ECE 298JA Evaluation of Exams Fall 2017

Jont B. Allen
University of Illinois at Urbana-Champaign

December 29, 2018

Abstract

Summary of grade analysis for ECE 537 for Fall 2018.

1 Summary of Student performance

Table 1 gives a summary of the grade distribution for the HW, two midterms and a final written report.

ID is a number to identify the student. HWs are the average of the home works for student ID. ExI and ExII are the scores for the two exams. From these two grade you should be able to identify your ID, since you can match the exam grades to the ID.

Rpt is the grade for the report. The mapping of the grade for the report the letter grade is 100=A+, 90=A, 80=A/A-, 70=A-, and so forth. A score of 20=C+.

The four scores (HW, ExI, ExII, Rpt) were weighted by [5, 22.5, 22.5, 50] %, result in the final grade given in the second to last column, which was then converted into the final letter grade, as shown.

Table 1: *Table of raw data for the homework, Exams I, II, and the final report. The first four columns were weighted by 5% (HW), 22.25 % (Exams I, II) and 45The last two columns are the final numerical and letter score, based on the four column weighted average average. Each student is identified by a numerical ID.*

ID	HWs	ExI	ExII	Rpt	Final	Letter
2	93	78	98	90	90	A+
7	93	93	96	80	88	A+
6	52	63	87	100	87	A+
5	75	70	66	100	85	A+
3	79	83	91	70	79	A
9	39	56	50	90	71	A
8	68	93	86	50	69	A
1	82	60	81	60	66	A
11	86	67	76	40	57	A-
4	65	67	72	20	45	B
10	36	30	61	30	38	B-

Histograms of grade distributions: Figure 1 shows the distribution of the final grades. For example there are four A+ grades in the range from 82-100, and four A grades in the range from 60-81 and one A- with a score of 57.

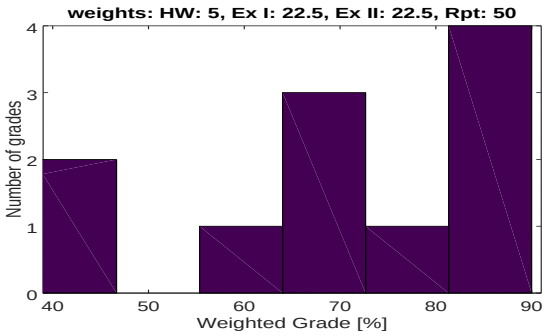


Figure 1: *This figure shows the histograms of the final grades. The figure title specifies the weights on each of the four components: HW (5%), Exams I, II (22.25%), Final written report (50%).*

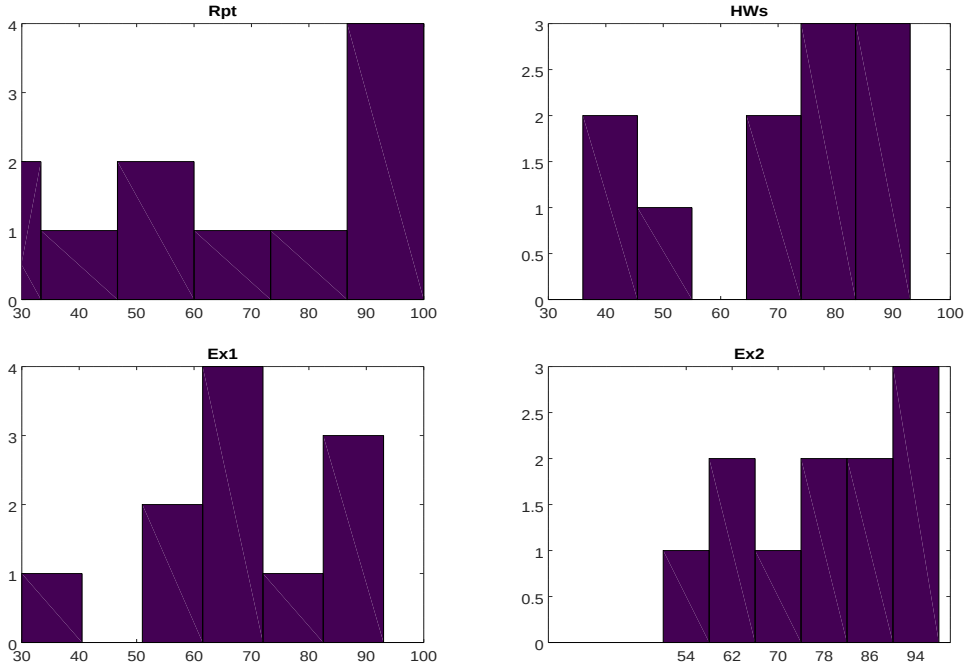


Figure 2: *This shows the distributions for the final report, the Home works, and Exam I and II. The grades for the report had 4 high grades, above 85 (A+). The rest were spread uniformly from 20 (C+) to 80 (A).*

Justification for the relative weights: The logic behind the weights is as follows: The HW is only weighted as 5% because the exams were based on the HW. If you didn't get the HW at the outset, it didn't hurt you that much. But by the time the exam was given, you needed to master them. Thus the exams were weighted 22.5% each. The Final paper has a weight of 50% because this is your opportunity to prove that you really got the material under control. Many of you did get it under control, and proved it with a masterpiece of writing. As graduate students you need to know how to write, and thus I'm testing you on your ability to communicate. After all, that's what ECE 537 is about, human communication.

The following sections contain five of the best final reports. Following that there is a 1 pages summary of each of the six remaining reports (e.g., pages 1-10).

To navagate to the next range of report grades, search for the key words *FINAL REPORT:*

1 FINAL REPORT: A+

This section consists of two reports.

FUNDAMENTALS OF HUMAN SPEECH PROCESSING

BY

SARAH YOONJI SHIM

FINAL EXAM

Submitted in fulfillment of the requirements
for the final examination of ECE 537: Fundamentals of Speech Processing
in the department of electrical and computer engineering of the
The University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Professor:

Professor Jont B. Allen

Contents

I. Psychoacoustics and the Human Ear

Psychophysics refers to the branch of psychology that studies the relationship between the physical world ($\Phi - domain$) and the mental interpretation ($\Psi - domain$) of it. Within the study of psychophysics, psychoacoustics is the research that concerns with the relationship between the physical phenomena and the human perception of sound. To study psychoacoustics, we must first understand what a sound is. A sound is a form of energy that is caused by the pressure variation of molecules in a medium that is picked up by our ears. For example, when one plays a violin, the vibrating chords generate pressure waves in the air. If the waves are picked up by a person nearby, the pressure waves are "perceived" by him/her as a specific sound (hopefully as a sound of a violin). Sound can also travel through other mediums – i.e., solid and liquid.

In the 1790s, a French scientist and mathematician Jean Baptiste Fourier (1768-1830) proved that any periodic waveform could be expressed as the sum of sinusoidal waves of frequencies that are integer multiple of a fundamental frequency. This idea was later defined as the Fourier Series, and represents a periodic function as the sum of simple sinusoidal signals. A simple sinusoidal wave is composed of an amplitude and frequency information. The amplitude of a signal gives information on the amount of pressure in the wave, and the frequency provides information on the number of cycles of a wave per second. In the ($\Psi - domain$), the perception of the amplitude sound sound is called the loudness, and the perception of the frequency is called a pitch. The human ear can pick up signals from the frequency of 20-20 kHz (nearly 10 octaves). In addition to loudness and pitch, a timbre of a sound refers to the perception of the quality of the signal. Timbre helps to identify and distinguish between different sound sources that produce sound at same pitch and loudness. It is the timbre that allows us to identify sound at middle A (440 Hz) played by a violin from middle A played by a trumpet.

In this chapter, to fully understand the basics of psychoacoustics, we will first explore the anatomy and physiology of the human ear. This will allow us to understand how a sound wave is picked up and processed into neural information. Afterward, we will identify the primary components associated with the cognitive processing of a sound. With a basic understanding of how humans sense and perceive sound, we will explore the history and introduce the intellectual giants that shaped the field of psychoacoustics. We will conclude this chapter by presenting different sound phenomenon such as masking.

1.1. The human ear

The human ear is an acoustomechanical organ that collects and processes sound. The organ is divided into three major regions: the outer, the middle, and the inner. Each area plays an essential role in transforming the physical input signals into neural data. As shown in Fig. 1, the outer region refers to the air-filled input of the ear and is composed of three main components. The pinna is the outermost part of the ear. It is responsible for increasing the collection of signals by acting as a horn and is responsible for reducing the radiation reactance of the ear. The signal that enters the pinna is led into a cylindrical canal called the external canal or the meatus. The meatus can be thought of as a transmission line that is terminated by the tympanic membrane or the eardrum. The tympanic membrane refers to a stiff, inwardly-directed tissue that facilitates hearing by transmitting sound wave from the air to the middle region of the ear.

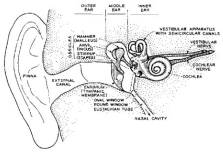


Figure 1. Basic anatomy of the human ear (from course textbook).

The middle region is an air-filled system that acts as air to fluid impedance transformer and amplifier. It is composed of three ossicle bones. The malleus or the hammer is connected to the tympanic membrane and moves as the membrane vibrates. The end of the hammer is attached to the incus or anvil which mechanically transfers the movement to the stapes or stirrup. The stapes is seated in an oval window and is retained via an annular ligament. The stapes transmits sound vibration from the incus to the oval window which is a membrane-covered opening of the inner region.

Unlike the outer and the middle regions of the ear, the inner region is filled with fluid. The acoustomechanical impedance of fluid is much higher than that of air. Thus, for efficient transmission of sound energy, the ossicle bones of the middle region provide a step-up impedance transformation and increases the force by x1.3. Also, the volume displacement of the stapes is a function of frequency. At low frequencies, the combined elasticity of the drum controls the stirrup motion. This means that the system acts as a spring, with the stapes displacement proportion to the

pressure of the eardrum. However, as the frequency increase, the stirrup displacement begins to diminish in amplitude and lag in phase – exhibiting a low-frequency characteristic.

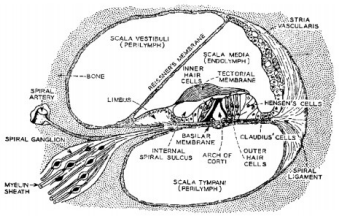


Figure 2. Anatomy of the cross-section of the cochlea (from course textbook).

The inner region is a fluid-filled system that acts as a neural impulse transducer. It is composed of the cochlea. The cochlea is a spiral cavity that is responsible for transforming sound wave into neural and spacial information. As shown in Fig. 2, it is divided into three chambers: the scala vestibuli, which is in charge of space perception, the scala tympani which, along with the scala media, is responsible for sound transduction. The scala vestibuli and the scala tympani are filled with a colorless liquid called the perilymph. This liquid has a viscosity of about double that of water and has a specific gravity of about 1.03.

The scala tympani is lined with the basilar membrane. This membrane is a critical element that is responsible for picking up frequencies of the input signal. In the 1950s, George Von Bekesy (the winner of the 1961 Nobel Prize for physiology and medicine) demonstrated that when a sound wave stimulated the cochlea from a human cadaver, the basilar membrane exhibited a motion of a traveling wave, traveling from the base to the apex. The wave presented maximum amplitude at certain places along the membrane, and this was later found to be dependent on the signal's frequency. Bekesy discovered that the higher frequency sounds exhibited maximal amplitude at the base of the cochlea, whereas the lower frequency sounds exhibited maximum amplitude toward the apex of the cochlea. The approximate logarithmic arrangement of frequency along the basilar membrane is called the Frequency-to-place (or Tonotopic) map (Fig. 3).

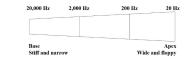


Figure 3. Tonotopic map of the basilar membrane.

A shown in Fig. 4, the critical bandwidth is the measure of cochlear bandwidth at a given basilar membrane region. For example, if a tone of1 kHz is heard, the area on the membrane tuned to 1 kHz will respond. However, in addition, the part of the membrane tuned to 950 Hz may also respond if it is within the critical bandwidth.

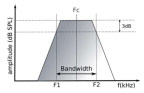


Figure 4. Amplitude vs. frequency plot showing the critical bandwidth between frequency F1 and F2.

In addition to the basilar membrane, the organ of Corti is a critical element of hearing. This sensitive element of the inner ear can be thought of as the body's microphone. It is responsible for transducing the mechanical motion of the membrane into neural activity. As shown in Fig. 5, the organ of Corti contains about 30,000 sensory receptive cells called the hair cells on which the auditory nerves are connected to. Distinctive parts of the basilar membrane that are tuned to specific frequency carried by the input signal move up and down and push on the tectorial membrane. As the tectorial membrane moves it shears against the inner and outer hair cells and this displacement cause activation of connected neurons.

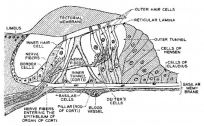


Figure 5. Cross section anatomy of the organ of Corti (Courtesy of Prof. Jont Allen)

The hair cells are stretched along the cochlear duct and are positioned on the ridge of the basilar membrane. On each segment of the basilar membrane lies three outer hair cells and one inner hair cell. In humans, there are approximately 12,000 outer hair cells and approximately 3,500 inner hair cells. The "hairs" consist of a series of cilia and one long stereocilium. Deformation of the cilia due to the shearing of tectorial membrane places tension on the attached protein filament. The hair cells have a resting membrane potential of approximately -65 mV compared to that of the endolymph. When the hair cell depolarizes, ions enter the cell and causes the release of neurotransmitters.

One of the fascinating aspects of the basilar membrane and the outer hair cells is that it is non-linear. This seemingly "flawed" behavior is, in fact, a critical feature that enables the broad dynamic range of human hearing. The inner hair cells of the cochlea hold a dynamic range of less than 50 dB. However, the non-linearity of the basilar membrane allows an increase of 4 dB in regions where the sound is most active. Such non-linear characteristic increases our dynamic range of hearing to 120 dB.

1.2. Neural signals of the auditory system

The sensory receptor cells of the ear communicate with the brain via a bundle of nerve cells called the neuron. To further understand the neural data transmission of the auditory system, we must have a good understanding of what a neuron is. A neuron is a specialized cell that transmits nerve impulses. It can be thought of as a highway on which information travels on. Similar to the highway system, there are different types of neurons to relay information all around the body – this will be covered in detail in a later section.

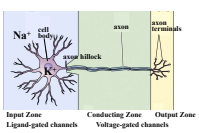


Figure 6. Primary structure of a neuron (from prof. Wickesberg's lectures).

As shown in Fig. 6, a neuron is composed of three major zones. The input zone is composed of the cell body and the dendrites. Dendrites are branched extension of the neuron along which receives signals from other neurons. The impulses received by the dendrites are transmitted to the cell body. Neurons presumably only have two states, the active and the inactive. As shown in Fig. 7, when the electrical input surpasses the threshold, an action potential occurs, and the neuron produces an electrical pulse that travels through the axon. During an action potential, the neuron is either in the

excitatory postsynaptic potential (ESPS) state or the inhibitory postsynaptic potential (IPSP) state. The conducting zone is composed of an axon which is a long, slender projection of a neuron. When an action potential occurs, the axon conducts electrical impulses away from the cell body and to the output zone. The output zone is composed of axon terminals that are specialized to release the neurotransmitters of the presynaptic cells.

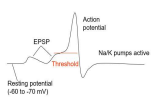


Figure 7. Action potential diagram (Courtesy of Prof. Wickesberg).

Neural information activated by the hair cells are sent out from the cochlea to the brain through the auditory nerve. The auditory nerve fibers are narrowly tuned at high frequencies and broadly tuned at low frequencies. Such characteristic was deduced by the plotting frequency response of the fiber on a tuning curve. Figure 8 shows the tuning curve of six different fibers in the auditory nerve of a cat. As shown in the figure, the tuning curve measures threshold loudness of a signal with a specific frequency that excites the fiber. Thresholds of auditory nerve fibers vary regularly with low threshold fibers on the pillar side of the inner hair cell and high threshold fibers on the modiolar side of the inner hair cell. An experiment also revealed that the pillar side of the inner hair cell exhibits high spontaneous activity rate whereas the modiolar side of the inner hair cell exhibits low spontaneous activity rate.

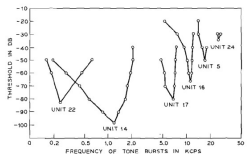


Figure 8. Tuning curve of six different fibers in the auditory nerve of a cat (Courtesy of Prof. Jont Allen).

Neural data from the auditory nerve fibers to the brain contains three primary information: tuning, timing, and threshold. The place theory of pitch perception, which explains the various resonant frequency distribution along the basilar membrane, provides the tuning information. In addition, the auditory nerve fibers preserve temporal information at low frequencies below 4kHz. According to the Volley theory proposed by Ernest Wever and Charles Bray in 1930, groups of neurons of the auditory system respond to a sound by firing action potentials slightly out of phase with each other. Each neuron can produce a standard electrical pulse of about a millisecond duration and are desensitized for a period of approximately one to three milliseconds. Thus, as seen in Fig. 9, when the action potentials that are out of phase combines, more frequency of sound can be encoded and sent to the brain to be analyzed. This gives timing/ temporal information to the neural data. Furthermore, the wide range of thresholds of the auditory nerve fibers gives threshold information to the neural data.

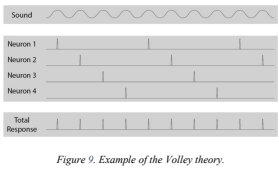


Figure 9. Example of the Volley theory.

Information from the auditory nerve is sent to the cochlear nucleus. The cochlear nucleus is the obligatory nucleus in which all ascending information from the cochlea forms synaptic connections with the auditory brain. It is divided into two regions: ventral cochlear nucleus and dorsal cochlear nucleus. The dorsal cochlear nucleus is a layered structure with different types of cells. The layers of the cell contain axons of the granule cells and a substantial layer of fusiform cells. Granule cells convey information about the position of the pinna to the molecular layer of the dorsal cochlear nucleus. Fusiform cells are narrowly tuned cells that compensate for the motion of the pinna. Each cell contains two sets of dendrites. One set extends into the molecular layer that is perpendicular to the paths of the granule cells, and the other extends into the deep dorsal cochlear nucleus and contact the auditory nerve fibers. The fusiform cells have complex, nonlinear characteristics similar to the Purkinje cell in the cerebellum. It is important to note that these layers are only found in animals that can move their ears. In addition to the fusiform and the granule cells found in the dorsal cochlear nucleus, the tuberculoventral neurons are narrowly tuned, nonlinear cells found in the deep layer of the dorsal nucleus and projects to the ventral cochlear nucleus. Their main function is to suppress echoes by delaying frequency-specific inhibition.

From the cochlear nucleus, the information is sent to the superior olivary complex, inferior colliculus, and nuclei of the lateral lemniscus via bushy cells, stellate and fusiform cells, and

octopus cells respectively. The superior olive is responsible for the localization of sound. It uses two cues to compute sound localization: the difference in the arrival time of the sound at each ear, and the intensity difference between the two ears caused by the head. The superior medial olive computes the time difference between the ears. It responds better to low frequency sounds, and the cells respond at specific interaural time difference (ITD). The superior lateral olive computes the intensity difference between the ears. It responds better to high frequency. The superior lateral olive gets excited by sound at ipsilateral ear by the spherical bushy cells and gets inhibited by sound at the contralateral ear by the globular bushy cells.

The auditory nerve fibers consist of many different cells. Primarily, the bushy cell is a second order neuron named after its bush-like, short dendrites. The bushy cell is divided into two types: the spherical bushy cell and the globular bushy cell. Unlike the bushy cell, the stellate cell has dendrites that run parallel to the path of the auditory nerve fibers. In addition, the octopus cell has two or three large dendrites that extend perpendicularly across the paths of many auditory nerve fibers. In addition to its distinct physical characteristics, the cells in the auditory nerve fibers exhibit different functions and characteristics as outlined in Table 1.

Table 1.Characteristics of different types of cells found in the auditory nerve (from prof. Wickesberg's lectures).

Name	Spherical Bushy Cell	Globular Bushy Cell	Stellate Cell	Multipolar Cell	Octopus Cell
Dendrite Characteristics					
Tuning	Narrow	Narrow	Narrow	Broad	Broad
Intrinsic Membrane Properties	Nonlinear	Nonlinear	Linear	Linear	Nonlinear
PST Histogram	Primary-like	Primary-like w. notch	Chopper	Onset-chopper	Onset
Function	Preserve Timing Information	Preserve Timing Information	Integrate loudness in critical band	Overall loudness	Synchrony integration across freq.
Project to	Superior Olivary Complex	Superior Olivary Complex	Inferior Colliculus	Cochlear nucleus on other side	Lateral Lemniscus

Stimuli received at the two ears interact both at the medulla and the midbrain levels. The final stage of the pathway is the primary auditory cortex. Two most important areas of the brain in psychoacoustics are the Wernicke's area and the Broca's area (Fig. 10). The Wernicke's area involves speech analysis, and the Broca's area involves speech production.

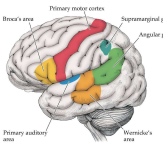


Figure 10. Primary parts of the brain responsible for audio processing.

1.3 History of psychoacoustics research

During the last few centuries, a series of consecutive historical discoveries fostered a great momentum to the advancement of auditory technology. While audio technologies are readily found in today's modern society (from a hearing aid to high-definition Bluetooth audio headphones), the origin of the technology traces itself all the way back to 350 BC, Greece. Greeks were fascinated by music and a Greek philosopher, Aristotle was the first person to suggest that the movement of air carries sound. In addition to Aristotle, Pythagoras studied the relationship between pitch and the length of strings on musical instruments. However, it wasn't until the 17th century that the relationship between the vibratory frequency and pitch was confirmed by Robert Hooke (1635 – 1703).

In 1846, a German physiologist Ernst Heinrich Weber published that the just-noticeable difference (JND) of an intensity (ΔI) is proportional to the physical intensity I and is always constant:

$$\frac{\Delta I}{I} = 0.1 = Weber's fraction$$

This law holds for floating point conversion. An ear is a floating-point converter with the mantissa containing information on the dynamic range, and the exponent with information on the level. Less than a decade later, the intensity JND was formally defined as "the relative signal level for detection 75% of the time" by L.L. Thurston (known as the father of decision theory model) in 1927, and by David Green in 1965.

In the 1860s, after Weber's law was published, his student, Gustav Theodor Fechner (also known as the father of psychology), applied Weber's law to loudness. The intensity JND (ΔI) is a measure of an internal noise (σ_L). According to the Fechner's hypothesis, the loudness JND is constant and do not depend on the intensity or the loudness ($L(I) \propto log(I)$). He concluded that the total change

in loudness between two loudness L_1 and L_2 are found by counting the number of JND. This is called the counting formula and is:

$$N_{JND} = \int_{L_1}^{L_2} \frac{1}{\Delta L} dL = \frac{L_2 - L_1}{\Delta L}$$

Fechner's idea was that the loudness L is proportional to the number of JND step N. In his hypothesis, Fechner made two assumptions: $\Delta L \propto L$ (the Weber's law), and that the internal noise ($\Delta L \propto \sigma_L$) is constant.

However, both of these hypotheses were debunked in 1928 when a Hungarian mathematician, Frigyes Riesz disproved Weber's law and rendered Fechner's law useless (however, Weber's law still holds for wideband noise intensity discrimination (Miller, 1947)) for pure tone signals. Riesz experimented with two beating tones that are 3 Hz apart (1kHz masker vs. 1.003kHz probe) to prove that Weber's law does not hold at the perception of higher intensities- that intensity discrimination improves at higher intensities. This deviation of Weber's law is known as the near miss of Weber's law. The near miss to Weber's law results from the fact that the internal noise is proportional to loudness JND and is not independent of loudness L. In 1997, Allen and Neely showed that the noise relationship exhibits Poisson-like behavior:

$$\Delta L(L) \propto \sqrt{L}$$

In the 1920s, Georg von Bekesy, a Hungarian biophysicist and the winner of the Nobel Prize in Physiology and Medicine, unveiled that the basilar membrane model using traveling wave. He discovered that the cochlea is analogous to a dispersive transmission line where the different frequency components that make up the input signal travel at different speed along the basilar membrane. This caused isolation of each frequency components of the input signal long different places on the basilar membrane.

Although it was clear that sound had a magnitude (loudness) and frequency (pitch), it was not easy to accurately manipulate and control sound. However, with inventions of electroacoustic devices like the Campbell's wave filter in 1911, quantitative research of psychoacoustics proliferated in the 1920s and 30s, especially at the Bell Laboratories.

Bell Lab was a prominent institution both in business and in science. The lab was found in 1925 as the research branch of the original American Telephone & Telegraph (AT&T) company. In 1925, engineers at the Bell lab, C. E. Lane, with his college, R. L. Wegel, implemented a transmission line model of the cochlea (Fig. 11). The stapes input pressure P_1 is at the left, with the input velocity V_1, as shown by the arrow. Mass of the fluids of the cochlea is modeled with inthe ductor l_n and the stiffness of basilar membrane is modeled with capacitor C_n.

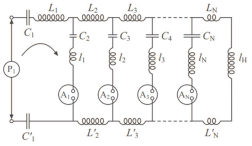


Figure 11. Wegel and Lane's electrical equivalent circuit.

In 1933, Harvey Fletcher and Wilden Munson experimentally determined that loudness can be summed up. If intensities of two different sound equal ($I_1 = I_2$, so that $L(I_1, f_1) = L(I|\ \ |2, f_2)$, then the resulting loudness will be double that of $L(I_1, f_1)$ and $L(I_2, f_2)$. Following this research, psychophysicist Stanley Smith Stevens (1906-1973) proposed the relationship between the magnitude of a physical stimulus and the intensity perceived by the human ear. The general form of this law is:

$$L(I) = I^{\frac{1}{3}}$$

In 1961, Donald D. Greenwood correlated the position of the hair cells in the inner ear to the resonant frequencies that stimulate their corresponding auditory neurons. This is known as the Greenwood cochlear map. The derivation of the cochlear map was based on counting the critical bands. The number of the critical band, $N_{CB}$is calculated by integrating the critical band density over both frequency and placement, and equating the two results:

$$N_{CB} = \int_0^{Xcr} \frac{1}{\Delta_x(X)} dX = \int_0^{Fcr} \frac{1}{\Delta_f(f)} df$$

Greenwood also verified that the critical bandwidth in Hz is:

$$\Delta_f(X) \propto 10^{\frac{K_{ab}(X)}{10}}$$

In 1966, Green and Swets applied to signal detection theory to psychoacoustics in their book, "Signal Detection Theory and Psychophysics."

1.4 Auditory Masking

In a real environment, there are multiple sound sources present. The auditory masking effect occurs when the perception of one sound is affected by the presence of another sound. There are two major classes of masking in psychoacoustics: neural masking and dynamic masking. Neural masking characterizes the internal noise associated with loudness (neural representation of the internal noise). Unlike the neural masking, dynamic masking is strictly cochlear. Dynamic masking is mainly due to the non-linearity of the outer hair cell. The masking has two forms: simultaneous and non-simultaneous masking. The temporal masking or non-simultaneous masking occurs when a masker makes target sound inaudible after the stimulus as shown in Fig. 13. The forward masking is mostly due to the nonlinear outer hair cell processing, which has a substantial effect over a long time.

Figure 12. The plot of temporal masking.

Dynamic masking includes the upward spread of masking effect as well as the two-tone suppression effect. Figure 12 plots are masking result of 1kHz tone at different levels of sound. We observe that threshold is maximum at the masker frequency and decreases as we go farther from the center. However, as masker tone amplitude increases, a higher level of masking occurs at higher frequencies compared to that of lower frequencies. In other words, masking is much stronger at a higher frequency as masker amplitude increase and the curve becomes much shallower in the higher frequencies than in the lower frequencies. This phenomenon is known as the upward spread of masking. As the masker frequency increases, the masking pattern becomes increasingly compressed, meaning, high-frequency maskers are only effective over a narrow range of frequencies close to the masker frequency whereas the low-frequency maskers are an effect over the wide frequency range. Also, two-tone suppression refers to a nonlinear property of the cochlea in which the total neural firing rate in the region most sensitive to a probe tone is reduced by the addition of a second (suppressor) tone at a different frequency.

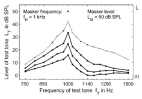


Figure 13. Level of the threshold of masking vs. frequency sweep for 1kHz masker frequency.

II. Information Theory

2.1 Information, entropy, and channel capacity

Information theory is the study of quantification, storage, and communication of information. Information is defined as the value of knowledge (regardless of whether it "factual" or not). The information of a message is quantitatively calculated as the inverse of its probability of occurrence and is expressed as:

$$I(x) = log\left(\frac{1}{p(x)}\right) = -log(p(x))$$

An excellent example of this is how much "impact" the information has when a person gets hit by an airplane versus when a person gets hit by a raindrop. The probability of a person getting hit by a plane is infinitesimally small. Therefore, the information has way more value than when a person gets hit by a raindrop (unless he/she lives in the heart of the Sahara Desert).

In mathematics, the expected value of a set *x*, *E(x)*, is the long-run average value of repetition of the experiment it represents:

$$E(x) = \sum_k P_k x_k \text{ given that } \sum_k P_k = 1$$

Entropy refers to the measure of randomness and is the expected value of information:

$$H(x) = E(I_k)$$

It is interesting to note that there are multiple ways of representing entropy. Some of the prominent ones include:

Nats: $I_k = log_e(x) = ln(x)$

Hartleys: $I_k = log_{10}(x)$

Bits: $I_k = log_2(x)$

Entropy is most popularly represented in bits:

$$I_k = log_2\left(\frac{1}{P_k}\right) = -log_2(P_k)[bits]$$

$$H(x) = E\left(log_2\left(\frac{1}{P_k}\right)\right) = \sum_k P_k log_2\left(\frac{1}{P_k}\right)[bitspersymbol]$$

It is important to note that the entropy is given by the average of information and is maximum when all possibilities of outcomes equal and minimum when there is only one possible outcome. The information rate is the average entropy per symbol and is calculated by:

$$R = symbolrate * H(x)$$

Given that a discrete distribution has a probability function p_k, and second, discrete distribution have q_k, the relative entropy of p with respect to q is defined as:

$$H_r(p) = \sum_k p_k log_2\left(\frac{p_k}{q_k}\right)$$

In addition, the channel capacity is the tight upper bound on the rate at which information can be reliably transmitted over a communication channel. The channel capacity of the white bandlimited Gaussian channel is

$$C = Blog_2\left(1 + \frac{S}{N}\right)\left[\frac{bits}{s}\right]$$

Where B – channel bandwidth

S – Signal power

N – Noise power within channel BW

2.2 Claude Shannon

One of the biggest names in information theory is Claude E. Shannon in 1949. Shannon's work in information theory provided a solid mathematical foundation for analyzing the performance of a communication channel. His proposed channel model for communication as shown in Fig. 14. The source provides its message (X) to the transmitter through a lossless, error-free connection. The transmitter signal (S) is transmitted through a channel to the receiver, which receives the message (R) and gives it in a lossless manner to the destination. A noise (N) is introduced into the channel between the transmitter and the receiver and acts to change messages so that what is received differs from what is transmitted.

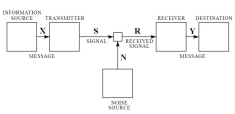


Figure 14. Schematic diagram of a Shannon's communication channel model.

However, the key to Shannon's definition of information is that the meaning of the message does not play any role in the model. Instead, information is strictly defined regarding the entropy *H*, computed over *N* discrete probabilities [p_n] of a set of binary symbols [0,1] as used by Morse code of dashes and dots. The use of bits in the Morse code allowed reduction of channel error rate since dash and dot can be made arbitrarily distinct by using an arbitrarily small BW.

2.3 Human Speech Recognition

Human speech recognition (HSR) is an interdisciplinary field that develops methodologies to enable the recognition of spoken language by machines. To successfully develop an HSR system, it is essential first to understand and model how speech is processed and recognized by humans. In the 1950s, an American psychologist George Armitage Miller introduced information theory to language modeling, following Claude Shannon. Figure 14 summarizes the Miller & Nicely's five channel speech recognition process by humans.

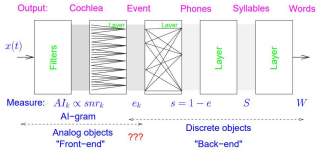


Figure 15. Five stage human speech recognition model from Miller & Nicely 1955.

The HSR system is composed of a discrete event that is processed via bottom-up, divide & conquer strategy. Humans recognize speech based on a hierarchy of context layers, and the entropy decreases as the model integrate context throughout the stages. The speech leaves the mouth and enters the ear and is processed by the cochlea as shown in the first block. During this process, the signal is broken into a filtered continuum of band-passed responses.

The articulation index (AI) is a tool used to predict the amount of speech that is audible to a person. The first articulation experiment was performed by Lord Rayleighs in 1908. The figure of AI ranges from zero to one and represents the proportion of the average speech signal that is audible. The output of the cochlea is characterized by a specific AI, and the average score is calculated by:

$$P_c(A) = 1 - e_{chanc}\ e_{min}{}^{AI}$$

In language processing, there are two indices: the speech transmission index (STI) and the speech intelligibility index (SII). The STI is an objective, physical measure of the quality of speech transmission. Similar to AI, the index ranges from 0 to 1, indicating the degree to which a transmission channel degrades speech intelligibility. Therefore, for perfectly intelligible speech will remain perfectly intelligible if it is transferred through a channel with an associated STI of 1. The closer the index value approaches zero, the more loss the information has. Unlike the STI,

which measures the ability of the channel to transport speech, the SII measures (from 0 to 1) the intelligibility of speech.

