Friedrich-Alexander-Universität Erlangen-Nürnberg

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Lab Course

Speech Analysis

International Audio Laboratories Erlangen

Prof. Dr.-Ing. Bernd Edler

Friedrich-Alexander Universität Erlangen-Nürnberg
International Audio Laboratories Erlangen
Am Wolfsmantel 33, 91058 Erlangen

bernd.edler@audiolabs-erlangen.de

International Audio Laboratories Erlangen
A Joint Institution of the
Friedrich-Alexander Universität Erlangen-Nürnberg (FAU) and
the Fraunhofer-Institut für Integrierte Schaltungen IIS

Authors: Prof. DSc. Tom Bäckström,
Johannes Fischer,
Alexandra Craciun,
Tobias Jähnel,
Arjola Hysneli,
Esther Fee Feichtner

Tutors:
Esther Fee Feichtner,
Ning Guo

Contact:
Johannes Fischer
Friedrich-Alexander Universität Erlangen-Nürnberg
International Audio Laboratories Erlangen
Am Wolfsmantel 33, 91058 Erlangen
johannes.fischer@audiolabs-erlangen.de

Lab Course

Speech Analysis

Abstract

This experiment is designed to give you an overview of the physiology of speech production. Moreover, it will give an introduction to tools used in speech coding, their functionality and their strengths, but also their shortcomings.

1 Motivation

Speech is the primary means of human communication. Understanding the main properties of speech communication tools, such as mobile phones, enables us to develop efficient algorithms. The purpose of this exercise to demonstrate challenges in analyzing the most basic and important properties of speech signals.

2 Speech signals

Speech signals are usually processed using a rudimentary speech production model based on the physiology of the speech production apparatus. In this context, we will focus on the physiological articulation and acoustics only and disregard the higher levels of abstraction, e.g. the linguistic content of speech. We define the following terminology with respect to articulation:

Phoneme is the smallest linguistic unit which may bring about a change of meaning. It is thus the fundamental building block of linguistics.

Phonation is the process where the components of the speech production apparatus produce sound. Some include only voiced sounds in the definition of phonations, but here we will use it for all speech sounds, i.e. also for unvoiced sounds.

Phone is a speech segment with distinct perceptual or physiological characteristics. Note that many different phones can be classified within one phoneme, such that the same phoneme can be realized as different phones depending on, for example, personal style, context, or dialect. It is entirely possible or even common that the same phone is mapped to different phonemes depending on its context.

In this context, we will consider the phones as the smallest distinctive acoustic units of a speech signal.

3 Physiology and Articulation

In short, on a physiological level, speech production starts in the lungs, which contract and push out air. This airflow can cause two types of effects. Firstly, the airflow can induce oscillation in the *vocal folds*, periodically closing and opening, such that the emitted airflow resembles a (semi-) periodic waveform. Secondly, the airflow can cause noisy turbulences at constrictions of the *vocal tract*. The oscillating or noisy waveforms then flow through the vocal tract, whose resonances shape the acoustic signal. These three components — oscillating vocal folds, turbulent noise in constrictions and acoustic shaping of the vocal tract — give the speech signal its defining characteristics.

Figure 1 illustrates the main parts of the vocal apparatus. The air flows through the *larynx* and the *glottis*, which is the orifice between the vocal folds. The airflow then proceeds through the

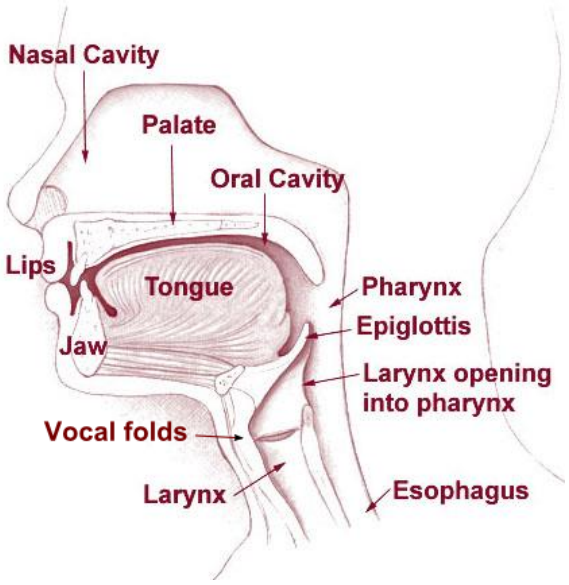


Figure 1: Human vocal apparatus used to produce speech. (Adapted from http://training.seer.cancer.gov/head-neck/anatomy/overview.html. This work is in the public domain in the United States because it is a work prepared by an officer or employee of the United States Government as part of that person's official duties under the terms of Title 17, Chapter 1, Section 105 of the US Code.)

pharynx, into the mouth between the *tongue* and the *palate*, between the teeth and is finally emitted through the lips. Sometimes air flows also through the *nasal cavities* and is emitted through the nostrils.

The most important excitation of the speech signal are the oscillations of the vocal folds. Given the right conditions, such as airflow speed and stiffness of the vocal folds, the airflow from the lungs brings the vocal folds to oscillate. When the airflow pushes the vocal folds open, they gain momentum. As the vocal folds open, air rushes through the glottis whereby the pressure drops. As a consequence, the vocal folds are not pushed out anymore but rather pulled back together, until they clash. As long as the airflow is constant, this process will continue in a more or less periodic manner.

