Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Speech Signal Processing

Prof. Dr.-Ing. Timo Gerkmann

Compilation date: April 3, 2018

Imprint:

Authors:	Prof. Dr.-Ing. Timo Gerkmann
Publisher:	Universität Hamburg
Edition:	Second edition (2017)
Layout:	Daryl Kelvasa

Learning objectives

- Speech production: How is speech produced by humans?
- Speech perception: How do we perceive speech signals?
- Speech synthesis: How can we produce speech synthetically?
- Speech analysis: What are the most important parameters of speech and how can we represent them?
- Speech coding: How can we code speech efficiently?
- Speech enhancement: How can we improve noisy speech?
- Speech recognition: How can we automatically rocognize speech by computers?

This Script is based on a transcription of the lecture *Speech Signal Processing* held by Prof. Dr.-Ing. Timo Gerkmann. Particular thanks goes to Daryl Kelvasa for his great job in putting my words, slides, and blackboard drawings into text, figures, and equations.

Timo Gerkmann, Hamburg 17.01.2018

Contents

1 — Introduction

Learning objectives

- Introduction
- Speech Production
- Source Filter Model
- Hearing

1.1 Introduction

Humans have evolved with the unique ability to use their lungs and vocal tracts to produce sounds that convey information. Being able to interpret these sounds allows us to pass down and impart knowledge and information about the world in which we live. Because speech is such an essential human trait, the advances made in speech signal processing have led to an exponential growth of our communication devices and have changed the world in which we live.

Already in 1939, Homer Dudley employed the source filter model of speech production in the development of his *Voder*. This machine produced artificial speech by employing an excitation signal that mimicked the signal produced by the lungs and the vocal chords. Band pass filters modeling the resonances of the vocal tract were used to filter the excitation signal in order to produce comprehensible speech. In present day, variations of this model are implemented in our cell phones and other speech transmission technologies. Digital signal processing plays a critical role in the efficient analysis, coding, transmission, enhancement and automatic recognition of speech.

Another benefit of signal processing is for people suffering from hearing loss. The input signal of a hearing aid must be digitally processed so that the output signal is comfortable for the user and fits the current listening situation. The usage of beamformers and different audio scenes allow for better speech comprehension in noisy environments while still being able enjoy music in a concert. Perhaps one of the most amazing speech processing technologies is the cochlear implant in which audio signals are converted into electrical impulses that directly stimulate the auditory nerve. The success of the cochlear implant in allowing the deaf to hear their first sounds is yet another remarkable achievement of speech signal processing.

The purpose of this class is to study some of the underlying processes involved in speech production and perception. This a priori knowledge will then be used to develop algorithms that allow us to digitally manipulate speech in order to perform certain tasks. These tasks are the foundation for the technologies that allow us to help the hearing impaired hear again and to communicate seamlessly with people across the world.

1.2 Speech Production

Speech production begins with the lungs which provide the airflow and therefore the energy required to produce speech. As this energy flows through the larynx, the vocal chords can either vibrate and periodically modulate the energy to produce a *voiced speech sound* or remain stationary to produce an *unvoiced speech sound*. This excitation signal then passes through the vocal tract in which different positions of the tongue, lips, jaw, etc., each correspond to unique resonances. The interaction of the excitation signal with the vocal tract produces the different sounds that are understood as speech. This was already noted by Isaac Newton in 1665 who wrote: "*The filling of a very deepe flaggon with a constant streame of beere or water sounds the vowells in this order w, u, ω, o, a, e, i , y*". The different levels of the liquid result in resonances that change our perception of the sounds created when filling the glass.

Figure 1.1, depicts the different anatomy involved in speech production. Airflow passes through the *trachea* which connects the *pharynx* and *larynx* to the lungs. The larynx contains the *glottis* and *vocal folds* that control the volume of speech and pitch of voiced speech by means of rapid vibrations. Once the modulated air leaves the pharynx, the mobility of the tongue and lips allow for fast changes in the geometry (and thus the resonances) of the vocal tract that produce the rapidly varying spectrotemporal properties that constitute different vocal sounds. The main two sections of the upper vocal tract are the oral and nasal cavities that are separated by the *velum*. An open velum allows the airflow to resonate in both oral and nasal cavities to produce speech that is commonly described as "nasal'.

Many unique speech sounds exist for the thousands of diverse languages spoken in the world. The most obvious distinction that can be made between all vocal sounds is to categorize them by their excitation. Vibrations of the vocal chords in voiced speech production create longer periodic vowel sounds. Short bursts of air corresponding to a noisy excitation are still filtered by the vocal tract to produce unvoiced speech sounds such as the fricative [sh]. *Plosives* are examples of unvoiced speech sounds in which there is a complete constriction of the vocal tract and then a sudden opening such as [k], [p], and [t]. Although the distinction between unvoiced and voiced speech is convenient for classification and modeling, many speech sounds contain characteristics of both excitations. These mixed excitation signals combine bursts of both types of excitations to produce sounds such as [v] or [z].

Figure 1.2 depicts time domain representations of the different speech sounds. The periodic excitation that produces a voiced speech sound is easily identified by a periodic structure in the temporal plot. The distance between two major peaks in the time-domain plot is called the *fundamental period*. This value represents the short amount of time during which the vocal chords open, close, and re-open. The airflow produced by unvoiced speech signals is more turbulent as it lacks this periodic modulation of the glottis. The result is an excitation with a more diverse spectral content which can be identified by the large amount of zero crossings in

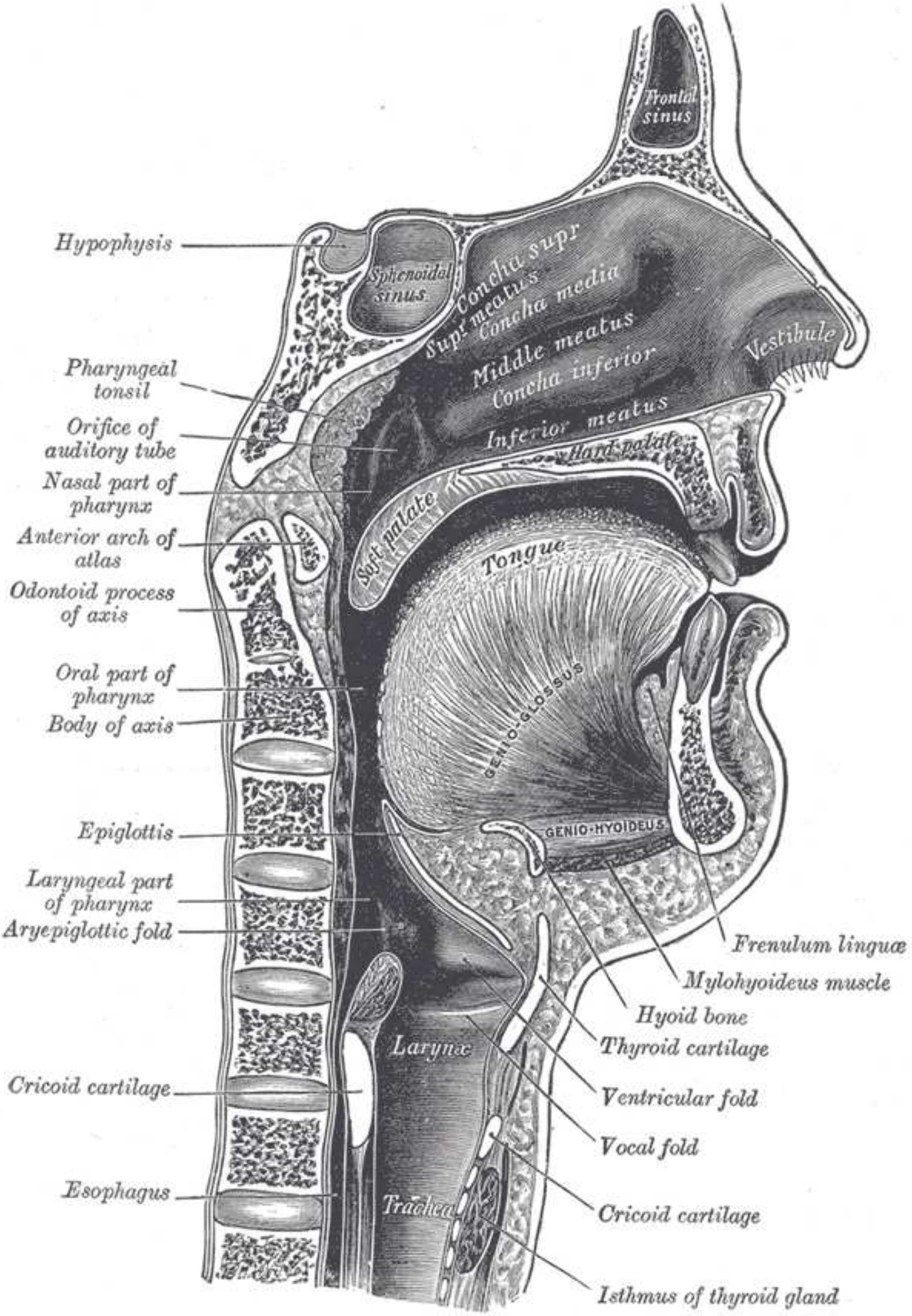
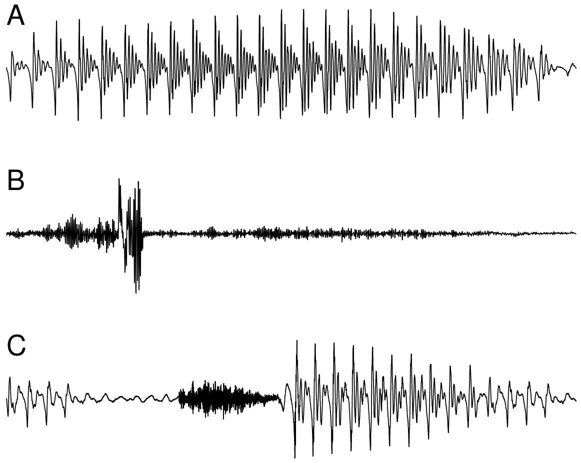