At the event level of the model, analog to digital conversion occurs. The recognition errors which are the result of event extraction labeling errors are shown in the second box and modeled by the articulation-band errors e_k:

$$e_k = e_{min}{}^{\frac{SNR_k}{K}}$$

where SNR is the signal to noise ratio between 0 and 1.

Using the above equation, we can find the total error:

$$e = e_1 e_2 * ... * e_K = e_{min}{}^{\frac{SNR_1 + SNR_2 + \cdots + SNR_K}{K}}$$

The speech SNR in dB measures the event error e_k, and thus the phoneme articulation can be calculated as:

$$s = 1 - e_1 * e_2 * ... * e_K$$

Phones, s, are defined as a sound such as a consonant (C) or vowel (V). A phoneme is a set of phones that form a meaning. The phoneme identification experiment was first conducted by George Campbell in 1910. In the nonsense, CVC syllable articulation model, the syllable, S is cubed of the phone, s. In addition, according to the Heuristic degree of freedom context model by Boothroyd, the word is calculated as:

$$W = 1 - (1 - S)^J$$

The performance of speech recognition can be tabulated as a confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the actual values are known. Each row of the matrix represents the instances of a predicted class, and the column represents the instance of the actual class. In 1955 Miller & Nicely established a natural phone hierarchical cluster confusion matrix.

The maximum a posteriori of speech can be estimated using the expectation-maximization (EM) algorithm. It is an iterative method between creating function of expectation of the log-likelihood and computing parameters that maximizes the expected log-likelihood. For example, if we would like to separate a given data into two Gaussian distribution by using an EM process, we would start by guessing the mean and standard deviation of each distribution of interest. Then, we would find the points where the ratio of the two distribution is one so that we can threshold the values into groups, depending on the region it is located at. We would then recompute the means and the standard deviations and repeat the process with the recomputed values. We will iterate through this process until the result converges.

2.4 Linear Prediction Coding (LPC)

Linear prediction coding is a method for signal source modeling in speech signal processing. It is often used as a formant extraction tool but also has wide application in other areas. Its main advantage comes from the reference to a simplified vocal tract model and removing redundancy in the signal. It is a useful method for encoding speech at a low bit rate. In principle, LPC tries to predict next point as a linear combination of the previous values by finding the coefficients of the n^{th} order linear predictor (FIR filter) that predicts the current value of the real-valued time series (s) based on past samples:

$$s_l = \sum_{j=1}^{N} a_j s_{l-j} + e_l$$

where $\{a_j\}$ are the n^{th} order linear predictor coefficients (LPC), and e_l is the residual prediction error. The predictor error can be thought of as the output of the prediction error filter shown in Fig. 16, where $P(z)$ is the optimal linear predictor, s_l is the input signal, and $\hat{s}_l$is the predicted signal. It is important note that the error signal e_l can be easily calculated by $e_l = (\hat{s}_l - s_l)$. $P(z)$ is found by taking the z-transform:

$$e(z) = [1 - P(z)]S(z)$$

$$P(z) = \sum_{j=1}^{N} a_j z^{-j}$$

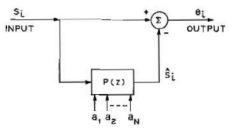


Figure 16. Model of basic linear prediction coding scheme.

III. Phonetics and the Human Speech

Communication via spoken language is one of the essential aspects that distinguishes humans from non-human species. While many animals communicate and exchange information using sounds, humans are unique in the complexity of the information that is conveyed using speech. Such ability allows humans to transfer complex range of ideas, thought, and emotions via words and expressions. Phonetics refer to the branch of linguistics that studies and classify human speech sounds. Humans produce speech by bringing air from the lungs to the larynx, a hollow muscular organ in the neck, via the diaphragm. In the larynx, vocal folds vibrate to make a sound and the articulators of the mouth, and the nose shapes the airflow.

Humans control vocal folds to make a sound. For example, when the sound "uh-oh" is made, the vocal folds close in the middle of the sound (after "uh-"). This stops the flow of air through the vocal tract. The quick silence in the middle of "uh-oh" is referred to as a glottal stop because the air is stopped completely, and the vocal folds close off the glottis. Another example is when the sound "haaaaaah" is made. When this happens, the vocal folds are open, and the air passes freely through the glottis. In addition, when a sound "aa" is made (as if the mouth is opened at the dentist), the vocal folds are close together and vibrates rapidly. After it goes through the larynx, the airflow continues into the nasal and the oral cavity. A velum is a part responsible for the selection of the air passage (between nasal and oral pathways).

This chapter provides an overview of phonetics and human speech production. We begin our study by acknowledging the works of researchers that significantly contributed to the field of speech. Then, we will explore how the human speech production mechanism is modeled. Transmission line theory, multi-port analysis, as well as signal processing methods will be introduced and will be used in our modeling.

Figure 17. Model of the human larynx that shows the thyroid cartilage (TC), the arytenoid cartilages (AC), and the vocal cords (VC) (from course textbook).

3.1 History

Before the discovery of the traveling waves, scientists and mathematicians were bewildered and curious about how humans can transmit and receive sounds. One of the first people in history to claim that sound travels in a form, or a wave was the Greek philosopher, Aristotle. Aristotle claimed that the quality of sound could stay constant and travel as far as the waves reached. Centuries later, in the 1600s, an Italian physicist, Galileo Galilei scientifically recorded for the first time, the relationship between the frequency of the wave to the pitch it produced. During his experiment, which he used a chisel to scrape against a brass plate, observed that the pitch of the screech varied directly to the spacing of the grooves that were created from the contact of the chisel and the brass plate.

Few years after Galileo's discoveries, a French mathematician, Marin Mersenne became the first person to record the speech of sound as it traveled through the air. After centuries of technological advancement, Mersenne's speed of sound was shown to have an error only less than 10%. Twenty years after Mersenne's discovery, a British scientist, Robert Boyle determined that for sound to travel, it has to go through a medium. To illustrate this concept, he placed a ringing bell inside a vacuumed glass jar. With the absence of an air source, the ringing bell was unheard, and he concluded that sound only travels through a medium. Couple years later, Sir Issac Newton proposed a simple formula for calculating the speed of sound in 1686. In his published work, the Principia Mathematica, Newton proposed the speed of sound through a medium: solid, liquid, or gas. Given the density of the medium ρ and the pressure acting on the sound (P), the speed of sound (c) was calculated as:

$$c = \sqrt{\frac{P}{\rho}}$$

A few years later, the flaws in Newton's work on the speed of the sound was realized by a French mathematician Pierre-Simon Laplace. Laplace found that Newton's formula neglected the influence of heat on the speed of sound by assuming that the process was isothermic. Laplace revised Newton's formula by adding an adiabatic constant (η) to Newton's pressure component:

$$c = \sqrt{\eta \frac{P}{\rho}}$$

The adiabatic process is a thermodynamic process that occurs without the transfer of heat of substances between the thermodynamic system and its surroundings- the energy is transferred to the surroundings only as work. The adiabatic constant is measured by taking a ratio between the specific heat c_p of pressure (holding pressure constant during one cycle of the wave) versus the specific heat c_v of the volume. The "constant" holding of the heat is possible because the temperature diffusion is slower than the acoustic cycle of between 30 to 30kHz. Today, the speed of sound is calculated to be traveling in $343\frac{m}{s}$, with the adiabatic compressibility of air

$\eta_o P_o = 1.4E5Pa$ and the density of air $\rho_o = 1.2\frac{kg}{m^3}$. In addition, it is important to take temperature into consideration in calculating the speed of a sound in a medium because of the particle density dependence on the temperature. This relationship is described by the static equation of absolute pressure and temperature:

$$\rho(P_o, T) = 1.29\left(\frac{P_o}{10^5}\right)\left(\frac{273}{T}\right)$$

From the equation, the particle density is inversely proportional to the temperature. Therefore, as the temperature rise, the particle density will decrease. From the speed of sound equation, the particle density is inversely proportional to the square of the speed. Therefore, we can deduce that the speed of sound is proportional to the square root of the temperature.

Research on sound waves continued throughout the 1700s and 1800s. In 1720, Daniel Bernoulli was one of the many prominent mathematicians in the Bernoulli family who won the ten prizes of the Paris Academy of Sciences- his feat equaled by only one other person of the century, his friend and rival Leonhard Euler. One day, due to an argument with Euler, Bernoulli became interested in the sound phenomena and later discovered that a closed organ pipe could produce odd harmonic sounds. He also discovered that the relative amplitudes of the harmonics are determined by pressure. Along with Bernoulli's success and contribution to speech analysis, Leonhard Euler developed the first solution to wave propagation of a string using partial differential equations. In addition, during the same time period, a French philosopher, Jean le Rond d'Alembert demonstrated the basic solution of the wave equation.

In 1759, on the recommendation of Euler and d'Alembert, Joseph-Louis Lagrange succeeded Euler as the director of mathematics at the Prussian Academy of Science in Berlin. During this time, he produced volumes of works that included creating of the calculus of variation and deriving the Euler-Lagrange equation for extreme of functional. A few years later, in 1790, a German mathematician and physicist, Carl Friedrich Gauss made an exceptional influence in many fields of mathematics and science. Even though he did not directly study acoustics, his laws have greatly contributed to the speech processing. For example, his laws can be used to relate acoustic power of sound source with sound intensity flow through any given surface. Many different sound-source shapes that are important in practical applications are analyzed by means of the Gauss' law. The invention of the Gaussian surface allows us to obtain a simple and straightforward method for calculating the sound intensity distribution analysis in space.

3.2 Model of Human Speech Production and Hearing

To model the production of speech, we capture the essence of the underlying physical mechanism by using a transmission line theory. We first simplify our model to a basic lossy cylindrical pipe as shown in Fig. 19a and build on the complexity of the model further. The model assumes that a signal travels as a plane wave, with sound pressure and volume velocity dependent on space (x). Pressure refers to the force applied perpendicular to the surface of an object per unit area over which the force is distributed. Sound pressure or acoustic pressure is the local pressure deviation from the ambient (average or equilibrium) atmospheric pressure, caused by a sound wave. Mechanisms of lossy cylindrical pipe can be represented in the circuit model shown in Fig. 19b.

Sound pressure and volume velocity are analogous to voltage and current. Therefore, the loss of dx of the tube can be represented in terms of impedance. Similar to a "loss," electrical impedance is the measure of the opposition that a circuit presents to a current when a voltage is applied. It is a vector quantity consisting of two independent scalar (one-dimensional) phenomena: resistance and reactance. Resistance denoted R, is a measure of the extent to which a substance opposes the movement of electrons among its atoms. On the other hand, reactance, denoted X, is an expression of the degree to which an electronic component, circuit, or system stores, and releases energy as the current and voltage fluctuates with each AC cycle.

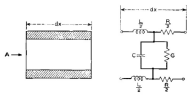


Figure 18. The reactance of the transmission line is modeled using inductor and capacitor, L and C respectively and the resistance of the line is modeled using the resistor and conductor, R and G respectively.

The passive component values of the per-unit-length constants are found by calculating:

$$L = \frac{\rho_o}{A} \qquad R = \frac{s}{A^2}\sqrt{\frac{\omega\rho_o\mu}{2}} \qquad C = \frac{A}{\rho_o c^2} \qquad G = \frac{s(\eta-1)}{\rho_o c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho_o}}$$

where A denotes the area of the tube, S denotes the circumference of the tube.

For an in-depth analysis of the transmission line behavior, we will first introduce the idea of multi-port parameters. As shown in Fig. 19, an n-port network has $2n$ terminals that are grouped in pairs to form n-ports. The i^{th} port has voltage, V_i, between its terminal and current I_i, flowing into one of the terminals and out of the other terminal. The transmission line model is a 2-port system comprised of linear network elements (i.e., linear resistor, capacitor, inductor). The model is also

time-invariant which merely means that the parameters of the network elements do not change with time. In a linear, time-invariant 2-port network, only four complex numbers are required to characterize its properties at one frequency completely. The parameters are interpreted in terms of input and output impedances (or admittances) and forward and reverse gains when a port is terminated with specific source and load impedances. The reference source and load impedances are constant for a given parameter set.

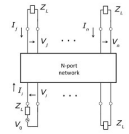


Figure 19. Basic n-port system representation with termination impedance Z_L.

The properties of the 2-port can be used to predict the properties of the 2-port in a system with arbitrary terminating impedances. Given the four complex numbers that comprise of a specific 2-port parameter set, it is possible to determine, for any combination of source and load terminations, the input and output impedance, the voltage and current gains, and the power gain. There are various ways to define the four complex numbers that describe a 2-port. The different possibilities arise from considering two of the four variables (V_1, I_1, V_2, I_2) to be dependent on the other two (independent) variables. Therefore, there are six possible choices of parameters. The four most commonly used parameter sets are Y, Z, ABCD, and S parameters. The ABCD parameters are defined such that the port one variables depend on the port two variables:

$$\begin{bmatrix} V_1(x,\omega) \\ I_1(x,\omega) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2(x+l,\omega) \\ I_2(x+l,\omega) \end{bmatrix}$$

Where x and ω are the distance and radian frequency respectively. The individual ABCD-parameters can be interpreted or calculated as follows:

$$A = \left.\frac{V_1}{V_2}\right|_{I_2=0}$$

$$B = \left.\frac{V_1}{I_2}\right|_{V_2=0}$$

$$C = \left.\frac{I_1}{V_2}\right|_{I_2=0}$$

$$D = \left.\frac{I_1}{I_2}\right|_{V_2=0}$$

Given the input and output voltage and current of the transmission line model of length l (V_1, I_1, V_2, I_2) in Fig. 20, the ABCD parameter is:

$$\begin{bmatrix} V_1(x,\omega) \\ I_1(x,\omega) \end{bmatrix} = \begin{bmatrix} cosh(\gamma l) & Z_o sinh(\gamma l) \\ Y_o cosh(\gamma l) & sinh(\gamma l) \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

where $\gamma = \sqrt{ZY}$ is the wave propagation constant.

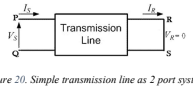


Figure 20. Simple transmission line as 2 port system.

In addition to a parametric analysis, transient analysis can be performed to analyze the transmission line by simulating an impulse wave (u) traveling along a transmission line with an open termination to model the glottis ($Z_g = \infty$) and resistive termination of Z_L to model the mouth. Fig. 21 illustrates 10 section division simulation of the transmission line.

Figure 21. Transient analysis of the transmission line that has been divided into N=10 sections. The traveling wave is denoted by u and the +/- indicates the direction of travel (+ being right and - being left).

The termination impedance (Z_L) can be found by using the pressure and velocity of the signal. The characteristic impedance of the transmission line is the ratio of the pressure and velocity of the signal:

$$Z_o = \frac{P_+}{U_+}$$

The reflectance (Γ) is a measure of the amount of reflected wave from the transmitted wave, thus is calculated by the ratio of the reverse and forward waves:

$$\Gamma = \frac{P_-}{P_+} = \frac{U_-}{U_+}$$

Using the Ohm's law, the termination impedance is the ratio of the pressure and velocity:

$$Z_L = \frac{P}{U} = P_{+}\frac{1}{U_{+-U_-\frac{P_-}{U_+}}}$$

Rearranging the above equation, we can find the reflection coefficient at the mouth (right end):

$$\Gamma_L = \frac{Z_L - Zo}{Z_L + Zo}$$

When the impulse hits u_{10}+, the amplitude will decrease by Z_L and reflect. Reflection coefficient at the glottis (left end) is 1 since the impedance is zero thus once the wave hits u_{10}-, the amplitude will not change and just bounce back in the + direction. If we let $Z_L = 2Z_o$, at the glottis: $r_g(t) = \delta(t)$ and at the mouth: $r_L(t) = \Gamma_L\delta(t)$ due to the reflection coefficient. Since $U_L = U_{L+-U_{L-=U_{L+(t-T_L)}}}$, we can get:

$$u_L(t,L) = u_{L+(t,L)-u_{L-(t,L)}}$$

$$\left(\delta(t) - Z_L(t)\right) * u_{L+(t,L)}$$

$$\frac{2}{3}u_{L+(t,L)}$$

$$\therefore u_L(t,L) = \sum\left(\frac{2}{3}\right)^n \delta\left(t - \frac{L}{c} - \frac{(n-1)2L}{c}\right)$$

$$u_L(t,L) = \sum\left(\frac{2}{3}\right)^n \delta\left(t - \frac{L}{c} - \frac{(n-1)2L}{c}\right)\vec{F}U_L(\omega,L) = \frac{2}{3}\frac{e^{-i\omega L/c}}{1-\frac{2}{3}e^{-2i\omega L/}}$$

The simple transmission line model can be extended by terminating it with a tube of difference size as shown in Fig 22. This model is referred to as the Helmholtz resonator, named after the German physicist Hermann von Helmholtz. The Helmholtz resonator consists of a smaller tube (the neck) and the larger tube (barrel) attached to it and has a single isolated resonant frequency. The model can be represented in a passive circuit using an inductor and a capacitor. The inductor has a mass $M = \frac{\rho_o l}{A_{neck}}$ and the capacitor has a compliance $C = \frac{V_{barrel}}{\eta P_o}$.

The resonant frequency of the model is calculated by setting the impedance to zero. Fig. xx explains the reasoning behind this. The electrical resonance occurs in an AC circuit when the two reactance which is opposite and equal cancel each other out ($X_L = X_c$). Thus in the Helmholtz resonator, the resonant frequency is calculated by:

$$X_L = X_C$$

$$sL = \frac{1}{sC}$$

$$\omega = \sqrt{\frac{1}{MC}} = \sqrt{\frac{1}{\frac{\rho_o l}{A_{neck}} * V_{barrel}}} = c\sqrt{\frac{A}{Vl}}$$

Converting radian frequency to frequency in Hz, we get:

$$f_{resonant} = \frac{c}{2\pi}\sqrt{\frac{A}{Vl}}$$

where c is the speed of sound in air. The isolated resonance of a Helmholtz resonator made it practical for the study of music and tones in the mid-19th century, before the invention of the electronic analyzers. When the resonator is held near the source of a sound, the air in it will begin to resonate; thus by listening to the tone of a musical instrument with the resonator, it was possible to identify the spectral component of a complex sound wave generated by the instrument.

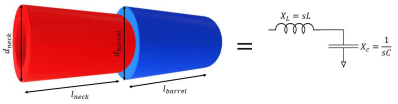


Figure 22. Simple model of Helmholtz's resonator with equivalent circuit model.

The complexity of the model can be added on further by introducing the topic of the horn with the varying area along the direction of propagation. Analysis of sound wave propagating in ducts of varying diameter like horns is suitably described by Weber's horn equation. The equation greatly simplifies the problem by reducing the problem from three to one dimension. This is achieved by assuming the crosswise uniform acoustic pressure field and averaging over a duct cross-section, decreasing the spatial dimension of he problem. For the one-dimensional scalar wave equation, the Wester Laplacian is defined as:

$$\nabla_r^2\varphi(r,t) = \frac{1}{A(r)}\frac{d}{dr}A(r)\frac{d}{dr}\varphi(r,t) = \frac{1}{c_o^2}\frac{d^2}{dt^2}\varphi(r,t) \leftrightarrow \frac{s^2}{c_o^2}P(r,s)$$

where $\varphi(r,t) \leftrightarrow P(r,s)$ is the acoustic pressure in time and frequency domains respectively. It is important to note than the operating frequency much lie below the critical value of

$$f_c = \frac{c_o}{2d}$$

for the equation to be valid, where d is the horn diameter and c is the speed of sound. Table 2 outlines the properties of N-1, 2, and three horns from *An Invitation to Mathematical Physics and Its History* by professor Jont. Allen.

Table 2. Table of horn properties for one, two, and three-dimensional horns.

N	Name	radius	Area/A_o	$F(r)$	$\mathcal{P}^{\pm}(r,s)$	$\varrho^{\pm}(r,t)$	$Y^a_{rad}/\mathcal{Y}$
1D	uniform	1	1	0	$e^{\pm\kappa(s)r}$	$\delta(t \mp r/c_o)$	1
2D	parabolic	$\sqrt{r/r_o}$	r/r_o	$1/r$	$H_2^{\pm}(-j\kappa(s)r)$	—	$\frac{-jrH_1^+}{H_0^+}$
3D	conical	r	r^2	$2/r$	$e^{\pm\kappa(s)r}/r$	$\delta(t) \pm \frac{c_o}{r}u(t)$	$1 \pm c_o/sr$

Even though the equation shows remarkable evidence of ingenuity and physical insight of horn modeling, the Webster Horn equation lacks an explanation of why the pressure may be assumed to be uniform, what the error of the approximation is, and what happens near the source or the load plane.

In 1962, a Polish neuroscientist, Jozef J. Zwislocki first modeled the cat ear. His work later built upon by the work of Lynch et al. in 1982. A simplified model of the model is illustrated in Fig. xx. The middle ear is modeled as a transmission line, terminated with capacitor C_{ad} which models the stiff ness and R_c which represents damping. The radiation impedance is represented by the resistance R_{rad} and inductor L_{rad}. As discussed in section 1.2, the pinna acts as a radiation impedance transformer. Therefore, it is represented by the pinna-horn transformer.

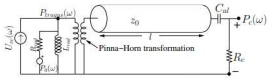


Figure 23. A simplified model of the ear canal and the middle ear (diagram from HW03).

Given that the radius of the canal is 3.75mm, the canal area is $\pi r_c^2 = 442mm^2$. Therefore, the characteristic impedance of the transmission line is $z_o = \rho c/\pi r_c^2 = 9.2M\Omega$. If we assume that the cochlear load resistor R_c is equal to twice the characteristic impedance of the ear canal, and that the impedance of the damping capacitance C_{ad} equals the load resistance at frequency of the 800Hz, we can calculate the capacitance value by setting $Z_{ad}=1/sC_{ad}=2z_o$ at frequency 800Hz, achieving the value of C_{ad}=10.8pF.

The free field transfer function of the middle ear can be calculated with the ratio of the cochlear pressure P_c/P_0. There are several methods of calculations. However, the problem can be easily solved with simple ABCD matrix introduced in the previous section. The system in the chain matrix form is:

$$\begin{bmatrix} P_o \\ U_o \end{bmatrix} = \begin{bmatrix} 1 & \frac{sL_eR_{rad}}{(sL_e+R_{rad})} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sinh\left(\frac{\omega}{c}l\right) & z_o\cosh\left(\frac{\omega}{c}l\right) \\ \frac{1}{z_o}\cosh\left(\frac{\omega}{c}l\right) & \sinh\left(\frac{\omega}{c}l\right) \end{bmatrix} \begin{bmatrix} 1 & sC_{ad} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Z_c & 1 \end{bmatrix} \begin{bmatrix} P_c \\ U_c \end{bmatrix}.$$

I would like to conclude this report with a sincere thank you to Professor Jont Allen for his teachings and support throughout ECE 537: Fundamentals of Speech Processing. As an electrical engineering student with no background in biology, I had faced some unique challenges in ECE 537. However, every time I had hard time understanding the course materials, Professor Allen had shown great generosity and patience with me. In addition, his passion for the subject directly reflected in his teachings during every class and made it enjoyable to show up to his classes and learn.

There are immense amount of materials covered in the class and to briefly summarize the whole material in a short 25 page max. report does not do it justice (I tried to simplify as much material as possible but still ended up with an extra page of a report so I apologize). ECE 537 was hands-down, the most interesting class I have taken during my time at the University of Illinois and I would definitely recommend this course to my colleagues.

One biggest take away that I apply directly to my everyday life is that I have completely stopped blasting loud music into my ears, damaging my poor hair cells ☺

ECE 537: Speech Processing Fundamentals
FINAL

Rebecca Chen

Fall 2018

1 Acoustics of Speech and Hearing

Speech, or sound, is transmitted through the air as a pressure wave. The source of a pressure wave of speech is the air from the lungs being transmitted through the glottis and vocal tract. A distinct speech sound, such as a consonant or a vowel, is called a *phone*. The sound wave then travels through the air and reaches the pinna, then travels through the middle ear and reaches the cochlea, which transforms the air pressure signal into electrical signals which are carried by auditory nerves to the brain. In the first section, we present the details of the propagation of sound as a wave, and we model sound waves as current travelling through a transmission line.

1.1 Speed of sound

The *spend of sound* through air is given by

$$c_0 = \sqrt{\frac{\eta_0 P_0}{\rho_0}}, \tag{1}$$

where η_0 is the adiabatic index, or the stiffness constant, P is pressure, and ρ is density. The stiffness constant is the ratio of the specific heat of air at constant pressure to the specific heat of air at constant volume. For air, $\eta \approx 1.4$, $P_0 = 1$ atm $= 101\,325$ Pa, and $\rho_0 = 1.225 kg/m^3$ at sea level.

Air density is dependent on temperature:

$$\rho = \frac{P_0}{RT}, \tag{2}$$

where R is the gas constant, which is $287 J/(K \cdot kg)$ for dry air. As temperature increases, air expands, and as the volume of the air increases, density decreases. If we consider a column of air, an increase in temperature causes the air to expand vertically. The mass of air particles in the column remains the same, so the force of the mass on any cross-sectional area of the column remains the same. Therefore air pressure is independent of temperature.

1.2 Measuring sound

The decibel (dB) is used to quantify ratios between two measurements. It allows us to compare values on a logarithmic scale and is suitable for comparing sound levels. The number of decibels (N_{dB}) between A_1 and A_2 is given by:

$$N_{dB} = 10 \log_{10} \frac{A_1}{A_2}. \tag{3}$$

For comparing squared values, we can see that:

$$N_{dB} = 10 \log_{10} \frac{A_1^2}{A_2^2} = 20 \log_{10} \frac{A_1}{A_2}. \tag{4}$$

The sound power level (SWL), sound pressure level (SPL), and sound intensity level (SIL) are three different quantities used to measure sound. SPL is given by $20 \log_{10}(P/P_{ref})$ dB, where $P_{ref} = 20\text{e-}6$ Pa. SIL is given by $10 \log_{10}(P/P_{ref})$ dB, where $P_{ref} = 10^{-6} W/m^2$. SWL is given by $10 \log_{10}(W/W_{ref})$ dB, where $W_{ref} = 10^{-12} W$. Sound pressure is inversely proportional to the distance from the source, sound intensity is inversely proportional to the squared distance, but power does not change over the distance from the source. The relationships between the three are given by:

$$W = \frac{I}{A} \tag{5}$$

and

$$I = \frac{|P^2|}{\rho_0 c}. \tag{6}$$

1.3 Scalar wave equation

The acoustic wave equation describes sound as a pressure wave:

$$\nabla^2 P(x,t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} P(x,t). \tag{7}$$

To derive Equation 7, we start with Newton's law, which says that force equals mass times acceleration ($F = ma$). A negative pressure gradient acts as the force accelerating a volume of air, so we have:

$$-\nabla P(x,t) = \rho_0 \frac{\partial}{\partial t} u(x,t), \tag{8}$$

where u is the volume velocity (volumetric flow rate), measured in m^3/s.

We also have Hooke's law, which says that the force required to deform a spring equals the stiffness multiplied by the distance the spring is deformed ($F = kx$). The acoustic complement is:

$$-\frac{\partial}{\partial t} P(x,t) = \eta_0 P_0 \nabla \cdot u(x,t) \tag{9}$$

We combine (8) and (9) by taking the divergence of (8):

$$-\nabla \cdot \nabla P(x,t) = \rho_0 \frac{\partial}{\partial t} \nabla \cdot u(x,t), \tag{10}$$

and then substituting for $\nabla \cdot u(x,t)$ from (9):

$$\nabla^2 P(x,t) = \frac{\rho_0}{\eta_0 P_0} \frac{\partial}{\partial t} P(x,t), \tag{11}$$

which gives us the wave equation:

$$\nabla^2 P(x,t) = \frac{1}{c_0^2} \frac{\partial}{\partial t} P(x,t) \tag{12}$$

1.4 Webster horn equation

Webster found a solution for Equation 12 for sound propagation in a horn by assuming that sound wave-front traveled in one direction (Figure 1), thus simplifying a three-dimensional problem into a one-dimensional problem. His solution is referred to as the Webster horn equation.

To derive the Webster horn equation, we begin by integrating (8) over iso-pressure surface S, where $\nabla P = 0$:

$$-\int_S \nabla P(x,t) \cdot dA = \rho_0 \frac{\partial}{\partial t} \int_S u(x,t) \cdot dA. \tag{13}$$

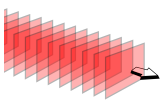


Figure 1: Plane wave traveling in one direction (Image source: Wikipedia)

We assume that pressure is uniform along the wave-front and define the average pressure:

$$P(x,t) := \frac{1}{A(x)} \int_S p(x,t) \hat{n} \cdot dA, \tag{14}$$

where $\hat{n}$ is the unit perpendicular vector to S. $\nabla P = \frac{\partial P}{\partial x} \hat{n}$, so the left-hand side of (13) becomes $\frac{\partial P(x,t)}{\partial x}$. The term $\int_S u(x,t) \cdot dA$ on the right-hand side is the definition of the volume velocity (or volumetric flow rate), or $\nu(x,t)$. Thus we have

$$-\frac{\partial}{\partial x} P(x,t) = \frac{\rho_0}{A(x)} \frac{\partial}{\partial t} \nu(x,t) \Longleftrightarrow \mathcal{Z}(x,s) \mathcal{V}. \tag{15}$$

This is Ohm's Law with impedance (in the frequency domain) $\mathcal{Z}(x,s) V$ on the right-hand side, where the impedance $\mathcal{Z}(x,s) = s\rho_0/A(x) = sM(x)$, and $M(x) = \rho_0/A(x)$ is the per-unit-length mass density of air.

Integrating (9) over the volume V between two iso-pressure surfaces x and $x + dx$ gives us:

$$-\int_V \nabla \cdot u \, dV = \frac{1}{\eta_0 P_0} \frac{\partial}{\partial t} \int_V p(x,t) dV. \tag{16}$$

We use the definition of average pressure to get:

$$\frac{\partial \nu}{\partial x} = -\frac{A(x)}{\eta_0 P_0} \frac{\partial P}{\partial t} \Longleftrightarrow -\mathcal{Y}(x,s) \mathcal{P}. \tag{17}$$

This is again Ohm's law with the admittance $\mathcal{Y}(s,x)$ on the right-hand side. In the Fourier domain, $\mathcal{Y}(s,x) = sA(x)/\eta_0 P_0 = sC(x)$, where $C(x) = A(x)/\eta_0 P_0$ is the per-unit-length compliance of the air.

In terms of $M(x)$ and $C(x)$, the speed of sound is

$$c_0 = \sqrt{\frac{stiffness}{mass}} = \sqrt{\frac{1}{C(x)M(x)}} = \sqrt{\frac{\eta_0 P_0}{\rho_0}} \tag{18}$$

Combining (15) and (17) and using the Webster Laplacian

$$\nabla_r^2 P(r,t) = \frac{1}{A(r)} \frac{\delta}{\delta r} A(r) \frac{\partial}{\partial r} P(r,t) \tag{19}$$

gives us Webster's horn equation:

$$\frac{1}{A(r)} \frac{\partial}{\partial r} A(r) \frac{\partial}{\partial r} P(r,t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} P(r,t), \tag{20}$$

where r indicates the distance along the axis of wave propagation, and c_0 is the speed of sound given by (1).

The Webster horn equation is a valid approximation only for certain frequencies, which we will see in the next section. Notice that (20) is equivalent to the scalar wave equation (Equation 7).

We can formulate equations (15) and (17) in matrix form:

$$-\frac{d}{dr} \begin{bmatrix} \mathcal{P}(r,\omega) \\ \mathcal{V}(r,\omega) \end{bmatrix} = \begin{bmatrix} 0 & sM(r) \\ sC(r) & 0 \end{bmatrix} \begin{bmatrix} \mathcal{P}(r,\omega) \\ \mathcal{V}(r,\omega) \end{bmatrix}, \tag{21}$$

where $M(r) = \rho_0/A(r)$ and $S(r) = A(r)/\eta_0 P_0$. To obtain (12), we take the partial derivative of the top equation and substitute for $\frac{\partial}{\partial r} \mathcal{V}$ and $\mathcal{V}$ using the top and bottom equations.

1.5 D'Alembert's solution to the wave equation

Solutions of the form $f(x - ct)$ are solutions to the wave equation (Equation 12). Since the horn equation in a second-order differential equation, it has two pressure eigenfunctions, $\mathcal{P}^+(r,s)$, corresponding to the forward-traveling wave, and $\mathcal{P}^-(r,s)$, corresponding to the backward-traveling wave. The general solution can always be written as the superposition of these two pressure functions, in the form of

$$P(x,t) = F(x - ct) + G(x - ct). \tag{22}$$

We can derive the relationship between $\mathcal{P}^+(r,s)$ and $\mathcal{P}^-(r,s)$.

