Tutorial on Phonetics and Speech Analysis

true true

Document compiled 02 Dec 2024 16:48

Contents

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The existing (outdated) full tutorial is still available at https://resources.lab.hum.uu.nl/resources/phonetics/index.html .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

TBA

Recommended textbooks

This tutorial is intended to be used in addition to one or more textbook(s) in Phonetics, to which this tutorial will provide additional background knowledge. Some excellent textbooks in Phonetics are those by Rietveld and van Heuven (2009) (in Dutch), Johnson (2011), Ladefoged and Johnson (2015), Reetz and Jongman (2020), and Zsiga (2024).

Details

License

This work is licensed under the *GNU GPL 3* license (for details see https://www.gnu.org/licenses/gpl-3.0.en.html).

Citation

TBA

Technical details

TBA

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.

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 chapter 2 we'll learn how to measure, record and store a sound wave.) Here, we jump ahead and present a graphical representation of a sound wave: the **oscillogram**. (We need the oscillogram to explain important properties of sound waves.)

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

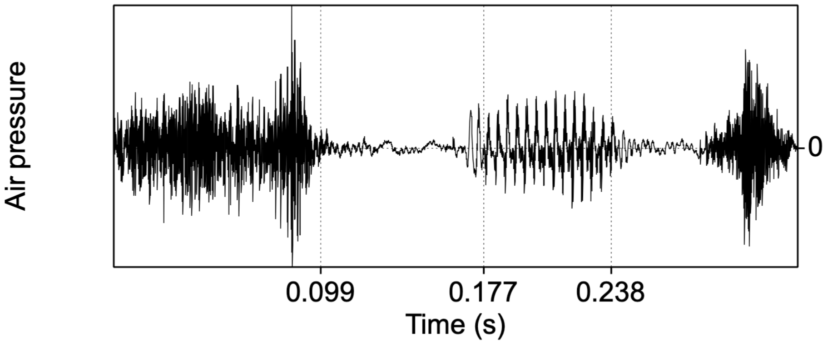

Figure 1.1: Oscillogram of the word *speech*, with boundaries between segments.

[3] 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.

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.

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[4].

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.

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)[5]. Each of these periods

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

[5] In older texts you may find 'cycles per second', abbreviated 'cps'.

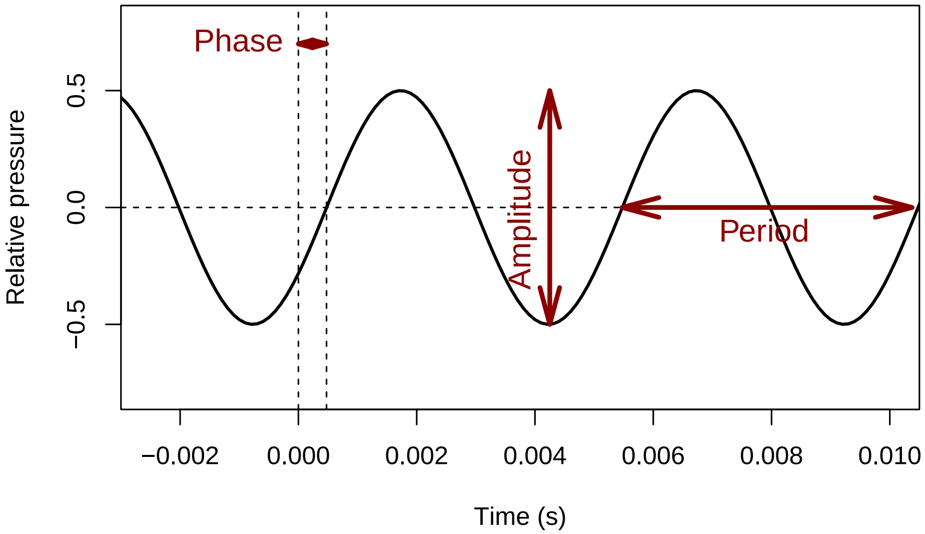

Figure 1.2: Oscillogram of a periodic sound (in black), with indications of the key properties frequency (1/period, see below), amplitude, and phase. The oscillogram is recorded over time (along the horizontal axis), at a fixed position in space.