Figure 1.1: Anatomy of the vocal tract. [10]



A) voiced B) unvoiced C) transition vowel-plosive-vowel

Figure 1.2: Time domain signals of voiced and unvoiced speech.

the time domain plot.

Speech sounds convey meaning, however the complex nature of speech does not allow for a simple one-to-one mapping of the speech sound to an intended meaning, thus making it necessary to develop a system to classify these sounds in a linguistic sense. A *phone* is defined as the smallest speech segment with distinct physical or perceptual properties. These distinctions can be obvious as in the difference between the [k] in ”cat” and the [p] in ”pat”, or not so obvious as in the [p] in ”pin” and the [p*] in ”spin”. The former pair may sound similar, however they contain spectrotemporal differences that make them distinguishable. Despite these differences, both words are still comprehensible after interchanging the two sounds, thus constituting a *phoneme*. Notice that this is not the case for the former pair of phones. Phonemes are the smallest segments of speech that can change the meaning of a word. Different realizations of a phoneme are called *allophones* meaning that one allophone is one of the many possible phones that can constitute a phoneme. The words ”cat”, ”kit”, ”school”, ”skill” all contain the phoneme [k] but are pronounced differently due to co-articulation effects from the different vowel transitions. These different phones would therefore form a set of allophones for the phoneme [k].

The large number of allophones that belong to a given phoneme can be explained by *co-articulation*. The position of the vocal tract cannot be changed instantaneously, and the result is a smooth transition of the tongue from one position to the next. The [n] in the word ”hen” is produced by placing the tongue behind the front teeth to create what is known as an *aveolar* sound. However, the [n] in the word ”tenth” is followed by the dental sound [th] that results in it being pronounced more dentally. This difference in articulation distinguishes the two as different phones, however they are both allophones of the same phoneme [n].

Different parts of the vocal tract are used to generate the different phonemes that constitute the thousands of languages spoken across the world. Natural human languages have between 10 and 80 phonemes that can be characterized by the manner in which they are articulated: voiced/unvoiced and place of articulation. The place of articulation is defined by the position of

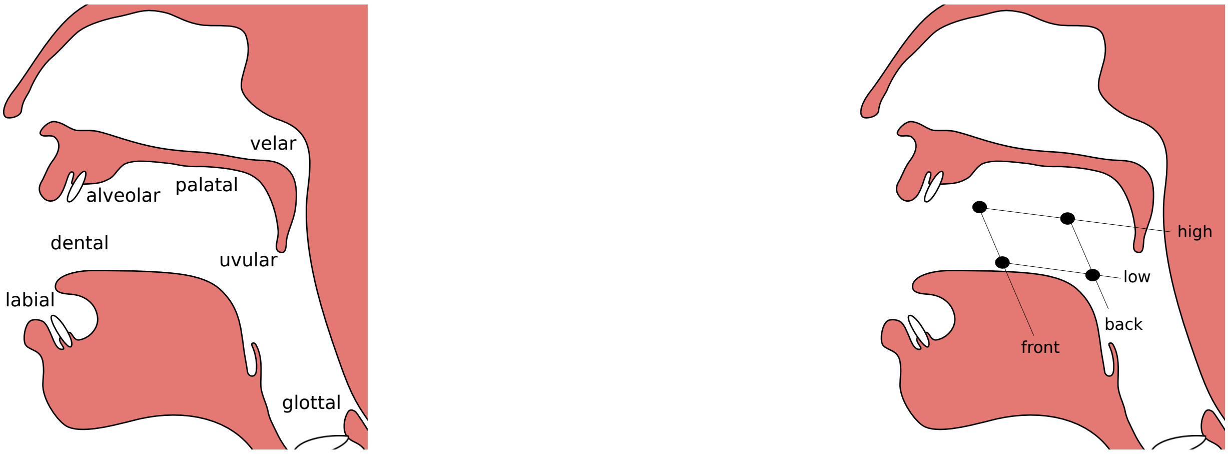


Figure 1.3: Left: Schematic drawing representing the place of articulation corresponding to phoneme production. Right: Tongue positions to characterize the production of the cardinal vowels. [9]

the tongue when the speech sound is produced. These different postions can be seen in Figure 1.3. These methods of distinguishing between phoneme production allow for a phonetic alphabet that are language independent. Figure 1.4 shows this phonetic alphabet for consonants distinguished by place and type of articulation. A language independent description of the vowel phonemes can also be created by characterizing vowel sounds by the relative position of the tongue when they are generated. The *primary cardinal vowels* are a set of phonemes created by a two-dimensional mapping (high/low and front/back) of tongue position in the oral cavity to the corresponding vowel sound. This mapping is used to characterize the respective phonemes shown in Figure 1.4. One axis depicts the positioning of the tongue from back to front, and a different axis depicts the opening of the mouth. Primary cardinal vowels can also be distinguished from the *secondary cardinal vowels* which are less common and more difficult to pronounce. The main difference between the primary and secondary cardinal vowels is the shape of the lips, which can be either open or round.

Prosody is another important characteristic of speech that encompasses the rhythm, stress, and intonation of an utterance. Although this course will place most of the focus upon the spectral content of speech, the temporal characteristics of speech that are described by prosody still carry important information. Differences in prosody can for instance constitute the difference between a question and a statement by simply raising the fundamental frequency at the end of the sentence.

”I’m going the *bank*”. (Spoken normally) $\Longrightarrow$ Declarative statement of intent.
”I’m going the *bank*”? (Pitch shifted high) $\Longrightarrow$ A question is being asked.

Emphasis on specific words can also greatly change the meaning of a sentence.

”Put the *green* ball on the table”. $\Longrightarrow$ There are multiple balls of different color.
”Put the green *ball* on the table”. $\Longrightarrow$ There are multiple objects that are all green.

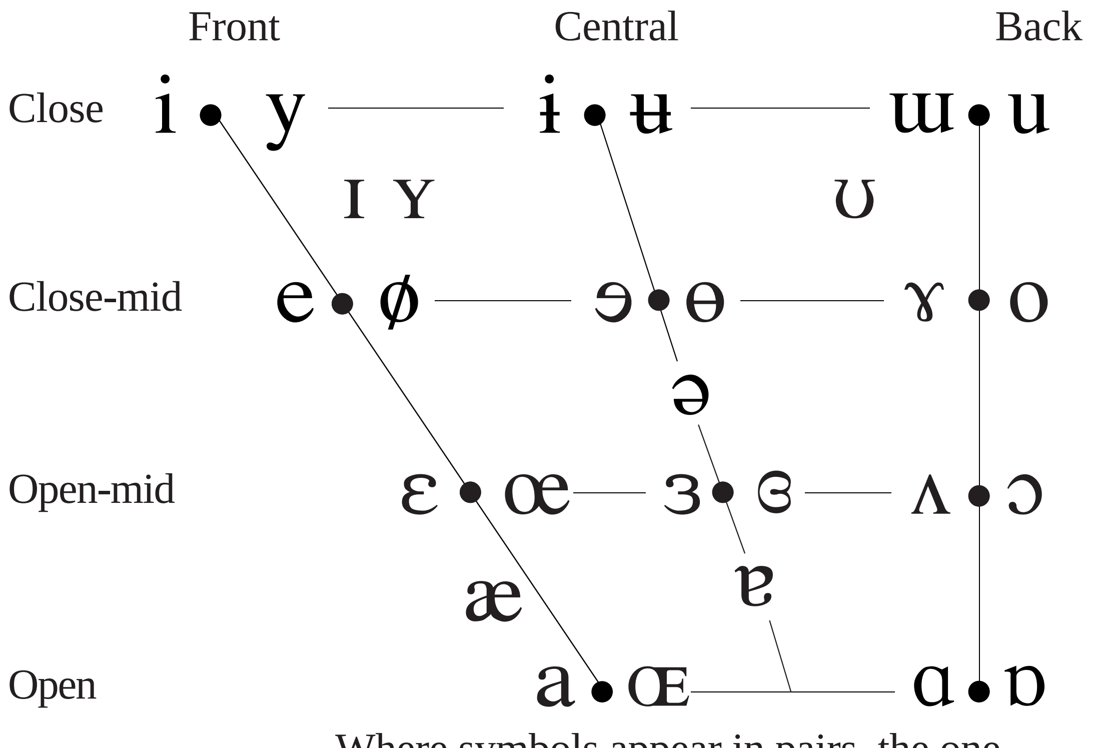
THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