The impedance looking into the horn is

$$Z_{in}(r,s) := \frac{\mathcal{P}(r,\omega)}{\mathcal{V}(r,\omega)} = \frac{\mathcal{P}^+ + \mathcal{P}^-}{\mathcal{V}^+ - \mathcal{V}^-} = \frac{\mathcal{P}^+}{\mathcal{V}^+} \left(\frac{1 + \mathcal{P}^-/\mathcal{P}^+}{1 - \mathcal{V}^-/\mathcal{V}^+} \right) = \mathcal{Z}(r) \frac{1 + \Gamma(r,s)}{1 - \Gamma(r,s)}. \tag{23}$$

$\Gamma(r,s)$ is called the reflectance and is given by

$$\Gamma(r,s) = \frac{\mathcal{V}^-(r,\omega)}{\mathcal{V}^+(r,\omega)} = \frac{\mathcal{P}^-(r,\omega)}{\mathcal{P}^+(r,\omega)}. \tag{24}$$

1.6 Acoustic horns

We can solve the Webster horn equation for different sizes and shapes of horns. For a cylindrical horn, the area is constant with respect to area, so Equation 20 simplifies to

$$\frac{\partial^2}{\partial r^2} P(x,t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2}. \tag{25}$$

For a conical horn (or a radiating sphere, since a cone is just a segment of a sphere), the area is proportional to the squared radius ($A(r) \propto r^2$), so $\frac{\partial}{\partial r} A(r) \propto r$, and Equation 20 becomes

$$\frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} p = \frac{1}{c_0^2} \frac{\partial^2 P}{\partial t^2}. \tag{26}$$

(Note that for a sphere, $x = r$.) For an exponential horn,

$$A(x) = A_0 e^{2mx}. \tag{27}$$

Plugging this into the Webster horn equation:

$$\frac{1}{A_0 e^{2mx}} \frac{\partial}{\partial x} A_0 e^{2mx} \frac{\partial}{\partial x} P = \left[\frac{\partial x}{\partial x^2} + m \frac{\partial}{\partial x} \right] P = \frac{1}{c^2} \frac{\partial^2}{\partial x^2} P. \tag{28}$$

This is in the form

$$P'' + aP' = \frac{1}{c^2} P'' \Longleftrightarrow \frac{s^2}{c^2} \mathcal{P}. \tag{29}$$

The left hand side can be rewritten as

$$\frac{1}{e^{mx}} \left[ae^{mx} \frac{\partial}{\partial x} + e^{mx} \frac{\partial^2}{\partial x^2} \right] P. \tag{30}$$

So the solution takes the form

$$\mathcal{P}^{\pm}(x,\omega) = e^{-mx} e^{\mp i \sqrt{\omega^2 - \omega_0^2} x / c}. \tag{31}$$

Note that the equations used for the area are just an approximation of the iso-pressure surface area required in the Webster horn equation.

1.7 Transmission line method

Wave propagation through a horn can be modeled as an electrical current flowing through a transmission line. We model each unit length of a horn as a cascaded element of a 2-port transmission line (Figure 2).

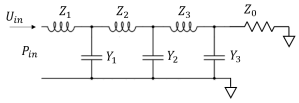


Figure 2: Cascaded transmission line elements

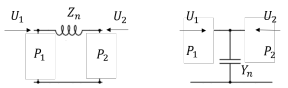


Figure 3: Left: Series element. Right: Shunt element

We can represent each element of the transmission line using a 2×2 matrix (ABCD-matrix). For a series element (Figure 3), the ABCD-matrix is

$$\begin{bmatrix} P_1 \\ V_1 \end{bmatrix} = \begin{bmatrix} 1 & Z(s) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_2 \\ -V_2 \end{bmatrix}. \tag{32}$$

For a shunt element (Figure 3), the ABCD-matrix is

$$\begin{bmatrix} P_1 \\ V_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ Y(s) & 1 \end{bmatrix} \begin{bmatrix} P_2 \\ -V_2 \end{bmatrix}. \tag{33}$$

To calculate the pressure and velocity through multiple elements of an acoustic transmission line, we can multiply the ABCD-matrices. For example, the unit-length horn (Figure 4) can be modeled as

$$\begin{bmatrix} P_1 \\ V_1 \end{bmatrix} = \begin{bmatrix} 1 & sL \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \begin{bmatrix} 1 & sL \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_2 \\ -V_2 \end{bmatrix} = \begin{bmatrix} 1 + s^2 LC & 2sL + s^3 L^2 C \\ sC & s^2 LC + 1 \end{bmatrix} \begin{bmatrix} P_2 \\ -V_2 \end{bmatrix}, \tag{34}$$

where the acoustic inertance $L = \frac{\rho_0}{A}$ and acoustic compliance $C = \frac{A_0}{\rho_0 c^2}$, and A is the area of the tube. If we make each inductance in Figure 4 $L/2$ and ignore higher order terms, the *ABCD*-matrix becomes

$$\begin{bmatrix} P_1 \\ V_1 \end{bmatrix} = \begin{bmatrix} 1 & sL \\ sC & 1 \end{bmatrix} \begin{bmatrix} P_2 \\ -V_2 \end{bmatrix}. \tag{35}$$

This approximation is true for $s < \sqrt{2} c_0$.

The input impedance (or Z_1) is given by

$$Z_1 = \frac{P_1}{V_1} = \frac{AP_2 - BV_2}{CP_2 - DV_2} = \frac{AZ_2 - B}{CZ_2 - D}. \tag{36}$$

For a uniform acoustic horn with a terminating load resistance, Z_1 refers to the characteristic impedance of the tube, and Z_2 refers to the load impedance. The characteristic impedance of a unit-length tube is given by

$$Z_0 = \frac{\rho_0 c_0}{A}, \tag{37}$$

where A is the area of the tube.

Figure 4: Simple LCL model of a unit-length horn

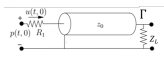


Figure 5: Piece of acoustic transmission line with reflectance Γ

For a section of transmission line of length l, (Figure 5) the ABCD-matrix is given by

$$\begin{bmatrix} P_1 \\ V_1 \end{bmatrix} = \begin{bmatrix} \cosh(j\omega l/c_0) & Z_0 \sinh(j\omega l/c_0) \\ Y_0 \cosh(j\omega l/c_0) & \sinh(j\omega l/c_0) \end{bmatrix} \begin{bmatrix} P_2 \\ -V_2 \end{bmatrix}, \tag{38}$$

where Z_0 and Y_0 refer to the characteristic impedance and characteristic admittance of the line, and

When there is a change in area of the horn along the length, reflectance of the wave occurs, resulting in the horn having poles and zeros at complex frequencies s_k where $\Gamma(r, s_k) = \pm 1$. The magnitude of the reflectance $|\Gamma|$ must be between 0 and 1 and is given by

$$\Gamma = \frac{Z_2 - Z_1}{Z_2 + Z_1}, \tag{39}$$

where Z_n is dependent on the area of the horn.

We can derive the input impedance of a transmission line in terms of reflectance:

$$Z = \frac{V}{I} = \frac{V^+ + V^-}{I^+ - I^-} = \frac{V^+}{I^+} \left(\frac{1 + V^-/V^+}{1 - I^-/I^+} \right) \tag{40}$$

$$= \frac{V^+}{I^+} \left(\frac{1 + \Gamma}{1 - \Gamma} \right) = z_0 \left(\frac{1 + \Gamma}{1 - \Gamma} \right). \tag{41}$$

In terms of characteristic impedance and load impedance, reflectance is

$$\Gamma = \frac{Z_L - z_0}{Z_L + z_0}. \tag{42}$$

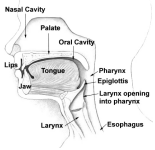


Figure 6: Vocal tract anatomy (Image source: http://my.ilstu.edu/)

Figure 7: Middle ear anatomy (Image source: Wikipedia)

If the horn terminates (a closed tube), $Z_L = \infty$ and $\Gamma = 1$. If instead the horn is open, $Z_L = 0$ and $\Gamma = -1$. If the horn is terminated in its own characteristic impedance ($Z_L = z_0$), then it is matched, and there is no reflectance.

1.8 Vocal tract

Figure 6 shows the vocal tract. The larynx is the source of sound for the vocal tract. The formants, spectral peaks of the sound spectrum, with the lowest three frequencies, correspond to the movement and position of the glottis, mouth, and lips. The vocal tract can be simply modeled as three uniform acoustic tubes in series, or three cascaded acoustic transmission lines of different radii (Figure 8). The velocity source is the lungs, modeled as a current source. The glottis is a vibrator that can be modeled as a Bernoulli oscillator. We can model the glottis as a resistor, so there will be a reflectance at the glottal end. We can also model the lips as a resistor, with its own reflectance. (There is also reflectance every time there is a change in area.) The lips can be modeled as a pulsating spherical radiator, where the radiation area is a semi-sphere.

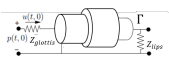


Figure 8: Model of the vocal tract

1.9 Ear canal simulation

Figure 7 shows the anatomy of the middle ear. The outer ear, or the pinna, acts as a transformer, which better matches the outside impedance to the characteristic impedance of the ear canal. It scales the pressure and volume velocity as follows:

$$P_{outside} = aP_{ec} \qquad U_{outside} = \frac{1}{a} U_{ec}, \tag{43}$$

where

$$a = \frac{A_{pinna}}{A_{ec}}. \tag{44}$$

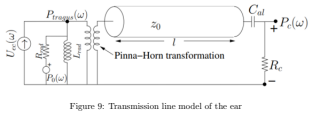


Figure 9: Transmission line model of the ear

This results in

$$Z_{outside} = a^2 Z_{ec}. \tag{45}$$

This allows us to hear lower frequencies. The sound pressure wave then travels through the ear canal, which is about 2.5 cm long, and causes the tympanic membrane (eardrum) to vibrate. The tympanic membrane is angled downward so that the incoming wave hits the top before it hits the bottom. This time delay helps the ear distinguish between different frequencies. The pressure wave is propagated through the ossicles - the malleus, incus, and stapes - then to the cochlea, which transmits signals of different frequencies to different neurons. These neurons form the auditory nerve, which transmits electrical signals to the brain.

The model of the middle ear has been studied by Zwislocki, Lynch, Rosowski, and others. Figure 9 shows a simplified model. We can calculate the cochlear impedance

$$Z_c = R_c + 1/sC_{ad} \tag{46}$$

and derive the reflection coefficient at the cochlea

$$\Gamma_c(s) = \frac{Z_c(s) - z_0}{Z_c(s) + z_0}. \tag{47}$$

Likewise, the radiation impedance looking out of the ear is given by

$$Z_{rad}(s) = \frac{sL_{rad} R_{rad}}{sL_{rad} + R_{rad}}, \tag{48}$$

where

$$R_{rad} = \rho c_0 / A_{rad}, \qquad L_{rad} = r_c \rho / A_{rad}, \tag{49}$$

and A_{rad} is the area of a half-sphere, given by

$$A_{rad} = 2\pi r_{rad}^2. \tag{50}$$

If we sample $\Gamma_c(s)$ and $\Gamma_{rad}(s)$ at a sampling frequency F_s, we can get $\Gamma_c(z^{-1})$ and $\Gamma_{rad}(z^{-1})$ in the form

$$\Gamma(z^{-1}) = \frac{b_1 + b_2 z^{-1}}{1 + a_2 z^{-1}} \tag{51}$$

We use

$$\Gamma_{rad}(z^{-1}) = \frac{-0.495 + 0.0105 z^{-1}}{1 - 0.516 z^{-1}}. \tag{52}$$

$$\Gamma_c(z^{-1}) = \frac{0.333 - 0.333 z^{-1}}{1 - 1 z^{-1}}. \tag{53}$$

To visualize a wave traveling through this system, we can derive the system response in the time domain by performing an inverse z-transform on $\Gamma_c(z^{-1})$ and $\Gamma_{rad}(z^{-1})$, giving us

$$y[n] = b_1 x[n] + b_2 x[n-1] - a_2 y[n-1], \tag{54}$$

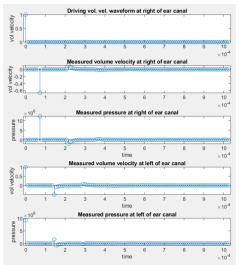


Figure 10: Impulse response of the middle ear model

where x is the input and y is the output. We simulate sound traveling through the ear canal in MATLAB, where the forward and backward velocity vectors $u_p(t,x)$ and $u_n(t,x)$ are stored as two N-dimensional arrays, with buffers for shifting the arrays. We apply the reflection boundary conditions at the ends of the ear canal. If we let x be an impulse, y will be the impulse response at the source, shown in Figure 10. It takes seven time steps for a wave to traverse the ear canal in our simulation, and we can see that the impulse reaches the cochlear end at steps 7,21,and 35, and it returns to the tragus at steps 14 and 28.

1.10 2-tube transmission line simulation of a vowel sound

We can simulate vowel sounds by using a 2-tube model of the vocal tract with several boundary conditions. Figure 11 shows the delay diagram for the two-tube model. First is the glottal boundary condition consisting of a velocity source in parallel with a flow resistance. Second is a lip boundary condition, consisting of a radiation impedance load. There is also a reflectance boundary condition where the tube changes area. The model parameters are defined in Figure 12. Section 1 corresponds to the glottal end and Section 2 corresponds to the mouth end. We synthesize four different vowel sounds: /i/ as in eve, /æ/ as in at, /a/ as in father, and /ə/ as in b/i/rd.

The average male vocal tract is L = 17 [cm] long. Assuming the speed of sound in air is 367 m/sec, it takes time T = L/c seconds to traverse the length L of the vocal tract. We use a sample rate of 44.1 kHz. The standard sampling rate (CD rates) 44.1 kHz has a sample period of dT = 1/Fs = 22.6757 μs. Sound travels dL = c/Fs [m], or 7.846 [mm] in 1 sample period. Thus we need M = 22 spatial samples at a time sampling rate 44.1 kHz to reach a length of 17 [cm].

As before, the forward and backward velocity vectors $u_p(t,x)$ and $u_n(t,x)$ are stored as two N-dimensional arrays, with buffers for shifting the arrays. The reflection coefficient at the glottis is $\Gamma_g = 0.95$. We create a glottal pulse using the pitch defined by $f_0(t) = 150 + 100t/1.5 + 25\sin(2\pi t)/1.5 + 0.2\sin(2\pi 20t)$. The glottal

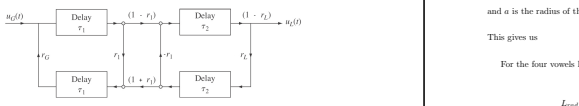


Figure 11: Diagram of two-tube delay line

Vowel	Section	Length [cm]	Area [cm^2]
/e/ve	1	9	8
	2	8	1
/m/t	1	4	1
	2	13	8
U/n/ther	1	9	1
	2	8	7
h/a/rd	1	17	6
	2	0	6

Figure 12: Dimensions of 2-tube model of vocal tract

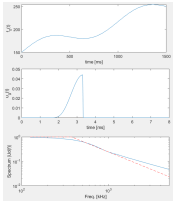


Figure 13: Top:$f_0(t)$. Center: $u_g(t)$. Bottom: $|U_d(f)|$

pulse is plotted in Figure 13.

The radiation impedance at the lips is given by

$$Z_{rad}(s) = \frac{sL_{rad}R_{rad}}{sL_{rad} + R_{rad}} = \frac{L_{rad}R_{rad}}{L_{rad}} \frac{s}{s + R_{rad}/L_{rad}}, \tag{55}$$

where

$$R_{rad} = 0.27\rho_0/a, \qquad L_{rad} = 0.459\rho_c/a^2,$$

and a is the radius of the lips, given by

$$a = \sqrt{A/\pi}.$$

This gives us

$$R_{rad} = 0.27\rho_0/a, \qquad L_{rad} = 0.459\rho_c/a^2,$$

For the four vowels labeled 1, 2, 3, and 4, this gives us

$$a_1 = 0.0056m;\ a_2 = 0.0160m;\ a_3 = 0.0149m;\ a_4 = 0.0138m;$$

$$L_{rad,1} = 56.89H;\ L_{rad,2} = 19.91H\ L_{rad,3} = 21.38H;\ L_{rad,4} = 23.09H;$$

$$R_{rad,1} = 63380000\Omega;\ R_{rad,2} = 7760000\Omega;\ R_{rad,3} = 8950000\Omega;\ R_{rad,4} = 10040000\Omega.$$

We plug in these values of L_{rad} and R_{rad} to get $Z_{rad}(s)$.

The characteristic impedances of the tubes is given by

$$Z_1 = \frac{\rho c}{A_1} \text{ and } Z_2 = \frac{\rho c}{A_2}.$$

The reflectance coffeficient at the lips is given by

$$\Gamma_{rad} = \frac{Z_{rad} - Z_2}{Z_{rad} + Z_2} \tag{56}$$

Substituting (55) into (56) gives us

$$\Gamma_{rad}(s) = \frac{s(LR - LZ_2) - RZ_2}{s(LR + LZ_2) + RZ_2} \tag{57}$$

If we put the B and A coefficients from (57) into MATLAB's bilinear() function with a sampling rate of 96 kHz, we get the time domain filter in the form

$$\Gamma_{rad}(z^{-1}) = \frac{b_1 + b_1 z^{-1}}{a_0 + a_1 z^{-1}}.$$

At the boundary where the area of the tube changes, we have another set of reflection coefficients. In the forward direction,

$$\Gamma_p = \frac{\rho c/A_2 - \rho c/A_1}{\rho c/A_2 + \rho c/A_1} = \frac{A_1 - A_2}{A_1 + A_2} \tag{58}$$

In the backward direction,

$$\Gamma_n = \frac{A_2 - A_1}{A_2 + A_1} = -\Gamma_1. \tag{59}$$

For vowels 1-4, this gives us:

$$\Gamma_{p,1} = 7/9;\ \Gamma_{p,2} = -7/9;\ \Gamma_{p,3} = -3/4;\ \Gamma_{p,4} = 0;$$

1.11 Room acoustics

Sound is reflected off walls at an incident angle and reflected off the wall (similarly to light). In the simplest model, the reflection angle is the same as the incident angle. If a sound source is at a distance d from a wall, the reflected sound can be modeled as an additional sound source at a distance d on the other side of the wall (Fig. 14).

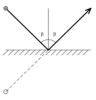


Figure 14: Reflection of sound off wall

2 Speech processing

2.1 Fourier and Laplace transform

The discrete Fourier transform of a signal is the projection of a vectorized signal F onto $E_\omega = \{e^{-j\omega_k t_k}\}_k$:

$$\mathcal{FT}\{F\} = F \cdot E_\omega = \sum_k f_k e^{-j\omega_k t_k}, \tag{60}$$

where $\cdot$ indicates a dot product. The Fourier transform is used to obtain the (real or imaginary) frequency components of (noncausal or causal) signals. As the time between samples $T_s \to 0$, this becomes the continuous Fourier transform.

The Laplace transform is the integral projection of a signal onto onto $E_s = e^{-st}$, given by

$$\mathcal{L}\{f(t)\} = \int_0^\infty f(t)e^{-st}dt. \tag{61}$$

The Laplace transform is used to analyze (causal) systems. The poles and zeros of the Laplace transform determine how the system operates on signals.

Both the Fourier transform and Laplace transform are linear.

2.2 Cepstral Analysis

Cepstral analysis can be used to detect pitch in speech signals. To get the cepstrum of a signal, the magnitude of its Fourier transform is taken, followed by a logarithm, and then the inverse Fourier transform is taken. After taking the log spectrum, the log-frequencies (quefrencies) corresponding to the harmonics of the speech become regularly spaced, and after taking the IFFT, there will be a peak at the fundamental frequency of the speech.

2.3 Short-time Fourier Transform

Short-time Fourier transform (STFT) is different method used to analyze the frequency components of speech over time. The Fourier transform is computed over sliding windows (e.g., Hamming window) of the signal, and these spectrums are added together. We call the length of the slide the hop size R. To get good reconstruction, R needs to be small enough. This method of filtering sections of the signal is called the overlap-add method. The spectrums can be plotted as a function of time to get the spectrogram (Figure 15). If a window has a constant overlap-add property at hop size R, the sum of successive DTFTs over time equals the DTFT of the entire signal. To see this, we begin with the discrete convolution of a signal $x[n]$ with the finite impulse response filter $h[n]$:

$$y[n] = x[n] * h[n] = \sum_{m-1}^{M} h[m] \cdot x[n - m]. \tag{62}$$

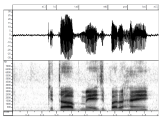


Figure 15: Spectrogram of speech

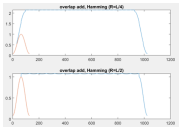


Figure 16: Hamming-128 overlap adds

If $x[n]$ is long, we can divide this convolution into shorter convolutions with segments of $x[n]$:

$$x[n] = \sum_k x_k[n - kL], \text{ where } x_k[n] = \begin{cases} x[n + kL] & n = 1, 2, ..., L \\ 0 & otherwise. \end{cases} \tag{63}$$

Then we can write:

$$y[n] = \left(\sum_k x_k[n - kL]\right) * h[n] = \sum_k (x_k[n - kL] * h[n]) = \sum_k y_k[n - kL]. \tag{64}$$

Of course, we use FFTs in practice to perform the convolutions.

To see how R is chosen, we observe that R is the downsampling factor of the convolution, since there is only one overlap every R samples. To avoid aliasing during reconstruction, R must be chosen so that $2f_{max}/R > f_c$, where f_c is the cutoff frequency of the window. By convention, this is usually the first zero-crossing of the window's spectrum beyond the main lobe. This gives us $R = L/4$ for Hamming and Hanning windows. (`https://ccrma.stanford.edu/~jos/parshl/Choice_Hop_Size.html`) Figure 16 shows overlap adds of Hamming-128 windows with different choices of R. If we zoom in to the overlap adds of a Kaiser window, we can see that the error $e(t)$ is periodic because the overlap adds are periodic with period R. At the ends, $|e(t)|$ is large because there is a ramp up to the flat band-pass filter.

2.4 Audio coding

Audio coding involves quantizing audio signals for compression. The encoder processes audio signals using signal processing techniques, and redundant or imperceptible information is thrown away to enable feasible

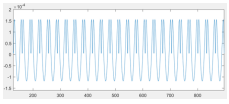


Figure 17: Overlap add error: $e(t)$, t=128:896

transmission and storage of audio. Speech and music are two types of audio whose differing characteristics necessitate different compression techniques.

2.4.1 μ-law quantizer

We can sample a continuous speech signal $S(t)$ with sampling period T to get $S[n]$. The value of each sample $s[n]$ must be quantized so the speech signal can be encoded at a certain rate. Figure 18 illustrates uniform and non-uniform quantization. $Q(x)$ is the quantization of x, and $q(x)$ is the quantization error $x - Q(x)$. Companding algorithms are used to reduce the dynamic range of the input signal. Since smaller signals are more sensitive to quantization noise than larger signals, we allocate more bits to smaller signals. Since perceived acoustic intensity level or loudness is logarithmic, we can compressing the signal logarithmically using μ-law companding. The equation of μ-law companding is given by:

$$Q(x) = sgn(x)\frac{\ln(1 + \mu|x|)}{\ln(1 + \mu)}, \ -1 \le x \le 1. \tag{65}$$

The μ-law algorithm was designed at AT&T in 1965, and it is used in 8-bit PCM digital telecommunication systems in North America and Japan (where $\mu = 255$) [Wikipedia]. The bandwidth of land-lines is around 3 kHz (0.4-3.4 kHz).

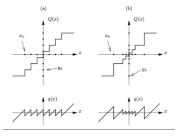


Figure 18: Uniform and non-uniform quantization

We can visualize a quantized sine wave in Fig. , which we quantize with a non-uniform quantizer in MATLAB (simple lossy codec from HW 6). When we reduce dynamic range by 1/2, we are encoding with 9 bits per speech sample. When we reduce dynamic rage to 1/3, we are encoding with 6 bits per symbol. Notice that the quantization is finer at smaller amplitudes. More quantization causes more distortion in the perceived sound.

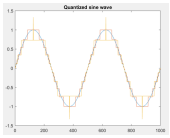


Figure 19: Quantized sine wave (16 (blue), 9 (red), 6 (yellow) bits per sample

2.4.2 Linear predictive coding

Linear predictive coding (LPC) is a technique used for encoding speech at low bit rates. Linear prediction assumes that the current sample signal can be estimated using a linear combination of the previous k values:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k s[n - k]. \tag{66}$$

The prediction error is:

$$e[n] = s[n] - \sum_{k=1}^{K} a_k s[n - k]. \tag{67}$$

The coefficients a_k's are computed by minimizing

$$\sum_{n=1}^{N} |e[n]|. \tag{68}$$

In speech, a_k's are slowly varying, so they are assumed to be constant over each syllable (about 20 ms), which is much lower than the sampling frequency (about once every 100 μs). Instead of sending $s[n]$, we send $a_1, ..., a_k$ at the lower syllabic rate and $e[n]$ at the speech sampling rate. However, the variance of $e[n]$ is much lower than at $x[n]$, so we can encode it using fewer bits.

If we take the z-transform of (67), we get:

$$E[z] = S[z] - \sum_{k=1}^{K} a_k z^{-k} S[z] = \left(1 - \sum_{k=1}^{K} a_k z^{-k}\right) S[z], \tag{69}$$

or

$$S[z] = \frac{1}{1 - \sum_{k=1}^{K} a_k z^{-k}} \cdot E[z]. \tag{70}$$

This is an all-pole filter.

In our speech production model, we had a tube with an impulse train input $i[n]$ at the glottis and a radiation impedance at the mouth. The speech formants corresponded to the areas of the tube. We model each formant with a 2-pole filter in the form $\frac{1}{1 - a_1 z^{-1} - a_2 z^{-2}}$, and we can cascade multiple filters to model more formants. We model the radiation impedance as a first-order filter in the form $\frac{1}{1 - bz}$. Taking the product of these filters, we will have a q-th order filter:

$$S[z] = \frac{1}{1 - \sum_{k=1}^{q} a_k z^{-k}} \cdot I[z]. \tag{71}$$

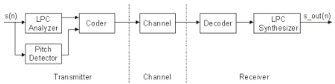


Figure 20: Linear predictive coding diagram

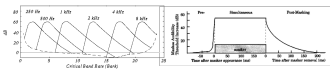


Figure 21: Left: Frequency masking. Right: Temporal masking

The a_k's give us the position and bandwidths of the formant resonances. Compare (70) and (71). If we choose K to be q, we can choose a_k to correspond to the formant characteristics for each syllable duration. We can encode the excitation pulses $I[z]$ with a certain pitch and amplitude, which can detect (Fig. 20). Because these do not vary quickly in speech, we can also encode these at the syllabic rate. In order to further reduce the bit-rate, we can also encode $e[n]$ at this lower rate. In summary, encoder combines the LPC analyzer output (the prediction coefficients, which characterize the formants) and the pitch detector output (the pitch and amplitude information). The pitch detector also outputs a voiced/unvoiced flag.

The decoder takes the pitch, amplitude, and voiced/unvoiced information and feeds it into an impulse/noise generator. The synthesizer factors the prediction coefficients into second-order terms to obtain formant coefficients, then filters the impulse using these coefficients.

LPC is used in voice codec of the the GSM standard for mobile communications. It is also used in other lossless audio codecs such as MPEG-4 ALS, FLAC, and SILK [Wikipedia].

2.4.3 MP3

MP3 is an audio codec that works by discarding parts of the audio that the average human ear cannot hear. These include signal amplitudes that are below the cochlea's threshold sensitivities, which are frequency-dependent. This method is commonly referred to as perceptual coding, or psychoacoustic modeling.

In addition to the minimum amplitude threshold, frequency and temporal masking in the cochlea can also occur. Frequency masking occurs when a loud frequency masks simultaneous neighboring frequencies of lower amplitudes. In other words, the neighboring frequencies must have a higher amplitude than normal for the ear to hear them. Temporal masking occurs when a loud sound masks neighboring frequencies before and after it occurs (see Fig. 21). This is referred to as perceptual transform coding.

After the audio signal is analyzed using a filterbank (with 32 subbands), it is quantized using perceptual transform coding and an optimum (nonlinear) frequency quantization. It is then compressed using a Huffman code.

3 Psychoacoustics and Anatomy

3.1 Cochlear anatomy

The cochlea is a spiral chamber filled with fluid that acts as a physiological filterbank. A in Fig. 22 depicts the *basilar membrane* (BM), and B depicts a cross section of the cochlea. The BM is a structure that separates two fluid-filled tubes - the *scala vestibuli* and the *scala timpani*. The BM is most stiff at the base and least stiff at the apex. Lower frequencies travel further along the basilar membrane and cause maximal vibration closer to the apex of the cochlea, while higher frequencies cause maximal vibrations at the base. As the ossicles vibrate, they send mechanical waves down the fluid in the cochlea. Within the organ of Corti (in pink), there are *hair cells* with cilia that shear against the *tectorial membrane* when the BM vibrates. There are about 30,000 hair cells along 35mm (or one every 12 μm). When the cilia bend, the hair cells fire and excite the cochlear nerve fibers. The firing rate is related to the vibration amplitude. Figure 23 shows Wegel and Lane's nonlinear cochlear model. In their 1924 paper, they determine how much tones 200 to 3500 Hz could mask tones of 150 to 5000 Hz. For a given masking frequency n, the slope increases from zero through nearly 1.0 for a frequency near n, then more slowly, approaching about 3 to 4 for the highest frequencies measured. They show that the masking is greatest for tones nearly alike, and that when the masking tone is loud it masks tones of higher frequency better than those of frequency lower than itself. When the masking tone is weak, there is little difference. Figure 24 shows some cochlear tuning curves (Kiang *et al.*, 1965). The steepest slope corresponds to the frequency at which the nerve is most sensitive.

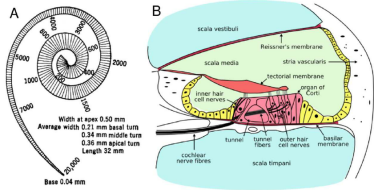


Figure 22: Cochlea and basilar membrane

3.2 The auditory neural pathway

The neuronal signals travel from the cochlear nerve fibers to the brain to be decoded as speech or some other auditory signal. Figure 25 shows a neuron. Chemical signals are sent between neurons in the form of neurotransmitters. The *dendrites* of one neuron receives multiple inputs from the axons of preceding neurons. The chemical signal is converted to an electrical signal within the neuron, and the signal travels from the dendrites to the *axon*. Within the cell body or soma, electrical potentials are graded, and in the axon, electrical potentials are impulses (e.g. logical 0/1), called *action potentials*. When the summation of input signals from the dendrite exceed a certain threshold, an action potential is sent through the axon. Some neurons have Myelin sheaths, which act as insulators along the axon. This speeds up the rate at which

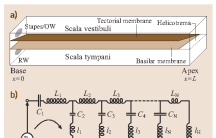


Figure 23: Wegel and Lane's nonlinear cochlear model (1924)

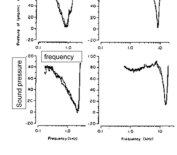


Figure 24: Cochlear tuning curves

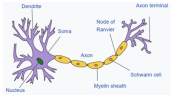


Figure 25: Neuron (Wikipedia)

impulses are propagated – impulses "jump" from one node of Ranvier to another, reducing current leakage across the neuron membrane. The region between connecting neurons is called a synapse.

The primary auditory pathway (shown in Fig. 26) carries information from the cochlea to the primary auditory cortex in the brain via a bundle of nerve fibers called the auditory nerve. The auditory nerve first synapses at the cochlear nuclei, which is in the brainstem. Each cochlear nucleus is divided into two parts: the *dorsal cochlear nucleus* (DCN) and the *ventral cochlear nucleus* (VCN). The VCN is further divided into anterior and posterior sections (AVCN and PVCN respectively). Neurons within each section form a tonotopic mapping.

There are five types of cells that convey information from the VCN: stellate cells, bushy cells (globular and

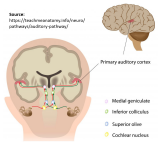


Figure 26: Neuronal pathway

spherical), octopus cells, and multipolar cells. The *stellate cells* have dendrites that run parallel to the auditory nerve fibers. Each stellate cell receives many inputs from only a few nerve fibers – each responds to only a narrow band of *frequencies*. Stellate cells have a linear response to depolarizations and hyperpolarizations. They produce a regular train of action potentials when depolarized and are thus named "chopper" cells. The number of action potentials in a train encodes the loudness of each frequency. Stellate cells project to the *inferior colliculus*. *Bushy cells* receive input from only a few nerve fibers and preserve *temporal* information. Bushy cells localize sound and have a nonlinear response to depolarization. They are hard to depolarize and produce only one or two action potentials for long depolarizations. Bushy cells project to the *superior olivary complex*. *Octopus cells* have two to three large dendrites which run perpendicular to the auditory nerve fiber. They integrate information across frequencies and detect synchrony – they are activated by synchronous activity across many nerve fibers and project to the *lateral lemniscus* (located between the superior olive and the inferior colliculus). *Multipolar cells* have dendrites that extend across a large number of nerve fibers and project to the cochlear nucleus on the other side of the brainstem. Octopus cells fire only at the onset of a broad-band stimulus. They fire with very high temporal precision and are thought to be important for extracting timing information. Figure 27 gives summary of the VCN neurons.

In animals that can move their ears, the *dorsal cochlear nucleus* is a layered structure. There are three main types of cells in the DCN: fusiform cells, granule cells, and tuberculoventral cells. *Fusiform cells* integrate information through two sets of dendrites which receive information from the two layers: the outer molecular layer and the deep layer. The outermost molecular layer contains *granule cells* which receive other types of sensory information, such as information about the location of the head and ears. *Tuberculoventral neurons* in the deep layer of the DCN project to the VCN. They inhibit cells that receive inputs from the same auditory nerve fibers that innervate them, thus suppressing echoes. A summary of cells in the DCN is given in Figure 28.