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[6]. 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).

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

[6] More exactly, the period *is* 0.0046 s.

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 will be discussed in §XXX below.

1.8.3 Intensity

1.8.3.1 Decibels

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 1.2

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?

No answer provided

...So... think this through...

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).

$$\lambda = c/f$$

$$\lambda = cT$$

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.

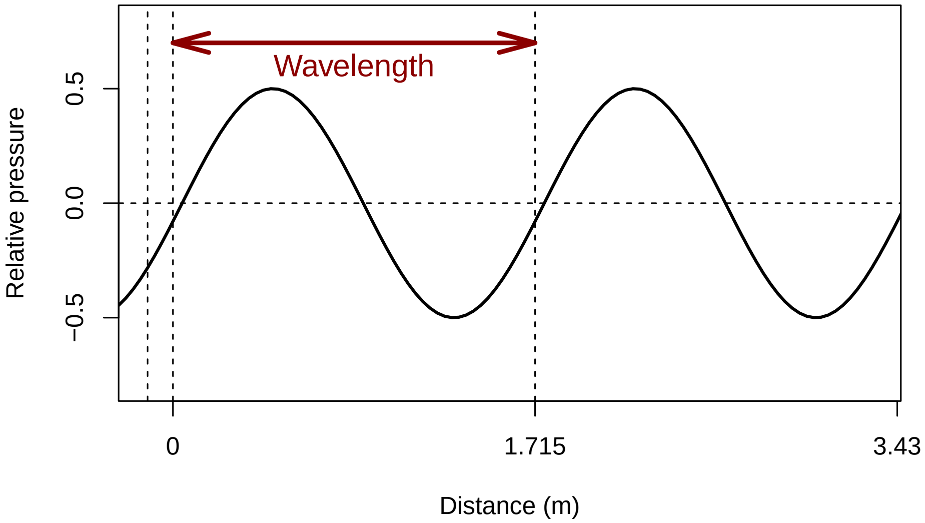

Figure 1.3: Snapshot of the pressure of a sound wave in air, varying with distance from the sound source (along the horizontal axis). The snapshot was taken at a fixed moment in time, with c=343 m/s in the air.

TODO adjust wavelenght plot

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[7].
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.

[7] 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).

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 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. 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.

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 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 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 working; 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 number of bits used to describe the amplitude value of the sample: the *amplitude resolution*. These are 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. This means that the source sound may not contain any (components with) frequencies above $f_s/2$, the so-called 'nyquist frequency'. In practice this is guaranteed by *low-pass* filtering the source sound (see §4.2), 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.

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

[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.

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, 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:

- 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.

[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.

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 §XXX) 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].

TODO crossref formants.

- 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.

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

[10] *Frequency* and *formant* measurements may be suspect if the corresponding *wavelength* is a multiple of one of the dimensions of the room (see §1.8.5); or if the reported frequency is below 100 Hz; or if the reliability of the frequency measurement is low.

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.

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.

[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.*

2.5 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 §2.6.2.)

2.6 How to manipulate 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 × 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

TODO add crossref intensity SPL

`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 §XXX). `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.

2.6.1.3 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. Due to the workings of the ear, this 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.

2.6.2 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, FFT, fourier analysis, harmonics, fundamental, overtone, component, timbre,

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' of a complex sound, these frequency components are being estimated.

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 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

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

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.