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

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

VOWELS



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

Figure 1.4: Top: Chart of international phonemes and their corresponding places of articulation. Bottom: Chart of tongue positions and the corresponding cardinal vowels. [11]

Information about the emotional state of the speaker is also carried in prosodic cues. Yelling generally instills a sense of urgency in the listener. Perhaps there is some emergency that the speaker is warning about. This reaction in completely different from that produced when the same utterance is whispered.

”I’M BEING FIRED”. (yelled) ⟹ The speaker is expressing outrage.
”I’m being fired”. (whispered) ⟹ This is private information for just the listener.

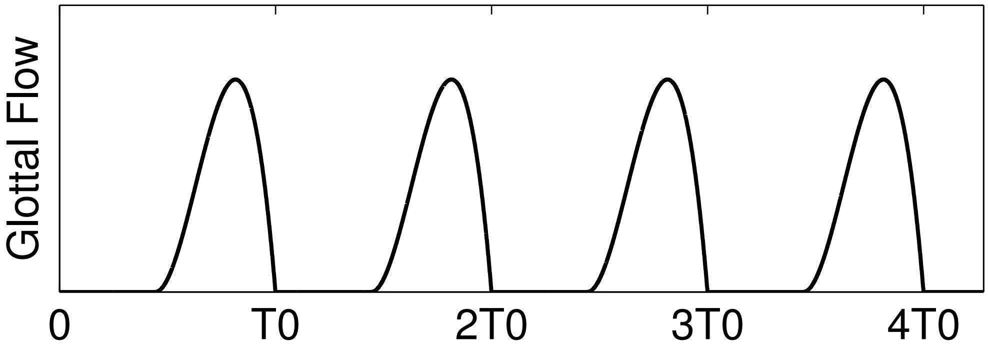


Figure 1.5: Glottal flow as a function of time.

1.3 Source Filter Model

Section 1.2 introduced the underlying anatomical mechanisms behind speech production. The lungs produce energy, in the form of airflow passing through the larynx. This airflow is either modulated by the oscillating vocal chords to produce voiced speech or allowed to flow into the vocal tract to produce unvoiced speech. The current positioning of the jaw, lips, tongue, etc., spectrally shapes this energy to produce the final speech sound. This process can be represented formally by the *source-filter model* so that is can be analyzed and computationally reproduced. In this model, the post-larynx airflow is the source, and the resonance frequencies of the vocal tract constitute a mathematical filter. A very critical assumption of the model is that these two processes remain independent of each other.

Due to the fundamental differences in voiced and unvoiced speech, the ”source” part of the model requires two separate methods for modeling the excitation signal. Voiced speech is produced from a modulated excitation signal that results from the periodic opening and closing of the vocal chords. This periodicity is shown in Figure 1.5 as the glottal flow behind the larynx as a function of time. The process begins with a closed glottis and as energy is produced by the lungs, the glottis is forced to open due to an increased pressure that builds up at the base. The opening of the glottis allows this pressurized air to escape and Bernoulli's principle states that the increased airflow results in a decrease in pressure. Because the vocal chords are under constant tension, this decrease in pressure allows them to snap shut and begin the process again. The result is a voiced excitation that is modulated at a fundamental period corresponding to this periodicity. To model this mathematically, a time-domain pulse train (or Dirac comb) can be used to model the opening and closing of the glottis by using a peak to peak distance corresponding to the fundamental period, T_0.

The turbulent airflow of unvoiced sounds lacks this periodicity, so it is better described by a random process. Figure 1.6 shows the time course of a white Gaussian signal, its histogram, and spectrum. The spectrum of the white Gaussian noise is flat meaning that it is made up of an equal

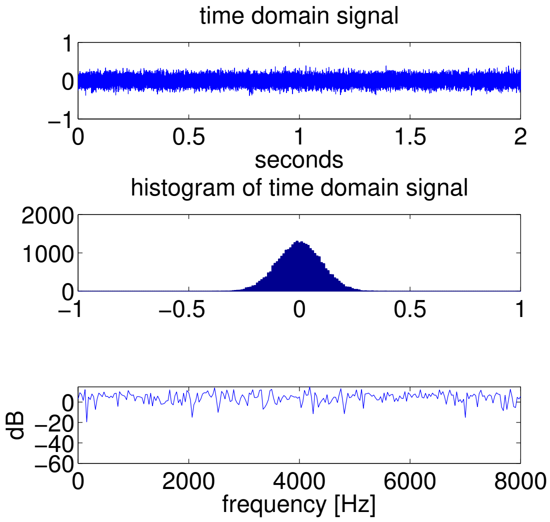


Figure 1.6: (Top) Time, (Middle) statistic, (Bottom) and frequency domain representations of white Gaussian noise.

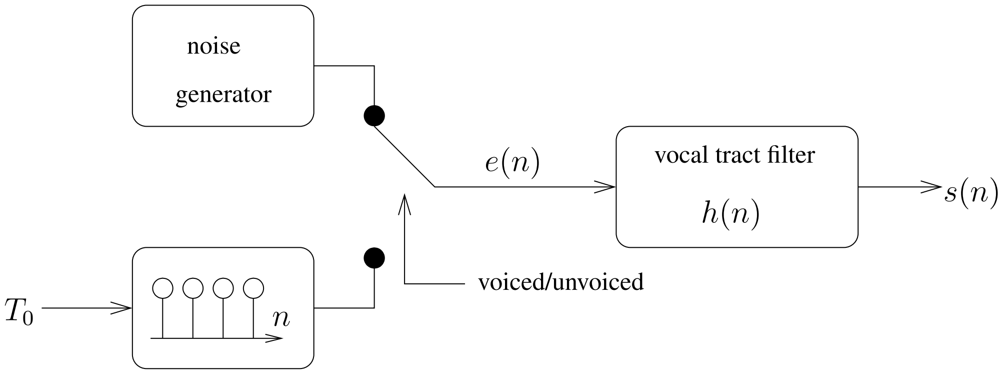


Figure 1.7: Simplified model of speech production.

distribution of all frequencies. This is analogous to ”white” light in optics, which also consists of an equal distribution of frequencies corresponding to all the colors of the visible spectrum. To model an unvoiced speech signal, a noise generator can be used to randomly choose numbers from a Gaussian distribution. The two forms of excitation can now be described by the simple model in Figure 1.7. It can be seen that it is necessary to have a switch that chooses between the two forms of excitation. This implies that the model requires some form of detector that can determine the type of excitation that is present in the current speech signal. Of course, this model is limited in that it cannot produce the mixed excitation signals that were introduced in Section 1.2. One can then imagine more complicated models in which a mixed excitation is produced by a weighted summation of the two excitation signals.

The vocal tract can now be modeled as a filter through which the chosen excitation signal passes. This is done by greatly simplifying the complexity of the vocal tract to the tube model in Figure 1.8. One section corresponds to the nasal cavity in which the output is the nose, whereas the bottom represents the oral cavity in which energy escapes through the mouth. The velum regulates the passage of air between the oral and nasal cavities, and this can be modeled by a simple switch. Much like a woodwind instrument, the shape and length of the tube segments corre-

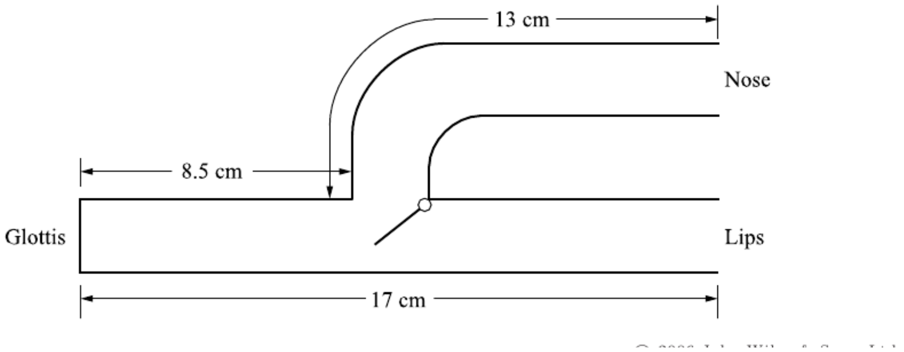


Figure 1.8: Simplified tube model of the vocal tract.[7]

spond to resonance frequencies that concentrate the energy of the excitation signal into certain frequency ranges called *formants*. These resonance frequencies can be mathematically described as a filter in what is referred to as the *vocal tract transfer function*. By changing this vocal tract transfer function over time, synthetic speech production can be achieved.