3.3 Loudness and the JND

Psychophysics describes how humans perceive physical magnitudes. Loudness is the way we perceive amplitude. Every time we hear a sound of the same intensity, we perceive it a little differently. The intensity JND (just noticeable difference) is a measure of this internal perceptual fluctuation. The loudness JND varies with the physical amplitude of sound. Ernst Heinrich Weber (1795–1878) conducted psychoacoustic experiments and showed that the JND is proportional to the initial stimuli: $\Delta I \propto I$ (Weber's Law 1846). His student Gustav Theodor Fechner suggested that

$$\frac{(JND)\Delta I}{I} = constant,$$

Characteristics of Ventral Cochlear Nucleus Neurons

	Spherical Bushy cells	Globular Bushy cells	Stellate cells	Multipolar cells	Octopus cells
Tuning	Narrow	Narrow	Narrow	Broad	Broad
Intrinsic Membrane Properties	Nonlinear	Nonlinear	Linear	Linear	Nonlinear
PST Histogram	Primary-like	Primary-like with notch	Chopper	Onset - Chopper	Onset
Function	Preserve timing information	Preserve timing information	Loudness - integrate in critical band	Loudness – overall (AGC?)	Synchrony – integrate across freq.
Project to - send info	Superior Olivary Complex	Superior Olivary Complex *	Inferior Colliculus	Cochlear nucleus on other side	Lateral Lemniscus

Figure 27: Summary of neurons in the ventral cochlear nucleus

Characteristics of Dorsal Cochlear Nucleus Neurons

	Fusiform cells	Granule cells	Tuberculoventral neurons
Tuning	Narrow/broad sidebands	?	Narrow
Intrinsic Membrane Properties	Nonlinear	?	Nonlinear
PST Histogram	Pauser - Buildup		Many different shapes
Function	Compensate for motion of pinna	Convey info about position of pinna	Suppress echoes
Project to - send info	Inferior Colliculus	Molecular layer of DCN	Ventral cochlear nucleus

Figure 28: Summary of neurons in the dorsal cochlear nucleus

or that loudness is proportional to the number of JNDs. The number of JNDs is given by

$$N_{JND} = \int_{L_1}^{L_2} \frac{dL}{\Delta L} = (L_1 - L_2)/\Delta L = \int_{I_1}^{I_2} \frac{dI}{\Delta I}. \tag{72}$$

He showed that perceived loudness is proportional to logarithm of the physical intensity:

$$L(I) \propto \log \frac{I}{I_0} \text{ Fechner's Law (1860)},$$

and suggesting counting JNDs. In 1927, Thurstone defines the JND as the "relative signal level for detection 75% of the time." In 1928, Robert R. Riesz shows that Weber's Law does not hold for pure tones and establishes the near-miss to Weber's Law. In 1933, Fletcher and Munson plot perception of loudness against the frequency of tones, resulting in the Fletcher-Munson curves of equal loudness. They plot the energy of

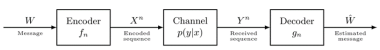


Figure 29: Shannon's communication channel

sound at different frequencies required for humans to consider two frequencies equally loud and show that two equally loud tones played together sound twice as loud. They suggest that

$$L(I) \approx I^{1/3} \text{ Steven's Law (1860)}.$$

In 1947 G.A. Miller describes the wide-band JND. In 1966, Green and Swets apply signal detection theory to psychoacoustics. In 1997, Allen and Neely reanalyze the Fletcher-Munson curves to find a Poisson-like relationship. They show

$$\Delta L(L) \approx \sqrt{L}.$$

4 Human speech recognition and information processing

4.1 Concepts from information theory

Some concepts from information theory are useful for analyzing speech. The field of information theory was founded by Claude E. Shannon in 1948 in his paper entitled "A Mathematical Theory of Communication." Figure 29 shows the fundamental elements of a communication channel. Shannon defined the information content of a random variable X as:

$$I(X) = -\log P(X), \tag{73}$$

where $P(X)$ is the probability mass function of X. *Entropy* measures the amount of uncertainty involved in the value of a random variable. The entropy of a discrete random variable X is defined as:

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log P(X)]. \tag{74}$$

If the possible values of X are $x \in \mathcal{X}$ then

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x). \tag{75}$$

When the base of the logarithm in the above equations is 2, the units of entropy and information are in *bits*. The entropy *rate* of a data source is equal to the average number of bits per symbol needed to encode it. Entropy gives an upper bound on the lossless compression of a data source. In a good (i.e. short) code, less common symbols are encoded with longer codewords and more common symbols are encoded with shorter codewords. The average length of a code C is given by:

$$L(C) = \sum_{x \in \mathcal{X}} P(x)\ell(x), \tag{76}$$

where $\ell(x)$ is the length of the codeword used to encode the symbol x. Shannon found experimentally that the information rate of the English language is between 0.6 and 1.3 bits per character. Shannon also analyzed the Morse code. Intuitively, the Morse code assigns shorter codewords to more common alphabet letters (such as e and a) and longer codewords to less common alphabet letters (such as x and z). Although the Morse code is decent, it does not achieve the optimal entropy bound. The Huffman code is an example of an optimal code that achieves this bound, and it is used in mp3 coding. Below are some important events in communications.

1832	The Morse code is invented
1924	Nyquist comes up with his rate theorem
1928	Hartley tries to formulate a theory of the transmission of information
1939	Dudley invents the vocoder, demonstrating that intelligible communication does not require a bandwidth at least as wide as the bandwidth of the signal
1948	Shannon publishes landmark papers defining the field of information theory

Table 1: Important events in communications

In addition to characterizing the message source, we can also characterize the channel using a conditional probability distribution $p(y|x)$, where x is a transmitted symbol and y is the received symbol. Shannon defines the *conditional entropy* $H(Y|X)$ as

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log P(y|x). \tag{77}$$

The channel capacity is the highest information rate that can be achieved with an arbitrarily small probability of error, defined as:

$$C = \sup_{P_X(x)} I(X; Y), \tag{78}$$

where $I(X;Y)$ is the mutual information of X and Y, given by $I(X;Y) = H(Y) - H(Y|X)$. Thus if the entropy of Y given X ($H(Y|X)$) is large, that is, there is a lot of uncertainty what the received Y will be when X is transmitted, then the channel capacity will be small.

This concept can be applied to confusion matrices (see Figure 30). The rows indicate the actual phoneme and the columns indicate the perceived phenome. The matrix values are counts of each event, where correctly perceived phonemes land on the diagonal. If we normalize the values the confusion matrix, we can get joint and marginal probabilities and calculate the channel capacity of speech.

	f	v	s	z	ʃ	ʒ	ɤ	NR	Total
f	**99**	*2*	*11*	*3*	*2*		*1*	1	120
v	*3*	**108**		*2*	*2*	*1*	*1*	2	120
s	*3*	*1*	**108**	*1*	*2*	*3*	*1*	1	120
z	*2*	*1*	*1*	**105**	*3*	*5*	*2*	1	120
ʃ	*2*	*2*	*1*	*2*	**107**	*1*	*3*	2	120
ʒ	*1*	*1*	*3*	*6*	*2*	**103**	*2*	2	120
ɤ	*3*	*1*	*2*	*1*	*4*	*2*	**103**	3	120
Total	113	116	126	120	122	115	113	12	840

The numbers in bold letters refer to the hits, and those in italic letters, to the errors.
Note: NR = non-response

Figure 30: Perceptual-auditory confusion matrix

The conditional probabilities are analogous to the transition probabilities of a Markov chain, where the states are phonemes or some other subdivison of speech (see Fig. 31). One can construct a confusion matrix from a Markov chain and vice versa if we assume speech is a Markov process.

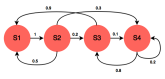


Figure 31: Markov chain state diagram

4.2 Articulation index

Harvey Fletcher (1884-1981) invented the hearing aid and is considered the father of stereophonic sound. He conducted the oil drop experiment measuring the charge of an electron, commonly attributed to his advisor Robert Millikan. Fletcher showed that speech features are usually spread over a wide frequency range, and developed a statistical articulation index to approximately quantify the quality of a speech channel by using nonsense phones. He proposed that different frequency bands had different articulation indices. He used 20 critical bands of speech (to match the 20mm of the basilar membrane) and measured articulation as

$$s = 1 - e_{total} = 1 - e_1 e_2 \dots e_{20}, \tag{79}$$

where e_i is the error for the ith band. Jont Allen refers to this as parallel processing of speech, in which the entire speech signal is correctly perceived as long as one band of the speech signal is correctly perceived. In 1947 French and Steinberg conducted numerous experiments during WWII to determine how factors such as intensity of speech, intensity of noise, echo, reverbration, hearing, etc., affected the articulation index. In 1951 George Miller analyzed human speech communication with Claude Shannon's theory of information and in 1955 he analyzed phone deceding as a function of the speech-to-noise ratio using confusion matrices.

4.3 Signal-noise separation using expectation-maximization

Expectation-maximization (EM) is an algorithm used to find the underlying distribution of some data. The best distribution is determined by the maximum-likehood rule, that is, we find the distribution that gives us the highest probability of generating our data. We first make some assumptions about the underlying distribution (such as a mixture of Gaussians) and then find the best parameters of the underlying model using EM. Given a dataset $\mathbf{X}$, we use $\mathbf{Z}$ as a latent variable indicating membership in one of a set of groups. The steps are as follows:

1. Initialize the parameters θ to some random values.
2. *Maximization step:* Compute the probability of each possible value of $\mathbf{Z}$, given θ, choose $\mathbf{Z}$ to be the maximum likelihood estimators of the distribution described by θ
3. *Expectation step:* Use the $\mathbf{Z}$ from the previous step to compute a better estimate for the parameters θ.
4. Iterate steps 2 and 3 until convergence.

White noise can be modeled as a zero-mean Gaussian signal and speech can be modeled as some other Gaussian or Poisson distribution. Since the two types of signals original from different distributions, they can be separated.

2 FINAL REPORT: A

This section consists of two reports.

ECE 537, Fall 2018
Final Project

Justin Sconza
sconza@illinois.edu

December 18, 2018

Introduction

The intended audience for this report is its own author in the future and the purpose of this report is to serve as a reminder of what was learned in ECE 537 in the Fall of 2018.

1 Basic Acoustics

The propagation of sound through a medium can be modeled as a series of masses interconnected by springs. The stiffer the spring, the faster the "push-back", meaning the quicker the next mass moves in response to the spring action between itself and its neighbor. In other words, the stiffer the medium, or the less compressible it is, the faster sound will move through it. Contrariwise, the heavier the masses, the harder it is to get any inertia going, meaning for the same amount of push, a heavier mass gets going less. In other words, the heavier the medium, or the more dense it is, the slower sound will travel through it.

There's a connection between Hooke's Law and capacitors and a separate connection between Newton's Law and inductors. Hooke's Law, commonly given as $F = kx = k\int vdt$, can also be expressed as $\int vdt = \frac{1}{k}F$, which says that a spring will remember, or integrate the motion it experiences in response to a force received, accumulating a distance compressed or expanded. It will store this distance compressed or expanded as its state and give it back in the form of a new, proportional force when allowed to do so. Similarly, $Q = CV$ or $\int Idt = CV$, which says that in response to an applied voltage, a capacitor integrates a flow of electrons in time, accumulating a charge. It stores this charge as its state and gives it back in the form of a new, proportional voltage. The phrase "push me, push you" applies to springs and also to capacitors.

According to Newton's Law, $F = ma$ or $\int Fdt = mv$, a mass remembers, or integrates the forces that were applied to it in time and stores this remembrance as its state in the form of a momentum. If for whatever reason, the forces applied to it cease, then it will continue moving in the direction imparted to it. Similarly, according to the formula $V = L\frac{dI}{dt}$ or $\int Vdt = LI$, when an inductor experiences an applied voltage, due to a current rushing through its coils, the inductor integrates this voltage in time and stores this energy as its state inside a magnetic field. If the applied voltage, and therefore the applied current, ceases, then it will discharge the stored energy in the form of a new, proportional current in the same direction as the applied current it had been experiencing. The phrase "push me, pull you" applies to masses and also to inductors.

These analogies prove extremely useful when modeling acoustic situations. In general, a change in pressure is analogous to a voltage drop and a change in volume velocity to current. From the following two expressions, which begin the derivation of the Webster-Horn equation, we draw many useful insights.

$$\frac{d}{dx}P = M\frac{d}{dt}U \tag{1}$$

$$\frac{d}{dx}V = C\frac{d}{dt}P, \tag{2}$$

where M is the mass $\propto \frac{density}{area}$, C is the compliance $\propto \frac{length}{area \cdot density}$, P stands for pressure and U for volume velocity. Equation 1 is Newton's Law for acoustics. It says that a mass of air will remember a changing pressure that accumulates upon it in time, store it and give it back as a volume velocity. Equation 2 is Hooke's Law for acoustics. It says that a flow of volume is accumulated in time, stored and given back as a pressure drop.

2 Simulation of a Transmission Line

Why study transmission lines in a course about speech processing? Because the physical mechanisms for human speech production as well as the inner workings of the ear can be modeled as one. Additionally, simulations of transmission lines using discrete delay lines not only verify the theories of speech production and hearing but also push these theories further. So it's quite natural to begin a course on speech processing by studying a general-purpose transmission line.

In contrast to the familiar lumped elements, such as resistors, capacitors or inductors, where impedance, $Z(s)$, is only a function of frequency, s, a transmission line is a distributed element, meaning that its characteristic impedance, $z_0(x, s)$, is not only a function of frequency but also of position, x, along its length, L. In general, the characteristic impedance at x varies inversely with the area, $A(x)$. At first, however, we will assume that the area is constant for the full length, L, leading to a simpler, constant characteristic impedance. Later on, we can relax this assumption.

Speaking from a very high-level perspective, inside a transmission line, a signal bounces back and forth, between the ends. At a junction between one transmission line and another or between a transmission line and a termination impedance, some of the signal will bounce back, and the other part will pass through, depending on the load impedance, Z_L, at the junction. We can think of Z_L, or we can think about the percentage of the signal that will reflect, called the reflectance, $\Gamma \in [0, 1]$, and the corresponding remainder that will transmit, $\tau \in [0, 1]$. In other words, Z_L and Γ, both functions of frequency, s, form two sides of the same coin. And in combination with the characteristic impedance, z_0, these two sides of the same coin determine each other according to the following formulae.

$$Z_L = z_0\frac{1+\Gamma}{1-\Gamma} \tag{3}$$

$$\Gamma = \frac{Z_L - z_0}{Z_L + z_0} \tag{4}$$

where $z_0 = \rho c/A$, the subscript L stands for "load" and everything is a function of x and s.

All of this is leading to one thing: whenever we have a junction, we have a filter. And the transfer function for this junction filter is given by the reflectance, Γ, in Equation 4.

For our initial, constant-area transmission line, the parameters of interest are the characteristic impedance, z_0, and the termination impedances at $x = 0$ and $x = L$. At first, the "filters" at each termination will simply be scalars. Later on, we will allow the area to vary with x, thereby inducing an x-dependent $z_0 \propto 1/A(x)$, and we will include lumped elements having state, such as capacitors and/or inductors, in the termination impedances, thereby inducing an s-dependent Z_L.

For our first model, we will look at the vocal tract as a constant area tube. The glottis will be a sealed input at $x = 0$, having $Z_g = \infty$, and the mouth will be the output at $x = L$, having $Z_m = r_m = 2z_0$. The input will be an impulse and we will look at the volume velocities at either end of the tube. Figure 1 serves to illustrate a continuous schematic of this first model.

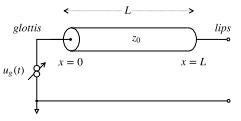


Figure 1: Continuous Model #1

Our job now is to map the continuous schematic in Figure 1 to something we can actually implement. As we said above, from a high-level perspective, a transmission line is a tube with signals bouncing back and forth between its ends. So, naturally, the best way to model a continuous transmission line using discrete tools is with two digital delay lines, one going forward and the other in reverse. At the input, a source and some backward-traveling, reflected amount will combine, then shuffle along with each sample clock toward the output, where some of it will pass beyond and some of it will turn around, shuffling its way back toward the input.

We'll take our sample rate to be $F_s = 44,100$ Hz, the tube will be $N = 10$ samples long and the speed of sound will be $c = 345$ m/s. As we said above, the impedance at $x = L$ is $Z_m = 2z_0$, meaning the reflectance at $x = L$ will be $\Gamma = \frac{2-1}{2+1} = \frac{1}{3}$. And since the impedance at $x = 0$ is infinite, the reflectance at $x = 0$ will be $100\% = 1$. Hence, as intended, the junction "filters", given by the reflectances, will be simple scalars at first. Figure 2 serves to illustrate the discrete model we will be implementing.

Intuitively, what will happen is the following. The input is $u_g = [1, 0, 0, 0, 0, 0, \ldots]$. The first 1 in $u_g[0]$ will combine with a value of 0 at sample clock 0 and then take $N = 10$ sample clocks to arrive at the output at $x = L$. At this point, $\Gamma = 1/3 = 33.33\%$ of the arriving 1 will reflect, and move its way back toward the input, arriving at $x = 0$ exactly 10 sample clocks later, and at the 20^{th} sample clock overall. At this point, the value of $1/3$ arriving at $x = 0$ will combine with $u_g[20] = 0$ and start heading toward $x = L$. And so it goes, on and on.

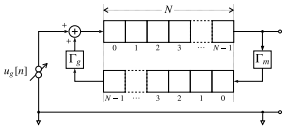


Figure 2: Discrete Model #1

In terms of the code, the language is Python and the delay lines are implemented as circular buffers. Everything else is straightforward. With all of the above in mind, here is a breakdown of the code, starting with the setup.

```
12  # sample rate
13  Fs = 44100
14  T = 1.0/Fs
15
16  # speed of sound
17  c = 345
18
19  # transmission line length in samples, then meters
20  N = 10
21  L = N*c*T
22
23  # length of one segment of transmission line in meters
24  dx = L/N
25
26  # impedances and reflection coefficients
27  z_0 = 50
28  z_R = 2*z_0
29  Gamma_L = 1
30  Gamma_R = (z_R - z_0)/(z_R + z_0)
31
32  # circular buffers for forward and backward traveling volume velocity waves
33  V_p = zeros(N)
34  V_n = zeros(N)
35  idx = 0
36
37  # length of the simulation
38  N_steps = 4*N
39
40  # driving function
41  drive = zeros(N_steps)
42  drive[0] = 1.0
43
44  # vectors for plotting
45  Vr_save = zeros(N_steps)
46  Vl_save = zeros(N_steps)
47
48
```

The setup is self-explanatory, but necessary to present in order for the rest to be readable. Let's take a look at the main processing loop.

```
49  for n in range(N_steps):
50
51      # grab the oldest sample that's about to exit the right side of V_p
52      out_p = V_p[idx]
53
54      # grab the oldest sample that's about to exit the left side of V_n
55      out_n = V_n[idx]
56
57      # put the new drive sample into the left side of the tube
58      V_p[idx] = drive[n] + Gamma_L*out_n
59
60      # put the new reflecting sample into the right side of the tube
61      V_n[idx] = Gamma_R*out_p
62
63      # right net V: sample leaving minus newest reflection
64      Vr_save[n] = out_p - Gamma_R*out_p
65
66      # left net V: newest input minus reflection leaving
67      Vl_save[n] = drive[n] - Gamma_L*out_n
68
```



```
69      # increment the circular buffer index
70      idx += 1
71      if idx == N:
72          idx = 0
```

The comments above each line in the main processing loop explain in plain English what's going on. Instead of moving the data within an array, a circular buffer moves the pointer, which is much more efficient in terms of optimizing CPU performance, although in this exact scenario, such streamlining hardly makes a difference.

At any point along a transmission line, the total pressure is the sum of the forward and backward going pressures, whereas the total volume velocity is the difference of the forward and backward going pressures. With this in mind, in Figure 3 we have the simulation volume velocity results.

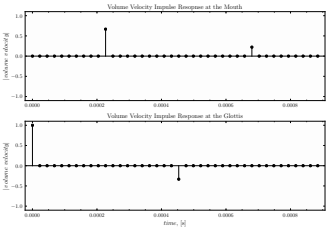


Figure 3: Volume Velocity Impulse Response Results for Simulation #1

Looking at Figure 3, we have exactly what we intuitively expected above. The non-zero start of the impulse, $u_g[0] = 1$, is immediately seen at the input at the 0^{th} sample clock. Then, nothing happens until sample clock 10 when $\Gamma = 33\%$ of $u_g[0]$ is seen at the output. At sample clock 20, the net volume velocity at $x = 0$ is 0 from the input signal minus the value of $1/3$, which has made its way all the way back to $x = 0$. And so it goes.

3 Simulation of the Middle Ear

For our next simulation, we will model the middle ear as a transmission line. Figure 4 is a continuous schematic of this second model.

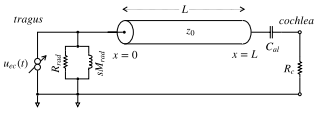


Figure 4: Continuous Model #2

Looking at Figure 4, because the load impedances at each end contain reactive elements, the reflectance filters, determined by the combination of Z_L and z_0, will no longer be simple scalars, but functions of complex frequency, s. And since our simulation is digital, we will use the bilinear transform to convert the continuous transfer functions to their z-domain equivalents.

In terms of the discrete schematic, it's exactly the same as in simulation #1, since the reflectance filters were already generalized. For completeness, however, we present the discrete schematic for simulation #2 in Figure 5.

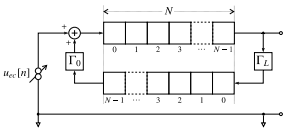


Figure 5: Discrete Model #2

Our goal is to build up progressively more complicated transmission line models, and so, to go through the reasoning for each lumped element in Figure 5 would be a distraction, especially since their explanations are described in detail in the homework. Therefore, let's move on to deriving the reflectance filters and converting from the s- to z-domain, using the bilinear transform.

Following the general form for a parallel impedance, the radiation filter at $x = 0$ is given by $Z_{rad} = \frac{sL/R}{2L+sL}$. For the reflectance filter at $x = 0$, Γ_0, we have

$$\begin{aligned}\Gamma_0 &= \frac{\frac{sL}{2L+sL} - z_0}{\frac{sL}{2L+sL} + z_0} \\ &= \frac{sLR - (sL + R)z_0}{sLR + (sL + R)z_0} \\ &= \frac{s(R - z_0)L - z_0R}{s(R + z_0)L + z_0R}\end{aligned} \tag{5}$$

And the series parallel impedance at $x = L$ leads to the filter transfer function $Z_c = \frac{1}{sC_{ul}} + R_t$. For the reflectance filter at $x = L$, Γ_L, we have

$$\begin{aligned}\Gamma_L &= \frac{\frac{1}{sC_{ul}} + R_t - z_0}{\frac{1}{sC_{ul}} + R_t + z_0} \\ &= \frac{1 + sC_{ul}(R_t - z_0)}{1 + sC_{ul}(R_t + z_0)} \\ &= \frac{s(R_t - z_0)C_{ul} + 1}{s(R_t + z_0)C_{ul} + 1}\end{aligned} \tag{6}$$

After plugging in the values for each lumped element derived in Homework #3, the magnitude responses, $|\Gamma(0, j\omega)|$ and $|\Gamma(L, j\omega)|$, are shown in Figure 6.

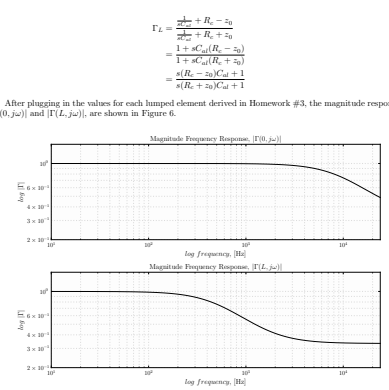


Figure 6: Reflectance Magnitude Frequency Responses at $x = 0$ and $x = L$ for Simulation #2

Our job now is to use the bilinear transform to obtain the corresponding IIR filter coefficients from the above s-domain filters. Generally speaking, the bilinear transform maps complex variables from planes to circles and back. In our case, we are mapping an s-domain function on the $(\sigma, j\omega)$ plain to a z-domain function inside, on, or outside the unit circle. The mapping from s to z is done by letting $s- > 2F_s\frac{z-1}{z+1}$. And the general form of the resulting z-domain IIR filter is given by $\Gamma = \frac{b_0+b_1z^{-1}}{a_0+a_1z^{-1}}$, where the b's are the feed-forward coefficients and the a's are the feedback coefficients. At $x = 0$, we have

$$\begin{aligned}\Gamma(0, s)\Big|_{s=2F_s\frac{z-1}{z+1}} &= \frac{2F_s\frac{z-1}{z+1}(R - z_0)L - z_0R}{2F_s\frac{z-1}{z+1}(R + z_0)L + z_0R} \\ &= \frac{2F_s(z-1)(R - z_0)L - (z+1)z_0R}{2F_s(z-1)(R + z_0)L + (z+1)z_0R} \\ &= \frac{\left[2F_s(R - z_0)L - z_0R\right]z - \left[2F_s(R - z_0)L + z_0R\right]}{\left[2F_s(R + z_0)L + z_0R\right]z - \left[2F_s(R + z_0)L - z_0R\right]} \\ &= \frac{\left[2F_s(R - z_0)L - z_0R\right] + \left[-2F_s(R - z_0)L - z_0R\right]z^{-1}}{\left[2F_s(R + z_0)L + z_0R\right] + \left[-2F_s(R + z_0)L + z_0R\right]z^{-1}} \\ &= \frac{b_{00} + b_{01}z^{-1}}{a_{00} + a_{01}z^{-1}}.\end{aligned}$$

And in code, we would divide each coefficient by a_{00} in order to get the transfer function in proper form. Running the numbers in Python, using the values derived in homework #3, and with a sample rate of $F_s = 69$ kHz, after normalizing, we arrive at

$$\Gamma(0, z^{-1}) = \frac{-0.54 - 0.08z^{-1}}{1 - 0.38z^{-1}}$$

Likewise, at $x = L$, we have

$$\begin{aligned}\Gamma(L, s)\Big|_{s=2F_s\frac{z-1}{z+1}} &= \frac{2F_s\frac{z-1}{z+1}(R - z_0)C + 1}{2F_s\frac{z-1}{z+1}(R + z_0)C + 1} \\ &= \frac{2F_s(z-1)(R - z_0)C + (z+1)}{2F_s(z-1)(R + z_0)C + (z+1)} \\ &= \frac{\left[2F_s(R - z_0)C + 1\right]z - \left[2F_s(R - z_0)C - 1\right]}{\left[2F_s(R + z_0)C + 1\right]z - \left[2F_s(R + z_0)C - 1\right]} \\ &= \frac{\left[2F_s(R - z_0)C - 1\right] + \left[-2F_s(R - z_0)C + 1\right]z^{-1}}{\left[2F_s(R + z_0)C + 1\right] + \left[-2F_s(R + z_0)C + 1\right]z^{-1}} \\ &= \frac{b_{L0} + b_{L1}z^{-1}}{a_{L0} + a_{L1}z^{-1}}.\end{aligned}$$

And in code, we would divide each coefficient by a_{L0} in order to get the transfer function in proper form. Running the numbers in Python, using a sample rate of $F_s = 69$ kHz, after normalizing, we arrive at

$$\Gamma(L, z^{-1}) = \frac{0.35 - 0.30z^{-1}}{1 - 0.95z^{-1}}$$

Probably the most important parameters of a filter are its poles and zeros. Rather than go through the gory details of figuring out actual numbers for the poles and zeros, especially because it was done in the homework already, perhaps its better to just explain the process, which is really rather straightforward. In either domain, s or z, the poles zeros are found by setting the numerator to zero and solving for s or z, respectively. And the zeros are found by setting the denominator to zero and solving for s and z, respectively. Because each of these filters are single pole, single zero systems, in the s-plane, both the zero and the pole will lie on the σ-axis, where $j\omega = 0$, and in the z-domain, both the zero and the pole will lie where $\omega = 0$. This is because when the coefficients of a filter are real, the poles and zeros must come in complex conjugate pairs, and since there's just one pole and one zero, the only complex conjugate pair that's possible is setting the imaginary parts equal to zero.

For the sake of brevity, especially since the details of deriving each lumped component's value was already done in the homework, it's worthwhile to compare the main processing loop for simulation #2 with #1. Therefore, the following code snippet is of the main processing loop handling the actual sample clockwork moving the machinery of the delay lines.

```
97  for n in range(N_steps):
98
99      # ---------- output and reflection filter at x = 0 ---------------------- #
100
101     # grab the oldest backward-going sample that's about to exit at x = 0
102     out_0 = vn_buffer[idx]
103
104     # run the filter for a new input to the forward-going wave
105     in_0 = Gamma_0[0]*out_0 + \
106             Gamma_0[1]*history_0[0] + \
107             Gamma_0[3]*history_0[1]
108
109     # move the filter history for next time
110     history_0[0] = out_0
111     history_0[1] = in_0
112
113     Gamma_0_ir_save[n] = in_0
114
115     # ---------- output and reflection filter at x = L ---------------------- #
116
117     # grab the oldest forward-going sample that's about to exit at x = L
118     out_L = vp_buffer[idx]
119
120     # run the filter for a new input to the backward-going wave
121     in_L = Gamma_L[0]*out_L + \
122             Gamma_L[1]*history_L[0] + \
123             Gamma_L[3]*history_L[1]
124
125     # move the filter history for next time
126     history_L[0] = out_L
127     history_L[1] = in_L
128
129     Gamma_L_ir_save[n] = in_L
130
131     # ------------ put the inputs in on both sides of the tube ---------------- #
132
133     # put the input into the tube at x = 0
134     vp_buffer[idx] = drive[n] + in_0
135
136     # put the input into the tube at x = L
137     vn_buffer[idx] = in_L
138
139     # net volume velocity at x = L: sample leaving minus newest reflection
140     v_net_L = out_L - vn_buffer[idx]
141     vL_save[n] = v_net_L
142
143     # net volume velocity at x = 0: newest input minus reflection leaving
144     v_net_0 = vp_buffer[idx] - out_0
145     v0_save[n] = v_net_0
146
147     # net pressure at x = L and x = 0 (not using these at the moment...)
```

```
148         p_net_L = out_L + vn_buffer[idx]
149         pL_save[n] = z_0 * p_net_L
150
151         p_net_0 = vp_buffer[idx] + out_0
152         p0_save[n] = z_0 * p_net_0
153
154         # ------------ radiation and cochlear filters ----------------------------- #
155
156         # radiation pressure requires a convolution, do the filter now
157         p_rad = z_rad[0]*v_net_0 + \
158                 z_rad[1]*history_rad[0] + \
159                 z_rad[3]*history_rad[1]
160
161         history_rad[0] = v_net_0
162         history_rad[1] = p_rad
163
164         prad_save[n] = p_rad
165
166         # cochlear pressure at the output
167         p_c = R_c * v_net_L
168         pc_save[n] = p_c
169
170         # ------------- increment the circular buffer index ------------------------ #
171
172         idx += 1
173         if idx == N:
174             idx = 0
```

The main thing to note with the above code, in comparison with the code for simulation #1, is that IIR filters are used for the reflectance instead of simple scalars. The impulse responses at $x = L$ and $x = 0$ are interesting to compare with the first simulation. In Figures 7 and 8, we have the impulse responses and their corresponding spectra, respectively.

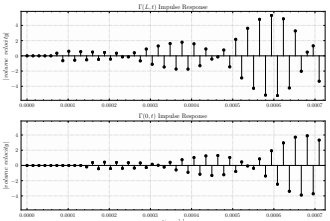


Figure 7: Reflectance Impulse Responses at $x = L$ and $x = 0$ for Simulation #2

Comparing Figures 3 and 7, we see that by making the reflectance filters frequency dependent, the impulse response begins to resemble a sinnsoidal waveform. And looking at the magnitude spectra in Figure 8, we see that for all intents and purposes, the spectra are flat. There is a peak around $f = 30$ kHz, but that's above human hearing so it's not too important.

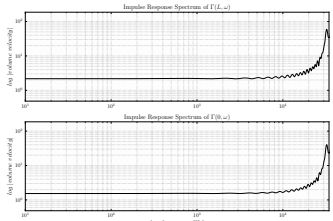


Figure 8: Spectra of Reflectance Impulse Responses at $x = L$ and $x = 0$ for Simulation #2

4 Vocal Tract Simulation

For the third and final simulation, we will vary the area of the tube in one discrete step along its length. In other words, there will be an L_1 having A_1 and an L_2 having A_2. While this isn't exactly a continuously varying area, it's the best first step toward modeling a more realistic physiological transmission line, such as the vocal tract or the middle ear.

The purpose of simulation #3 is to generate synthetic speech samples of four vowel utterances using a 2-tube model of the vocal tract. The two constant-area, cylindrical tubes will have lengths, L_1 and L_2 and areas A_1 and A_2, respectively. The model will be driven by the glottis, at $x = 0$, and will terminate at the lips in a radiation impedance at $x = L$. An illustration of the model is presented in Figure 9.

Between this new model and the previous two, the main difference is that there is now a junction of two constant-area transmission lines. This means that in going from tube 1 into tube 2, we will have a forward set of reflection and transmission coefficients. And in traveling from tube 2 back into tube 1, we will have a complementary set of reflection and transmission coefficients.

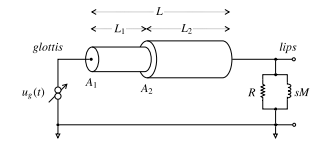


Figure 9: Continuous Model of Speech Synthesizer for Simulation #3

Rather than rederive the actual numbers used, it's more important to focus on the concepts. In terms of their closed form expressions, we can compute the various reflection and transmission coefficients using the formula for reflectance, $\Gamma = \frac{Z_L - z_0}{Z_L + z_0}$, where Z_L is the impedance seen as a load from the perspective of the traveling wave and z_0 is the characteristic impedance of the tube in which the wave is currently traveling. For the positive going wave, at the junction of the two tubes, the load impedance is $\rho c/A_1$ and the characteristic impedance is $\rho c/A_1$. The transmission coefficient is simply $\tau = 1 - \Gamma$.

$$\begin{aligned}\Gamma_p &= \frac{Z_L - z_0}{Z_L + z_0} \\ &= \frac{\frac{\rho c}{A_2} - \frac{\rho c}{A_1}}{\frac{\rho c}{A_2} + \frac{\rho c}{A_1}} \\ &= \frac{A_1 - A_2}{A_1 + A_2}\end{aligned}$$

$$\tau_p = 1 - \Gamma_p$$

And for the negative going wave, the situation is just reversed in terms of the areas and in terms of which tube is seen as the load by the traveling wave.

$$\begin{aligned}\Gamma_m &= \frac{Z_L - z_0}{Z_L + z_0} \\ &= \frac{\frac{\rho c}{A_1} - \frac{\rho c}{A_2}}{\frac{\rho c}{A_1} + \frac{\rho c}{A_2}} \\ &= \frac{A_2 - A_1}{A_2 + A_1}\end{aligned}$$

$$\tau_m = 1 - \Gamma_m$$

Again, for the sake of brevity, we won't go into the details of determining each of the reflection and transmission coefficients, especially since they were determined in the homework already. It's more important to focus on the higher-level idea of building up simulations using progressively more complicated, and therefore more realistic delay lines and filters as models for transmission lines and loads.

