments at their default (standard) values. Click `OK`.
- Make sure that you can see the Spectrogram in the SoundEditor, by selecting the option `Spectrogram > View Spectrogram`. This setting is saved from session to session, so you may have to switch `on` the spectrogram view.
- If you are happy with the spectrogram in the SoundEditor panel, you can then choose `Spectrogram > Paint visible spectrogram...` in order to draw that spectrogram in the Picture window.
- You can also `Extract` and then `Save` the spectrogram from the SoundEditor.

6.5 How to read a spectrogram

(with Clizia Welker)

This section intends to provide some hints on how to read a spectrogram. The skill of reading a spectrogram is highly useful whenever you need to *segment* (cut up) a speech signal into syllables or phones, and/or need to *transcribe* (label) the words or sounds in a speech recording. For more detailed explanations, please refer to your textbooks on phonetics (for recommended textbooks see §0.1), and see Chapter XX on segmenting and labeling speech. Also consult the valuable instructions and background on how to read a spectrogram by Rob Hagiwara.

TODO link to chapter on segmenting and labeling

While by observing an oscillogram it is only possible to identify broad phonetic classes, the spectrogram provides us with sufficient information to determine the phones. It is possible to deduce the phones (with good chances of success) by exploiting information about

- the frequency values of the formants,
- the energy pattern throughout the spectrogram,
- the formant transitions (from vowel to consonants and from consonants to vowels).

Except for the formant transitions (see below), these features may be found in a spectrum too, but spectrograms allow us to inspect these features over time.

In order to read a spectrogram, we need to remember that:

- the blacker regions are frequency regions with a high degree of energy (§6.1),
- the black band at the bottom of the spectrogram displays the f_0 in voiced parts; it may be difficult to distinguish fundamental f_0 from formant F1 (especially in closed vowels, which have a low F1 close to f_0 [2]),
- vowels are characterised by a clear formant pattern (§5.2), visible as horizontal black bands,
- consonants (especially voiceless ones) do not have a clear formant pattern, but the distribution of energy throughout the spectrum helps us identifying them,
- the vertical bands at regular time intervals in the broad-band spectrogram correspond to the glottal pulses (cycles of opening and closing of the vocal folds).

Vowels may be distinguished through their formant values, especially through F1, F2 and F3; consult a vowel diagram in your textbook.

Dipthongs (such as in the Dutch words *bij*, *bui* and *bouw*) are composed of two different vowel sounds, with a gradual movement of the formants from their initial to final frequency values.

For **voiced consonants** in a CV context, formant frequencies change relatively rapidly. The *F2* value in the vowel from which the formant travels in this formant transition is often called the 'F2 locus', and this locus is correlated with the place of articulation of the consonant. The F2 transition is also due to the fact that acoustic resonators (tube sections) are uncoupled at the moment of closure or frication, but become coupled as the mouth opens. For labial consonants the F2 locus is around 700 Hz, for alveolar consonants it is around 1800 Hz, and for velar consonants it is around 2700 Hz.
In CV context, rapid F1 transitions are related to the degree of jaw opening, with F1 increasing from zero to its vowel value during opening of the mouth.
In VC contexts we see similar patterns as in CV contexts, but reversed in time.

While **voiceless consonants** are not characterised by formants (even in an intervocalic context), their place of articulation may be assessed by looking at which frequency regions display a high concentration of energy. Moreover, in CV context, the rapid F2 transitions may seem to travel from the same F2 locus as

[2] If speaking or singing at high fundamental f_0, the f_0 may even be higher than F1, thus rendering F1 inaudible.

for corresponding voiced consonants with the same place of articulation, albeit with formants less visible in transitions involving voiceless consonants.

The distinction between voiced and voiceless plosive consonants largely depends on the **voice onset time (VOT)**, "defined by the time elapsed between the release of the stop consonant constriction (sometimes called the "burst") and the onset of periodicity in the following voiced segment" (Rubin, 2022). The release or burst is visible as a brief noise segment; the following voiced segment typically has initial formant transitions (see above), and visible glottal pulses or visible harmonics (see above).

In producing **alveolar** consonants, the vocal tract is divided into two smaller cavities: a large back cavity, and a small front cavity between the alveolar obstruction and the outside air. Due to its small size, the front cavity yields resonances in the higher frequency regions, as we can see in the initial consonant [s] in Fig.6.2. In **palatal** consonants (such as the final consonant in the same spectrogram), the place of artiulation is more backward, the front cavity is larger, and the so-called 'spectral centre of gravity' in the resulting consonant is correspondingly lower.