Speech sounds produced by oscillations of the vocal folds are called *voiced* sounds and the process of uttering voiced sounds is known as *voicing*. This manner of articulation is called *sonorant*.

Figure 2 illustrates the vocal folds and the glottis viewed from above. Here the vocal folds are seen in their abducted or open position, where air can freely flow through them.

Unvoiced speech excitations are produced by constricting or even stopping the airflow in some part of the vocal tract, such as between the tongue and teeth, tongue and palate, between the lips or in the pharynx. This manner of articulation is thus known as *obstruent*, since airflow is obstructed. Note that these constrictions can occur concurrently with a voiced excitation. However, speech sounds with only an unvoiced excitation are known as *unvoiced* sounds. A constriction causes the airflow to go into a chaotic regime, which is effectively a turbulent mode. It is characterized by random variations in airflow, which can be perceptually described as noise.

Obstruent articulations where the airflow is obstructed, but not stopped, are called *fricatives*. When the airflow is temporarily stopped entirely, only to be subsequently released, it is known as a *stop*, and when the stop is released into a fricative, it is an *affricative*. Some of the main forms of articulation are listed in Table 1.

Finally, important characteristics of speech signals are defined by the shape of the vocal tract. The different shapes give the tube distinct resonances, which determine the differences between vowels. These resonances are known as *formants* and numbered with increasing frequency, such

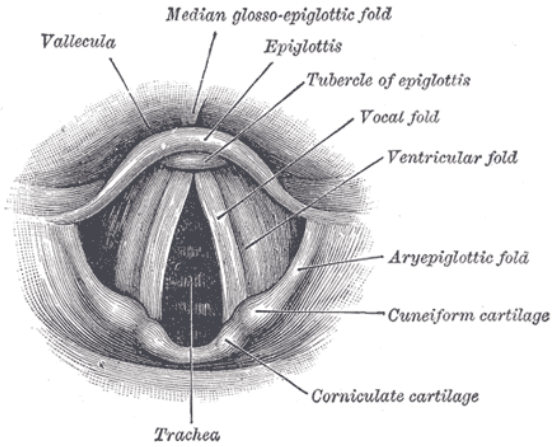


Figure 2: A view on the glottis from above. (This faithful reproduction of a lithograph plate from Gray's Anatomy, a two-dimensional work of art, is not copyrightable in the U.S. as per Bridgeman Art Library v. Corel Corp.; the same is also true in many other countries, including Germany. Unless stated otherwise, it is from the 20th U.S. edition of Gray's Anatomy of the Human Body, originally published in 1918 and therefore lapsed into the public domain.)

Table 1: Manners of articulation. Observe that several of these manners can be active at the same time.

Obstruent – airflow is obstructed
Stop – airflow is stopped, also known as *plosives*
Affricative – airflow is stopped and released into a fricative
Fricative – continuous turbulent airflow through a constriction
Sonorant – vocal folds are in oscillation
Nasal – air is flowing through the nose
Flap/Tap – a single contraction where one articulator touches another, thus stopping airflow for a short moment
Approximant – articulators approach each other, but not narrowly enough to create turbulence or a stop
Vowel – air is flowing freely above the vocal folds
Trill – consonants with oscillations in other parts than the vocal folds, such as the tongue in /r/.

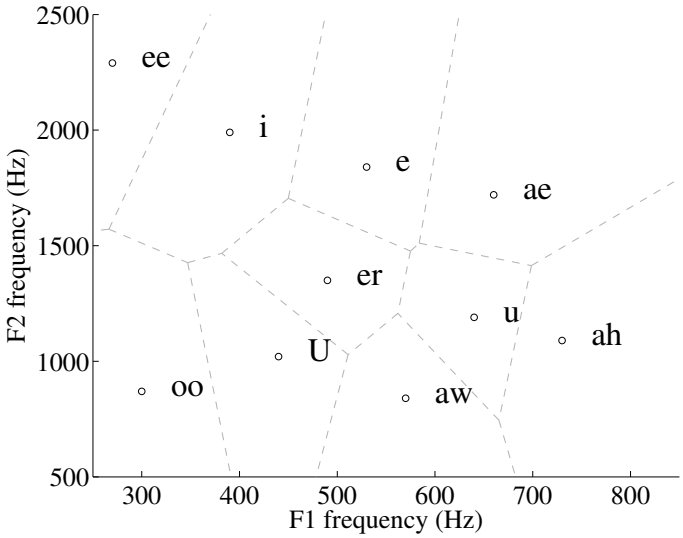


Figure 3: The distribution of vowels with respect to the two first formants, F1 and F2, averaged over 76 male English speakers. The dashed lines depict the approximate regions where phones would be classified to the corresponding phoneme.(Formant frequencies extracted from [1]).

that the first formant F1 is the resonance with the lowest frequency. The two first formants, F1 and F2, are from a linguistic point of view the most important, since they characterize the vowels. Figure 3 illustrates the distribution of English vowels on the axes of F1 and F2. We can see here that the formants are rather evenly distributed on the 2-dimensional plane. It is well-known that vowels are identified mainly based on F1 and F2, and consequently, they have to be well separated in the 2-dimensional plane in order to be easily identified. Conversely, would a language have vowels close to each other, they would most likely in time shift frequency such that they become more easily to identify, as people attempt to pronounce clearly and avoid misunderstanding.