The discrete schematic of the continuous model in Figure 9 is presented in Figure 10.

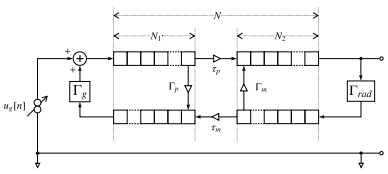


Figure 10: Discrete Model of Speech Synthesizer for Simulation #3

Looking at Figure 10, each of the two tubes is decomposed into a forward- and backward-going buffer, represented as a typical C-style array. At the junction of the two tubes, there are four coefficients to consider: the reflection and transmission coefficients of the forward traveling wave, Γ_p and τ_p, and the reflection and transmission coefficients of the backward traveling wave, Γ_m and τ_m, respectively.

At the glottis, where $x = 0$, the reflection coefficient, which, as derived in the homework, is simply a scalar number equal to $\Gamma_g = 0.95$, multiplies the output of the leftmost backward-traveling buffer, which is then summed with the next input sample and put into the leftmost forward-traveling buffer. At $x = L$, on the other hand, the radiation reflection, Γ_{rad}, is a single pole, single zero bilinear filter, the coefficients for which are determined by finding the s-domain reflectance and using the bilinear transform to map the transfer function from the s- to z-domain, as we did with simulation #2.

All things considered, the only real difference between the block diagram for this model and the previous one is that here we have a junction of two constant-area tubes, which leads to four variables to store each iteration of the main processing loop. In terms of the code, it's really just as simple as multiplying the samples entering and leaving the internal delay lines by the appropriate scalar reflection/transmission coefficients.

5 Linear Predictive Coding

According to the source-filter model for speech production, there is 1) a glottal signal, which is either an impulse train corresponding to voiced speech or noise corresponding to unvoiced speech, and 2) a vocal tract filter, which smooths and shapes the glottal signal into the speech waveform that a listener can understand. The glottal signal, being either an impulse train or noise, has a flat spectrum and the vocal tract filter has distinct resonances, which is modeled as an all-pole filter. For our purposes, let's call the glottal signal $e[n]$, the vocal tract filter coefficients, $h[n]$, and the resulting speech, $x[n]$, which is the only signal among the three that we can actually measure. In other words, we have $x[n] = h[n] * e[n]$, and both $e[n]$ and $h[n]$ are unknown. And in the frequency domain, by the convolution theorem, this relationship transforms to $X(z) = H(z)E(z)$.

The core idea behind linear predictive coding, as applied to speech, is to extract the key parameters associated with the source-filter model of speech, i.e., $e[n]$ and $h[n]$. For our pusposes, we will only focus on

discovering the filter coefficients in $h[n]$ since pitch detection combined with voiced/unvoiced determination would require a much longer and perhaps less relevant discussion.

For the sake of a tractable model and also because it works quite well, we assume the vocal tract to be an all-pole filter. In terms of setting up the math, although not required, for the sake of completeness, it's generally known that 3 peak frequencies in the vocal tract, or formants, are enough to discriminate most vowels. Thus setting the number of poles to $p = 6$ will work well for a speech signal. In terms of the memory structure of an all-pole filter, there are only feedback coefficients in $h[n]$. The general form for $x[n]$ will therefore be

$$x_n = e_n + \sum_{k=1}^{p} a_k x_{n-k}, \tag{7}$$

where the indices have been notated using subscripts for easier legibility, and where $a_k, k \in [1, p]$ are the p filter coefficients in h. In Figure 11, we have a schematic block diagram of Equation 5, where the delay elements are contained within a typical C-style array.

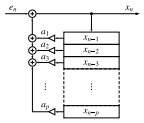


Figure 11: Schematic Block Diagram for an All-pole LPC Filter

Rearranging Equation 5 and solving for the glottal signal, e_n, we have

$$e_n = x_n - \sum_{k=1}^{p} a_k x_{n-k}. \tag{8}$$

As a vector of values, e_n is either a train of impulses separated by many zeros compared to its length, or a random noise signal. In either case, it is assumed that we have already done the proper pitch/voiced/unvoiced detection so that e_n is known. In the voiced case, e_n will indeed be a train of impulses, where the number zeros separating each impulse is the pitch period in samples. In general, we don't want to include more than about two pitch periods in a chunk of samples. Otherwise, we risk more than one pitch period detected, which goes against the models intended functionality. Therefore, in the case of a voiced pitch period, compared to the vocal tract filtered signal, x_n, the impulse train e_n is essentially a zero vector. In other words, Equation 5 is basically zero if an actual pitch is detected in the glottis.

With this in mind, we seek to minimize e_n, often called the residual since it's the difference of the actual filtered sample, x_n, and the estimated filtered sample, $\sum_{k=1}^{p} a_k x_{n-k}$. Our goal is to find the a_k filter coefficients. As stated above, it's assumed that we have already performed pitch/voiced/unvoiced detection, so we know e_n.

In general, there are two ways of approaching the problem of minimizing e_n. One way seeks to minimize the mean-squared error, $\|e_n\|^2$. And the second way is via least squares, leading to a "best fit" solution. Let's take a look at each approach and see how they compare. In both cases, it's assumed that we have

$N \gg p$ samples, which for most practical sample rates, is easily the case. Figure ?? provides a visualization of the situation.

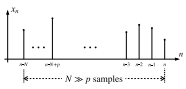


Figure 12: Block of $N \gg p$ Samples for LPC Processing

5.1 Minimizing the MSE

The mean-squared error, or MSE, of e_n is a function of the filter coefficients, a_k, and is defined as $\mathbb{E}\left[\|e_n\|^2\right] = \mathbb{E}\left[\left(x_n - \sum_{k=1}^{p} a_k x_{n-k}\right)^2\right]$. To minimize the MSE in terms of its p-dimensional argument, $a_k,\ k \in [1, p]$, we take the partial derivative of the MSE with respect to each filter coefficient, a_k, and set this partial derivative to zero as follows.

$$\begin{aligned}\frac{\partial}{\partial a_l} MSE &= \frac{\partial}{\partial a_l} \mathbb{E}\left[e_n^2\right] \\ &= \frac{\partial}{\partial a_l} \mathbb{E}\left[\left(x_n - \sum_{k=1}^{p} a_k x_{n-k}\right)^2\right] \\ &= \mathbb{E}\left[\frac{\partial}{\partial a_l}\left(x_n - \sum_{k=1}^{p} a_k x_{n-k}\right)^2\right] \\ &= \mathbb{E}\left[2\left(x_n - \sum_{k=1}^{p} a_k x_{n-k}\right)\frac{-\partial}{\partial a_l}\sum_{k=1}^{p} a_k x_{n-k}\right] \\ &= -2\mathbb{E}\left[\left(x_n - \sum_{k=1}^{p} a_k x_{n-k}\right)\frac{\partial}{\partial a_l}\sum_{k=1}^{p} a_k x_{n-k}\right] \\ &= -2\mathbb{E}\left[\left(x_n - \sum_{k=1}^{p} a_k x_{n-k}\right)x_{n-l}\right] \\ &= -2\mathbb{E}\left[x_n x_{n-l} - \sum_{k=1}^{p} a_k x_{n-k} x_{n-l}\right] \\ &= -2\left(\mathbb{E}\left[x_n x_{n-l}\right] - \mathbb{E}\left[\sum_{k=1}^{p} a_k x_{n-k} x_{n-l}\right]\right) \\ &= -2\Big(r_x[l] - a_k r_x[\ell - k]\Big)\end{aligned} \tag{9}$$

where $r_x[\cdot]$ is the autocorellation of our N-long chunk of the speech signal x_n, evaluated at sample lag $\cdot$, and both $\ell, k \in [1, p]$. To minimize the quantity in Equation 9, we set it equal to zero to obtain the following.

$$\begin{aligned}r_x[\ell] &= a_k r_x[\ell - k],\ \ell, k \in [1, p] \\ \begin{bmatrix} r_x[0] \\ r_x[1] \\ \vdots \\ r_x[p] \end{bmatrix} &= \begin{bmatrix} r_x[0] & r_x[1] & r_x[2] & \dots & r_x[p] \\ r_x[1] & r_x[0] & r_x[1] & \dots & r_x[p-1] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_x[p] & r_x[p-1] & r_x[p-2] & \dots & r_x[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \\ \vec{r}_x &= \mathbf{R}\vec{a}\end{aligned} \tag{10}$$

where we have made use of the symmetry of the autocorrelation by making positive any negative indices in $\mathbf{R}$.

Finally, to obtain the filter coefficients, we use the Levinson-Durbin algorithm to invert $\mathbf{R}$ to obtain $\vec{a} = \mathbf{R}^{-1}\vec{r}_x$. In summary, the $MMSE$ approach looks at the problem in terms of the autocorrellation of an N-long frame of the input signal. To actually implement this technique, an overlap-add algorithm would be used to handle the overlapping frames, but that's not the focus of the explanation of LPC here.

5.2 The Least-squares Approach

Getting to Equation 10 was relatively straightforward. Using the Levinson-Durvbin algorithm to invert $\mathbf{R}$, however, is not very intuitive. In comparison to the least-squares method, it's more efficient and so if optimization is a concern, the Levinson-Durbin algorithm should be used. If optimization is not a concern, however, the least-squares approach is much easier to follow.

If we have $N \gg p$ samples, which is easily done, then we have $N - p$ equations with only p unknowns. In other words, the system is overdetermined and we can find an optimal, "best-fit" solution for the vocal tract filter coefficients. If our DSP while loop is at sample location n, and we have N past samples at our disposal, then our system of equations is the following set of many Equation 5's.

$$\begin{aligned} x_{n-N+p} &= e_{n-N+p} + \begin{bmatrix} x_{n-N+p-1} & x_{n-N+p-2} & \dots & x_{n-N+p-p} \end{bmatrix} \vec{a} \\ &\vdots \\ x_{n-2} &= e_{n-2} + \begin{bmatrix} x_{n-3} & x_{n-4} & \dots & x_{n-p-2} \end{bmatrix} \vec{a} \\ x_{n-1} &= e_{n-1} + \begin{bmatrix} x_{n-2} & x_{n-3} & \dots & x_{n-p-1} \end{bmatrix} \vec{a} \\ x_n &= e_n + \begin{bmatrix} x_{n-1} & x_{n-2} & \dots & x_{n-p} \end{bmatrix} \vec{a} \end{aligned} \tag{11}$$

where $\vec{a} = \begin{bmatrix} a_1 & a_2 & \dots & a_p \end{bmatrix}^T$ is the filter we are solving for. Rewriting Equation 11 in matrix form, we have

$$\begin{bmatrix} x_{n-N+p} \\ \vdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} e_{n-N+p} \\ \vdots \\ e_{n-2} \\ e_{n-1} \\ e_n \end{bmatrix} + \begin{bmatrix} x_{n-N+p-1} & x_{n-N+p-2} & \dots & x_{n-N+p-p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n-3} & x_{n-4} & \dots & x_{n-p-2} \\ x_{n-2} & x_{n-3} & \dots & x_{n-p-1} \\ x_{n-1} & x_{n-2} & \dots & x_{n-p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

Putting this matrix equation into concise vector notation, we have

$$\vec{x} = \vec{e} + \mathbf{X}\vec{a} \tag{12}$$

Rewriting Equation 5.2 for the residual, *vece*, which is the *extremely* sparse, pitched, voiced, impulse train, we have

$$\vec{e} = \vec{x} - \mathbf{X}\vec{a} \tag{13}$$

Now this is where we take a hard turn toward linear algebra. The vector subtraction in Equation 5.2 can be interpreted as follows. If our modeled filter has perfectly matched the actual speech signal, then $\vec{x}$ is 100% equal to the glottal signal, $\vec{e}$ plus the modeled filter convolution of the past speech with $\vec{a}$. In this case, 5.2 will be exactly $\vec{0}$.

In fact, we already expected $\vec{e}$ to be approximately zero for most samples, since we are dealing with about two pitch periods, meaning two non-zero samples with many, many zero samples in between. Compared to the filtered output, which is rich in feedback, $\vec{e}$ is intuitively, very near $\vec{0}$.

But in reality, we will never have a perfect match and $\vec{e}$ is only approximately the zero vector. So, we'll simply ignore any component of $\vec{e}$ that lies in the column space of $\mathbf{X}$ and treat this more convenient residual as our vector, $\vec{e}$. So, by construction, $\vec{e}$ is orthogonal to every column in $\mathbf{X}$. In other words, $\mathbf{X}^T\vec{e} \approx \vec{0}$ in the following series of equations.

$$\begin{aligned} \vec{e} &= \vec{x} - \mathbf{X}\vec{a} \\ \mathbf{X}^T\vec{e} &= \mathbf{X}^T\vec{x} - \mathbf{X}^T\mathbf{X}\vec{a} \approx \vec{0} \\ \Rightarrow \mathbf{X}^T\vec{x} &= \mathbf{X}^T\mathbf{X}\vec{a} \\ \Rightarrow \vec{a} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{a} \end{aligned} \tag{14}$$

where $\mathbf{X}^T\mathbf{X}$ is guaranteed to be invertible since no two columns of the overlapping portions of our real-world speech signal, $\vec{x}$ will ever be the same, nor will they be linear combinations of the others. And there we have it. If optimization is not a concern, we will simply perform three matrix multiplications, one matrix inversion and one matrix-vector multiplication to obtain the filter coefficients, $\vec{a}$.

5.3 LPC Summary

The main idea behind LPC for speech processing is to take a speech signal, chop it up into frames, which contain samples in the time-domain that are extremely redundant, discover the source-filter parameters and transmit those instead of the actual samples. As an example, $x_3$3 of some speech signal is probably only different from $x_3$2 in the fourth decimal place for a decent sample rate and a floating-point converter. So instead of transmitting all of these highly redundant samples, we transmit the filter coefficients, a_k, the pitch period if the source is voiced or the variance of the noise if the source is unvoiced, placing the burden of reconstruction on the receiver's end. Overall, it's very efficient for speech signals.

6 Short-Time Fourier Transform, Overlap-Add & Speech Coding

Speech is burst-y, in the sense that it comes in short packets of either wide-band transients called consonants, or longer, harmonic segments called vowels. The FFT of a speech signal lasting the duration of many bursts, so to speak, or phones, is of little use. In taking such a transform, while we may discover all of the frequencies that occurred throughout all of the words contained in the signal, we will know nothing about *where* those frequencies occurred in time and *how they transitioned* in time.

The solutions is to take many short transforms, on the order of milliseconds, throughout the duration of the longer skignal, which could be on the order of seconds or even much longer. In this way, we generate frames of frequency-domain information, good for a short duration of time, that likely only contains one consonant or one vowel and not much more. This method of short transforms covering a longer duration signal is known as the Short-Time Fourier Transform or STFT for short.

Conceptually, speaking this is nice, but at as they often say, the devil is in the details. Actually implementing the STFT effectively, especially if we aim to do any frequency-domain modifications, is not as straightforward.

As we said above, speech is burst-y, so we need to make sure we don't include too much time-domain information in our FFT frames, so as to avoid including more than one phone per FFT. On the other hand, if we make it too short, we won't have enough frequency-domain samples to work with. A good compromise for sample rates of 44,100 Hz or 48,000 Hz is to use 256 samples of time-domain information and then zero-pad the 256 samples to 512 for a longer FFT with slightly higher resolution. At a sample rate of 44,100 Hz, 256 samples amounts to about 6 ms of signal duration, which likely only includes one vowel or consonant. More generally speaking, a good time duration for the window is around 10 ms, leading to a window length in samples of L samples $= T$ seconds $\times F_s \frac{\text{samples}}{\text{second}}$.

6.1 Boxcar Bandwidth

Which window to use on the time-domain frame is an important question. By the convolution theorem, multiplication in time transforms to convolution in frequency. The transform of the boxcar window is

$$\begin{aligned} W_{\text{boxcar}}[k] &= \sum_{n=0}^{L-1} 1 e^{-j\omega_k n} \\ &= \frac{1 - e^{-j\omega_k L}}{1 - e^{-j\omega_k}} \end{aligned} \tag{15}$$

Setting the numerator in Equation 15 to zero will yield the zeros of the boxcar window's transform. We have

$$\begin{aligned} 1 &= e^{-j\omega_k L} \\ e^{-j2\pi k} &= e^{-j\omega_k L} \\ 2\pi k &= \omega_k L \\ \omega_k &= \frac{2\pi}{L} k,\ k \in [0, L-1]. \end{aligned} \tag{16}$$

In other words, the transform of the boxcar will have zeros at every sample location of the L-long FFT.

Setting the denominator in Equation 15 to zero will yield the poles of the boxcar window's transform. We have

$$\begin{aligned} 1 &= e^{-j\omega_k} \\ e^{-j2\pi k} &= e^{-j\omega_k} \\ 2\pi k &= \omega_k \\ \omega_k &= 2\pi k. \end{aligned} \tag{17}$$

The transform of the boxcar has a single pole at DC, which will cancel the zero at DC. Therefore, the first zero of W_{boxcar} is at the first bin location, where $k = 1$. And so, by taking into consideration the negative frequencies, the full bandwidth is 2 bins wide. In terms of analog frequencies, the bandwidth, B, for the boxcar is $2 \times \frac{F_s}{L}$. The following magnitude FFT illustrates the case where $L = 64$.

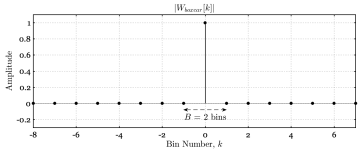


Figure 13: Magnitude FFT of the Boxcar Against Bin Number

What this means is that if we window our time-domain input with the boxcar function, we will convolve its transform with the ideal delta spikes lying at the detected frequencies. During the convolution, if any of the delta spikes are closer than 2 bins apart, they'll be smeared together.

6.2 Hamming Bandwidth

By performing the same analysis on a Hamming window, we can begin to see the trade-offs involved in improving the window's bandwidth. The Hamming window is given by $w[n] = 0.54 - 0.46\cos(\frac{2\pi}{L}n), n \in [0, L-1]$. The FFT of the Hamming window is

$$\begin{aligned}
W_{Hamming}[k] &= \sum_{n=0}^{L-1}(0.54 - 0.46\cos(\omega_1 n))\, e^{-j\omega_k n} \\
&= \sum_{n=0}^{L-1}\left(0.54 - \frac{0.46}{2}\left(e^{j\omega_1 n} + e^{-j\omega_1 n}\right)\right) e^{-j\omega_k n} \\
&= 0.54\sum_{n=0}^{L-1} e^{-j\omega_k n} - \frac{0.46}{2}\sum_{n=0}^{L-1}(e^{j\omega_1 n} + e^{-j\omega_1 n})e^{-j\omega_k n} \\
&= 0.54\sum_{n=0}^{L-1} e^{-j\omega_k n} - \frac{0.46}{2}\sum_{n=0}^{L-1}(e^{-j(\omega_k-\omega_1)n} + e^{-j(\omega_k+\omega_1)n}) \\
&= 0.54\frac{1-e^{-j\omega_k L}}{1-e^{-j\omega_k}} - \frac{0.46}{2}\left(\frac{1-e^{-j(\omega_k-\omega_1)L}}{1-e^{-j(\omega_k-\omega_1)}} + \frac{1-e^{-j(\omega_k+\omega_1)L}}{1-e^{-j(\omega_k+\omega_1)}}\right)
\end{aligned} \tag{18}$$

Ignoring the 0.54 and $\frac{0.46}{2}$ in Equation 18, in the first tem, we have the same pole/zero situation at DC as we had with the Boxcar, and in the second term, we have additional pole/zero combinations at the $\pm k$-th bins. These two new terms have zeros at every sample point of our FFT since the L in $\omega_k \pm \omega_1 = \frac{2\pi}{L}(\omega_k \pm \omega_1)$ will cancel with the L in the numerator, and $e^{-j(\omega_k-\omega_1)} = 1,\ \forall k$. And these two new terms have single poles at $\omega_k = \pm\omega_1$, the first positive and negative bin locations. The first zero for the Hamming is therefore at the second bin above and below DC, where $k = \pm 2$. In other words, the total bandwidth for the Hamming window is $B = 4\frac{F_s}{L}$, which is twice as wide as that of the boxcar.

The following magnitude FFT illustrates the case where $L = 64$.

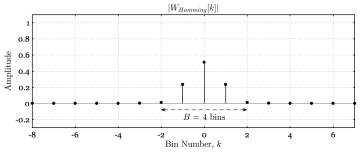


Figure 14: Magnitude FFT of the Hamming Against Bin Number

When we convolve our time-domain input's transform with the Hamming window's transform, we smear a 4 bin wide "hump", or main lobe, across the ideal delta spikes lying at the detected frequencies in our signal. Because the Hamming's main lobe is 2 bins wider than that of the boxcar, we need 2 times more sample points in our FFT in order to resolve the same frequency content. As long as all detected delta spikes are 4 or more bins apart, the Hamming window will work the same as the boxcar.

6.3 Time-Bandwidth Product of a Window

Naturally, one might ask, why choose a window that forces a higher sample rate in the frequency domain? In other words, why would we use the Hamming window instead of the boxcar when the Hamming requires twice as many FFT samples for the same frequency resolution as the boxcar? The answer has to do with the edges of the window. The boxcar causes a false detection of high frequencies which show up only because of the abrupt drop from 1 to 0 at the time-domain edges. So the boxcar causes aliasing. And to improve upon this, the Hamming smoothly approaches $0.54 - 0.46 = 0.08$ at its edges, which obviously isn't zero, but it's certainly much better than the ends of the boxcar. This smoother transition yields virtually no aliasing in the frequency-domain.

One might then ask, can we do better than the Hamming? And the answer is yes, with a Kaiser window. How could we know that the Kaiser performs better than the Hamming without actually implementing two versions of the STFT? We would like some measure of a window's performance.

This single number is known as the time-bandwidth product of a window,γ, equal to the time-duration of a window, T, multiplied by its bandwidth, B. In terms of the boxcar, $\gamma_{boxcar} = \frac{L}{F_s} \times 2\frac{F_s}{L} = 2$. And as for the Hamming window, $\gamma_{Hamming} = \frac{L}{F_s} \times 4\frac{F_s}{L} = 4$.

6.4 Overlap-Add Error Function

In order to achieve the best possible reconstruction during the synthesis portion of the STFT process, we would like to ensure that the DTFT of the entire signal, $X(\omega)$, is exactly or at least approximately equal to the sum of the shorter DTFT's, $X_r(\omega)$. Speaking more generally now, we'll let the signal, $x[n]$, and the window, $w[n]$, run for an infinite amount of sample clocks, n. We will also say that the hop size between frames is R. In this case, we have

$$\begin{aligned}
\sum_r X_r(\omega) &= \sum_r\sum_n x[n]w[n-rR]e^{-j\omega n} \\
&= \sum_n x[n]e^{-j\omega n}\sum_r w[n-rR] \\
&= \alpha\sum_n x[n]e^{-j\omega n} \\
&= \alpha X(\omega),
\end{aligned} \tag{19}$$

where we are seeking α to be a constant, ideally unity, and at least approximately equal to a constant if not exactly. In the case where $\alpha = 1$, then indeed the sum of shorter DTFT's will equal the DTFT of the entire signal, and as long as we take proper care during any modifications in the frequency-domain, we can achieve the best possible synthesis during the STFT process.

When α approximately equals any constant, then we say the window, $w[n]$ has the Constant Overlap-Add (COLA) property at hop-size R. In other words, if $\alpha = \sum_r w[n-rR] = \text{const} \Rightarrow w \in COLA(R)$. In the more general case, it can shown that,

$$\sum_r w[n-rR] \approx \frac{W(0)}{R},$$

where $W(0) = \sum_n w[n]$ is the DC value of the window's transform and R is the hop size, or frame rate. Consequently, we can define an error function $e[n] = \left(\frac{R}{W(0)}\sum_n w[n-rR]\right) - 1$.

7 Information Theory, Fake Shakespeare & the EM Algorithm

7.1 Key Contributors to the Theory of Information

Information theory is a probabilistic study of the quantification, storage and communication of information, which, incidentally, has nothing to do with meaning. Information is typically encoded in numbers of bits and a key measure in the theory is entropy, which quantifies the amount of uncertainty contained in the information.

The most famous contributor to the theory of Information is Claude Shannon, who is often, and arguably erroneously, credited with single-handedly creating the theory from scratch in a series of groundbreaking papers published in the late 1940s and early 1950s. Some of the lesser credited contributors, however, leading up to Shannons work, are Samuel F.B. Morse (of telegraphy fame), Harvey Fletcher and Ralph Hartley.

7.1.1 Samuel F.B. Morse and the Theory of Information

Morse's key insight was to realize that in order to make an efficient code for transmitting English with only dots, dashes and spaces, one needs to exploit the frequencies of letter occurrences. The fact is some letters happen more than others, and so their representation should be simpler, which is how Morse designed his code. Shannon took this idea and ran miles with it.

7.1.2 Havery Fletcher and the Theory of Information

Shannon's model for how information is communicated between transmitter and receiver is very similar to Fletcher's idea. Fletcher was interested in hearing and came up with a model for how speech is converted from energy in the air to phones in the brain. It's important to note that Fletcher did not consider the intelligibility of sounds, or what sort of meaning their concatenation carried, but only the identification of them, what he called articulation.

The model works as follows. Inside the ear, the cochlea acts as a filter bank on incoming acoustic signals. The filterbank contains twenty bandpass filters with overlapping bandwidths and center frequencies, ω_k. The output of each filter is the error in the k-th band, ϵ_k. From each of these errors, a probability, s, of correctly detecting a phone is produced. Fletcher's key insight was how to connect s to the 20 ϵ_k's.

The formula for ϵ_k is given by

$$\epsilon_k = c_k^{SNR_k}, \tag{20}$$

where SNR is the signal-to-noise ratio for the speech filtered by the k-th bandpass filter and $c_k \in (0, 1)$ is some constant.

The formula for s is given by

$$\begin{aligned}
s &= 1 - \epsilon_1\epsilon_2\epsilon_3\ldots\epsilon_{20} \\
&= 1 - c_1^{SNR_1}c_2^{SNR_2}c_3^{SNR_3}\ldots c_{20}^{SNR_{20}} \\
&= 1 - c^{SNR_1+SNR_2+SNR_3\ldots SNR_{20}} \\
&= 1 - c^{\sum_{k=1}^{20} SNR_k},
\end{aligned} \tag{21}$$

[1] where $c \in (0, 1)$ is some constant which encapsulates all of the c_k's at once.

If any of bandpass filters in the cochlea detect non-negligible energy, the SNR for that filter will be significantly greater than 1. Therefore, the output for that filter will be $\epsilon_k = c_k^a \approx 0$ since $0 < c_k < 1$ and $\alpha \gg 1$. And because the ϵ_k's multiply each other, their product will go to zero, no matter what any of the other filter outputs happen to be. Looking at Equation 21, this will cause $s \to 1$, indicating a high probability of the detection of a phone by the brain.

If, on the other hand, none of the bandpass filters in the cochlea detect significant energy, their SNR's will be less than 1. Therefore, the output for all of the filters will be $\epsilon_k = c_k^\beta \approx 1$ since $\beta \approx 0$. Now the product in Equation 21 will be twenty numbers, all approximately equal to 1, which is also approximately equal to 1, causing $s \to 1 - 1 = 0$, indicating a low probability of the detection of a phone by the brain.

7.1.3 Claude Shannon and the Theory of Information

Like Fletcher, Shannon is not concerned with meaning, and he makes this very clear in the first paragraph of his paper. What he cares about are the sets from which all pieces of communication are drawn. For example, sentences are composed of words. These words are drawn from the set of English words and each of them has an empirical frequency of occurrence or probability. It doesn't matter that a string of these words has any meaning, but only that the words are drawn from a known set, having a known distribution.

Let's say we have a message, $\{x_1, x_2, x_3, \ldots, x_N\}$. Each of the x_i's is a word drawn from the set of English words and each fo them has probability $\{p_1, p_2, p_3, \ldots, p_N\}$. Empirically speaking, when we are using actual data, each of these words will have probability, $\bar{p} = \frac{N_k}{\sum_k N_k}$, where N_k is the number of times the k-th word has appeared in the dataset. In this way, we have a sensible probability distribution since all the $\bar{p}$'s sum to 1.

Shannon defines the information density as $\mathcal{I} = \frac{1}{p_k}$. In units of bits, the information density is given by $\mathcal{I}_{bits} = \log_2\frac{1}{p_k} = -\log_2 p_k$. Finally, the entropy of this information is given by the average information density (usually in bits), expressed as

$$\begin{aligned}
\mathcal{H} &= \mathbb{E}[\mathcal{I}_{bits}] \\
&= \sum_k p_k \log_2\frac{1}{p_k}
\end{aligned} \tag{22}$$

In terms of the set of all messages, if the k-th message has a low probability, p_k, then the information density, $\mathcal{I}$, is high. As an example, the chances of winning the lottery are pretty slim, and so if it happens, there's a lot of information contained in that event.

If all of the p_k's are equal, then the entropy is maximum. For example, if a novel is written in such a way that all of its words occur an equal amount of times, then the novel has maximum entropy, or maximum unpredictability. In other words, it probably doesn't make any sense.

In this way, Shannon has effectively combined Morse and Fletcher into one cohesive theory of Information. From Morse he takes the idea that some pieces of the language happen more often than others, and in order to maximize the channel capacity, communication should exploit this natural characteristic of language and information. And like Fletcher, he chooses to completely ignore the problem of meaning and its understanding.

7.2 A Program for Generating Fake Shakespeare

In Shannon's first paper on Information Theory, he outlines how to generate successive pairs of symbols by giving the "digram" probabilities $p(i, j)$,i.e., the relative frequency of the digram, or in this case word pair, i j.

7.2.1 Discovering the Distribution

From Shakespeare's "As You Like It", all of the "regular" words, as in not including proper nouns, or theatrical directives, etc., were put into a list in the order in which they originally appeared. From this list, a second list was created containing the word pairs appearing in "As You Like It." For example, if the first list of words as they originally appeared contained "The cow jumped over the moon," then the word pairs in the second list would be "The cow," "cow jumped," "jumped over," "over the," and "the moon." From this second list of word pairs as they originally appear, a third list was created with only the unique word pairs. Finally, each of the unique pairs was searched in the original list of word pairs for a number of occurrences. From this number of occurrences, a probability of occurrence was generated by simply dividing the number by the total number of word pairs in the original text.

In mathematical terms, the procedure just described begins with a set of N word pairs, $\{x_1, x_2, x_3, \ldots, x_N\}$. From this set, another set of occurrences for each word pair, x_i, is computed, $\{N_1, N_2, \ldots, N_N\}$. Finally the probability of each word pair is given by

$$p(x_i) = \frac{N_i}{N} \tag{23}$$

7.2.2 Using the Distribution

Once we have a valid probability distribution of word pairs, PDF $= p(x_i)$ (all the values sum to 1), we integrate it to form the discrete cumulative distribution function, or CDF $= P(x_i) = \sum_i^n x_i$. Now, in order to generate fake Shakespeare, we need to be able to "draw" samples from this empirically determined distribution. Because by construction, the CDF has y-axis values from 0 to 1, we start by generating a random number between 0 and 1, y_i. We then find the x_i such that $P(x_i) = y_i$. This x-axis location on the CDF is the word pair we are after. At this point, we have "drawn" from our empirically determined distribution of word pairs and we simply append the drawing to our running generation of fake Shakespeare. Graphically, the situation is as follows.

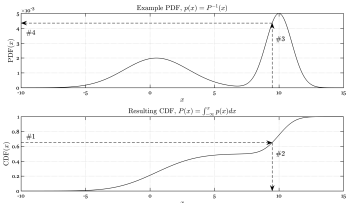


Figure 15: Graphical Explanation of Drawing from an Empirical Distribution

Looking at Figure 15, and with the above description of the process, in order to draw sample word pairs from our empirically discovered distribution, we

#1 Generate a random number, $y_i \in [0, 1]$, along the y-axis of the CDF.

#2 Find the corresponding x_i on the x-axis of the CDF.

#3 The desired word pair will be the same x_i on the PDF.

#4 If we need or want it, the probability of this word pair is the corresponding y-axis value of the PDF, $p(x_i)$.