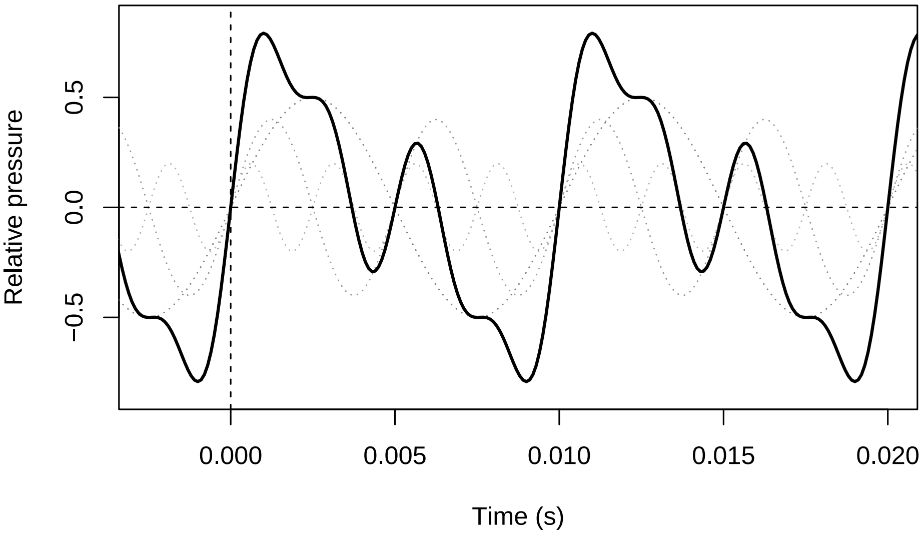

Figure 3.1: Oscillograms of three sinewave sounds and their resulting complex periodic sound.

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.

3.1.2 Octave

An octave refers to a fundamental frequency ratio of $1 : 2$ or $2 : 1$, that is, doubling or halving of the fundamental frequency. It is the distance between 12 semitones (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. (As an aside: doubling the frequency means halving the wavelength, and vice versa, see §1.8.5).

3.2 Spectrum

An unchanging 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*.

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 horizontal axis)[2]. Hence, a spectrum shows the frequency and amplitude of each component.

TODO Praat instruction box

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[3]. Aperiodic 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.

[2] Here, the phase of the frequency components is ignored.

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

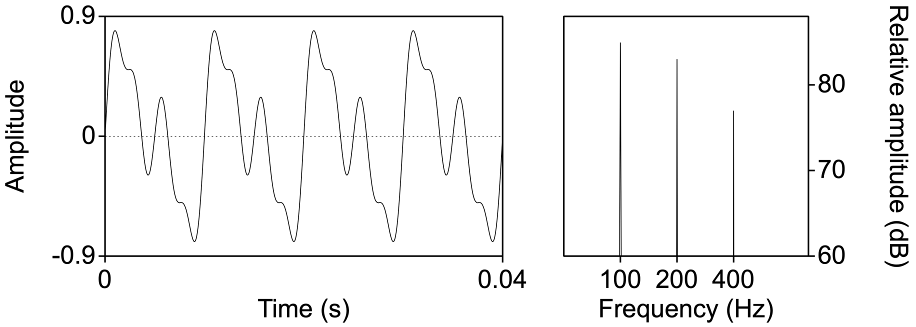

Figure 3.2: Oscillogram (left) and spectrum (right) of a complex periodic sound.

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[4], 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, 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 was calculated over a finite amount of time[5], 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.

[4] 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.

[5] If you would listen to white noise for a VERY long time, then all frequency components would indeed be equally strong.

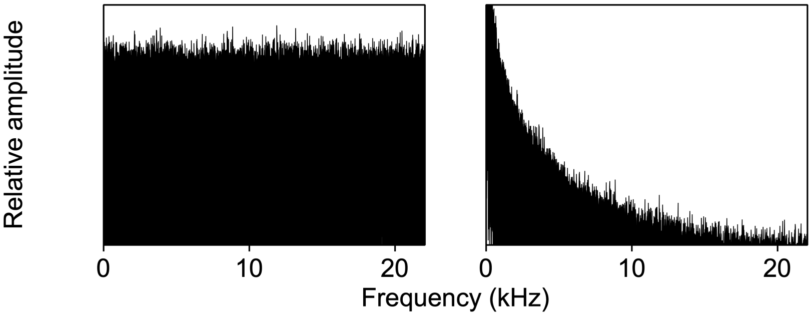

Figure 3.3: Spectra of white noise (left) and of brown noise (right), with a linear frequency axis (in kHz).