Figure 4 illustrates the prototype shapes for English vowels. Here the characteristic peaks of formant frequencies are depicted, corresponding to the resonances of the vocal tract [1].

The importance of the two first formants is further demonstrated by the fact that we have well-known non-technical descriptions for vowel characteristics, which can be intuitively understood. Specifically, vowels can be described on the axes of closeness (closed vs. open) and backness (front vs. back). The standard form vowel diagram representing backness on the horizontal and closeness on the vertical axis, is depicted in Figure 5.

Observe that both the vowel diagram, as well as the F1 and F2 frequencies are unique for each language. For females and children, the frequencies are shifted higher in comparison to males, while the closeness and backness remain constant.

4 Phonemes

4.1 Vowels

As described before, vowels are sonorant phonations, that is, the vocal folds exhibit a periodic excitation and the spectrum is shaped by the vocal tract resonances (the formants). The two first formants define the vowels and their average locations are listed in Table 2. The third formant is less important, but is essential to reproduce natural sounding vowels.

Table 2 lists each vowel[1] with their corresponding symbol in the International Phonetic Alphabet

[1]This is a representative list of vowels, but in no way complete. For example, diftongs have been omitted, since for our purposes they can be modelled as a transition between two vowels.

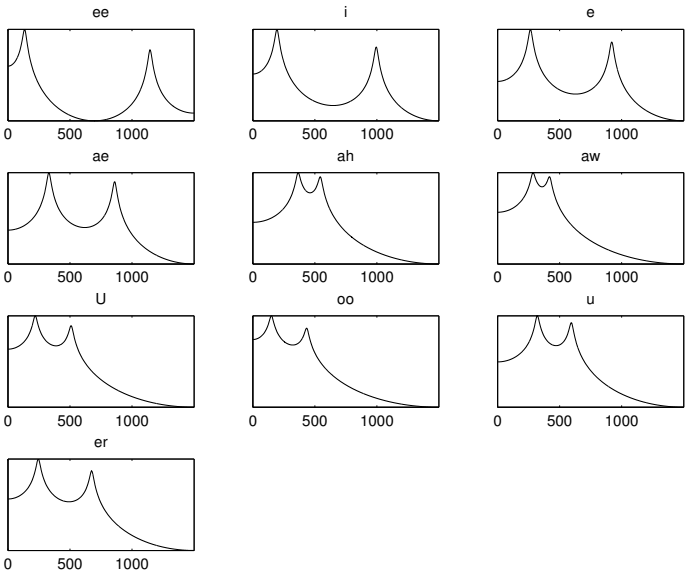


Figure 4: Illustration of prototype spectral envelopes for English vowels, showing the characteristic peaks of the first two formants, F1 and F2, averaged over 76 male English speakers, depicted on a logarithmic magnitude scale. (Formant frequencies extracted from [1]).

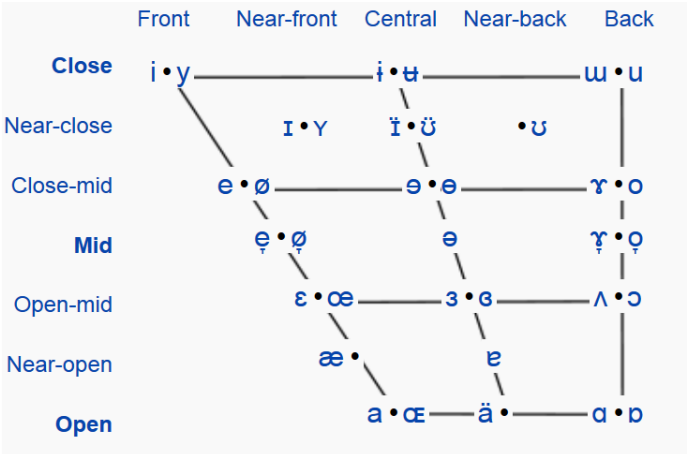


Figure 5: IPA Vowel diagram illustrating the vowel backness and closeness. From Wikipedia http://en.wikipedia.org/wiki/Vowel_diagram.

(IPA), as well as the symbol (or combination of symbols) using the Speech Assessment Methods Phonetic Alphabet (SAMPA) set. The latter has been designed for easy application on computers with only ASCII characters.

4.2 Consonants

In principle, all phonemes which are not vowels are consonants. However, note that linguists have more precise definitions. In the following, we will present the most important consonant groups, which correspond to the manners of articulation presented in Table 1.

4.2.1 Stops

In stops, the airflow through the vocal tract is completely stopped by a constriction and subsequently released. A stop thus always has two parts, a part where air is not flowing and no sound is emitted, and a release, where a burst of air causes a noisy excitation. In addition, stops are usually combined with a subsequent vowel, whereby the transition begins practically from the start of the burst.

4.2.2 Fricatives and Affricatives

Fricatives are consonants where the airflow is partially obstructed to cause a turbulent noise, shaped by the vocal tract. Affricatives begin as a stop, which later releases into a fricative.

4.2.3 Nasals, Laterals and Approximants

Most common nasals such as /n/ and /m/, and laterals and approximants such as /l/, /w/ and /r/ (without the trill) are sonorants, that is, the vocal folds are oscillating. For nasals, the air flows through the nose (at least partially) instead of the mouth. While this mode of phonation seems very different, some theoretical differences aside, we can still model it with the same approach as we model vowels. The nasal cavities form a tube similar to the vocal tract and can thus be modelled by a filter.