7.2.3 Matlab Implementation

The following Matlab code was written to discover the distribution of word pairs.

```
% digrams will be A row1*row2, row2*row3, row3*row4, etc...
A_digrams = {};
for idx = 2:n_words-1
    A_digrams{end+1} = strcat(A(idx),{' '},A(idx+1));
end

% sorted digrams in 'as you like it'
A_digrams = A_digrams';
A_digrams_sorted = sort(A_digrams);
n_digrams = length(A_digrams_sorted);

% unique digrams in 'as you like it'
A_digrams_unique = unique(A_digrams, 'stable');
n_unique_digrams = length(A_digrams_unique);

freqs_digrams = zeros(n_unique_digrams,1);
for i=1:n_unique_digrams
```



```
    digram = A_digrams_unique(i);
    freqs_digrams(i) = sum(ismember(A_digrams_sorted,digram)) / n_digrams;
end

% sort the frequencies of words in descending order and grab original idx's
[freqs_digrams_sorted,idxs_digrams] = sort(freqs_digrams,'descend');

% reorder A_digrams_unique according to descending freq idxs
A_digrams_ranked = A_digrams_unique(idxs_digrams);                % 2a
```

The top 20 most likely word pairs are simply the first 20 entries of the vector called `A_digrams_ranked` above. They are the following.

Rank	Word	Rank	Word
1	in the	11	is not
2	it is	12	to the
3	of the	13	out of
4	of a	14	if you
5	you have	15	you to
6	the forest	16	to you
7	you are	17	in love
8	of my	18	and the
9	enter and	19	he is
10	to be	20	is the

Table 1: Top 20 Most Likely Word Pairs

The following Matlab code was written to use the empirically discovered distribution in order to generate the actual fake Shakespeare text.

```
% need lots of digrams in order to get good 2d statistics going
digrams_per_sentence = 10;
num_sentences = 1000;
FS_digrams = {};
probs = zeros(num_sentences*digrams_per_sentence,2);

chars_per_line = 50;
char_counter = 0;

% 2g
% inverse of CDF is the PDF
% CDF y-axis is [0,1]
% get a y-axis value randomly with rand(1)
% find x-axis location corresponding to this CDF value
% that x-axis location is the x-axis location for the PDF
% the x-axis location is the digram we're after
CDF = cumsum(freqs_digrams);

fprintf("\n    ");

char_counter = 0;
for digram_counter = 0:num_sentences*digrams_per_sentence-1

    sentence_counter = mod(digram_counter,digrams_per_sentence);
```



```
    % sample the distribution
    random_drawing = rand(1);
    digram_idx = find(CDF >= rand(1), 1, 'first');

    % get the digram and also store it in digrams array
    digram = A_digrams(digram_idx);
    FS_digrams{end+1} = {digram};
    probs(digram_counter+1,1) = digram_idx;
    probs(digram_counter+1,2) = freqs_digrams(digram_idx);

    % capitalize first letter of first digram in every sentence
    if sentence_counter == 0
        digram(1) = upper(digram(1));
    end

    % put a space after every digram but the last one
    if sentence_counter < digrams_per_sentence-1
        digram = [digram, ' '];
    end

    % put a period and space after every sentence
    elseif sentence_counter == digrams_per_sentence-1
        digram = [digram, '. '];
    end

    fprintf(digram);

    % wrap the line if char_counter overflows
    char_counter = char_counter + length(digram);
    if char_counter >= chars_per_line
        fprintf("\n");
        char_counter = 0;
    end
end

fprintf("\n\n");

FS_digrams = FS_digrams';
```

The following is an example of the "fake Shakespeare" generated by the above algorithm.

Him of you will my good ourselves and what care than that lives your brother is not phebe in such. Take from and pitiless it is we did must look much bound me up plants with some man take from. You with you he his practices to find and a dost call practise against that and what for you. The end that he perfect is that your comes an the beginning show here her usurping me ladies fetch that. Ungarter'd your the fleeces them with can hardly renders me prodigal portion to thy old robin my lord pride fell. Change for by thy hour ago and shining and didst please you you will pray you th shoulder poor dappled. Well but wrestled it next village her patience the forest who you so he you wrestle his banquet withal that. Woman in cannot recompense this hand to stay pity that god made to thee you off truth in wrestling is. A traitor of her say many that hated whom sir her virtues fashion bequeathed so villainous nut is innocent nose. Your fair not with but that of my on thyself but take even till with his follow the was seeking. Deserve well phebe phebe you that of churlish leave to injury for swear by nor none blush and upon my. Not to thee rosalind those that the wrestling not rosalind a moment a man lands and have better poor old. A moment brother gain must slender circle i'll by him may rest beards if what is did not have a. A beggar figure which duke's wrestler dost call be bawd then the cry 'holla of a coz ay therefore my.

The code written for discovering the distribution was used on the output of the fake Shakespeare generated above. The top 20 word pairs from the fake Shakespeare are presented now.

Rank	Word	Rank	Word
1	in love	11	in the
2	of my	12	enter and
3	here in	13	it is
4	practise against	14	of his
5	you wrestle	15	you are
6	no more	16	you here
7	of the	17	out of
8	to the	18	and yet
9	of a	19	she would
10	they are	20	find will

(a) Fake Shakespeare

Rank	Word	Rank	Word
1	in the	11	is not
2	it is	12	to the
3	of the	13	out of
4	of a	14	if you
5	you have	15	you to
6	the forest	16	to you
7	you are	17	in love
8	of my	18	and the
9	enter and	19	he is
10	to be	20	is the

(b) Original

Table 2: Top 20 Most Likely Word Pairs

Comparing Tables 2 and 1, we can see that indeed, the algorithm generating the fake Shakespeare is drawing from the same distribution as that obtained empirically above.

The following two plots are of the distributions of the original word pairs (digrams) and of the word pairs contained in the fake Shakespeare. These plots serve as a quick, visual reassurance that the fake Shakespeare is indeed drawing from the distribution obtained above.

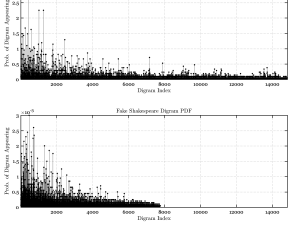


Figure 16: Relative Frequency of Digram vs. Index Into Digram List

7.3 The Expectation Maximization Algorithm

The idea behind the EM algorithm is the following. In our case, we are given a 1d array of data points. These data points have been drawn from two different Guassian "hats", so to speak, or distributions. Each "hat" has an associated mean, μ, and standard deviation, σ, and from these parameters, the Guassian distribution can be completely described.

As they say, a picture says a thousand words. The three plots in Figure 17 provide a visual explanation of the data we are working with when trying to discover the two sets of parameters from the combined samples. Looking at the figure, the first two plots are x_i from two unknown distributions. We draw a number from each hat and place it along the x-axis at a location corresponding to its value. The third plot in Figure 17 is what we actually have: a combination of the two unknown distributions provided as one mixed up distribution.

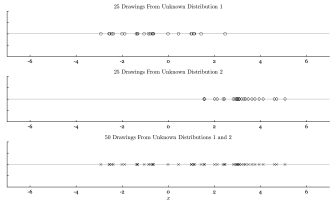


Figure 17: Two Unknown Distributions Provided as One Combination Distribution

Now, our job is to figure out what the parameters of these two Guassian "hats" are with only the raw array of data to work with. The algorithm starts out by guessing the μ's and σ's of each distribution and then proceeds to iterate through a two-step process, the E and M steps.

In the E step, we get the y-axis values, or the likelihoods, of the two guessed-at distributions for each x-axis data point we have been given. We then convert these likelihoods to relative proportions at each sample location. For example, if one of the data points given to us is x_i, our two guessed-at distributions will have y-axis values, y_1 and y_2. And so the relative proportions at this datum's sample location, x_i, will be $y_1/(y_1 + y_2)$ and $y_2/(y_1 + y_2)$.

In the M step, we find the center of mass that our x-axis data values have under each of the guessed-at distributions. A center of mass corresponding to a particular guessed-at distribution will be the dot product of the data points with the relative proportions for that distribution, divided by the sum of all the relative proportions for that distribution. This center of mass will be one of the μ's determining the distributions in the E step that we use to find the relative proportions. Likewise, in the M step, we also update the σ's for the next E step.

Again, a picture says it all. In Figure 18, we have a visual explanation of one iteration of the EM

algorithm. In the first two plots of Figure 18, the guessed distributions are in dotted lines, the guessed averages are circles along the x-axis, and the diamonds are the center of masses due to the actual data we are dealing with as they line up with the guessed distributions. And so in the third plot of Figure 77, we see the updated, guessed distributions move inward and tighten up, so to speak. In other words, the averages and standard deviations are approaching the best estimate of the unknown parameters.

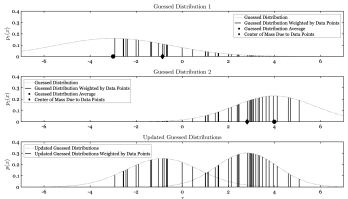


Figure 18: A Visualization of the E-step in the EM Algorithm

And this goes on for a while until the updated parameters aren't really changing that much. What we have at the end of the process is a good approximation of the defining parameters of the two Guassian distributions from which the data was originally drawn from, and their associated mixture ratios.

The following Matlab code was written to peform the EM algorithm described above on the data provided in homework #7.

```
14  num_components = 2;
15  num_data_points = length(q);
16
17  mean_guess = [1,8];
18  std_guess = [0.5,2];
19  mix_guess = [0.5,0.5];
20
21  num_iterations = 100;
22  for iteration = 1:num_iterations
23
24  % ------------- e step ----------------------------------------------- %
25
26  % sample the current dist. at locations in q, 2 column vectors result
27  likelihoods = mix_guess .* normpdf(q, mean_guess, std_guess);
28
29  % get relative proportions at each sample location
30  weights = likelihoods ./ sum(likelihoods,2);
31
32  % ------------- m step ----------------------------------------------- %
33
34  % sum of each weight column
```



```
35  total_weight = sum(weights,1);
36
37  % update overall component mixing proportions
38  mix_guess = total_weight ./ length(q);
39
40  % dot product of each weight column with q / sum of each weight column
41  mean_guess = (weights' * q) ./ total_weight';
42  mean_guess = mean_guess';
43
44  % diffs = [each q - mean_guess(1), each q - mean_guess(2)];
45  diffs = repmat(q,1,2) - repmat(mean_guess,length(q),1);
46  weighted_diffs = (weights .* diffs.^2) ./ total_weight;
47  std_guess =  sum(weighted_diffs,1);
48  std_guess = sqrt(std_guess);
49
50  end
```

The following output is printed to the command window as a result of running the algorithm:

```
*********************************
MATLAB DETERMINED PARAMETERS

Component 1:
---------------------------------

Mixing Proportion:      0.495180
Mean:                   2.951110
Standard Deviation:     2.035641

Component 2:
---------------------------------

Mixing Proportion:      0.504820
Mean:                   1.032308
Standard Deviation:     1.025217


*********************************
EM DISCOVERED PARAMETERS

Component 1:
---------------------------------

Mixing Proportion:      0.494818
Mean:                   2.952265
Standard Deviation:     2.035556

Component 2:
---------------------------------

Mixing Proportion:      0.505182
Mean:                   1.032551
Standard Deviation:     1.025631
```

Conclusion

This class was easily in the top three of my favorite ECE classes. I've dedicated all of my extra electrical engineering time to audio, whether it be artistic, as in music, or practical, as in speech and hearing, and ECE 537 was a *tour de force* of everything I'm into. What I'm saying is that I loved it. On a separate but related personal note, beginning on day one of the semester, I began the process of taking out a mortgage, finding a house in Chicago and by the last two weeks of the semester, moving from Urbana to Chicago. To

say the very least, trying to juggle everything was an extreme challenge. But because I found the topics so fun, I didn't have to "make time" for ECE 537. I wholeheartedly enjoyed working on the problems. My only regret is that I wasn't able to spend more time than I did on everything. Thank you for such a great class! I feel like I've only scratched the surface.

ECE 537 Final Report

December 19, 2018

1 Overview

This class contains three parts: Acoustics/VT modeling/Signal processing; Psycho-acoustics; Human Speech Recognition/Information Processing. I will review the history of the three parts respectively, summarize the content of each class including paper review and go over all the homework this semester.

2 Review of the history

2.1 Acoustics/VT modeling/Signal processing

People started to explore Acoustics since ancient Greek. According to legend, Pythagoras (c.570 - c.495 BC) discovered that musical notes could be translated into mathematical equations when he passed blacksmiths at work one day and heard the sound of their hammers clanging against the anvils. Aristotle (384 - 322 BC) claimed that the quality of sound will be unchanged and will travel as far as the waves reaches.

In 17th century, Galilei, Galileo (1564 - 1642) was one of the first to understand sound frequency. By scraping a chisel at different speeds, he linked the pitch of the sound produced to the spacing of the chisels skips, a measure of frequency. Marin Mersenne (1588 1648), French mathematician, was the first person to record the speed of sound as it travels through air in the year of 1640. His measurement had an error around than 10 percent which was impressive considering the lack of technology at that time. Later in the century Robert Hooke (1635 1703), an English physicist, first produced a sound wave of known frequency, using a rotating cog wheel as a measuring device. By 1660 the Anglo-Irish scientist Robert Boyle had improved vacuum technology to the point where he could observe sound intensity decreasing virtually to zero as air was pumped out. Boyle then came to the correct conclusion that a medium such as air is required for transmission of sound waves. Based on Boyles law, Newton, Sir Issac (1642 1726) formulated an empirical law of cooling, made the first theoretical calculation of the speed of sound, and introduced the notion of a Newtonian fluid.

In 18th century, Daniel Bernoulli (1700 - 1782) was a Swiss mathematician and physicist and was one of the many prominent mathematicians in the Bernoulli family. He is the first believed (in the 1733- 1742 time frame) that any acoustic vibration could be expressed as a superposition of simple modes (sinusoidal vibrations). Leonhard Euler (1707 - 1783) was a Swiss mathematician and engineer. The speed of sound is calculated from the relativistic Euler equations, if relativistic effects are important. Jean le Rond dAlembert (1717 - 1783) was a French mathematician, mechanician and physicist. DAlemberts formula for obtaining solutions to the wave equation is named after him. The wave equation is sometimes referred to as dAlemberts equation. Joseph-Louis Lagrange (1736 - 1813) indicates a mistake made by Newton, obtains the general differential equation for the motion, and integrates it for motion in a straight line. He also points out a lack of generality in the solutions previously given by Brook Taylor, DAlembert,

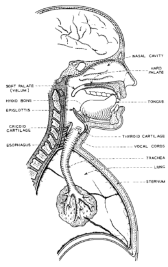


Figure 1: Schematic diagram of the human vocal mechanism

as well. Transmition line is a model that we simplified to know what happened in our ear and a easier way to notation.

After that, we move from uniform tubes of constant crossection to the topic of horns, which have an area that is changing in the direction of wave propagatation. This is a very important topic to communication acoustics.In three dimensions the basic acoustic equations are based on two laws, Newtons second law of conservation of momentum

$$\nabla P = -s\rho_0 U \tag{1}$$

$$\nabla \cdot U = -\frac{s}{\eta P_0} P \tag{2}$$

where $P(\omega)$ is the pressure, $U(\omega)$ is the vector particle velocity, with s the Laplace frequency $s = sigma + j\omega$, 0 the density of air,$eta = c_p/,v$, P_0 is the static pressure of air. We refer to the ratio of pressure to particle velocity as the specific acoustic impedance in [Rayls], and the pressure over a volume velocity is the acoustic impedance in [acoustic ohms].

One of a special case of conical horn is spherical acoustics,

$$\nabla_r P = -s\rho_0 U_r \tag{3}$$

The similar transform applies to velocity,

$$\nabla_r U = \frac{1}{r^2}\frac{dr^2 U_r}{dr} = -\frac{s}{\eta P_0} P \tag{4}$$

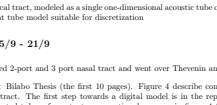


Figure 4: (a) The vocal tract, modeled as a single one-dimensional acoustic tube of varying cross-sectional area and (b) an eight tube model suitable for discretization

3.3 Lecture 15/9 - 21/9

This week, we learned 2-port and 3 port nasal tract and went over Thevenin and Norton theory.

Paper review: Bilabo Thesis (the first 10 pages). Figure 4 describe concatenarte acoustic tube model of the vocal tract. The first step towards a digital model is in the representing the tube as a series of N concatenated tubes of constant cross-sectional areas as in figure 4 (b) ($N = 8$). The tubes are assumed to be of equal length $delta$; if L is the total length of the vocal tract, we have $N\delta = L$. As δ becomes small, the shape of the approximation of the series of the tubes will converage to that of continuous vocal tract shown in Figure4(a).

3.4 Lecture 24/9

In this lecture, we reviewed Fourier Series, Fourier Transform, Laplace Transform, Z Transform, The Discrete Time Fourier Transform (DTFT), discrete Fourier transform (DFT), fast Fourier transform (FFT).

The Fourier transform (FT) decomposes a function of time (a signal) into the frequencies that make it up, in a way similar to how a musical chord can be expressed as the frequencies (or pitches) of its constituent notes. The Fourier transform of a function of time is itself a complex-valued function of frequency, whose absolute value represents the amount of that frequency present in the original function, and whose complex argument is the phase offset of the basic sinusoid in that frequency.

The Discrete Time Fourier Transform (DTFT) is the member of the Fourier transform family that operates on aperiodic, discrete signals. The best way to understand the DTFT is how it relates to the DFT. To start, imagine that you acquire an N sample signal, and want to find its frequency spectrum.

DFT converts a finite sequence of equally-spaced samples of a function into a same-length sequence of equally-spaced samples of the discrete-time Fourier transform (DTFT), which is a complex-valued function of frequency.

A fast Fourier transform (FFT) is an algorithm that computes the DFT of a sequence. The best-known FFT algorithms depend upon the factorization of N, but there are FFTs with O(N log N) complexity for all N, even for prime N. Many FFT algorithms only depend on the fact that $e^{-2\pi i/N}$ $e^{-2\pi i/N}$ is an N-th primitive root of unity, and thus can be applied to analogous transforms over any finite field, such as number-theoretic transforms. Since the inverse DFT is the same as the DFT, but with the opposite sign in the exponent and a 1/N factor, any FFT algorithm can easily be adapted for it.

and Euler.

In 19th century, Johann Carl Friedrich Gauss (1777 - 1855) was a German mathematician and physicist, Practical application of Gauss' law in acoustics is not a very well known method. However, any inverse square law behavior can be formulated in the way similar to Gauss' law, which allows us to extend the same principle to sound waves propagation. Pierre-Simon Laplace (1745 - 1827) was a French scholar whose work was important to the development of engineering, mathematics. He formulated Laplace's equation, and pioneered the Laplace transform which appears in many branches of mathematical physics, a field that he took a leading role in forming. Jean-Baptiste Joseph Fourier (1768 - 1830) was a French mathematician and physicist born in Auxerre and best known for initiating the investigation of Fourier series and their applications to problems of heat transfer and vibrations. Hermann von Helmholtz (1821 1894) was a German physician and a pioneer in the scientific study of human vision and audition. He coined the term "psychophysics," to capture the distinction between the measurement of physical stimuli and their effect on human perception. Oliver Heaviside (1850 1925) was an English self-taught electrical engineer, mathematician, and physicist who adapted complex numbers to the study of electrical circuits, invented mathematical techniques for the solution of differential equations (equivalent to Laplace transforms), reformulated Maxwell's field equations in terms of electric and magnetic forces and energy flux, and independently co-formulated vector analysis. Strutt, 3rd Baron Rayleigh 1842 30 June 1919), was a British scientist who made extensive contributions to both theoretical and experimental physics. He studied and described transverse surface waves in solids, now known as "Rayleigh waves". He contributed extensively to fluid dynamics, with concepts such as the Rayleigh number.

In 20th century, David Hilbert (1862 - 1943) was a German mathematician. Hilbert dedicated himself to the study of differential and integral equations. His work had direct consequences for important parts of modern functional analysis. In order to carry out these studies, Hilbert introduced the concept of an infinite dimensional Euclidean space, later called Hilbert space. Harvey Fletcher (1884 - 1981) was an American physicist. Known as the "father of stereophonic sound," he is credited with the invention of the 2-A audiometer and an early electronic hearing aid. He was an investigator into the nature of speech and hearing, and made contributions in acoustics, electrical engineering, speech. Harry Nyquist (1889 - 1976) was a Swedish-born American electronic engineer who made important contributions to communication theory. A lot of terms are named after him. Nyquist rate: sampling rate twice the bandwidth of the signal's waveform being sampled; sampling faster than this rate assures that the waveform can be reconstructed accurately. Nyquist frequency: half the sample rate of a system; signal frequencies below this value are unambiguously represented; plus Nyquist filter, Nyquist plot, Nyquist ISI criterion, Nyquist (programming language)Nyquist stability criterion and so on. Hendrik Wade Bode (1905 - 1982)was an American engineer, researcher, inventor, author and scientist. As a pioneer of modern control theory and electronic telecommunications he revolutionized both the content and methodology of his chosen fields of research. Homer W. Dudley (1896 - 1980) was a pioneering electronic and acoustic engineer who created the first electronic voice synthesizer for Bell Labs in the 1930s and led the development of a method of sending secure voice transmissions during World War Two. Sir Richard Arthur Surtees Paget, (1869 - 1955) was a British barrister and amateur scientific investigator, who specialised in speech science and the origin of speech. Following the publication of his book on these topics, Human Speech, in 1930, Paget worked for the remaining decades of his life on a new type of signing system for the deaf, which became the Paget Gorman Sign System. Still more revealing and more informative is the technique of high-speed photography, pioneered by Farnsworth, in which moving pictures are taken at a rate of 4000 frames/sec, or higher. Claude Elwood Shannon (1916 - 2001) was an American mathematician, electrical engineer, and cryptographer known as "the father of information theory". James Loton Flanagan (1925 - 2015) was an electrical engineer. He has worked in voice communications, computer techniques, and electroacoustic systems. At Bell Laboratories he was the department head of the Acoustics Research Department for many years, and managed and supported work such as James E. West's invention of the electret microphone, Bishnu S. Atal's work on speech coding, David Berkley and Gary Elko's work on acoustics, Jont Allen and Joe Hall's work on psycho-acoustics.

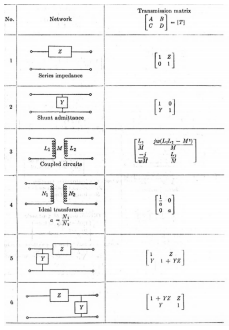


Figure 2: This figure summarize the equation series impedance, shunt admittance, coupled circuits, ideal transformer

where the subscript r indicates the radial component of the pressure gradient and velocity divergence. we define the per-unit length series acoustic impedance,

$$Z(r,s) = s\frac{\rho}{A(r)} \tag{5}$$

and the per-unit length shunt acoustic admittance,

$$Y(r,s) = s\frac{A(r)}{\eta P_0} \tag{6}$$

$$\frac{d}{dr}\begin{bmatrix} P(r,\omega) \\ U(r,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & Z(r,s) \\ Y(r,s) & 0 \end{bmatrix}\begin{bmatrix} P(r,\omega) \\ U(r,\omega) \end{bmatrix} \tag{7}$$

Thus the spherical wave solution may be expressed as a Webster (conical) horn equation. If one reduces this to a second-order equation in pressure, the classic Webster horn equation results. This is a scaled form of the conical horn (it is a horn because the per-unit-length impedance Z and admittance Y are a function of r), with an angle that subtends 4π

3.5 Lecture 28/9 - 1/10

In the lecture, we reviewed the history of acoustics which I discussed before. We went over all three homework for the preparation of midterm.

3.6 Lecture 3/10

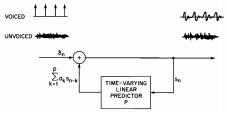


Figure 5: Block diagram of a functional model of speech production based on the linear prediction representation of the speech wave.

We talked about Linear prediction of speech.

Paper review: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave by B. S. Atal and SUZANNE L. Hanauer. The figure 5 describe the basic idea of how to construct the output of prediciton. The linear predictor P, a transversal filter with p delays of one sample interval each, forms a weighted sum of the past p samples at the input of the predictor. The output of the linear filter at the nth sampling instant is given by

$$s_n = \sum_{k=p} pa_k s_{n-k} + \delta_n \tag{14}$$

where the "predictor coefficients" a_k account for the filtering action of the vocal tract, the radiation, and the glottal flow; and δ_n represents the nth sample of the excitation.

We now use this property of the speech wave to determine the predictor coefficients. Define the prediction error E_n as the difference between the speech sample s_n and its predicted value

$$\hat{s}_n = \sum_{k=p}^{p} a_k s_{n-k} \tag{15}$$

E_n is then given by

$$E_n = s_n - \hat{s}_n = s_n - \sum_{k=p}^{p} a_k s_{n-k} \tag{16}$$

We define the mean-squared prediction error E^2_{an} as the average of E^2_n over all the sampling instances n in the speech segment to be analyzed except those at the beginning of each pitch period. The predictor coefficients a_k of 14 are chosen so as to minimize the mean-squared prediction error. The same procedure is used to determine the predictor parameters for unvoiced sounds, too.

2.2 Psycho-acoustics

Harvey Fletcher (1884 - 1981) was an American physicist.Known as the "father of stereophonic sound," he is credited with the invention of the 2-A audiometer and an early electronic hearing aid. He was an investigator into the nature of speech and hearing, and made contributions in acoustics, electrical engineering, speech. In this course, we focus on his contribution on speech perception. He showed that speech features are usually spread over a wide frequency range, and developed the articulation index to approximately quantify the quality of a speech channel. He also developed the concepts of equal-loudness contours, loudness scaling and summation, and the critical band.

In 1924, an important paper by Wegel and Lane was published. They conduct experiment of masking of one pure tone.

2.3 Human Speech Recognition/Information Processing

George Ashley Campbell (1870 - 1954) was an American engineer. He was pioneer in developing and applying quantitative mathematical methods to the problems of long-distance telegraphy and telephony. His most important contributions were to the theory and implementation of the use of loading coils and the first wave filters designed to what was to become known as the image method. Georg von Bksy (1899 1972) was a Hungarian biophysicist born in Budapest, Hungary. In 1961, he was awarded the Nobel Prize in Physiology or Medicine for his research on the function of the cochlea in the mammalian hearing organ. Gustav Theodor Fechner (1801 - 1887) was a German philosopher, physicist and experimental psychologist. An early pioneer in experimental psychology and founder of psycho-physics, he inspired many 20th century scientists and philosophers. He is also credited with demonstrating the non-linear relationship between psychological sensation and the physical intensity of a stimulus via the formula which became known as the WeberFechner law, and I will discuss it later section.

3 Review of all the lecture

3.1 Lecture 29/8 - 31/8

The first week talked about the course information in general. I was surprised when I reviewed my note that this course mentioned Shannons theory of information at the very beginning. Even though we did not cover the details, still, we can have a basic outline of it. In the first week, we discussed entropy in details. Entropy representing the unavailability of a system's thermal energy for conversion into mechanical work, often interpreted as the degree of disorder or randomness in the system. We also learned the how the human generate sound and the path for sound transmission in 1. And we derived the speed of sound, which we also covered in the homework. We learned the definition of Consonants that are sounds produced with a constriction at some point in the vocal tract and the definition and the definition of Vowels that are speech sounds with no narrow constriction in the vocal tract. They are usually voiced (produced with vocal fold excitation), though they may of course be whispered.

3.2 Lecture 5/9 - 15/9

This week introduced ABCD matrix methods that makes circuit analyze much simpler and the concept of transmission line we used it though the whole semester. The following table contains all the circumstances. At first, I cannot understand why the circuit can be related to acoustics. After reading the paper "Wave model of the cat tympanic membrane" by Jont Allen and Pierre Parent, I finally got the answer that indeed our ear contains millions of transition line and we need to derive reflection coefficient

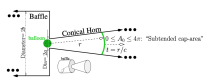


Figure 3: Experimental setup showing a large pipe on the left terminating the wall containing a small hole with a balloon, shown in green. At time t = 0 the ballon is pricked and a pressure pulse is created. The baffel on the left is meant to represent an inf long tube having a very large radius compared to the horn input diameter 2a, such that the acoustic admittance looking to the left ($A/\rho_0 cwithA \longrightarrow$ inf), is very large compared to that looking into the horn, $Y_a(a,s)$ 13. At time t = b/c the outbound pressure pulse $\delta(t\delta/c)/r$ has reached radius b.

The functions Z and Y define the acoustic characteristic impedance (resistance) that depends on the radius r;

$$Z_0(r) = \sqrt{\frac{Z}{Y}} = \frac{\sqrt{\rho_0\eta P_0}}{A(r)} = \frac{\rho_0 c}{A(r)} = 1/Y_0(r) \tag{8}$$

and a wave propagation function, closely related to the sound speed by

$$\gamma = \sqrt{ZY} = s/c \tag{9}$$

Equation 17 may be re-expressed as a single equation in pressure as

$$\frac{d^2}{r^2}rP = \gamma^2 rP \tag{10}$$

Following d'Alembert (1747), the solution to this equation in spherical coordinates, corresponding to a spherical symmetry, is given by the sum of an outbound $P^+(r,\omega)$ and an inbound $P(r,\omega)$ wave.

$$P(r,\omega) = P^+_\omega\frac{e^{-sr/c}}{r} + P^-_\omega\frac{sr/c}{r} \tag{11}$$

where $P^+_\omega(a) \Leftrightarrow p^+(t)$ and $P^-_\omega(a) \Leftrightarrow p^-(t)$ are the Fourier source strengths and a is the radius corresponding to the source location (the radius the waves are launched from). When $P^+_\omega(a) = 1$ and $P^-_\omega(a) = 0$, the resulting outbound pressure wave is a Dirac delta function.

Substituting the expression for the pressure into ?? results in an expression for the radial component of the particle velocity

$$U_r = \frac{1}{Z}\frac{\partial}{\partial r}P^+_\omega\frac{e^{-sr/c}}{r} + P^-_\omega\frac{e^{sr/c}}{r} = Y^+P^+ - Y^-P^- \tag{12}$$

corresponding to out and inbound velocity waves $U^\pm_r = Y^\pm P^\pm$. From the above definition the acoustic radiation admittances at $r = a$ for outbound (Y_+) and inbound (Y) waves are

$$Y^\pm(a,s) = \frac{A\rho s^2}{\rho\eta P_0} \pm \frac{A\rho s^2}{spa} \tag{13}$$

The method of applying a bilinear Z, FIR, IIR filter in Matlab build in function is introduced in our homework 3.

Paper review: "Wave model of the cat tympanic membrane" by Jont Allen and Pierre Parent 10/9

3.7 Lecture 5/10

The short-time Fourier transform (STFT), is used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on each shorter segment. One then usually plots the changing spectra as a function of time.

3.8 Lecture 10/10

First we reviewed several equations.

Webster Horn Equation:

$$\frac{d}{dr}\begin{bmatrix} P(r,\omega) \\ U(r,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & Z(r,s) \\ Y(r,s) & 0 \end{bmatrix}\begin{bmatrix} P(r,\omega) \\ U(r,\omega) \end{bmatrix} \tag{17}$$

We used to deal with plane waves. In this lecture, we used spherical bessel function that is a solution to the spherical Bessel under Spherical coordinates differential equation.

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - \alpha^2)y = 0 \tag{18}$$

$$x^2\frac{d^2y}{dx^2} + 2x\frac{dy}{dx} + [x^2 - n(n+1)]y = 0 \tag{19}$$

solution is j_n, y_n which is the same as Bessel function J_α, N_α.

$$j_n(x) = (-x)^n(\frac{1}{x}\frac{d}{dx})^n\frac{\sin(x)}{x} \tag{20}$$

$$y_n(x) = -(-x)^n(\frac{1}{x}\frac{d}{dx})^n\frac{\cos(x)}{x} \tag{21}$$

we did mention head shadowing effect but not in details.

3.9 Lecture 12/10

Professor Mark Hasegawa-Johnson taught this class on speech coding. There are two kinds of waveform coding: Pulse code modulation(PCM), Differential PCM(DPCM). Pulse code modulation (PCM) is the name given to memoryless coding algorithms which quantize each sample of $s(n)$ using the same reconstruction levels $\hat{s}_k$, k = 0,..,k, regardless of the values of previous samples. The reconstructed signal $\hat{s}(n)$ is given by

$$\hat{s}(n) = \hat{s}_m \quad (s(n)\hat{s}_m)^2 = min(s(n)\hat{s}_k)^2 \tag{22}$$

Uniform PCM is the name given to quantization algorithms in which the reconstruction levels are uniformly distributed between $S_m ax$ and $S_m in$. Suppose that a signal is quantized using B bits per sample. If zero is a reconstruction level, then the quantization step size δ is

$$\delta = \frac{S_{max} - S_{min}}{2^B - 1} \tag{23}$$

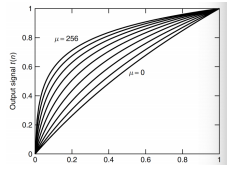


Figure 6: μ -law companding function, μ = 0,1,2,...256

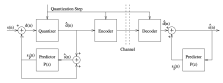


Figure 7: Model of differential PCM

Assuming that quantization errors are uniformly distributed between $\delta/2$ and $\delta/2$, the quantization error power is

$$10log_{10}E[e^2(n)] = 10log_{10}\frac{\delta^2}{12} \approx = constant + 20log_{10}(S_{max} - S_{min}) - 6B \quad (24)$$

Companded PCM is the name given to coders in which the reconstruction levels $\hat{s}_k$ are not uniformly distributed. Such coders may be modeled using a compressive nonlinearity, followed by uniform PCM.A typical example is the μ-law companding function which is given by

$$t(n) = S_{max}\frac{log(1+\mu|s(n)/S_{max}|}{log(1+\mu)}sign(s(n)) \quad (25)$$

Figure 6 shows the changes of output signal under different μ.