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 §2.6.1.3 for more.

3.3.3 ADSR envelope

discuss here ??

impulse has brief A and D, low S, R n/a

Chapter 4

Filtering

Chapter keywords: TBA

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.

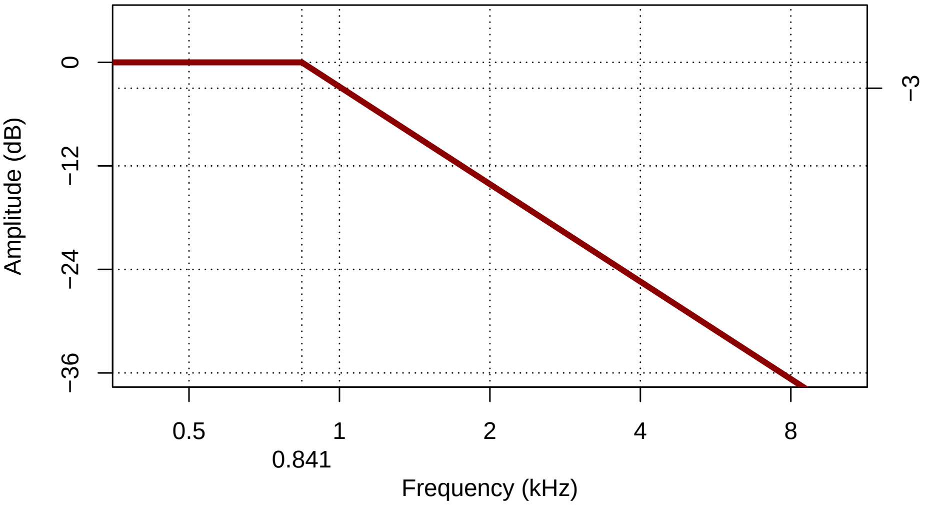

Figure 4.1: Frequency characteristic of a low-pass filter, with cutoff frequency 1000 Hz, and slope of -12 dB per octave.

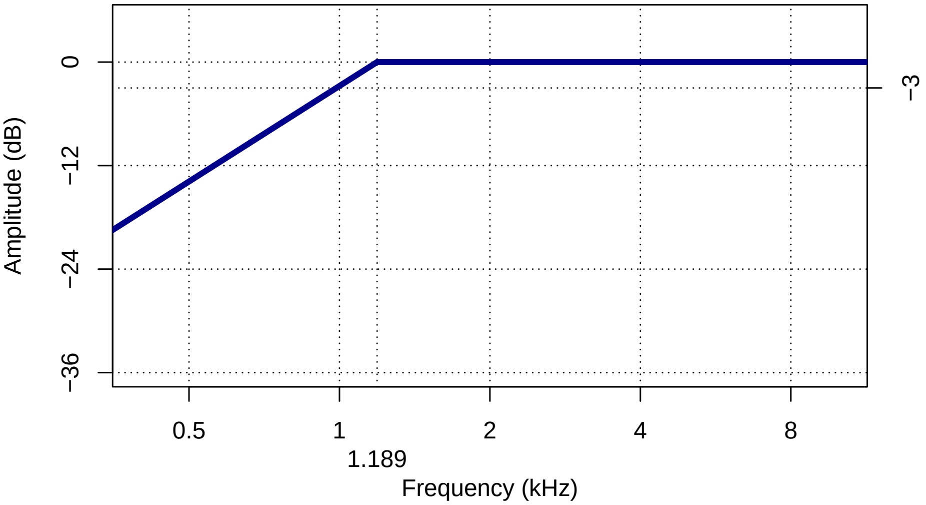

Figure 4.2: Frequency characteristic of a high-pass filter, with cutoff frequency 1000 Hz, and slope of -12 dB per octave.