4.2.4 Trills

Trills, such as a rolling /r/, are characterized by an oscillation of some other part of the human production apparatus than the vocal folds. Most commonly, they are produced by the tongue, but can also be produced by the lips. Modeling such an oscillation is not easily encompassed in our model, but by extending the fundamental frequency model to include very slow oscillations, we obtain at least a basic functionality. The impulse train then models the oscillations of the tongue, the noise input models the associated turbulent noise and the shaping effect of the vocal tract is modelled by the linear predictive filter (Figure 6).

Note, however, that in most accents of English trills are not used, but approximants are used instead.

5 Intonation, Rhythm and Intensity

The linguistic content of speech is practically always supported by variations in intonation, intensity and speaking rythm. Here, intonation refers to the time contour of the fundamental frequency, rhythm to the rate at which new phonemes are uttered and intensity to the perceived loudness of the speech signal (closely related to the energy of the signal). By varying the three factors, we can communicate a variety of para-linguistic messages such as emphasis, emotion and physical state.

For example, the most important word of a sentence (or other segment of text) is pronounced in most languages with a high pitch and intensity, as well as at a slow speed. This makes the

Vowel		Formant (Hz)			Examples
IPA	SAMPA	F1	F2	F3	
i	i	290	2300	3200	cit**y**, s**ee**, m**ea**t
y	y	280	2150	2400	*German:* **ü**ber, R**ü**be
ɨ	1	290	2200	2500	ros**e**'s
ʉ	}	330	1500	2200	r**u**de
ɯ	M	330	750	2350	*Irish:* c**ao**l
u	u	290	595	2390	thr**ou**gh, y**ou**, thr**ew**
I	I	360	2200	2830	s**i**t
Y	Y	400	1850	2250	*German:* f**ü**llt
ʊ	U	330	900	2300	p**u**t, h**oo**d
e	e	430	2150	2750	*German:* G**e**nom, M**e**than, B**ee**t
ø	2	460	1650	2100	*French:* p**eu**
ə	@	500	1500	2500	**a**bout, **a**rena
ɘ	@\	420	1950	2400	*Dutch:* **i**k
ɵ	8	520	1600	2200	*Australian English:* b**i**rd
ɤ	7	605	1650	2600	*German:* m**ü**ssen
o	o	400	750	2000	*German:* **O**fen, R**o**man
ɛ	E	580	1850	2400	b**e**d
œ	9	550	1600	2050	*German:* H**ö**lle, g**ö**ttlich
ɜ	3	560	1700	2400	b**i**rd
ɞ	3\	580	1450	2150	*Irish English:* b**u**t
ʌ	V	700	1350	2300	r**u**n, w**o**n, fl**oo**d
ɔ	O	540	830	2200	l**a**w, c**au**ght, **a**ll
æ	{	770	1800	2400	c**a**t, b**a**d
ɐ	6	690	1450	2300	*German:* od**er**
a	a	800	1600	2700	h**a**t
œ	&	570	1550	1800	*Swedish:* h**ö**rt
ɑ	A	780	1050	2150	f**a**ther
ɒ	Q	650	850	2000	n**o**t, l**o**ng, t**a**lk

Table 2: Formant locations of vowels identified by their International Phonetic Alphabet (IPA) symbol as well as the computer readable form SAMPA. (From
http://en.wikipedia.org/wiki/Formant
http://en.wikipedia.org/wiki/Table_of_vowels
http://www.linguistics.ucla.edu/people/hayes/103/Charts/VChart/
http://en.wikipedia.org/wiki/International_Phonetic_Alphabet_chart_for_English_dialects) .

IPA	SAMPA	Examples
b	b	**b**uy, ca**b**
d	d	**d**ye, ca**d**, **d**o
ð	D	**th**y, brea**th**e, fa**th**er
ʤ	dZ	**g**iant, bad**g**e, **j**am
f	f	**ph**i, ca**ff**, **f**an
g	g	**g**uy, ba**g**
h	h	**h**igh, a**h**ead
j	j	**y**es, **y**acht
k	k	s**k**y, cra**ck**
l	l	**l**ie, s**l**y, ga**l**
m	m	**m**y, s**m**ile, ca**m**
n	n	**n**igh, s**n**ide, ca**n**
ŋ	N	sa**ng**, si**nk**, si**ng**er
θ	T	**th**igh, ma**th**
p	p	**p**ie, s**p**y, ca**p**
r	r	**r**ye, t**r**y, ve**r**y (trill)
ɹ	r\	**r**ye, t**r**y, ve**r**y (approximant)
s	s	**s**igh, ma**ss**
ʃ	S	**sh**y, ca**sh**, emo**t**ion
t	t	**t**ie, s**t**y, ca**t**, a**t**om
tʃ	tS	**Ch**ina, ca**tch**
v	v	**v**ie, ha**v**e
w	w	**w**ye, s**w**ine
z	z	**z**oo, ha**s**
ʒ	z	equa**ti**on, plea**s**ure, vi**si**on, bei**g**e

Table 3: Table of consonants used in English.
(From http://en.wikipedia.org/wiki/Help:IPA_for_English
http://en.wikipedia.org/wiki/X-SAMPA)

important word or syllable *really* **stand** `out` from its background, thus ensuring that the important part is perceived correctly.