In differential PCM (DPCM), each sample $s(n)$ is compared to a prediction $s_p(n)$, and the difference is called the prediction residual $d(n)$. d(n) has a smaller dynamic range than $s(n)$, so for a given error power, fewer bits are required to quantize d(n). Common sub-types of DPCM include: (1)Sigma-Delta coder: s[n] is upsampled by a factor of 8 or 16, then d[n] is quantized using only one bit per sample and inside an A/D; (2) Adaptive differential PCM (ADPCM): G.726 ADPCM is frequently used at 32 kbps in land-line telephony. The predictor in G.726 consists of an adaptive second-order IIR predictor in series with an adaptive sixth-order FIR predictor. Figure 7 shows the preceedure that applies DPCM.

$\Delta I \propto \sigma_I$:

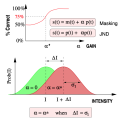


Figure 10: Theory of Signal Detection

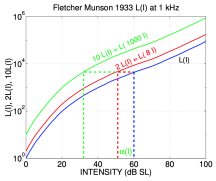


Figure 11: Fletch Munson 1933 L(I) at 1kHz

3.11.3 Basic model of observer

3.12 Lecture 19/10

Review paper: Nonlinear Cochlear Signal Processing and Masking in Speech Perception by Jont Allen

Each inner-hair-cell voltage is a low-pass-filtered representation of the detected inner- hair-cell cilia displacement. Each hair cell is connected to many neurons, having a wide range of spontaneous firing rates and thresholds.

As shown in 13, there are typically three (occasionally four) outer hair cells (OHCs) for each inner hair cell (IHCs), leading to approximately 12 000 OHCs in the human cochlea. Outer hair cells are used for intensity dynamic-range control.

During the following discussion it is necessary to introduce the concept of a one-port (two- wire) impedance. Ohms law defines the impedance as

Impedance = effort/flow

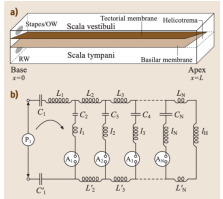


Figure 15: On the upper panel (a) see the basic 2-D box model of the cochlea. The Base (x=0) is the high-frequency end of the cochlea while the Apex(x = L) carries the low frequencies. The lower panel (b) is the 1924 Wegel and Lane electrical equivalent circuit. The model is built from a cascade of electrical sections.

amount of masking by tones of frequency 200 to 3500 was determined for frequencies from 150 to 5000 per sec. The magnitude of a tone is taken as the logarithm of the ratio of its pressure to the threshold value, and masking is taken as the logarithm of its threshold value with masking to that without. The curves of masking as function of magnitude are approximated straight lines as a rule except for rounded feet, of slopes intersecting the magnitude axis at minimum masking magnitude. For a given masking frequency the slope increases from zero through nearly 1.0 for a frequency near, then more slowly, approaching about 3 to 4 for the highest frequencies measured. The intercept is small or zero below, then increases rapidly, approaching the value 3 for high frequencies. Except when the frequencies are so close together as to produce beats, the masking is greatest for tones nearly alike. When the masking tone is loud it masks tones of higher frequency better than those of frequency lower than itself. When the masking tone is weak, there is little difference. If the masking tone is introduced into the opposite ear, no appreciable masking occurs until the intensity is sufficient to reach the listening ear through the bones of the head. At intensities considerably above minimum audibility, there is no longer a linear relation between the sound pressure and the response of the ear. Data are given showing combinational tones resulting from this non-linearity when two tones are simultaneously introduced in the ear. The presence also of subjective overtones in a loud tone accounts for the large amount of masking of tones higher than itself by a loud masking tone.

Dynamics of inner ear. The data on masking together with Knudson's data on frequency sensibility are interpreted in terms of the dynamical theory of the cochlea which ascribes its frequency selectivity to a passing of vibrations along the basilar membrane and a shunting through narrow regions of the membrane at points depending on the frequency. Conjectured curves are given for a few single frequencies of the amplitude of vibration of this membrane as a function of the distance along it.

Masking refers to a (typically negative) perceptual interaction between sounds. Masking studies are useful for telling us about the selectivity of the auditory system, and how we process complex sounds and sound environments. Masking can be energetic or informational.

3.10 Lecture 15/10

This lecture is an introduction of loudness Psychophysics. The intensity of JND is a mystery. Each time we hear the same short tone pulse, we hear it with a different loudness. The intensity JND Δ I is a measure of this intern perceptual fluctuation (noise) given by σ_L. The relation between intensity and loudness is as the following figure.

Figure 8: The relation of intensity and loudness

Note: the loudness JND ΔL (the difference of the mean value)is the proportional to the $\sigma_L(L)$ (the standard deviation of loudness). And the loudness JND is proportional to the internal "loudness noise".

The are several laws about the intensity. Unfortunately, most laws are flaws, making intensity more mysterious. Sometimes, people believed a certain law and abandoned it several years later.

3.10.1 Weber's Law

Weber's law is proposed by Weber in 1846. He showed by his experiment that $\Delta I \propto I$ where I is the physical intensity, and ΔI is called JND. As a result,$\Delta I/I$ **is called the Weber Fraction**. And his law stated that Weber Fraction is constant. This law sometimes has a good estimation under some circumstances. For example,**a floating point converter obeys Weber's law. And for fix point, $\sigma_I = \Delta I$ is a constant.** Plus, our ear is a floating point converter.

However, this law is flaw. **In 1947, Miller proposed that wide band noise intensity discrimination.** In 1928, Riesz established the near-miss to Weber's law for tones. Riesz found out **that Weber fraction is not constant for pure tones.** Logically, Weber's law cannot be right. If intensity is linear while the observers are not. Observers do not have access to the intensity. The only have access to the loudness. Therefore, $\Delta I/I$ remaining a constant seems impossible even though there were some good estimation.

3.10.2 Near-miss to Weber's law by Riesz

Riesz used two beating tones 3Hz apart for this measurement(i.e. 1kHz masker and a low-level 1003 Hz probe) **Near-miss to Weber's law results from the fact that the internal noise $\sigma_I = \Delta I$ is not independent of L.** In 1997, Professor Allen and Neely prove that the noise is Poinsson-like:

$$\Delta L(L) \approx \sqrt{L} \quad (26)$$

3.10.3 Fechner's Hypothesis (1860)

Fechner is called the father of psychophysics. Fechner's hypothesis (or postulate) was that the loudness JND $\Delta L(I)$ is constant. He assumed that the total change in the loudness

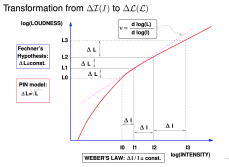


Figure 12: Transformation form $\Delta I(I)$to$\Delta L(L)$

In an electrical system the impedance is the ratio of a voltage (effort) over a current (flow). In a mechanical system it is the force (effort) over the velocity (flow). For linear time-invariant causal (LTIC) systems (i.e., an impedance), phasor notation is very useful, where the tone is represented as the real part (Re) of the complex exponential

$$e^{(2\pi ft+\phi} = cos(2\pi ft+\phi) + isin(2\pi ft+\phi). \quad (3.2)$$

The symbol denotes equivalence. It means that the quantity to the left of is defined by the quantity on the right. More specifically, impedance is typically defined in the frequency domain using Laplace transform notation, in terms of a damped tone

$$Ae^{\sigma t}cos(2\pi ft+\phi) = ARe e^{st+i\phi} \quad (3.3)$$

excitation, characterized by the tones amplitude A, phase and complex Laplace frequency $s = \sigma + i2\pi f$. When a function such as Z(s) is shown as a function of the complex frequency s, this means that its inverse Laplace transform $z(t) \stackrel{L^{-1}}{\leftarrow} Z(s)$ must be causal. In the time domain, the voltage may be found from the current via a convolution with $z(t)$. Three classic examples of such impedances are presented next. Example 3.1: The impedance of the tympanic membrane (TM, or eardrum) is defined in terms of a pure tone pressure in the ear canal divided by the resulting TM volume velocity (the velocity times the area of TM motion) [3,34, 35]. The pressure (effort) and volume velocity (flow) referred to here are conventionally described using complex numbers, to account for the phase relationship between the two. Example 3.2: The impedance of a spring is given by the ratio of the force F(f) to velocity V(f)=sX(f) with displacement X

$$Z(s) = F/V = K/s = 1/sC$$

where the spring constant K is the stiffness, C the compliance, and s is the complex radian frequency. The stiffness is represented electrically as a capacitor (as parallel lines in Fig. 3.2b). Having $s = \theta + i2\pi f$ in the denominator indicates that the impedance of a spring has a phase of π /2 (e.g., 90). Such a phase means that when the velocity is cos $(2\pi ft)$, the force is sin $(2\pi ft)$. This follows from Hookes law

$$F = KX = K\,V/s.$$

From Newtons law F = Ma where F is the force, M is the mass, and acceleration a(s) = sV (s) (i.e., the acceleration in the time domain is $dv(t)/dt$). The electrical element corresponding to a mass is an inductor, indicated in Fig 15 by a coil. Thus for a mass $Z(s) = sM$. From these relations the

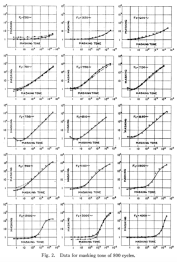


Fig. 2. Data for masking tone of 800 cycles.

Figure 16: modifying: The figure shows the amount of masking of various frequencies from 250 to 4000 cycles produced by an 800 cycle tone, plotted as a function of the magnitude of this masking tone. For example, the first curve shows the amount of masking as already defined, of 250 cycles plotted as a function of the magnitude of an 800 cycle masking tone. Each plotted point represents the average of four observations taken in succession at one time. In all figures, the masking tone is designated by F1 and the masked by F2.

3.14 Lecture 2/11 - 11/7

Prof. Wickesberg comes and gave three lectures on Auditory Pathway. Sound is the vibration of air molecules. A tuning fork produces a sinusoidal vibration. The number of cycles each second is the frequency (Hertz). We hear frequencies between 20 Hz and 20,000 Hz.

Pitch is the perceived frequency of a sound. Figure 18 and figure 19 shows how neuron receive signals.

Frequency response is usually plotted as a tuning curve at how loud does a tone at a specific frequency have to be to excite an auditory nerve fiber. Auditory nerve fibers are narrowly tuned at high frequencies and broadly tuned at low frequencies.

Stellate cells have dendrites that run parallel to the path of auditory nerve fibers. Each stellate cell receives many inputs from only a few nerve fibers and responds only to a narrow band of frequencies.

Octopus cells have two or three large dendrites that extend perpendicularly across the paths of many auditory nerve fibers and are ideally situated to integrate information across frequency.

Multipolar cells in figure 24 have dendrites that extend across a large number of auditory nerve fibers. They also project to the DCN, the AVCN and to the cochlear nucleus on the other side of the

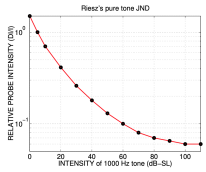


Figure 9: Intensity of 1000Hz tone

between two intensities I_1 and I_2 may be found by counting the number of JNDs. The "counting formula" is:

$$N_{JND} \equiv \int_{L_1}^{L_2}\frac{dL}{\Delta L} = (L_2 - L_1)/\Delta L \quad (27)$$

There are two assumptions in Fechner's law: (1) $\Delta I \propto I$, i.e. Weber's law; (2) the internal noise $\Delta L = \sigma_L$ is constant. Those two assumptions give Fechner's "law" :

$$L(I) \propto log(I) \quad (28)$$

There is a major breakthrough, however, incorrect.

3.11 Lecture 17/10

Continue with the same topic as the last class

3.11.1 Theory of Signal Detection

L.L Thure stone 1927 and later David Green 1965: Formally define the intensity JND as "the relative signal level for detection 75 percentage of the time"

3.11.2 Loudness Additivity

Fletcher and Monson 1933 showed that loudness adds **(1) adjust I_2 so that:** $L(I1, f1) = L(I2, f2)$ **(2) Two equally load tones played together are twice as loud:** $L(I1, I2, f1, f2) = 2L(I1, f1)$ Find gain $\alpha(I)$ such that $L(\alpha I_1, f1) = 2L(I1, f1)$ Result: α is about 9dB (actually it depends on intensity)

Fletcher and Munson's loudness growth data based on loudness additivity is now called:

Steven's law $L(I) = I^{\nu}$ **with ν = 1/3**

Loudness vs intensity for 1, 2, and 10 equally loud components:

magnitude of the impedance of a spring decreases as 1/f , while the impedance magnitude of a mass is proportional to f. The stiffness with its -90 phase is called a lagging phase, while the mass with its +90 phase is called a leading phase.

Different points along the basilar membrane are represented by the cascaded sections of the lumped transmission line model of Fig 15. The position X along the model is called the place variable and corresponds to the longitudinal position along the cochlea. The series (horizontal) inductors (coils) denoted by L_k represent the fluid mass (inertia) along the length of the cochlea, while the shunt elements represent the mechanical (acoustical) impedance of the corresponding partition (organ of Corti) impedance, defined as the pressure drop across the partition divided by its volume velocity per unit length

$$Z_p(s, X) = K_p(X)/s + R_p(X) + sM_p \quad (29)$$

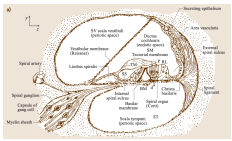


Figure 13: All the major structures of the cochlea.The three chambers are filled with fluid. Reissners membrane is an electrical barrier and is not believed to play a mechanical role.

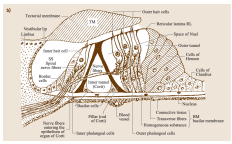


Figure 14: The right panel (b) shows the inner and outer hair cells, pillar cells and other supporting structures, the basilar membrane (BM), and the tectorial membrane (TM)

3.13 Lecture 24/10

Paper Review: WegelLane24.pdf Auditory masking of one pure tone by another. Using an air damped telephone receiver supplied with current with a proper combination of two frequencies, as source, the

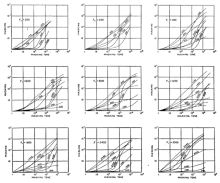


Fig. 3. Masking for tones in same ear.

Figure 17: This figure shows some of the corrected curves from Fig2 reproduced on common axes. Curves for masking tones of 200, 300, 400, 600, 1200, 1800, 2400 and 3500 cycles are also included in this figure. A tone of a frequency much below the masking tone is not perceptibly so when the tones are very loud. A tone of much higher frequency than the masking tone is not perceptibly masked for the lower range of intensities, but at a rather definite high intensity masking occurs perceptibly and quickly becomes very great as the masking tone is increased.

brainstem.

3.15 Lecture 10/11

Introduction to Information Theory

Information theory is quantitative measure of information. It describes the relation between information and its probability Most unexpected events give maximum information.

Information is inversely proportional to its probability of occurrence and is continuous function of its probability.

$$I(x) = -logp(x) \quad (30)$$

where p(x) is the probability of occurrence of x and I(x) is Information.

Take source S delivers messages as $X = x_1, x_2, x_3, x_4, x_m$ for example: the probability of x is p x $= p_1, p_2, p_3, p_m$ corresponding to message X. In this case, x1 occurs N*p1 times. Single occurrence of x1 conveys information = $-logp1$ and Np1 times occurrence conveys information, therefore is $-N_p1 * logp1$

Total Information is $-N_p1logp1 - N_p2logp2. - Np_mlogp_m$

Average information = $1/N * TotalInformation$

The entropy of H(x) is $\sum_{i=1}^{m} p_i log(p_i)$ Apparently, complete probability scheme $\sum p_i = 1$

There are several units for I(x): bits, nats, hartley. Their relationship are as the followings:

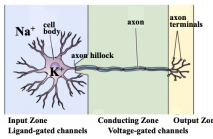


Figure 18: The figure shows the chemicals exchanges when our neuron receiving signals

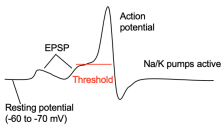


Figure 19: The figure shows the magnitude of a neuron receiving signals

$I(x) = -log_2 p(x)$ bits;$I(x) = -ln p(x)$ nats and $I(x) = -log_{10} 0p(x)$ hartley. And entropy H(x) = bits per symbol or bits per message

3.15.1 Mutual Information

Mutual information gain through the channel. First, x_i is transmitted with a probability of $p(x_i)$. As a result, the initial uncertainty of x_i is $-log(p_i)$. After that, a y_i is recieved. And the final uncertainty of transmission of x_i is $-log p(x_i/y_i)$. Mutual information system has the following relationships:

$$I(x_i, y_j) = -log p(x_i) + log p(x_i/y_j) \tag{31}$$

$$I(x_i, y_j) = log p(x_i/y_j)/p(x_i) \tag{32}$$

I(X,Y) = average $I(x_i, y_j)$ for all values of I and j.

$$I(X,Y) = \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) log p(x_i/y_j)/p(x_i) \tag{33}$$

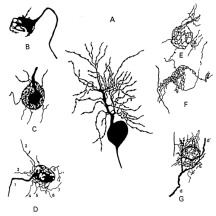


Figure 22: Spherical bushy cells (A) in anterior AVCN receive endbulbs of Held from only a couple of auditory nerve fibers (B,C,D). Globular bushy cells in posterior AVCN receive multiple inputs from just a few auditory nerve fibers (E,F,G). Each bushy cell responds only to a narrow band of frequencies.

The basic equations of sound propagation is as the following:

$$\frac{d}{dx}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & Z(x,s) \\ Y(x,s) & 0 \end{bmatrix}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} \tag{40}$$

where $Z = s\rho_0/A$ and $Y = sS/\eta P_0$, with A is the area of the tube. The above equation can be transformed as a second order equation in terms of the pressure P. The formulas are as the followings:

$$P' + ZU = 0 \tag{41}$$

$$U' + YP = 0 \tag{42}$$

Take the partial derivative of x for the both equations and gives:

$$P'' + ZU' = P'' - ZYP = 0 \tag{43}$$

Since the wave equation is

$$\frac{\partial^2 P}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P}{\partial x^2} \tag{44}$$

Therefore, $\omega^2/c^2 = ZY$.

The following part of homework is mainly about the defination of Joules and intensity.

A flash bulb puts out 50 Joules and it lasts for 20 μ s It will consume $50J/(20 \times 10^{-6}) = 2.5 \times 10^{-6}$. The displacement magnitude (RMS) of air particle in pm is even smaller than a single Hydrogen atom, while such small number is still meaningful because of a collective result. The pressure at 0dB SPL is

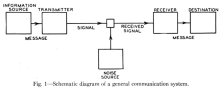


Fig. 1—Schematic diagram of a general communication system.

Figure 25: An information source which produces a message or sequence of messages to be communicated to receiving terminal. A transmitter which operates on the message in some way to produce a signal suitable for transmission over the channel. The channel is merely the medium used to transmit the signal from the transmitter to receiver. The receiver ordinarily performs the inverse operation of that done by the transmitter, reconstructing the message from the signal. The destination is the person for whom the message is intended.

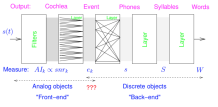


Figure 26: Model block diagram summary of speech recognition by humans. At the top of each block is a label that attempts to identify the physical operation, or a unit being recognized. The labels below the boxes indicate the probability measure defined at that level. See the text for the discussion of objects, at the very bottom. The speech $s(t)$ enters on the left and is processed by the cochlea (first block), breaking the signal into a filtered continuum of band-passed responses.

4.2.3 Defination of some basic acoustics

Pharynx is the cavity encompassed by Larynx and the nasal cavities; Hyoid is related ot the hyoid bone, a U shaped bone at the root of the tongue that support tongue muscles; Palate is the roof the mouth; Mandible is the lower jaw; Velum is soft palate or the back of the top of the mouth; Alveolar ridge is back ridge along the teeth; viscosity is the measure of resistance in a fluid, much like ohm.

4.2.4 Wave equations

French scientist D'Alembert first derive his solution to the wave equation in 1747. He solution is:

$$f(t - x/c) + g(t + x/c) \tag{50}$$

Starting from the basic transmission line equations with an area function given by $A(x) = A_0 e^{(2mx)}$, we can derive the corresponding Webster Horn Equation. From the basic definition:

$$\frac{\partial^2 p}{\partial t^2} + 2m\frac{\partial p}{\partial x} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{51}$$

From the second derivative equation, we can get the solution just by observation. The solution is $P(x,s) = e^{mx} e^{\sqrt{m^2 + s^2}\, x/c}$

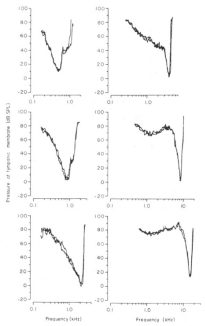


Figure 20: The figure shows the pressure of tympanic membrane versus frequency.

The relationships between information and entropy are as the followings:

$$I(X,Y) = H(X) - H(X/Y) = H(X) + H(Y) - H(X,Y) \tag{34}$$

3.15.2 Channel capacity

Channel capacity is maximum of Mutual information that may be transmitted through the channel. We cannot minimize H(Y/X) or H(X/Y) because they are the properties of channel. However, we can maximize H(x) and H(Y). If all the messages are equi-probable, H(X) or H(Y) is the largest.

$$C = I(X,Y)_m ax = H(X)_m ax - H(X/Y) \tag{35}$$

$$Efficiency_{\%} = I(X,Y)/C \tag{36}$$

Figure 23: octopus cells

20 μPa. And the characteristic impedance of air is $\rho_0 c = 407$ MKS-Rayls. Therefore, the displacement is $|X_l| = p/\omega z_0 = 25/\pi$ pm.

Assume a person is speaking at an intensity of I_0 66 dB-SPL at 1 meter. 66 dB-SPL is equal to a pressure of $20 \times 10^{-6} \times 10^{66/20} = 0.04$ according to the intensity to pressure equation.

$$I_0 = |P^2|/\rho_0 c = 0.04^2/407 = 3.93 \times 10^{-6} [Watts/m^2] \tag{45}$$

The area is a sphere with radius = 1. Therefore, the total speech power is about $16\pi \approx 50$ [μ W]

In Spherical coordinates, intensity I_0 can be varied as the following:

$$I(\theta, \phi) = I_o cos(\theta/2) cos(\phi/2) \tag{46}$$

from which θ is the angle in the horizontal plane and ϕ in the vertical plane. We can calculate total power P by integral of the I_0 over the area:

$$P = I_0 r^2 \times \int_{-\pi/2}^{\pi/2} cos(\phi/2) cos(\phi) d\phi \times \int_{-\pi}^{\pi} cos(\theta/2) cos(\theta) d\theta \tag{47}$$

Plug in all the numbers, the total power is about 60 % of that unchanged I_0.

4.2 Homework 2

This homework introduces basic units of acoustics, Larynx, wave equation and transmission line. Wave equatios, webster horn equations and reflectance would be mentioned for many times in the rest of semester.

4.2.1 Speech

The first part of this homework is the similar content of the homework 1. To light a 60 watt, we need get about 10 times of people on earth to talk at 20 dB-SPL at 1m. 20 dB SPL equals to 200

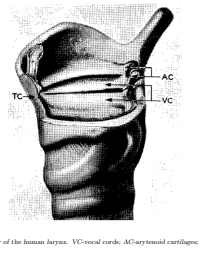


Figure 27: *Cut-away view of the human larynx. VC-vocal cords; AC-arytenoid cartilages; TC-thyroid cartilage*

4.2.5 Reflectance

Assuming characteristic impedance $z_0(x, s)$, the input impedance of a transmission line can be derived by

$$Z = P/U = \frac{P_+ + P_-}{U_+ - U_-} = z_0 \frac{1+R}{1-R} \tag{52}$$

From above equation, we can solve for R

$$R = \frac{Z_L - z_0}{Z_L + z_0} \tag{53}$$

Two transmission lines are in cascade, the first one having an area of a and the second one having an area of b, with lengths L_1 and L_2 respectively, terminated with a resistor $r = \rho c/A$.

$$R(s) = \frac{\frac{a\rho}{b} + \frac{a\rho}{a}}{\frac{a\rho}{b} + \frac{a\rho}{a}} e^{-s2L_1/c} \tag{54}$$

This is exactly the model of the last problem of Homework 1.

In this homework, we analyze the impedance of the following two circuit.

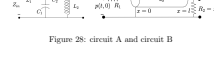


Figure 28: circuit A and circuit B

In Figure 28,

$$Z_i n = sL_1 + \frac{1}{sC_1}[(\frac{1}{sC_2} + sL_2) = sL_1 + \frac{(1 + s^2 L_2 C_2/C_1)}{s(C_1 + C_2 + s^2 L_2 C_1 C_2)} \tag{55}$$

Figure 21: stellate cells

3.16 Lecture 12/11

3.17 Lecture 26/11

Articulation is identification of "Nonsense" sound; Intelligibility is the identification of meaningful sound. cochlea has a lot of output. The nature of noise are: (1) coocstall party effect; 50 people talking at the same time; (2) noise generator like matlab 20 bands contain speech in the range below uniform distribution form 300Hz to 2kHz.(not discuss details) Give great credit to the team. They made a lot of stuff like microphones,filters, even computer(not the digital computer nowadays).

e_k assumes that exist. We come up a model: s(t) represents signal input; n(t) is the noise; Guassion assumption makes sense E_{max} (estimation) $x(t) = s(t) + n(t)$ is a model that estimate:

$$E_i = s(t1) + n = n(t2) + s \tag{37}$$

likelihood ratio is the ratio of two distributions = P_s/P_N

3.18 Lecture 7/12

The last two lectures went through all the homework and provided suggestions on writing a good report.

4 Review of all the homework

4.1 Homework 1

The first homework gives us a general introduction of basic acoustics and units.

It introduced the formula of the speed of sound that we used frequently through the whole semester.

$$c = \sqrt{\eta P_0 \rho} \tag{38}$$

The left is the speed of sound c and on the right are $\eta = c_p/c_v = 1.4$, the barometic pressure $P_0 = 10^5$ [Pa], and the density ρ_0 [kgm/ m^3]. In the formula, ηP_0 represents the compressibility of air, where P_0 is the weight of air above us that depends on humidity, wind and other variables rather than temperature. The static density in the formula ρ_0 depends on both pressure P and temperature T.

$$\rho(P_0, T) = 1.29 \frac{P_0}{10^5} \frac{273}{T} \tag{39}$$

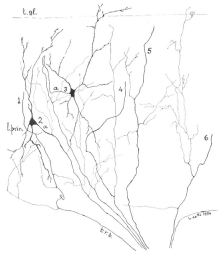


Figure 24: multipolar cells

$\times 10^{(}-6) = 2 \times 10^{-4}$ [Pa]. From HW1,

$$I_0 = |P^2|/\rho_0 c = 4 \times 10^{-8}/407 = 10 \times 10^{-10} [Watts/m^2] \tag{48}$$

And the power is $4\pi I = 1.26 \times 10^{-9}$ Watts. Therefore, the number of people is N = 60/1.26 $\times 10^{-9}$ = 5billion. If all the people are speaking at 60 dB-SPL, leading to create 100 times in pressure and therefore 10000 in power. Thus need 5 million people.

A phone is any speech sound, such as a consonant or a vowel. Phones that have meaning are called phonemes. On average, people speak 5 phones per second.

4.2.2 Anatomy of the vocal system

A picture of the Larynx is as the following: The purpose of the Larynx is the source of sound for the vocal tract. Larynx is just a few centimeter across.

Acoustic impedence in MKS acoustic ohms is defined as

$$1[\Omega] = 1[Pa - s/m] = 1[N - s/ms] \tag{49}$$

Transfering from impedence allows us to derive many equal unites. We can convert from m^3/Nt t0 [F] scale by 10^6 based on $C = V/\eta P_0$; $1\ kg/m^4 = 1$ Henrys based on $M = \rho l/A$; $Nt - s/m^5 = 10^{-6}\Omega$ based on $R = (l/A)\sqrt{\rho\mu}$.

For circuit A,

$$\begin{bmatrix} P_i \\ U_i \end{bmatrix} = \begin{bmatrix} 1 & sL_1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ sC_1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1/sC_2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1/sL_2 & 1 \end{bmatrix}\begin{bmatrix} P_2 \\ U_2 \end{bmatrix} \tag{56}$$

4.3 Homework 3

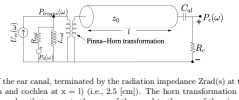


Figure 29: Model of the ear canal, terminated by the radiation impedance Zrad(s) at the tragus (x = 0), and by the eardrum and cochlea at x = l) (i.e., 2.5 [cm]). The horn transformation is the transformer that represents the concha, that converts the area of the canal to the area of the pinna/concha opening. This transformer ratio seems necessary to reduce the effective radiation mass seen by the ear canal, to improve the matching of energy between the canal and free-space. It does this by making the area defining Lrad larger by the transformer turns ratio. The effective turns ratio needs to be estimated from measured data

In Homework 3, we simulated the middle ear using a circuit model with transmission line.

As shown in the figure, the free field sound pressure, defined as Pi0() acts as a source in series with the radiation impedance $Z_{rad}(s, A_{rad})$, composed of a combination of resistance R_{rad} and mass L_{rad}, where

$$R_{rad} = \rho c/A_{rad} \quad L_{rad} = r_o \rho/A_{rad} \tag{57}$$

which represents the local stored energy field. The two impedances are in parallel, thus

$$Z_{rad}(s) = sL_{rad}R_{rad}/(sL_{rad} + R_{rad}) = (sL_{rad} + R_{rad}) = 1/Y_{rad}(s) \tag{58}$$

The radiation admittance for a sphere is similar in form (it differs in the area): $Y_{rad} = 1/Z_{rad} = \frac{A_0}{\rho c_0} + \frac{A_0}{sc}$ To model the middle ear we first need to simulate the transmission line by using dAlembert solution of the wave equation.

In the left of the circuit in figure 29, $Z_{rad}(s) = \frac{sL_{rad}R_{rad}}{sL_{rad}+R_{rad}}$ At the cochlear end we terminate the line with an impedance: $Z_c = R_c + 1/sC_{ol}$ Each impedance may be converted into a reflection coefficient, defined as: $R(s) = \frac{V_r}{V_i} = \frac{Z(s)-z_0}{Z(s)+z_0}$

Finally is the question of the boundary conditions. These are dealt with by defining the end reflection coefficients $R(L, s) = \frac{U_r(L,s)}{U_i(L,s)}$ $R(0, s) = \frac{U_r(0,s)}{U_i(0,s)}$

Each reflection coefficient R is given by: $R(s) = \frac{Z(s)-z_0}{Z(s)+z_0}$ We can simply derive the following equation by characteristic impedance of the transmission line, where Z(s) is the load impedance and $z_0 = \rho c/A$ is the . Taking the case of x = 1 we find

$$R(L, s) = \frac{R_c + 1/sC_{ol} - z_0}{R_c + 1/sC_{ol} + z_0} \tag{59}$$

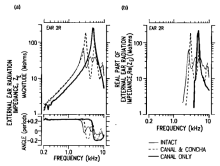


FIG. 10. Radiation impedance measured in one ear (2R), before surgical modification (intact), after removal of the pinna flange (canal and concha), and after removal of the concha (canal only). (a) Magnitude and angle of the impedance. (b) The real parts $\mathrm{Re}[Z_n]$'s of the impedances calculated from the data in (a). As was discussed in Sec. I, the accuracy of real-part estimation deteriorates as the impedance angle approaches 0.25 periods. The $\mathrm{Re}[Z_n]$'s in (b) are only plotted at frequencies greater than the low-frequency limit above which they rise monotonically to their maximum.

Figure 30: Z...

4.4 Homework 5

4.4.1 Transmission Line Simulation

Transforming the speech data into spectral form by Fourier analysis is a common way. Yet, this method do not provide a sufficiently accurate result of speech articulation. In this homework, we learned a new method to speech analysis and synthesis. We represent the speech waveform directly in the terms of time-varying parameters. Because we model the speech wave itself instead of its spectrum, we can completely avoid the several problems in the frequency domain method such as long speech segment leading to inaccurate analysis of rapidly changing speech and little information about the spectrum between pitch harmonics.

This homework is based on the paper "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave". We learn and apply that theory to our problem.

4.4.2 Model for parametric representation of the speech wave

The first step is to make a model of the signal. The excitation of the human vocal tracts, excited by a series of nearly periodic pulse, produces speech sound. In this method, we assume that vocal tract is a discrete time-varying linear filter and the changes of vocal tract shape can be viewed as a succession of the stationary shapes. Therefore, we define a transfer function in the z domain for the vocal tract so that we can represent the system by poles and zeros.

From Figure 31, we have predictor P with p delaus of one sample interval each. The output of

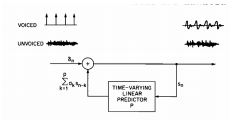


Figure 31: *Block diagram of a functional model of speech production based on the linear prediction representation of the speech wave. The input is the voiced sound produced by a pulse generator with adjustable period and amplitude. The unvoiced sound are produced by a white-noise source. The linear predictor P forms a weight sum of the past p samples at the input of the predictor. The output of the linear filter at the n sampling instant is* s_n

the linear fiter at the nth sampling instant is given by:

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + \delta_n \tag{60}$$

where the "predictor coefficients" a_k account for the fltering action of the vocal tract; and δ_n represents the nth samples of the excitation

The transfer function of the linear filter of Figure 31 is:

$$T(z) = 1/(1 - \sum_{k=1}^{p} a_k z_{-k}) \tag{61}$$

The linear filter has a total of p poles which are either real or occur in conjugate pairs. Additionally, the poles must be inside the unit circle because the linear filter need to be stable.

In order to represent the influence of the poles of the vocal tract transfer function adequately, the linear predictor memory must be equal tot twice the time required for sound waves to travel from the glottis to the lip. Normally, we assume vocal tract is 17cm in length so that the memory of the predictor should be roughly 1 msec to represent the poles (the speed of sound is roughly 340m/s). The corresponding value of p is 10 for a sampling interval of 0.1 msec plus another two poles required fir the glottal flow and the radiation added. $\mathrm{p} = 12$ is a rough estimate at sampling rate of 10kHz. In this homework problem, our sampling rate is 8kHz, which makes p smaller. p is a function of the sampling frequency f_s and is roughly proportional to f_s.