Nasal and lateral consonants are phonetically complex. They resemble vowels, but the vocal tract has additional cavities during the production of these sounds: for nasals, the nasal cavity, and for laterals, additional cavities under and aside of the tongue. The resulting sound has resonances due to these cavities, as well as anti-resonances due to the acoustic *coupling* between branching cavities.

For more tips and tricks and background about how to identify speech segments in a spectrogram, please consult your textbooks on phonetics, as well as the resources mentioned in the first paragraph of this section.

Chapter 7

Segmenting and labeling speech sounds

TODO expand keywords

Chapter keywords: phonetics, linguistic units, segments, sounds, coarticulation.

7.1 Introduction

Segmenting speech and labeling speech are at the core of phonetic analysis. In spoken language, speakers realize abstract linguistic sound units in the form of articulatory gestures and movements and positions. The articulatory movements from one sound to the next require complex motor coordination in space and in time. In turn, this strongly affects the speech stream: the resulting speech sounds are not clearly related to the intended or the canonical linguistic sound units. Instead, the articulatory commands and their acoustic effects are smeared out over time, affecting neigboring speech sounds: this is called **coarticulation**, and you can read more about this in your phonetics textbook. As a result, speech is not analogous to but very different from typed or printed text; speech is rather analogous to and more similar to cursive *handwriting*.

The complex relation between linguistic and spoken units is illustrated in Figure 7.1 which is a copy of Figure 2.1.

The SoundEditor in Fig.7.1 shows a fragment of speech, taken from the audio file named `10-18-17_Council_SLASH_10-18-17_Council_DOT_HD_DOT_mp3_00285.flac` from the *Peoples Speech* corpus at https://huggingface.co/datasets/MLCommons/peoples_speech. For details about that corpus, see Galvez et al. (2021).

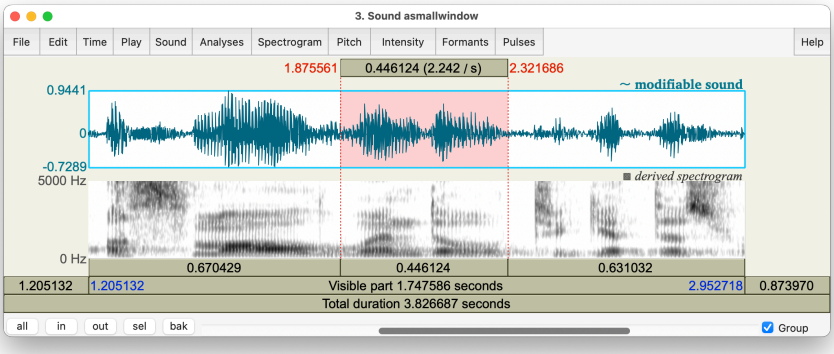

Figure 7.1: SoundEditor window, showing oscillogram and spectrogram of the fragment *...a small window to get this...*, with the word *window* highlighted.

Figure 7.1 illustrates the following three[1] characteristics of casual, fluent speech:

- Boundaries between speech segments are not clearly present, due to coarticulation, e.g. between the vowel and the final consonant of [sm] *small*.
- Some speech segments have hardly any acoustic-phonetic manifestation at all, e.g. the final consonant of *get*;
- Some speech segments are difficult to transcribe, e.g. the vowel in the last syllable of *window*: diphthong [w ndo] or monophthong [w ndo] ?

Subsequently, listener retrieve the indended units (sounds, words, meanings) from this continuous speech stream. In sum, segmenting and labeling speech addresses the core issue in phonetics, viz. how abstract linguistic units and meanings are related to their concrete physical realizations, both in production and in perception of spoken language.

In this chapter, we discuss **segments**, and thus we assume that speech sounds or segments have unambiguous boundaries. We also discuss **labels** (transcription), and thus we assume also that speech segments can be unambiguously labeled and identified as linguistic units. Nevertheless, we need to remember that actual speech sounds (speech segments, or phones) do not have a clear one-to-one relation with abstract linguistic units (phonemes).

[1] You may also notice the F2 locus of the [d] in *window*, at 2.118 s and ~1800 Hz, see §6.5, voiced consonants.

7.2 TextGrid objects

Praat has a special type of object for annotation, called a **TextGrid**. A TextGrid object may contain multiple layers, or tiers, of annotation. As a first example, a TextGrid of a single-speaker recording may have one layer of word-level transcription in regular orthography (spelling), a second layer of phone-by-phone phonetic transcription, and a third layer marking special moments of interest to the researcher, e.g. the moment of maximum intensity of a vowel. As a second example, a TextGrid of a dialogue recording may contain two layers of word-level transcription, one for each speaker in the dialogue. These examples illustrate two distinct types of layers or tiers within a TextGrid object:

- interval tiers containing time intervals (corresponding with e.g. words, phones, utterances); these intervals have boundaries and they may have labels, and
- point tiers containing time points (e.g. moments of intensity peaks); these points have a single time point (moment) and they may have labels.