Emotions are also often communicated by variations in these three parameters. I am sure the reader can imagine the speaking style which communicates anxiousness (rapid variations in the fundamental frequency F_0, high speed and intensity), boredom (small variations in F_0, low speed and intensity), sadness, excitement etc.

Sometimes especially intonation also plays an important linguistic role. For example, a sentence with a questions is, depending on language, often finished with a rapidly rising pitch, while a statement has a constant or sinking pitch. Thus the main difference between "Happy?" and "Happy!" is the pitch contour. Moreover, some languages use pitch contours to distinguish words. Such languages are known as tonal languages and they are especially common in Asia.

Homework Excercise 1

Speech Production

1. How are voiced sounds physiologically produced?
2. Which physiological part(s) of the speech production system gives vowels their characteristic features?
3. Which physical effects cause noise-like phonations?

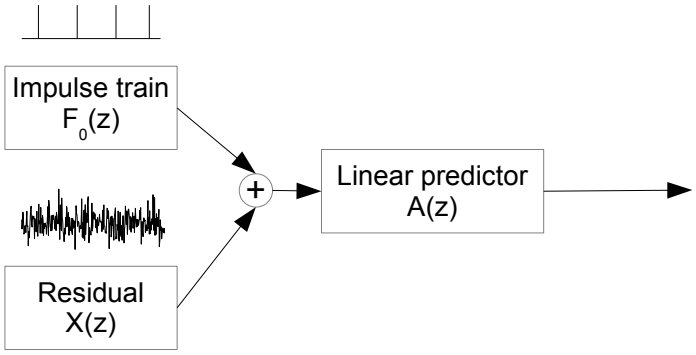


Figure 6: Simplified diagram of the dual excitation speech production model.

6 Introduction to this Lab course

The physiology of speech production described above is typically modeled by a so-called *dual excitation speech production model*, which is illustrated in Figure 6. The goal of this lab course is to extract the parameters needed to determine the dual excitation speech source model. The model consists of a filter $A(z)$, modeling the influence of the vocal tract. This filter is excited by an impulse train, resembling the fundamental frequency (F_0), as well as noise, resembling the residual (X).

Therefore, this lab will cover

- estimation of optimal analysis parameters such as frame length and model order,

- estimation of the vocal tract filter $A(z)$,
- application of linear predictive filtering.

The corner stone of this lab course are basic signal processing techniques, covering

- filters (finite impulse response (FIR), infinite impulse response (IIR)),
- filter representation (impulse response, coefficients, polynomial),
- transfer function and the Z-Plane.

If any of these terms is unfamiliar to you, please consider a short revision before the lab.

6.1 Windowing

Processing in this lab course is block-based, therefore the input signal has to be cut into pieces, the so-called *frames*. The frames for the processing of the signal should be of the length of the stationarity of speech (ca. 20 ms) to ensure constant statistic properties within one frame. At the same time, the framelength for the estimation of the autocorrelation should cover at least two periods of the fundamental frequency F_0 (100 - 400 Hz) and is typically longer. The longer frames for the estimation of the correlation need to be windowed by an appropriate window (e.g. Hamming), such that there are no discontinuities at the window borders. In contrast, the frames for the signal processing do not need windowing, since the linear prediction filter ensures continuity at the window borders. This is equivalent to applying a rectangular window of the size of the framelength, which only cuts the signal into frames. The two different windows are depicted in Figure 7.

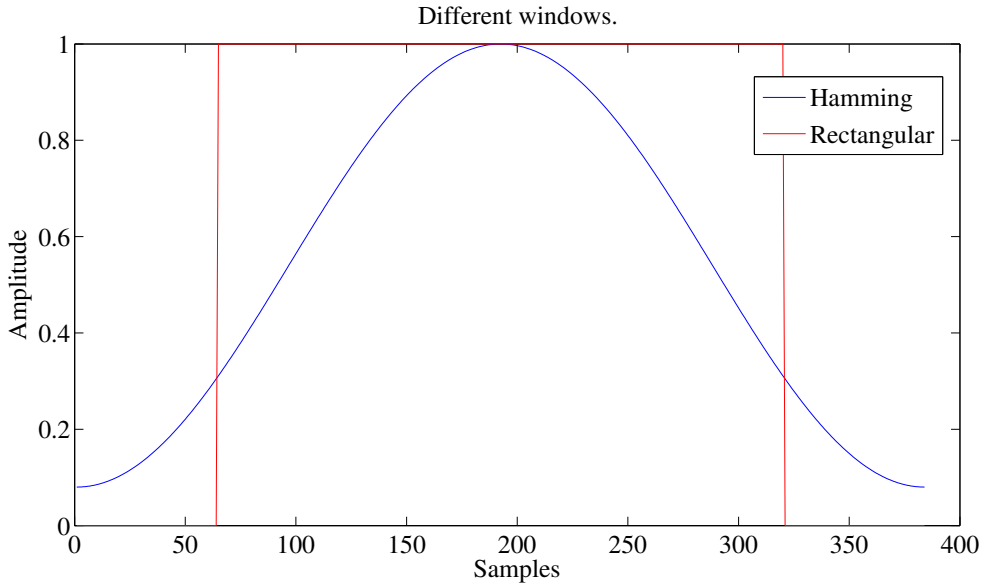


Figure 7: Illustration of the two windows.