Mathematically speaking, this filtering is represented by a convolution of the excitation signal and the vocal tract impulse response in the time domain or a multiplication of the excitation spectrum and vocal tract transfer function in the frequency domain (Figure 1.9). After the multiplication, the spectrum of the final speech sound is what would be seen upon a Fourier transform of the recorded speech. The spectrum of the excitation signal consists of the fundamental frequency and a set of harmonics at integer multiples of the fundamental frequency. The formants appear as spectral peaks in the transfer function that correspond to the resonances in the vocal tract. Formants contain important information because they are used by the brain to distinguish between different speech sounds. The influence of the excitation signal and the vocal tract are both present in the final signal. It is very important to understand that these are two independent signals in the model. The final signal in Figure 1.9 reveals that the formant *peak* is not always present in the final speech sound. This is because in a sense, the excitation signal samples the transfer function at discrete points, namely the fundamental frequency and its harmonics. As a consequence, the fundamental frequency and its harmonics are the only frequencies in the final signal. These are then spectrally shaped by the vocal tract transfer function which concentrates the energy of the harmonics into the corresponding formant frequency regions. When the fundamental frequency or a harmonic falls precisely at a formant peak, the transmitted energy can be maximized. This is often exploited by singers.

The fundamental frequency and formant frequencies of the vocal tract are two fundamentally different things and it is important to distinguish between the two. The fundamental frequency contains information with respect to intonation and prosody, however it carries no real information with respect to meaning. Also, the fundamental is a property of the excitation signal which is the active energy source of the modeled system. In contrast, the vocal tract transfer function is a passive filter and adds no new energy into the system. Instead, the filter allows the harmonics of the excitation signal to resonate within the formant frequency regions to produce a spectrally colored phoneme. In particular, the first two formants are essential for vowel recognition and can be used classify vowel phonemes as seen in Figure 1.10.

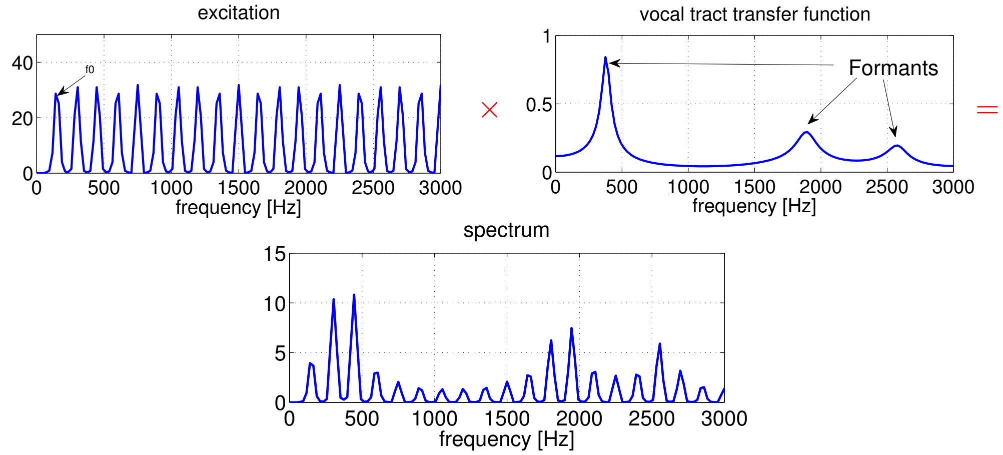
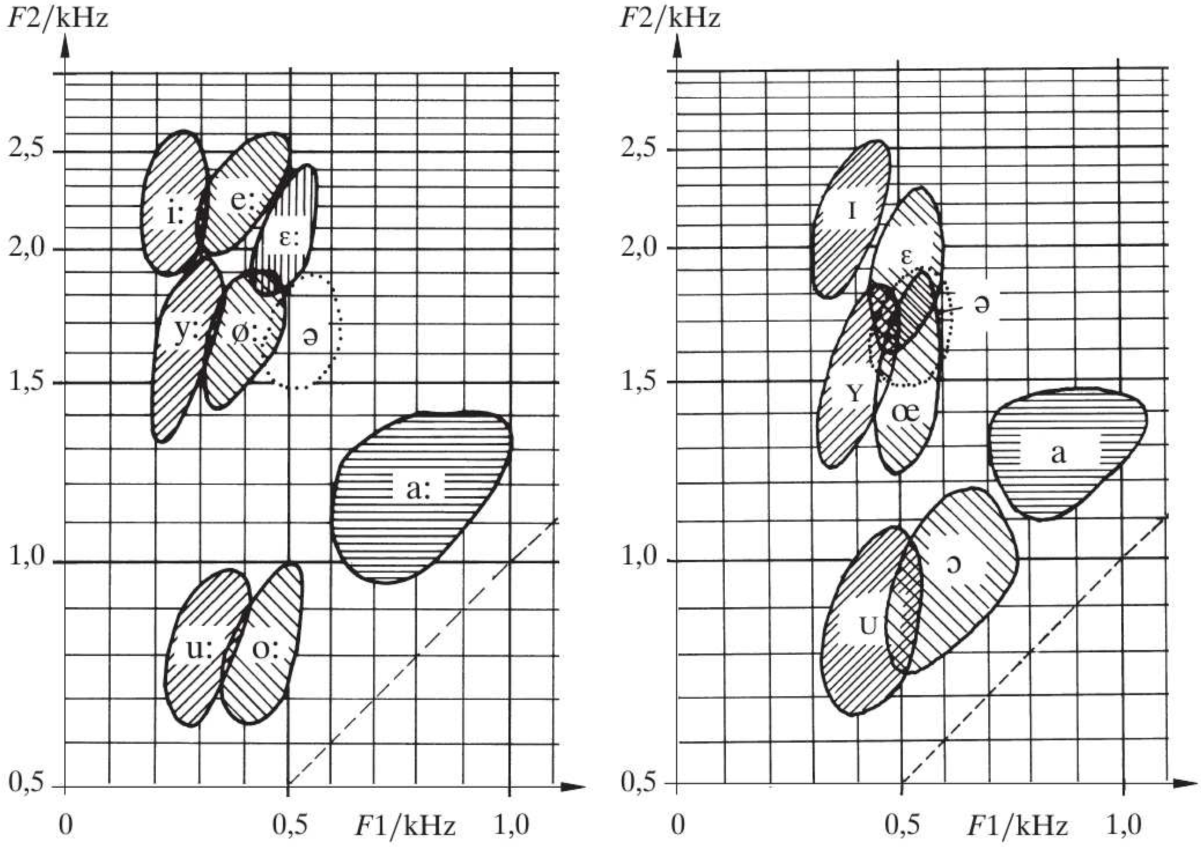


Figure 1.9: Top: Spectrum of voiced excitation signal. Middle: Vocal tract transfer function. Bottom: Spectrum of excitiation signal filtered by the vocal tract.

This section has introduced a formal model for the speech production process in which mathematical operations can be performed using a set of parameters to synthetically produce speech. To derive these parameters, it is necessary to compute whether a speech frame contains voiced or unvoiced speech, the fundamental period if the speech is voiced, and the vocal tract transfer function. A large part of this course will deal with extracting these parameters from real speech signals, so that they can be coded, transmitted, and then re-synthesized using this simplified model of speech production.



Quelle: Vary, Heute, Hess (1998): Digitale Sprachsignalverarbeitung, Teubner, Stuttgart

Figure 1.10: Phonemes as a function of first and second formant frequencies.[7]

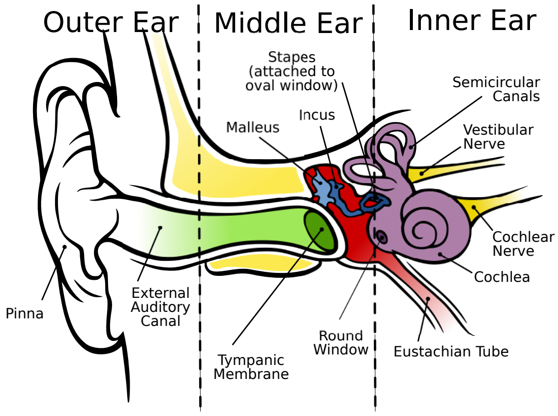


Figure 1.11: The peripheral auditory system. [12]

1.4 Hearing

The physiological components that are involved in the peripheral processes of human sound perception can be divided down into three main parts: the *outer ear*, the *middle ear* and the *inner ear*. The outer ear collects acoustic energy in the form of aerial compression waves and transfers it to the middle ear where it is converted to fluid waves within the inner ear. The inner ear then converts this information into action potentials that the brain can use to interpret the sound. These parts and their corresponding components are depicted in Figure 1.11.