As explained in the Praat tutorial on annotation (in the Help system), the easiest way to create a TextGrid matching a Sound recording is as follows:

- Select an input Sound object in the Praat Objects window.
- Choose Annotate > and then To TextGrid....
- In the subsequent menu (Sound to TextGrid), just type the names of the respective tiers, separated by blanks only, e.g. word phone. (You can later rename the tiers.)
- In the same menu, specify which of the tiers are point tiers, by typing the names (again) of the point tiers. Tiers not mentioned as point tiers will be created as interval tiers. (See above for the distinction between these two kinds of tiers.) Then click OK.
- The resulting TextGrid object is again added at the bottom of the list of objects. The TextGrid object has the same begin and end times as the Sound from which it was created, and it will have the tier(s) as specified, but there will be no information yet within the tier(s).
- You may Draw or Tabulate or interactively Query or Modify the contents of the TextGrid. (Choose Modify to rename tiers.)
- Remember to Save the TextGrid object if you wish.

7.3 How to segment and label speech sounds in `Praat`

In practice, it's easiest to set segment boundaries (segmentation) and to label those segments (labeling or transcription) at the same time. The boundaries and labels are kept in a TextGrid object.

The recommended workflow in `Praat` is as follows.

(1) Open the speech recording to annotate as a Sound object (see §sec:praatopen).

(2) Create a corresponding TextGrid (see §7.2 above).

(3) Select *both* the Sound and corresponding TextGrid objects; in order to select multiple objects from the list, press `Command` while selecting objects.

(4) Choose the button labeled `View & Edit`; this will open a TextGrid Editor window, which is quite similar to the SoundEditor window (see §2.6.2).

Figure 7.2 shows an example of a TextGridEditor, displaying the fragment *a small window to get this right* spoken by a male speaker of American English; the single word *window* is highlighted. In this example, the TextGridEditor shows the two tiers of annotation in this particular TextGrid: the `word` (interval) tier for orthographic transcription, and the `phone` (interval) tier for broad phonetic transcription. And in this illustration, `Praat` focuses on the `word` tier, as indicated by the yellow highlight; you can change the focus by clicking in a tier.

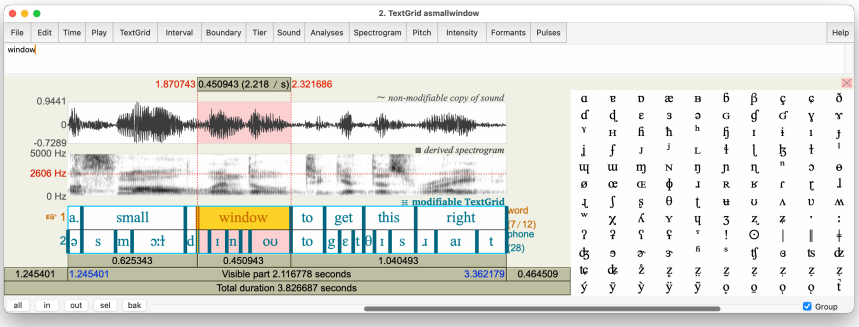

Figure 7.2: TextGridEditor window, showing oscillogram, spectrogram, and 'word' and 'phone' annotation tiers of the fragment *...a small window to get this right*; the word *window* is highlighted.

The SoundEditor in Fig.7.2 shows a fragment of speech, taken from the audio file named `10-18-17_Council_SLASH_10-18-17_Council_DOT_HD_DOT_mp3_00285.flac` from the *Peoples Speech* corpus at https://huggingface.co/datasets/MLCommons/peoples_speech. For details about that corpus, see Galvez et al. (2021).

(5) In the TextGridEditor:

(a) if you click in the oscillogram or spectrogram, then the vertical cursor moves to that position. Moreover, a *candidate boundary* appears in the annotation tiers; it is marked by a small circle or 'handle' at the upper end of the boundary mark in a tier. You may accept the proposed boundary by clicking on its 'handle'. By repeatedly placing boundaries, you can **segment** the continuous speech signal into consecutive discrete intervals within the annotation tiers.

(b) You may **copy** a boundary from one tier to another by clicking within a tier on the boundary to copy (e.g. from the `word` to the `phone` tier). Again, a *candidate boundary* appears in the other annotation tier(s) (if there are any other tiers, of course). Again, you may accept the proposed boundary by clicking on its 'handle'.

(c) You may **move** a boundary in the TextGridEditor as follows. Select the boundary by clicking on it, then drag it to an earlier or later position. It is recommended to place boundaries at zero crossings: to do so, choose `Boundary > Move to nearest zero crossing` (see §2.6.2 for background). (In order to move the *cursor*, choose `Sound > Move cursor to nearest zero crossing`).