6.2 LPC

6.2.1 Modeling the vocal tract.

Taking a second look at Figure 6, one notices that this speech production model consists of the filter $A(z)$, excited by uncorrelated noise and by a harmonic signal or a pulse train. The filter $A(z)$

models an approximation of the vocal tract by the so called tube model, which is illustrated in Figure 8.

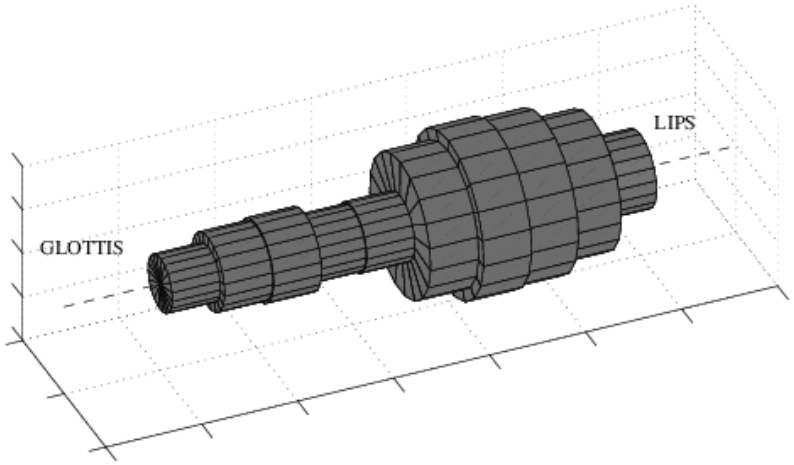


Figure 8: Illustration of the tube-model of speech production.

This tube model is analytically equivalent to a linear predictor of length M. The necessary predictor length M can be computed as

$$M = \frac{2 f_s L}{c}, \tag{1}$$

with c being the speed of sound, f_s being the sampling frequency and L being the length of the tube. The average length of the human vocal tract can be assumed to be between 14 and 17 cm. Such a linear prediction filter tries to estimate future values by a linear function of the previous samples

$$\xi_n = -\sum_{k=1}^{M} \alpha_k \xi_{n-k} + \epsilon_n, \tag{2}$$

with α_k being the prediction coefficients, ξ_n being the n-th input sample and ϵ_n being the residual, which is the unpredictable part of the signal.

Assuming a sufficiently high prediction order M (i.e. number of coefficients), it is possible to determine this tube model from the input signal and eliminate the correlation it introduced. Since the tube model is excited by white noise, the residual will ideally be uncorrelated and have a white spectrum.

Such a prediction filter can be derived by minimizing the residual ϵ_n in Equation 2. This minimization can be reformulated in matrix notation as

$$\mathbf{R}_{\mathrm{xx}} \mathbf{a} = [1, 0, \ldots, 0]^T, \tag{3}$$

where $\mathbf{R}_{\mathrm{xx}}$ is the autocorrelation matrix and $\mathbf{a} = [\alpha_0, \alpha_1, \ldots, \alpha_k]$ is the vector of prediction coefficients. Solving Equation 3 for $\mathbf{a}$ involves a matrix inversion, which is computationally complex. Therefore, the more efficient and stable Levinson-Durbin recursion algorithm should be used. This is available in Matlab as the function `levinson()`, which returns the coefficients $\mathbf{a}$ and requires the autocorrelation $\mathbf{R}_{\mathrm{xx}}$ starting at lag zero as input. However, calculating the correlation vector in Matlab normally results in a symmetric vector, also covering negative lags, as shown in Figure 9. Thus, everything before lag zero must be omitted.

The prediction filter can be interpreted as a whitening filter. The representation of such a filter in the Z-plane is given in Figure 12.

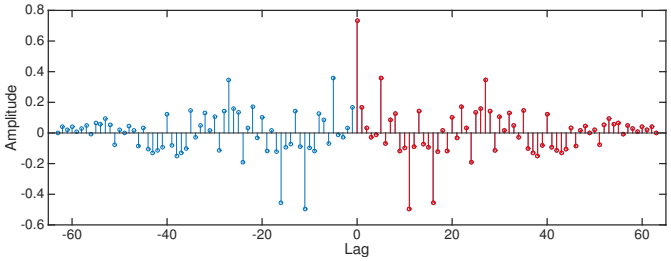


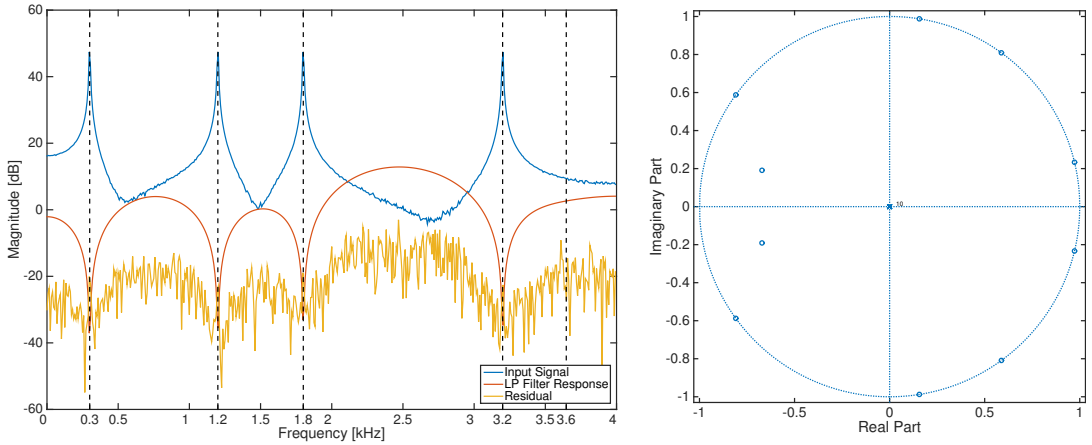
Figure 9: An example of a symmetric correlation vector, as produced by MATLABs `xcorr()` function. Only the red part should be used for the Levinson-Durbin recursion.