The outer ear consists of the *pinna*, the *ear canal*, and the *ear drum* or *tympanic membrane*. Sound pressure compression waves travel through the ear canal, where they transfer their energy to the tympanic membrane. This membrane is connected to the inner ear by the *malleus*, *incus*, and *stapes*. These small bones transfer energy from the larger area of the tympanic membrane to the smaller area of the *oval window* which is attached to the *cochlea*, a fluid filled spiral shaped cavity within the inner ear. The acoustic compression waves are therefore converted to fluid waves, that produce a traveling wave along the *basilar membrane* within the cochlea. The length of the basilar membrane contains points of varying stiffness that allow this wave to travel until it reaches a frequency dependent position where it can then resonate. The tectorial membrane and the organ of Corti are located on the basilar membrane as shown in Figure 1.13. The resonating basilar membrane creates sheering forces between these two components that result in the stimulation of hair cells which are innervated by auditory nerves. The auditory nerves then fire action potentials that the brain can interpret to perceive sound. Figure 1.12 shows that the peak of a high frequency tone would occur closer to the base of the basilar membrane, whereas the resonances of lower frequencies, would occur more toward the apex. This implies that hair cells close to the base encode the high frequency content of the signal, while the hair cells at the apical end encode the low frequency content of a signal. This tonotopic representation

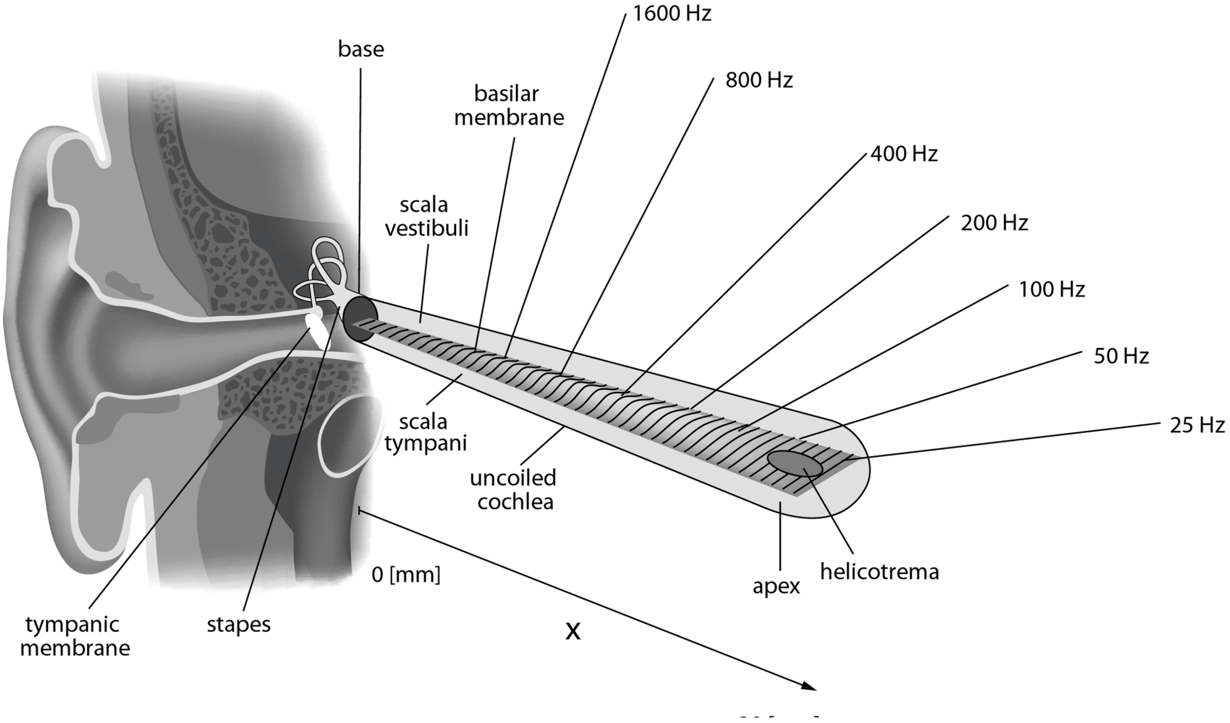


Figure 1.12: Frequency place coding along the cochlea. [15]

is a form of frequency place coding that results in a frequency analysis of the signal analogous to the techniques that we will be implementing in this course. This is also the reason why this frequency analysis is such an intuitive signal representation.

Figure 1.14 depicts the auditory sensation area for a normal hearing person as a function of frequency and level. It can be seen that more energy is required to perceive lower frequency sounds. Humans are able to perceive frequencies between 20 Hz to 20 000 Hz and begin to have problems with higher frequencies at older ages. Figure 1.14 also shows the threshold of pain that defines the sound level at which permanent hearing damage occurs when the listener is exposed to supra threshold sound for an extended period of time. Interestingly, speech formants correspond to a range of the auditory sensation area where the ear is most sensitive. This implies that the ear has evolved to be optimized for speech perception (or the other way around!). The hearing impaired, or those suffering from sensory neural hearing loss, maintain similar threshold levels of pain as normal hearing individuals, however their threshold of hearing is increased. A hearing aid could then be used to amplify softer sounds, but simply implementing linear amplification would result in output signals with levels beyond that of the threshold of pain. Compression algorithms can then be used in hearing aids, so that softer sounds are amplified more than louder sounds. However, this process significantly decreases the SNR and therefore requires some noise reduction to enhance the final signal.

It is interesting to note that there are different types of hearing losses that are associated with different methods of treatment. In a conductive hearing loss, the small bones within the middle ear can stiffen over time so that sound is not properly conducted to the inner ear. Sound is still perceivable, but attenuated, meaning that a hearing aid can be nicely used to amplify the sound and treat the condition. Conductive hearing losses are generally easier to treat than sensory

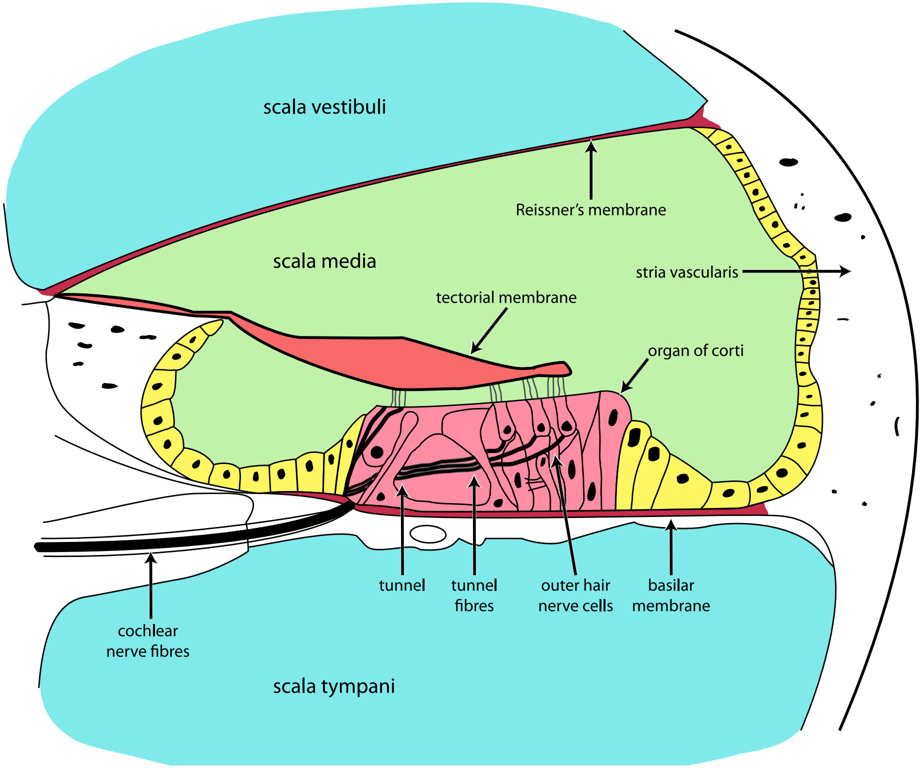


Figure 1.13: Cross section of the cochlea. [13]