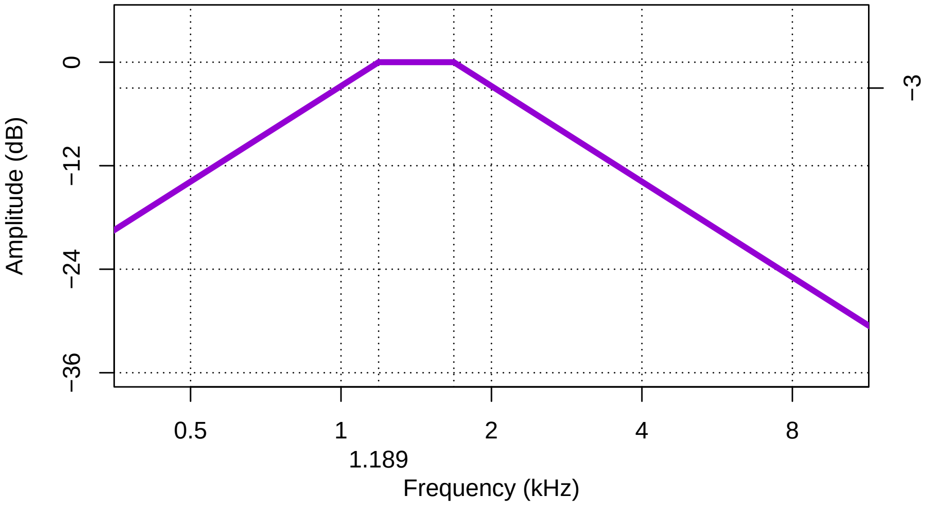

Figure 4.3: Frequency characteristic of a band-pass filter, with a pass band from 1 to 2 kHz (one octave), and with slopes of -12 dB per octave on both sides.

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

– 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*.

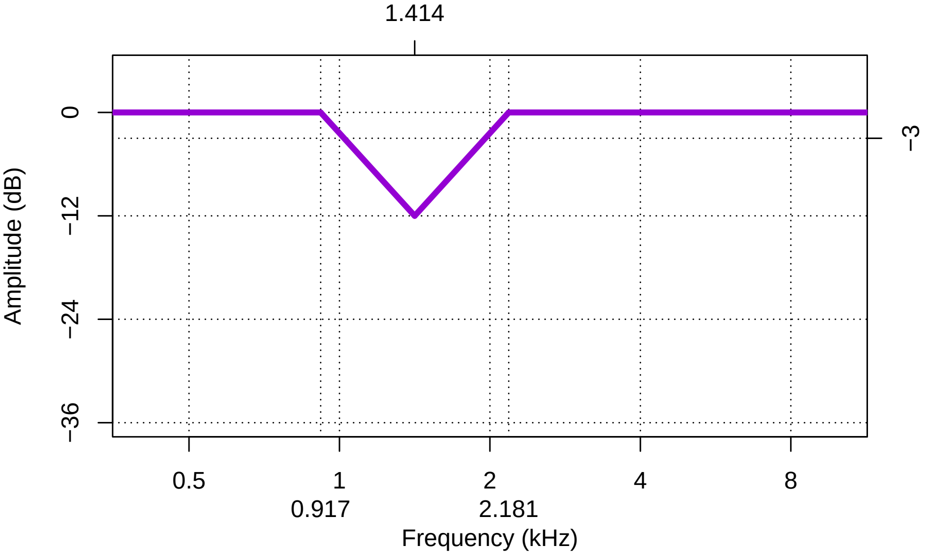

Figure 4.4: Frequency characteristic of a band-reject filter, with a reject band from 1 to 2 kHz (one octave), and with slopes of -24 dB per octave on both sides.

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 (§3.1.2), 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 (§3.1.2) 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

> TODO crossref source-filter, spectral slope of speech

For reasons that we will see later (in §REF), the spectrum of speech has a typical overall spectral slope of −6 dB/octave. On average, the amplitude of higher frequency components decreases by about −6 dB for each doubling of the frequency (§3.1.2). 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.3.3.

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. 70 Hz.
The resulting filtered output Sound object is again added at the bottom of the list of objects.

4.5 Time envelope

> TBA

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

Attack

Decay

Sustain

Release

Not only applicable to filter characteristics.

Bibliography

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

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.

Johnson, K. (2011). *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.

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.

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

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.