A measure describing how well the signal was predicted is the so-called prediction gain. This gain describes how much energy can be reduced by this prediction filter. The prediction gain is defined as follows:

$$\mathrm{PG} = 10\log_{10}\left(\frac{\sigma_s^2}{\sigma_n^2}\right), \tag{4}$$

with σ_s^2 being the power of the input signal and σ_n^2 being the power of the residual.

In order to better illustrate the behaviour of linear prediction, please have a look at Figure 10 and Figure 11. These figures show the input signal, the residual and the response of the linear prediction filter in the frequency domain. In addition, the filter coefficients are depicted in the z-plane. The input signal was chosen to consist of four sinusoids of same amplitude and frequencies of 300, 1200, 1800 and 3200 Hz. In addition, noise was added to achieve an SNR of approximately 50 dB. The model order in this scenario is four.



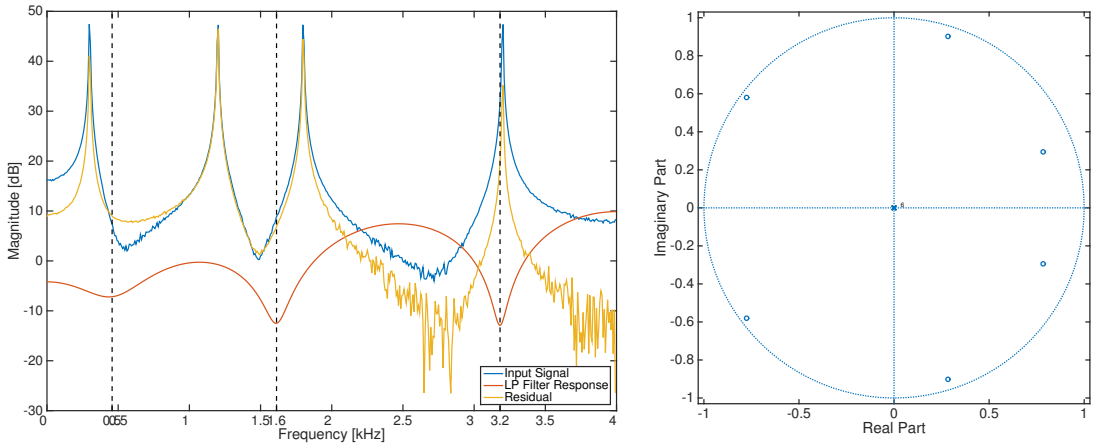
(a) (b)

Figure 10: Linear prediction in the frequency- and z-domains. Depicted for a prediction order of 10. Figure (a) depicts the input signal, the filtered signal and the linear prediction filter response in the frequency domain. The dashed lines illustrate the exact positions of the roots of the filter. Figure (b) shows the roots of the linear prediction filter on the z-plane.

In Figure 10a, the result of a predictor of order 10 is illustrated. It is clearly visible that the roots of the linear predictor are at the exact frequencies of the sinusoids. Therefore, the filtered signal, which in the case of linear prediction is the residual, appears rather white. The high suppression and small bandwidth of these filters can also be deduced from the z-plane shown in Figure 10b, as the roots are very close to the unit circle. As the prediction order was chosen higher than the degrees of freedom of the signal, the fifth root reflects the overall shape of the spectrum. As this

root is close to the origin, its effect is not really visible in the spectrum since it yields a very smooth filter.

In contrast to the upper scenario, Figure 11 shows the results if a smaller prediction order is used than the degrees of freedom. As the prediction order was chosen to be 6, this order is clearly lower than the degrees of freedom of the origin signal. As a predictor can be interpreted as a whitening filter, the roots try to cover the overall shape of the spectrum without having the degrees of freedom to suppress each sinusoid individually. Therefore, the residual signal is not as white as it was in the previous example. Figure 11a depicts this behavior, where we can see that the roots of the filter don't coincide with the frequencies of the sinusoids used to synthesize the signal. Moreover, Figure 11b illustrates that the roots are located closer to the origin, and therefore yield smoother filters.



(a) (b)

Figure 11: Linear prediction in the frequency- and z-domain. Depicted for a prediction order of 6. Figure (a) depicts the input signal, the filtered signal and the linear prediction filter response in the frequency domain. The dashed lines illustrate the exact positions of the roots of the filter. Figure (b) shows the roots of the linear prediction filter on the zplane.

6.2.2 Extracting the formants

In principle, when choosing $M = 6$, there is the chance that the roots of the filter are located at the first three format frequencies, as these should have the highest energy. It should be noted though, that in practice the order M should then be at least eight, in order to also cover a possible tilt of the spectrum.