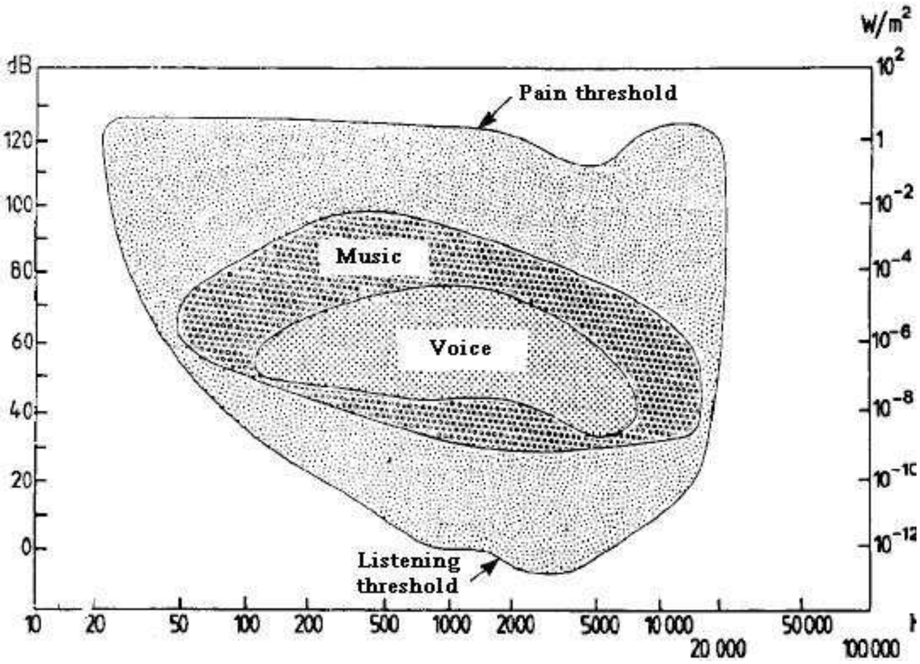
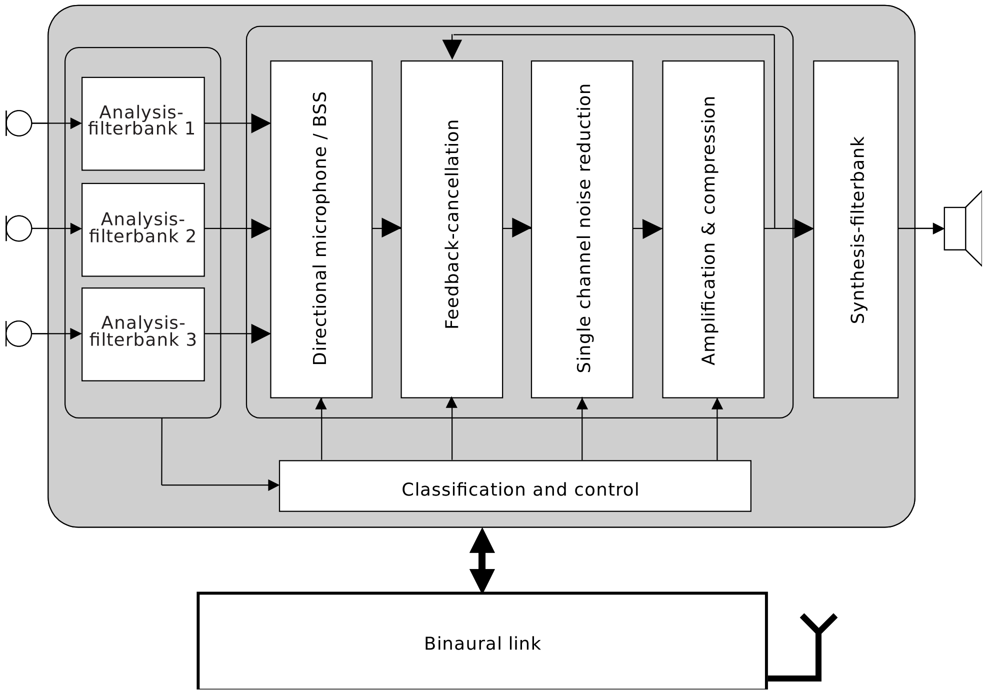


Figure 1.14: Audible range in SPL as a function of frequency. Frequency range of common sounds are color coded. [14]



Hamacher et al. in: Martin et al. (eds.), Wiley 2008

Figure 1.15: Flow chart depicting the processing stages of a hearing aid. [8]

neural hearing losses in which inner hair cells die and the sound cannot be transferred into a neural code. This can result from trauma induced by long term exposure to high level sound or simply by old age. New hair cells do not grow, nor do they regenerate, so that once the hair cells are damaged, they are gone forever. As a result, soft sounds are no longer perceivable, while at the same time the threshold of pain remains the same. This produces a reduced area in the auditory sensation map in Figure 1.14. Sensory neural hearing loss is also problematic because it is often accompanied by a decrease in frequency resolution of the human ear. Because the frequency resolution of the auditory perception is is not sufficient to separate speech from noise, one could theoretically still perform well on an auditory test, but have trouble perceiving speech in noisy environments. This is another application of noise reduction algorithms that can greatly aid the hearing impaired.

Current hearing aids commonly implement multiple microphones. Figure 1.15 depicts a multi-channel setup in which the multiple microphones are all connected to an analysis filter bank that implements a time-frequency analysis similar to the processing performed by the cochlea. Multiple microphones allow for a directional processing that can be used to attenuate sounds originating from a specific direction while sounds originating from a different direction remain transparent. Users can choose to whom they want to listen simply by looking at that person and placing them within the beam pattern. A single hearing aid has closely spaced microphones and results in a rather broad beam. However, binaural hearing aids contain microphones at both ears that can be used to form a more narrow beam. This system requires that information is transmitted between hearing aids which can be done wirelessly, however this requires an increased energy consumption. In the simplest case, control information, such as level control, could be transmitted in order to ensure that both hearing aids are in the same state, e.g. both have the same volume settings, and both apply the same enhancement scheme. Many hearing aids have auditory scene analyzers that use different processing algorithms for different auditory scenes. If one wanted to understand speech in a noisy environment, then they simply turn on the noise reduction in order to perceive speech better. This same noise reduction would not be so desired at a concert where one wanted to enjoy music. Finally, a feedback cancellation is employed in order to avoid an undesired whistling of the hearing aids, when signal components are amplified in a feedback loop.

2 — Pitch

Learning objectives

- Fundamental Frequency Estimation

Chapter 1 introduced the underlying physiological mechanisms behind speech production and sound perception. Furthermore, the source-filter model was also introduced as a simple model of speech production. The goal now is to take a given speech signal and extract the necessary parameters required for this model. This chapter will focus on creating a parametric representation of the excitation signal, or the ”source” in our model. The necessary parameters to accomplish this are: voiced or unvoiced judgment, the speech fundamental frequency, and the energy of the particular speech segment. Given these parameters, the excitation signal for an unvoiced speech segment can be modeled by white Gaussian noise while a voiced speech segment can be modeled by a pulse train with a pulse distance of the fundamental period, T_0, respectively.

Figure 1.9 shows that the excitation signal is readily identifiable in the spectrum of the vocal tract filtered signal in the form of the fundamental frequency and its harmonics. The fundamental frequency is an important parameter in speech signal processing that is required by many speech processing algorithms. It is therefore of great interest to learn how to efficiently and accurately extract this feature. Without it, speech is comprehensible, however the synthesized speech is monotonic and robotic. Furthermore, it is difficult to judge the intended emotion of the speaker and to distinguish between questions and statements. Also for speech enhancement, knowledge of the speech fundamental frequency can help to distinguish speech from noise. Noisy speech can be thought of as white noise filling up the gaps of a spectrogram. If the fundamental frequency is known, the noise could be attenuated in time-frequency bins where speech is not present, while the bins containing energy from the fundamental frequency could be preserved.

Very often in speech processing, the word pitch is used synonymously with fundamental frequency. However, it is important to realize that pitch is a perceptual quantity. The measure of pitch is therefore based on listening experiments with human subjects. This is in contrast to the fundamental frequency, which is a physical parameter, usually obtained using instrumental measures. The difference between pitch and fundamental frequency is best demonstrated in experiments where the loudness of a pure tone, a sinusoid with a constant fundamental frequency, is adjusted and subjects are asked to comment on the pitch of the tone. Subjects very often notice a change of pitch with a change in loudness, but a constant fundamental frequency [6]. This demonstrates that pitch is a qualitative phenomenon, however fundamental frequency is a quantitative, measurable parameter of sound.

The typical range of the fundamental frequency of voiced speech is from $40\,\mathrm{Hz}$ to $600\,\mathrm{Hz}$. $600\,\mathrm{Hz}$ is a bit on the high side and is something that would only really be seen in children. Typically, speech fundamental frequencies are around $100\,\mathrm{Hz}$ for male speakers, and about $200\,\mathrm{Hz}$ for female speakers.