(d) If you are not happy with a boundary, you can **remove** it by first selecting the boundary (click on it), then choose `Boundary > Remove`.

(e) If you click in an interval in a tier, then that interval is selected. You can **listen** to the selected interval, or the neigboring parts, using the buttons labeled with durations.

(6) Still in the TextGridEditor:
If you click in an interval in a tier, then that interval is selected. As soon as you type a character, or pick a phonetic symbol from the righthand chart, that symbol is used as the transcription **label** for the selected interval in the focused tier.
Phonetic symbols are explained in your phonetics textbook and in the Han (1999).

(7) Boundaries and labels are modified (added, moved, removed, etc) *immediately* in the TextGrid. If you are done, you may close the TextGridEditor window, and the contents of the TextGrid are kept in the TextGrid object that you've just edited.

(8) Back in the `Praat` Objects window:
you may `Draw` or `Tabulate` or interactively `Query` or `Modify` the contents of the TextGrid. (Choose `Modify` to rename tiers.)
The `Tabulate` command is convenient if you wish to export the TextGrid information as a dataset for further analysis. Each interval of each tier is written as a separate row in the dataset. After having saved the resulting table from `Praat`, you could use your statistical software on this tabulated-and-exported dataset, e.g. to count words, to compute average segment durations, etc etc.
(For statistical guidance, see this tutorial on Quantitative Methods and Statistics.

(9) Remember to `Save` the TextGrid object if you wish.

Chapter 8

Prosody

Chapter keywords: TODO .

TODO add text here

8.1 Pauses in speech

8.2 Durations

8.3 Pitch

8.4 Intensity

8.5 Stresses and accents

Bibliography

(1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Aalto, D., Malinen, J., and Vainio, M. (2018). *Formants*.

Beňuš, . (2021). *Investigating Spoken English: A Practical Guide To Phonetics And Phonology Using Praat*. Palgrave Macmillan.

Boersma, P. and Weenink, D. (2024). Praat: Doing phonetics by computer (version 6.4.23).

Dahl, P. H., Miller, J. H., Cato, D. H., and Andrew, R. K. (2007). Underwater ambient noise. *Acoustics Today*, 3(1):23–33.

Fant, Fant, G. (1970). *Acoustic Theory of Vowel Production: with Calculations Based on X-Ray Studies of Russian Articulations*. Mouton, The Hague, 2nd edition.

Fletcher, H. (1953). *Speech and Hearing in Communication*. Van Nostrand, New York.

Fry, D. (1979). *The Physics of Speech*. Cambridge University Press, Cambridge.

Galvez, D., Diamos, G., Ciro, J., Cerón, J. F., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., and Reddi, V. J. (2021). The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. arXiv:2111.09344 [cs].

Harris, C. M. (1971). Effects of humidity on the velocity of sound in air. *The Journal of the Acoustical Society of America*, 49(3B):890–893.

Henderson, C. (2023). *A Book of Noises: Notes on the Auraculous*. Granta, London.

Huygens, C. (1905). *Oeuvres complètes, publiées par la Société Hollandaise des Sciences*, volume 10. M. Nijhoff, La Haye.

Johnson, K. (2012). *Acoustic and Auditory Phonetics*. Blackwell, Malden, MA, 3rd edition.

Ladefoged, P. and Johnson, K. (2015). *A Course in Phonetics*. 7th edition.

Minnaert, M. (1970). *De natuurkunde van 't vrije veld*, volume 2. Thieme, Zutphen.

Nooteboom, S. and Cohen, A. (1984). *Spreken en Verstaan: Een nieuwe inleiding tot de experimentele fonetiek*. Van Gorcum, Assen, 2nd edition.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Reetz, H. and Jongman, A. (2020). *Phonetics: Transcription, Production, Acoustics, and Perception*. Wiley-Blackwell, Chichester, 2nd edition.

Rietveld, A. and van Heuven, V. (2009). *Algemene Fonetiek*. Coutinho, Bussum, 3rd edition.

Rubin, P. (2022). *Arthur Abramson*.

Shadle, C. H. (2010). *The Aerodynamics of Speech*, page 39–80. Wiley-Blackwell, Chichester, 2nd edition.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press, Cambridge, MA.

Tokuda, I. (2021). *The Source–Filter Theory of Speech*.

Xie, Y. (2024). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.40, https://pkgs.rstudio.com/bookdown/.

Xie, Y., Allaire, J., and Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida.

Zsiga, E. C. (2024). *The sounds of language: An introduction to phonetics and phonology*. Wiley-Blackwell, Chichester, 2nd edition.