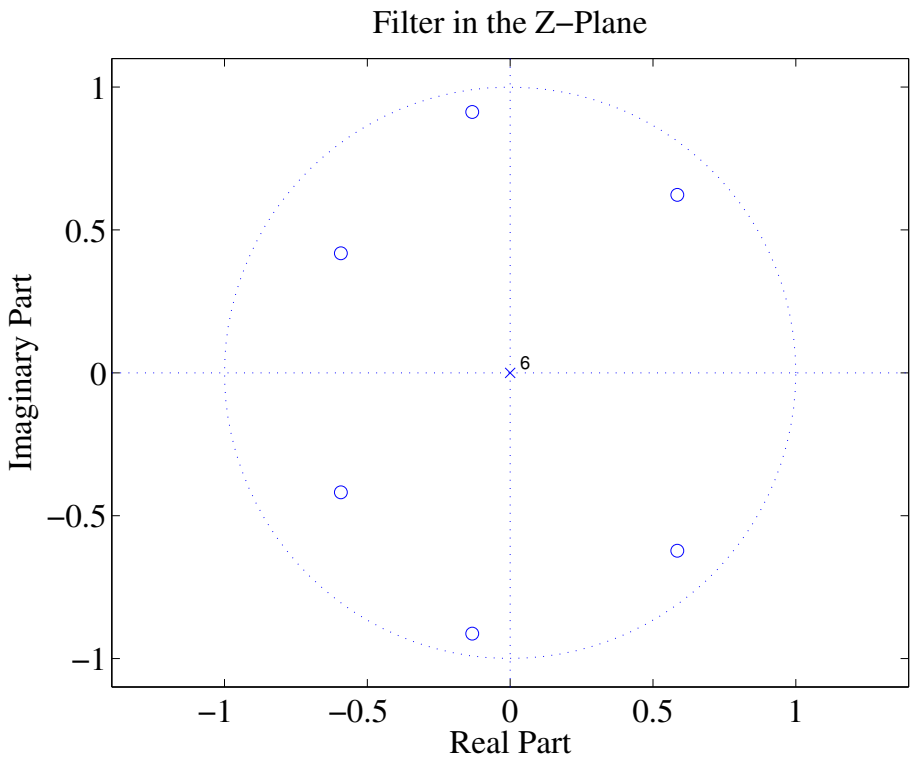


Figure 12: Example for a vocal tract filter in the Z-Plane.

7 Homework

Homework Excercise 2

Given the sampling frequency $F_s = 12800$ Hz calculate:

1. The framelength (in ms and samples) for the linear prediction, covering at least two periods of the fundamental frequency F_0.
2. The framelength in samples given the stationarity of speech.
3. Calculate the prediction order sufficient to model the human vocal tract as a tube model.
4. Assume you have a linear predictive filter of order M. What would happen to the prediction gain when you still increase the order of the predictor for the input signal being
   (a) a synthetic signal generated by filtering noise with an IIR-filter of order M,
   (b) pure speech, no background noise and
   (c) speech degraded by reverberation and background noise.

   Justify your answers.
5. Sketch the amplitude response of the filter given in Figure 12. (Coarsely, no calculations needed.)

8 The Experiment

In order to get a more concrete understanding of the tools used in speech coding, the following experiments should give you some practical insights. In the following, the steps which have to be done in the execution of the lab course will be described. An already existing script covering the framework is given in `speech_analysis.ipynb`. First complete the framework by performing the following tasks.

8.1 Completing your toolbox

Lab Experiment 1

In the first assignment **Toolbox** you will complete different functions that you will use later during the lab course. Each function is accompanied by an *assert* function used to test your code. These *assert* functions will throw an error in case your implementation is not working correctly. More detailed instructions can be found inline in the Jupyter Notebook **speech_analysis.ipynb**.

8.2 Visualization

Lab Experiment 2

The main focus of this lab course is getting familiar with different tools that are used in the field of speech coding. In order to gain a deeper understanding of these tools you are asked to plot different signals and filters, both in the time and frequency domain. More detailed instruction are given inline in the the code.

Good practice for figures:

- Every graphic should have a descriptive, yet short title.
- All axis need to be labeled. A sensible unit is indispensable. Time can be either displayed in **samples** or **seconds**, while the measures should be in a reasonable magnitude. Frequency should be displayed in **Hertz**, when possible.
- If multiple graphs are displayed in one figure, there must be a legend.
- Figures in the frequency domain should show powers, displayed in a logarithmic fashion.

8.3 Main Process

Lab Experiment 3

In this part of the lab course we will begin processing a speech file. After you filled in the correct parameters, calculated in your homework, you will load an audio file. This file is then resampled and later on frame-wise processed. More detailed instructions can be found as always in the Jupyter notebook.

Lab Experiment 4

Now that the processing is working. You should analyze different frames and interpret the graphs. Some interesting questions might be:

- What does the filter response look like in comparison to the signal in the frequency domain?
- What does it mean if all zeros of the filter are inside the unit circle?
- Which signal has the lowest energy: Input, the residual of lower or the residual of higher order?
- What kind of feature must the input signal have such that you would expect the highest prediction gain?

Lab Experiment 5

In the last part of the lab course we want to have a look on whether we can estimate the formant frequencies from the prediction filter.

- Why could this be possible?

Instead of loading the file *female_english_short.wav* you should now load the file *concatenated_phonemes.wav'*. This file consists of three phonemes: **i, ɛ and u** pronounced after another.

- Can you determine in which sequence they are without listening to the file?

References

[1] T. D. Rossing, *The science of sound*. New York: Addison-Wesley, 1990.