The *residual effect* is a phenomenon in which we perceive the fundamental frequency of a sound even when this fundamental frequency is not present in the signal. Applying a $300\,\mathrm{Hz}$ high-pass filter to a male speech signal with a fundamental frequency of $100\,\mathrm{Hz}$ still results in the perception of a $100\,\mathrm{Hz}$ fundamental frequency. A good example, that everyone is familiar with is traditional telephone speech. For historic reasons, only frequencies above $300\,\mathrm{Hz}$ are transmitted, meaning that the fundamental frequency is not present in phone speech. Still, we are able to distinguish between male and the female speakers because the fundamental frequency can be estimated by the distance between successive harmonics. In fact, if two pure tones at $200\,\mathrm{Hz}$ and $300\,\mathrm{Hz}$ are summed and played together, the perceived pitch is $100\,\mathrm{Hz}$. The time domain signal in Figure 2.1 reveals that the sum of the two tones, and the distance between peaks is $0.01\,\mathrm{s}$ (cf. upper right

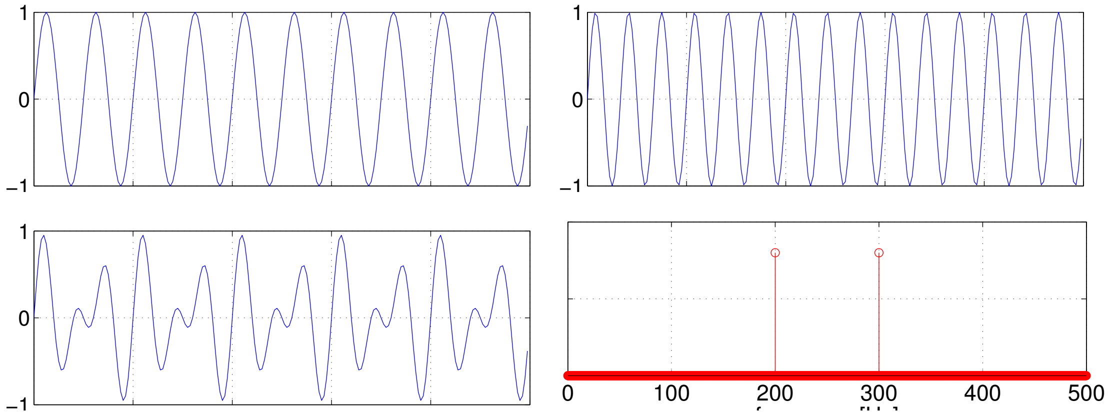


Figure 2.1: Example demonstrating the residual effect. Top: (Left) 200 Hz pure tone. (Right) 300 Hz pure tone. Bottom: (Left) Sum of the two harmonics reveals temporal pattern of fundamental harmonic at 100 Hz. (Right) Fourier transform of tone superposition shows no energy at fundamental harmonic.

of Figure 2.1), or 100 Hz in the frequency domain (cf. lower right of Figure 2.1). Performing a discrete Fourier transform on the time domain signal reveals no energy at 100 Hz because there are no 100 Hz components in the signal. However, the fundamental period is still 0.01 s (upper right of 2.1) and a lower tone is perceived.

In addition to frequency content below 300 Hz, the frequency content above 3400 Hz is also not transmitted in standard telephone speech (e.g. ISDN or GSM). As a consequence it is almost impossible to distinguish some phonemes, such as [s] and [f]. This is the reason why spelling alphabets are commonly used over the phone, e.g. "*c* like Charlie." Despite this, when heard in the context of flowing speech, the speech is still intelligible enough to justify the savings in bandwidth. Efficiently reducing signals down to the minimum bandwidth required to extract the required information is of great interest in order to save costs when transmitting multiple conversations at the same time.

2.1
Fundamental Frequency Estimation

The easiest method of measuring the fundamental frequency would be to take the time domain signal and measure the time between the periodic zero crossings or the periodic peaks. However, this value can vary because we often do not have perfect periodicity in the time domain signal. One could also imagine that there is noise in the signal, making it much more difficult to find the points at which the periodic structure repeats itself. In other words, this method is prone to errors and very difficult to automate. Fundamental period estimators based on this concept are not considered to be very robust. A much more robust estimator of the speech fundamental frequency is based on the the *autocorrelation function* (ACF) of a time domain signal.

Autocorrelation Function: formal definition

$$\varphi_{\mathrm{XX}}(\lambda) = \mathrm{E}(x(n)x^*(n+\lambda)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u\,v\,p_{x(n)x^*(n+\lambda)(u,v)\mathrm{d}u\mathrm{d}v} \tag{2.1}$$

As seen in Equation (2.1), the autocorrelation function is defined as the expected value of the product of a signal $x(n)$ and a shifted version of itself. If x would be a complex valued signal, then the product would be of the complex signal and the complex conjugate of a shifted version of itself. The complex conjugate is denoted by x^*. However, real-world physical time-domain signals are real-valued, such that the * can be removed.

The expected value is defined as the integral over these two signals multiplied by the joint probability density function of the two signals. This is a formal definition and in practice, the joint probability density function, $p_{x(n)x^*(n+\lambda}$ is generally not available, therefore it must be estimated by replacing the integral with a summation over realizations of the signal resulting in the simpler Equation (2.2).

Autocorrelation Function: estimated

$$\widehat{\varphi}_{\mathrm{xx}}(\lambda) = \frac{1}{N-|\lambda|}\sum_{n=0}^{N-|\lambda|-1} x(n)x^*(n+\lambda) \tag{2.2}$$

The ACF computes the average of the product of the signal and shifted versions of itself. This yields a measure of the self similarity of the signal. The ACF of a periodic time-domain signal at zero lag, $\lambda = 0$, results in the product of two identical signals on top of each other and the ACF is at it maximum value. All of the products would be positive and their sum results in a large positive value that estimates the power of the considered signal segment. If the original signal is shifted by half of a period, then the summation yields zero. The autocorrelation function of a periodic function is therefore also periodic because it begins at a maximum at $\widehat{\varphi}_{\mathrm{xx}}(0)$, decreases, and then achieves another maximum at lags corresponding to the fundamental period, i.e. at $\widehat{\varphi}_{\mathrm{xx}}(T_0)$. Because of the averaging by means of the sum in (2.2), detecting the second maximum $\widehat{\varphi}_{\mathrm{xx}}(T_0)$ yields a rather robust measurement of T_0 and thus of the fundamental frequency $f_0 = 1/T_0$.

It is also important to note that the Fourier transform of the autocorrelation function is called the *power spectral density* (PSD). This is formally defined as

Power spectral density

$$\Phi_X(f) = \sum_{\lambda=-\infty}^{\infty} \varphi_{\mathrm{XX}}(\lambda)\mathrm{e}^{-\mathrm{j}\Omega\lambda} \tag{2.3}$$

The power spectral density can reveal interesting properties of a signal that are not immediately

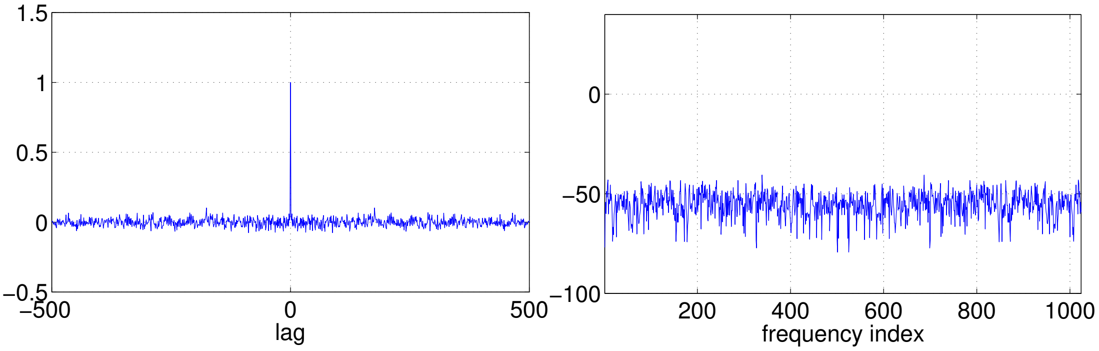


Figure 2.2: Left: Autocorrelation function of a Gaussian white noise signal. Right: Power spectral density of a Gaussian white noise signal.

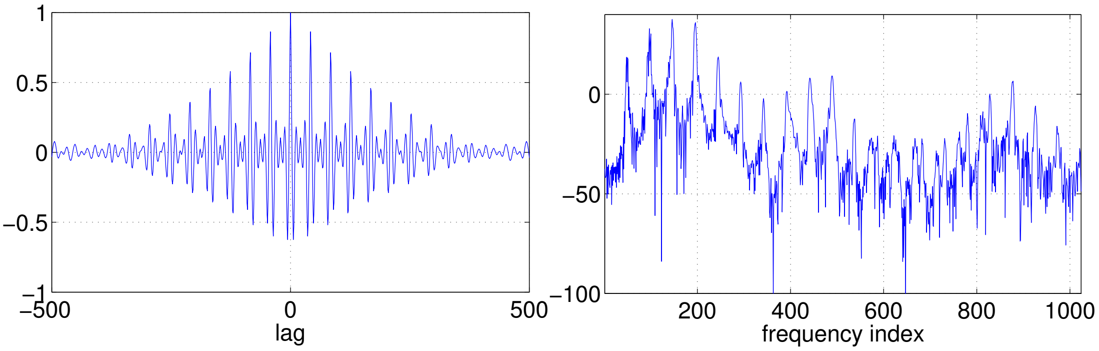


Figure 2.3: Left: Autocorrealtion function of a voiced speech segment. Right: Power spectral density of a voiced speech segment.

seen in the autocorrelation function. To visualize this, we can use the example of an uncorrelated signal. If a signal is *uncorrelated* that means that any two different samples are not correlated, or in other words, the autocorrelation function of a zero mean variable is zero everywhere except at $\varphi_{\mathrm{xx}}(0)$. This can be observed in Figure 2.2 which displays an *estimated* ACF and PSD of an uncorrelated Gaussian noise signal. Both representations reveal some fluctuation because the expected value was approximated by computing an average over a finite number of samples, i.e. we are showing the *estimated* autocorrelation function $\widehat{\varphi}_{\mathrm{xx}}(\lambda)$. From system theory, we will learn that the Fourier transform of a delta peak results in a flat spectrum. Thus, the PSD of an uncorrelated signal should be perfectly flat. Again, as in Figure 2.2 we show that the Fourier transform of the estimated ACF, also the PSD shows some fluctuations. If we were to average several PSD estimates, a flatter spectrum would be achieved. As an uncorrelated signal has a flat PSD, it is also referred to as a *white* signal. Therefore, uncorrelatedness and whiteness are often used synonymously, where the first describes the temporal characteristics of a signal while the latter describes the resulting spectral characteristics.

Speech signals, on the other hand, contain periodicities meaning that successive signal samples are correlated. The ACF of a periodic speech segment is shown in Figure 2.3. It reveals a peak at lag zero, $\widehat{\varphi}_{\mathrm{xx}}(0)$, and peaks at multiples of the fundamental period, $\widehat{\varphi}_{\mathrm{xx}}(kT_0)$. The PSD reveals a first peak at the fundamental frequency $\widehat{\Phi}_X(f_0)$, and then at integer multiples corresponding to the spectral harmonics, $\widehat{\Phi}_X(kf_0)$.

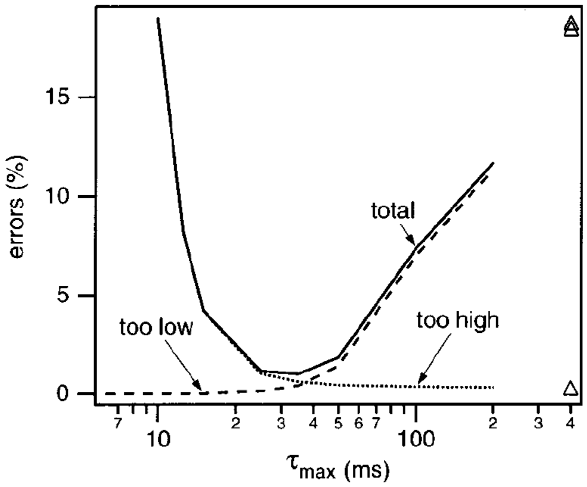


Figure 2.4: Fundamental frequency estimation error as a function of window length. [16]

A very important parameter in estimating the ACF is the window length, i.e. the number of samples, N in Equation (2.2) that are used to approximate the expected value in Equation (2.1). There is a certain trade-off between short and longer window lengths. Longer window lengths allow for multiple periods within one ACF window, thus making the estimation of the fundamental frequency more robust. However, the maximal applicable window length is limited because the fundamental frequency of speech is not stationary and changes over time. Again, note that speaking at a constant fundamental frequency would result unnatural sounding speech. $30\,\mathrm{ms}$ is a typical window length that is long enough to allow multiple periods of the fundamental period fit within one window, but short enough to follow changes in the fundamental period. At a fundamental frequency of 100Hz, there would be roughly three periods within the corresponding ACF window. This can be seen in Figure 2.4 which also shows the estimation error of the fundamental period.

There are also variants of this method, however many pitch estimators are still based on the ACF. A well known estimator, called YIN [16], is based on the following difference function

$$x(t) - x(t + T_0) = 0 \quad \forall t.$$

Similar to the ACF, we take a signal and the same signal shifted by a certain lag, however instead of multiplying, we subtract the two. For a perfectly periodic signal, this difference would be zero at lags corresponding to integer multiples of the fundamental period. To find the fundamental period T_0, we compute the square of this difference function and average over N samples in order to compute $d_{T_0}(t)$.

$$d_{T_0}(t) = \frac{1}{N - |\lambda|} \sum_{n=0}^{N-|\lambda|-1} (x(t) - x(t + T_0))^2.$$

The algorithm for the estimator would try to find the T_0 that minimizes $d_{T_0}(t)$. This method

Method	Error, (%)
ACF	10
Difference	1.95
YIN	0.50

Table 2.1: Percent error of some common fundamental frequency estimation algorithms.[16]

is very much related to the autocorrelation function in the use of shifted versions of the signal and the summation. If we now multiply out this square, we see that the $d_{T_0}(t)$ function actually consists of the autocorrelation function estimate at time $\widehat{\varphi}_{\mathrm{x(t+T_0)}}(0) = \widehat{\varphi}_{\mathrm{x(t)}}(0)$

$$d_{T_0}(t) = \widehat{\varphi}_{\mathrm{x(t)}}(0) + \widehat{\varphi}_{\mathrm{x(t+T_0)}}(0) - 2\widehat{\varphi}_{\mathrm{x(t)}}(T_0)$$

For a perfectly periodic signal, the autocorrelation function at $t = 0$ and the autocorrelation function at $t = t + T_0$ would yield exactly the same result. In that case, the two methods are identical. However in practice, this is not the case. Error measurements of some commonly implemented fundamental frequency estimators are shown in Table 2.1 and it can be seen that the ACF method produces a much larger error than YIN, which is another method that is based on the difference function.

List of Figures

Figure Sources

[7] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung*. Teubner, Stuttgart, 1998.

[8] P. Vary and R. Martin, *Digital speech transmission*. Wiley, 2006.

[9] Wikimedia Commons. (2007). Places of articulation. created by Ishwar. Permission: GNU FDL, CC-BY-SA-2.5, [Online]. Available: `https://commons.wikimedia.org/wiki/File:Places_of_articulation.svg`.

[10] ——, (1918). Saggital mouth. From Gray's Anatomy 1918 edition. Permission: Public Domain, [Online]. Available: `https://commons.wikimedia.org/wiki/File:Sagittalmouth.png`.

[11] International Phonetic Association. (2005). The international phonetic alphabet. Permission: Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA), [Online]. Available: `https://www.internationalphoneticassociation.org/content/full-ipa-chart`.

[12] Wikimedia Commons. (2009). Anatomy of the human ear. created by Chittka L, Brockmann. Permission: Creative Commons Attribution 2.5 Generic license., [Online]. Available: `https://commons.wikimedia.org/wiki/File:Anatomy_of_the_Human_Ear_en.svg`.

[13] ——, (2010). Cochlea-crosssection. created by Oarih. Permission: Creative Commons Attribution-Share Alike 3.0 Unported license., [Online]. Available: `https://commons.wikimedia.org/wiki/File:Cochlea-crosssection.svg`.

[14] ——, (2006). Audible. created by Booby. Permission: Public domain., [Online]. Available: `https://commons.wikimedia.org/wiki/File:Audible.JPG`.

[15] ——, (2008). Uncoiled cochlea with basilar membrane. created by Kern A, Heid C, Steeb W-H, Stoop N, Stoop R. Permission: Creative Commons Attribution 2.5 Generic license., [Online]. Available: `https://commons.wikimedia.org/wiki/File:Uncoiled_cochlea_with_basilar_membrane.png`.

[16] A. d. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

References

[1] A. d. Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[2] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung*. Teubner, Stuttgart, 1998.

[3] P. Vary and R. Martin, *Digital speech transmission*. Wiley, 2006.

[4] J. Deller, J. Hansen, and J. Proakis, *Discrete-time Processing of Speech Signals*. IEEE Press, 2000.

[5] R. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement - A Survey of the State of the Art*. Morgan & Claypool, 2013.

[6] W. B. Snow, “Change of pitch with loudness at low frequencies.,” *Journal of the Acoustical Society of America*, vol. 8(1), pp. 14–9, 1936.