To sum up, the predictor coefficients a_k, the pitch period, the rms values of the speech samples, and a binary parameter indicating voiced or unvoiced speech, provide a complete representation of the speech wave over a time interval (we assume that the vocal-tract shape is constant). In this homework, we readjust these parameters periodically every 5 msec.

4.4.3 Speech Analysis

In Figure 31, we conclude that, samples of voiced speech are linearly predictable in terms of the past p speech samples,except for one sample at the beginning of every pitch period. Therefore, we can apply this property of the speech wave to determine the predictor coefficients. Define the prediction error e_n as the difference between the speech sample s_n and its predicted value $\hat{s}_n$is:

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k} \tag{62}$$

e_n is:

$$e_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} \tag{63}$$

It follows from Figure 31 that, every sample of the voiced speech waveform can be predicted from the past sample values. Therefore, the positions of individual pitch pulses can be determined by computing the prediction error e, given by Eq.4 and then locating the samples for which the prediction error is large.

4.4.4 Conclusion

In this homework, I loop through the speech frames of 5 [ms] and take a total of 20 [ms] segment. $A(n) \equiv a_n(k)$. $a_n(k)$ is the coefficient. A vector has form $[1, a_1, a_2, ..., a_k]$, where $\mathrm{K} = 12$. Therefore, the figure I plot has 12 roots. After that, I plot the prediction applying LPC method and compare it with the original signal.

4.5 homework 6

In homework 6, we explore spherical waves and speech coding.

$$p(r,t) = f(ct - r)/r + g(ct + r)/r \tag{64}$$

$$f(ct - r)/r = \frac{p_0}{r}\delta(t - (r - r_0)/c), for r \geq r_0 else 0, \tag{65}$$

which has a Fourier transform

$$F(\omega, r) = \frac{p_0}{r}e^{-j(r-r_0)\omega/c} \tag{66}$$

4.6 homework 7

In homework 7, we explored how to calculate the entropy of speech and expectationmaximization (EM) algorithm. EM algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.

5 My advice

This is the first time that I took an graduate-level course. I am thrilled that our class content is closely related to research. I read more research paper than the sum of any other undergraduate class. I hope that before we discuss a certain paper, professor can remind us and assign the paper as homework. It is hard for me to follow the class if I do not read the paper before.

3 FINAL REPORT: A/A-

This section consists of one report.

impulse response hit one of the filters, the different frequency component in it will have different reflection phases. Therefore the reflected signal is no longer an impulse any more. Different frequency component will be lagged for different extent, and the reflected signal will be stretched. In time domain the reflected signal is a continuous signal that will decay to 0 over time. This simulation is a very precious opportunity for me to visualize how sound signal behave when it bounces between two s-domain filters.

First we calculate the radiation impedance of the pinna (left side of the transmission line). By having some assumed values, we get

$$Y_{rad} = \frac{A_{rad}}{sr_c\rho} + \frac{A_{rad}}{\rho c} = \frac{2\pi {r_c}^2}{sr_c\rho} + \frac{2\pi {r_c}^2}{\rho c} = \frac{2}{z_0}\left(\frac{92000}{s} + 1\right),$$

$$Z_{rad}(s) = \frac{1}{Y_{rad}} = \frac{1}{2}z_0\frac{s}{s + 92000},$$

where ρc is the acoustic impedance of the free space, A_{rad} is the effective radiation area of the pinna, which is assumed to be a half-sphere attached to the pinna. And the pinna is assumed to be a circle with radius of r_c. Then we can find the frequency response of this reflection filter, as shown in Fig. 2.

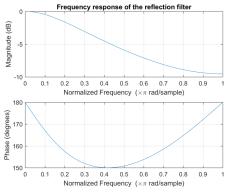


Fig. 2 Frequency response (magnitude and phase) of the pinna reflection filter

As we can see from Fig. 2, the reflection magnitude and phase is dependent on frequency. After obtaining the frequency-dependent reflection coefficient, we can then do the bilinear z transform on it. Only after we obtain the bilinear transform coefficient b(1), b(2), a(1) (normalized to 1), and a(2), we can setup the time-domain filter by difference equation:

$$y(m) = b(1)x(m) + b(2)x(m-1) - a(2)y(m-1).$$

The bilinear transform coefficients also depends on sampling frequency, and the sampling frequency is dependent how many sections we divide the ear canal into for simulation. The division process will be discussed later.

Then we can use similar processes to find frequency response and the bilinear coefficients of the middle ear. After we setup the time domain filters on both sides of the transmission line, we can incorporate them into our code to define what will happen if the sound velocity touches the ends of transmission line.

In the MATLAB program, total length of the canal from glottis to mouth is defined as L. The total length L is divided into N=10 samples, i.e. sections. Since there are 10 sections in total, there are 10 two-port systems connected to each other. Each 2-port system has a backward volume velocity vector U_n(N$_0$), defined on the left boundary, and a forward volume velocity vector, U_p(N$_0$), defined on the right boundary, where N$_0$ denotes the number of the section.

Sampling rate F_s is defined as $F_s = 1/T$, where T is the time that the sound (345 m/s speed) needs to spend to travel one section of the canal (tube, in this case). Since the question specify the sampling rate to be 44100 Hz, we can calculate that

L = 1/44100*10*345 = 0.0782 m = 7.82 cm

In my program, I took 100 samples (100 time steps) to simulate the wave propagation in the tube. The basic operation is that after one time step, all the forwarding vectors' value will shift to the right by one section, all the backward vectors' values will shift to the left by one section. The reflection filter operation takes place at the boundaries. At the pinna, not only the reflection is considered, but also the drive current is added up with the reflected vector.

Finally, we can save the time domain sound pressure at both ends, take FFT, and find the frequency-domain transfer function of the pinna-to-middle-ear transmission line, as shown in Fig. 3. This transfer function is very useful because with this information, by measuring the sound velocity at the pinna, we can determine how much sound power is delivered to the middle ear.

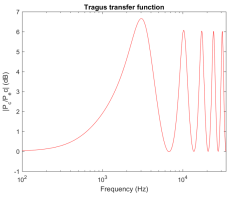


Fig. 3 The tragus transfer function. It is defined as the cochlea pressure (right end of the transmission) over pinna pressure (left end of the transmission line) in frequency domain

Vocal Tract Model

In this section, we extend the simulation method to a more complicated, 2-tube case. Specifically, we simulate human vocal tract as a 2-tube system. On its left end we generate a glottal velocity array as the initial voice source. When this glottal velocity propagates back and forth in the vocal tract, we need to construct the boundary conditions at the left end (glottis), the right end (lips), and the middle (interface between 2 tubes). Our 2-tube system is like a "voice box". On its right end (lips), the voice of vowels are sent out. With the same initial glottal velocity travelled inside 4 different 2-tube simulations of our vocal tract, it becomes 4 different vowels. I am glad that my vowel simulation files sound very close to the corresponding human pronunciation. Also, the formants F1, F2 I found are very close to the established results in the past paper.

There are 3 essential steps needed before coding. First, I found the 3 reflection coefficients at the glottis, the tube boundary, and the lips. Since there are four vowels we are supposed to simulate, there are four different sets of reflection coefficients in total. Here I take the first vowel /e/ as example.. Second, I found the glottal velocity $u_g(t)$ with modification of the provided code. This part will be skipped in this report because professor mentioned that the lab manual and the provided code had some wrong instruction on it. Third, I found new algorism at the boundary between two sections.

At the beginning, I will specify the parameters to divide and sample my vocal tract. I assume the sound speed in vocal tract is $c = 367\ m/s$, which is the sound speed in air at body temperature. Since the sampling rate we will use is the standard CD sampling rate $F_s = 44.1\ kHz$, one sample period will be $dT = 1/F_s = 22.6757\ \mu s$. In one sample period, the sound travels $dL = c/F_s = 8.321995\ mm$.

Our vocal tract is about 17 cm. Thus we need $N = 21$ spatial samples to reach a length of 17 cm. The total vocal tract length in our simulation will be $L = dL * N = 17.47619\ cm$. For the vowel /e/, we see that section one has length $L_1 \approx 9\ cm$ and section two has length $L_2 \approx 8\ cm$. Therefore, I will assign 11 spatial samples to section one and 10 spatial samples to section two. Thus,

$$L_1 = dL * 11 = 9.1542\ cm,$$

$$L_2 = dL * 10 = 8.3220\ cm.$$

Then I will find the reflection coefficients, and the corresponding z-transform. Since the area of tube 2 is $A = 1\ cm^2$, the corresponding radius is $a_2 = 0.0056419\ m$. So the radiation resistance and the radiation mass at lips are

$$R_{rad} = \frac{0.459\rho_0 c}{{a_2}^2} = 6.244686\ M\Omega$$

$$L_{rad} = \frac{0.27\rho_0}{a_2} = 56.470338\ kg/m^4$$

Where z_c, the characteristic impedance of section 2 is $z_2 = \rho_0 c/A_2 = 4.3306\ M\Omega$. Then the radiation impedance and the lip reflection coefficient can be found:

$$Z_{rad} = \frac{sL_{rad}R_{rad}}{sL_{rad} + R_{rad}} = \frac{s * 56.47 * 6.24M}{s * 56.47 * 6.24M} = 6.24M\frac{s}{s + 110583.471}$$

$$\Gamma_{lips}(s) = \Gamma_{rad}(s) = \frac{Z_{rad}(s) - z_2}{Z_{rad}(s) + z_2} = \frac{\left(\frac{R_{rad}}{z_2} - 1\right)s - 110583.471}{\left(\frac{R_{rad}}{z_2} + 1\right)s + 110583.471} = \frac{0.44199s - 110583.471}{2.44199s + 110583.471}$$

Then I use MATLAB to perform bilinear z-transform at 44.1 kHz, and get the coefficients $A = [1, -0.3215]$, $B = [-0.2197, -0.4588]$. Then, I can write out the lips' reflection filter in time domain:

$$y(m) = b(1)x(m) + b(2)x(m-1) - a(2)y(m-1)$$
$$or, y(m) = -0.2197x(m) - 0.4588x(m-1) + 0.3215y(m-1)$$

Here, if the lip area is time-varying, like in the real case, then the radiation impedance at the lips is time-varying. Therefore, the reflection coefficient at the lip boundary becomes time-varying. Then the recursion relationship in the time-domain will become time-dependent and have time-dependent coefficients:

$$y(m,t) = b_1(t)x(m) + b_2(t)x(m-1) - a_2(t)y(m-1).$$

After defining boundary condition at lips, I will find the reflection coefficient on the boundary of two tubes. The positive going wave that travel from section 1 to section 2, and the negative going wave that travels from section 2 to section 1 have reflection coefficients Γ_p and Γ_n, respectively.

$$\Gamma_p = \frac{z_2 - z_1}{z_2 + z_1} = \frac{\frac{1}{A_2} - \frac{1}{A_1}}{\frac{1}{A_2} + \frac{1}{A_1}} = \frac{1 - \frac{1}{8}}{1 + \frac{1}{8}} = \frac{7}{9} \qquad \Gamma_n = \frac{z_1 - z_2}{z_1 + z_2} = \frac{\frac{1}{8} - 1}{\frac{1}{8} + 1} = -\frac{7}{9}$$

At this stage, together with the provided glottis reflection coefficient $\Gamma_g = 0.95$, we have found all the reflection coefficients of the filters we need.

In the third preparation step, I will add a new algorism at boundary between two sections. Section 1 has 11 spatial samples (N1=11). Section 2 has 10 spatial samples (N2=10). Therefore, at the boundary between two sections, the next state of the positive going wave after (right next to) the boundary, should be equal to the current state of positive going wave at boundary, minus its reflected part, plus the reflected negative going wave at the boundary.

The algorism for negative going wave is similar, the next state of the negative going wave after (left next to) the boundary, should be equal to the current state of negative going wave at boundary, minus its reflected part, plus the reflected positive going wave at the boundary, namely:

$$U_{p_{next}}(N1+1) = U_p(N1) - \Gamma_p U_p(N1) + \Gamma_n U_n(N1+1);$$

$$U_{n_{next}}(N1) = U_n(N1+1) - \Gamma_n U_n(N1+1) + \Gamma_p U_p(N1);$$

Plug in the values, we have

$$U_{p_{next}}(N1+1) = U_p(N1) - \frac{7}{9}U_p(N1) - \frac{7}{9}U_n(N1+1);$$

$$U_{n_{next}}(N1) = U_n(N1+1) + \frac{7}{9}U_n(N1+1) + \frac{7}{9}U_p(N1);$$

After obtaining all these information, it is very straightforward to code. After the coding, we save the time-domain pressure at lips, it is very straight forward. It is equal to the positive velocity plus the negative velocity at the lips (x=L), multiplied by the characteristic impedance of tube 2. With drive current as an impulse, we save the time-domain lip pressure and take FFT of it. Then we get the frequency impulse response of this vocal box, as shown in Fig. 4.

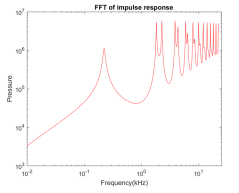


Fig. 4 Impulse response obtained from taking FFT of the time-domain impulse response. F1 is at around 210 Hz, F2 is around 2 kHz.

Finally, I change the drive current from impulse to the glottis velocity array we've got in preparation step, and write the lip pressure into sound file "Vowel_e.wav". It sounds reasonably close to the pronunciation of "/e/". Then I plot a 100ms waveform from middle of the sound in Fig. 5. We find that there is higher "concentration" at low amplitude.

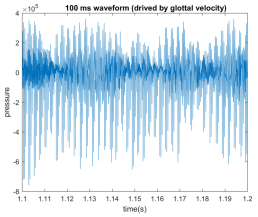


Fig. 5 100 ms waveform of the pressure at lips, with the glottis side driven by glottal velocity

To simulate the other vowels, the sections length, characteristics impedance, radiation impedance and reflectance all need to be modified. The form below presents all these parameters for different vowels.

Vowel	N1	z_2 (MΩ)	$\Gamma_{rad}(s)$	Bilinear z-transform	Γ_p	Γ_n
/e/	11	4.3306	$\frac{0.44199s - 110583.471}{2.44199s + 110583.471}$	A=[1,-0.3215] B=[-0.230,-0.459]	$\frac{7}{9}$	$-\frac{7}{9}$
/æ/	5	0.5413	$\frac{0.44199s - 39097.161}{2.44199s + 39097.161}$	A=[1,-0.6927] B=[-4.5e-4,-0.307]	$\frac{7}{9}$	$\frac{7}{9}$
/a/	11	0.6187	$\frac{0.44199s - 41796.623}{2.44199s + 41796.623}$	A=[1,-0.6749] B=[-0.011,-0.314]	$-\frac{3}{4}$	$\frac{3}{4}$
/ɔ/	21	0.7218	$\frac{0.44199s - 45145.513}{2.44199s + 45145.513}$	A=[1,-0.6534] B=[-0.024,-0.323]	0	0

I obtained pretty good results for the other three vowels, too, and the formants I found are also reasonably accurate.

Linear Predicting Coding (LPC)

The LPC filter analysis and synthesis is a very powerful audio compression technique. The LPC theorem proposes that, in a sampled speech, any sample can be expressed a linear combination of its previous p samples. The prediction of sample $\hat{x}(n)$ can be obtained by inputting the original speech sample series $x(n)$ into the transfer function $H(z) = -a(2)z^{-1} - a(3)z^{-2} - \cdots - a(p+1)z^{-p}$. If we translate $H(z)$ into time domain recursion relationship, we find that the value of $\hat{x}(n)$ is dependent on the previous p speech samples. This recursion relationship is

$$\hat{x}(n) = -a(2)x(n-1) - a(3)x(n-2) - \cdots - a(p+1)x(n-p)$$

The LPC filter, after calculating the predicted samples $\hat{x}(n)$, subtracts $\hat{x}(n)$ from the original speech sample $x(n)$ to get the error $e(n)$, as shown in Fig. 6. Therefore, in summary, LPC filter takes in the original speech signal $x(n)$, find the corresponding coefficient vector $A(z)$ that minimize the error, and output the error signal $e(n)$. The importance of these operations is, if we put the error signal $e(n)$ back into the reverse LPC filter, we can reconstruct the original speech $x(n)$.

The amazing part is, when the unvoiced part of the speech (noise) is put into LPC filter, the output $e(n)$ will also be noise, because noise is not predictable. On the other hand, if the voiced part of the speech is put into LPC filter, the output $e(n)$ will be the addition of noise and some "pulses". These pulses correspond to the human speech. Therefore, if we pass a speech with noise through LPC filter, zero the noise in $e(n)$, and use $e(n)$ and $A(z)$ to reconstruct the speech. The reconstructed speech will be less noisy than the original speech.

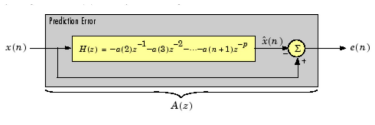


Fig. 6 LPC filter A(z) with order of p, the input and output of the LPC filter

In our homework, we use MATLAB code to realize an LPC filter, plot out the error signal $e(n)$, and do formant analysis. We first divide the provided speech "WhenAllElse8k.wav" into frames of 5ms. Since the sampling frequency of the speech is 8 kHz, each frame contains 40 samples. Then, we call a certain frame, together with its 3 previous frames, a "segment". To determine the LPC coefficient vector for one frame, we analyse the whole segment, which contains this particular frame and its 3 previous frames, with order of analysis p (ex. $p = 8$). Therefore, each frame will have a coefficient vector. In our case, there are 654 frames in total, and thus 654 coefficient vectors. Each vector has a dimension of $p + 1$.

After the coefficient vector is obtained for all frames, we input the original speech, frame by frame, into the time-varying LPC filter. The coefficient vector of LPC filter is time-varying because different frame have different coefficient vector. We save the filter state and swap to a new coefficient vector when encountering a frame boundary:

$$[out\_frame, filt\_state] = filter(A(frame\_number), 1, input\_frame, filt\_state);$$

The "$filt\_state$" saves the current filter state, so in the next loop (increment of "$frame\_number$") it can be loaded into the filter operation. In this way, we obtain the LPC filter output $e(n)$, frame by frame. With the LPC coefficient vectors, we can also perform formant analysis. According to Atal, B. and Hanauer, S.'s paper "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", we will need to find the roots of the polynomial function with coefficient $(1, a(2), a(3), \ldots a(p+1))$, i.e. the coefficient vector for the particular frame. Then we discard the redundant negative roots and use the equation

$$F_k = \left(\frac{1}{2\pi T}\right) Im(\ln z_k),$$

$$B_k = \left(\frac{1}{2\pi T}\right) Re\left(\frac{1}{\ln z_k}\right),$$

to find the formant candidates. Finally, I use the criteria that the formants should be greater than 90 Hz and have bandwidths of less than 400 Hz to find the formants in the particular frame.

First I tried with analysis order $p = 8$, the error signal goes high at around 1.3 s, as shown in Figure 7. This means that the LCP prediction does not work well in that region. I guess that the analysis order is not enough so that sometimes LCP coefficients are not good enough for prediction.

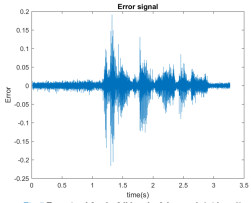


Fig. 7 Error signal for the full length of the speech (with p=8)

As shown in Figure 8, if we zoom into t = 0 – 0.3s, which corresponds to the unvoiced part of the original speech, we see noise, as we expected in the introduction part. There are some pulse-like low-magnitude spikes here, however, because the environmental noise has some sort of "pitches", which are recognized as pulses in error signal. The results here (Figure 8) is already much better than the case of the 48k-version audio, which I used at first. Then because of this problem, I switch to the 8k-version audio.

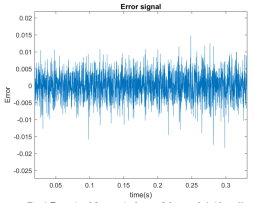


Fig. 8 Error signal for unvoiced part of the speech (with p=8)

As shown in Fig. 9, if we zoom into $t = 2 - 2.4s$, which corresponds to the voiced part of the original speech, we see the pulses, as we expected.

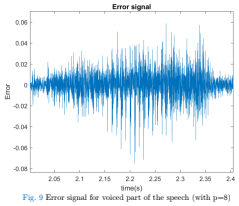


Fig. 9 Error signal for voiced part of the speech (with p=8)

Figure 10 is the scatter plot of formants (in kHz) versus time. This looks very similar to the formants plot in Atal, B and Hanauer, S.'s paper, except that there are some discontinuities in the lines. I notice that these discontinuities happen at the time when LCP error signal is goes very large. It makes sense because in those regions the LCP does not work well.

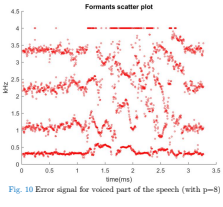


Fig. 10 Error signal for voiced part of the speech (with p=8)

Then I tried with higher analysis order p = 15. The high-error region is still there but the spikes reduces from 0.2 to 0.15.

In this lab, we go through the procedures of LCP filters with MATLAB. Previously, I read the paper on LPC analysis and synthesis but had very limited understanding on LPC technique. After this lab, and professor Allen's office hour, I think my understanding has greatly deepened. I understand how exactly LPC filter is implemented, what is the importance of error signal, how to reconstruct the original speech, and how to analyze formants based on the prediction coefficient vectors. I learned a lot from this lab.

Spherical waves

The general solution to the spherical wave equation can be written as

$$p(r,t) = f(ct-r)/r + g(ct+r)/r \ ,$$

where p is pressure, r is the radius, c is the speed of sound, and t is time. An outward traveling impulse solution, and its retrograde wave, in the time domain is

$$f(ct-r)/r = \frac{p_0}{r}\delta(t-(r-r_0)/c), \text{for } r \geq r_0 \text{ else } 0,$$

$$g(ct+r)/r = \frac{p_0}{r}\delta(t+(r-r_0)/c), \text{for } r \leq r_0 \text{ else } 0,$$

which has a Fourier transform

$$F(\omega,r) = \frac{p_0}{r}e^{-j(r-r_0)\omega/c} \ ,$$

$$G(\omega,r) = \frac{p_0}{r}e^{j(r-r_0)\omega/c} \ .$$

Since the retrograde wave propagates inwards, we assume that at t = 0 the pressure pulse is at $r = r_0$, therefore, for all the spatial points with $r = r_0$, their pressure as a function of time will be an impulse at t = 0. For the spatial points with $r < r_0$, the pressure as a function of time will be a delta function at $t = (r-r_0)/c$. The physical interpretation is that, the wave takes $t = (r-r_0)/c$ to reach the spatial point. The plotted pressure as a function of time is in Fig 11.

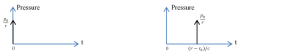


Fig. 11 The pressure as a function of time, left side is for spatial points with $r = r_0$, right side is for spatial points with $r < r_0$

If we take the integral of the delta function of the pressure function over time, we get p_0/r for the spatial points at a radius of r. Then the total energy passed by at certain radius will be

$$E = 4\pi r^2\left(\frac{p_0}{r}\right)^2/z_0 = 4\pi p_0^2/z_0 = 4\pi/410 = 0.03065\ J$$

We assume that the "intensity" here actually means the integral of intensity over that very short time that the pulse passes by, otherwise it will be a delta function. The intensity at radius of r will be

$$I(r)\Delta t = \frac{p_0^2}{z_0 r^2}\Delta t \text{, where } p_0 = 1 \text{ Pa}, z_0 = 410 \text{ Acoustic Ohm}$$

Now let us consider an inward-bound wave that starts at 1m radius. The wave solution is

$$g(ct+r)/r = \frac{1}{r}\delta(t+(r-1)/c), \text{for } r \leq 1 \text{ else } 0$$

, assuming that the pressure is 1 Pa at 1 meter at initial condition. When the wave reaches 1mm, the time and intensity will be

$$t_{1mm} = \frac{0.999\ m}{345\ m/s} = 0.0029\ s = 2.9\ ms$$

$$I_{1mm}\Delta t = \frac{1}{410 \times 0.001^2} = 2439\ J/m^2$$

If we assume that the delta function in reality is a rectangular pulse with duration of 1 second, then the intensity will be

$$I_{1mm} = 2439\ W/m^2$$

We found that for an inward-bounded retrograde spherical sound wave, the intensity is so large at small radius, and the time it takes for sound to travel there is so short. I think this property is very valuable. For example, it will be good to utilize this technique to break up kidney stones. The high intensity will guarantee that the kidney stones are destroyed, and the short time it takes guarantees the accuracy.

Speech Compression and Quantization

Quantization of signal level is unavoidable when we save the digital waveform for a speech. We can also reduce the speech file size by using a very simple lossy speech codec. It encodes and decodes the speech, reducing the number of quantization levels, i.e. reducing number of bits for representation. We will try to reduce the bit rate as much as possible providing that we can still understand the speech.

The original speech has 16 bit/sample. First we strip the sign and 15 bits left. Given the compression number C=2, we take the square root and 8 bits left, fix the numbers to be integers (quantization step), and then square it and put the sign back on. After these steps we successfully obtain 9 bits/sample speech. Why we reduce quantization level, i.e. do compression in this way? Because with this method, there are less quantization levels at high signal magnitude. For example, after the same magnitude after square root are fixed to integers: 1,2,3,4,5 and then squared back: 1,4,9,16,25, we can see that the level separation becomes larger and larger as signal magnitude goes higher. This is efficient because for speech the signal waveform concentrate in low-magnitude range. There are less samples with higher magnitude. Similarly, we can do C=5 and C=10 compression. The sample histogram for C=2, C=5 and C=10 are shown in Fig. 12. Due to compression, there are less and less available sample magnitude levels. The speech waveform comparison between original one and the one after C=5 compression is shown in Fig. 13. We can clearly identify the position of new quantization levels after compression. Surprisingly, at C=10, I can still recognize the speech.

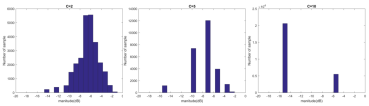


Fig. 12 Histogram for C=2, 5, 10 from left to right

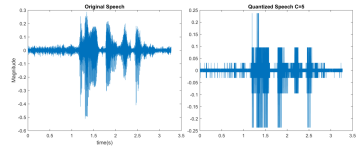


Fig. 13 The original speech waveform (left) and the waveform after C=5 compression (right)

II. Psychoacoustics

In this section, we will discuss about signal processing in cochlea, masking in speech perception, and human speech recognition. In the section of cochlear signal processing, we will introduce the basic electrical model of the cochlear, and the functions of different types of cells. In the section of masking in speech perception, we will focus on simultaneous masking (two tone suppression) and self-masking which is related to internal JND. In the section of human speech recognition, we will discuss the stages of human speech recognition, and the relationship among them.

Signal processing in cochlea

The simple model of cochlea is treated as an uncoiled long thin box, as shown in Fig. 14a [8]. The Base (x=0) is the high-frequency end of the cochlear (up to 20000 Hz), where it is stiff and narrow. The Apex (x=L) is the low-frequency end of the cochlear (down to 20 Hz), where it is wide and floppy. Helmholtz (1885) modeled the cochlea with a bank of highly tuned resonators selective to different frequencies in his book *On the Sensations of Tone*. Each string corresponds to a place X on the basilar membrane. Then Wegel and Lane, and Fletcher provided deeper insight into the model. The electrical equivalent circuit of cochlea, modelled by 1924 Wegel and Lane, is shown in Fig. 14b [8]. They also quantitatively showed how a high-level low frequency tone affects masks a second low-level higher-frequency tone by increasing its audible threshold. The details about the masking will be discussed in next section. Finally, G. von Bekesy (1924) found that the cochlea is analogous to a dispersive transmission line where signals with different frequencies travel with different speed along the basilar membrane.

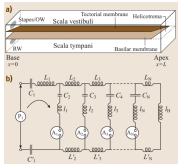


Fig. 14 (a) basic 2-D box model of the cochlea and (b) 1924 Wegel and Lane electrical equivalent circuit

Then the basilar membrane (BM) impedance, a one-port impedance will be discussed in order to have deeper understanding of the cochlear model. The BM impedance is

$$Z_p(s,X) = \frac{K_p(X)}{s} + R_p(X) + sM_p \ ,$$

where K_p is the spring constant, or stiffness, M_p is the mass, and R_p is the partition resistance. The stiffness dominates at base, but decreases exponentially along the cochlea. The mass is independent of place so it dominates at the apex, as shown in Fig. 15a [8]. In this figure of BM impedance vs. place X, it is also shown that for different frequency input signal, there is a local minimum, a *hole*, at different place. It means that different places X have different resonant frequencies. In other word, different place X is designated to receive sound of different frequency. This resonant frequency, as a function of X, is expressed as

$$F_{cf}(X) = \frac{1}{2\pi}\sqrt{\frac{K_p(X)}{M_p}} \ ,$$

which is also called cochlear map function, as shown in Fig. 15b [8]. The inverse function can be used to find location of the hole for different input frequency, as shown in Fig. 15a [8]. The derivation of the cochlear map function is first based on counting critical bands as shown by Fletcher and popularized by Greenwood. Then the direct observation of cochlear map in cat (cochlear length L=21mm) was made by Liberman and Dodds. The empirical formula to fit the data is

$$F_{cf}(X) = 456(10^{2.1(1-X/L)} - 0.8) \ .$$

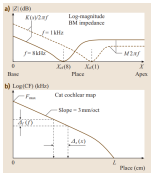


Fig. 15 (a) Impedance as function of location X, for two different input frequencies (b) Cochlear map of cat according to Liberman and Dodds. The critical bandwidth and the critical spread are related through the map

Next, we will examine the cochlear structure in the scales of tissues and cells. And discuss their functions. The basic functions of cochlear are to separate input sound into different overlapping frequency bands, as we discussed before, and to compress the large intensity range input into smaller electrical dynamic range of the inner hair cell. It is like a floating point converter.

The three major chambers in cochlea, as shown in Fig. 16 [8], are separated by Reissner's membrane and Basilar membrane (BM), and are all filled with fluids. Mechanically there are only two chambers as Reissner's membrane only provides electrical isolation of the scala media (SM). Inner hair cells (IHC) and outer hair cells (OHC) sit between BM and tectorial membrane (TM). The TM shears against the reticular lamina when BM moves up and down. It leads the cilia of IHC and OHC to bend.

After passing through the cochlear frequency band filters discussed before, a low-level pure tone has a narrow spread of excitations that excites the cilia of around 40 contiguous inner hair cells. As discussed before, the inner hair cell has a narrow bandwidth, as can be seen in the tuning curve, and it has a center that depends on the location along the BM. For each inner hair cell, there are typically three outer hair cells. Outer hair cells are used for intensity dynamic-range control, which is a form of nonlinear signal processing. Telephone speech is similarly compressed.

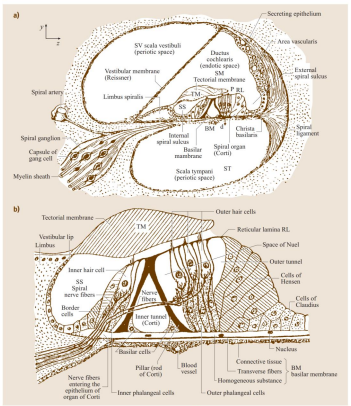


Fig. 16 (a) shows the structures of the cochlea and (b) zoom in to the Tectorial Membrane (TM), Basilar Membrane (BM), and the inner hair cells and outer hair cells in between them.

At the end of this section, we will focus to the different type of cells in the ventral cochlear nucleus and their functions. Sound, the vibration of air molecules, after the processing of cochlea, becomes oscillating potential to arrive the ventral cochlear nucleus. The Excitatory Postsynaptic Potential (EPSP) can create depolarization which excites the neuron cells to fire action potential, or spikes, while the Inhibitory Postsynaptic Potential (IPSP) can cause hyperpolarization current which inhibit the neuron cell to fire spike (even when no sound is presented, most neurons are active and fire spikes spontaneously).

The spike firing in neuron cells let us receive the loudness and frequency of the input sound signal. How? For an input signal with larger amplitude, the neuron cells will fire more frequently, i.e. spike rate increases. Our perception of loudness is actually the sum of all the spikes. If every neuron cell fires more rapidly, then of course the total number of spikes is larger. Thus we will perceive the sound to be louder. For the frequency (pitch) detection of input signal, Fig. 17 [8] shows the basic mechanisms. The left diagram shows a simple case while the input sound is low-level. All individual neurons fires periodically at around the positive peaks of the input tone. Then the aggregate results will be periodic with the same frequency as the input tone. The right diagram in Fig. 17 shows the more general case. Note that the result is like a rectifier of the input waveform, at IPSPs the neurons firing will be inhibited. Even the spontaneous firing when no sound is present will be inhibited. Therefore, the frequency information is coded very successfully.

The thresholds of auditory nerve fibers vary regularly. The low threshold fibers are on the pillar side of the inner hair cell and have high spontaneous rate. The high threshold fibers are on the modiolar side. The nerves responds to a limited frequency range, with the frequency center determined by its location in cochlea along BM, as discussed before. This is called tuning. The fibers at base are tuned to high frequencies and the fibers at apex are tuned to low frequencies.

There are different types of Ventral Cochlear Nucleus Neurons, whose properties are summarized in Fig. 18 [8]. **Stellate cells** fire regularly in response to tones. It is linear for both de- and hyperpolarization. It fires more rapidly to a higher intensity and sum EPSPs (not over a wide frequency range). They respond to tones with a chopper-like PST. **Bushy cells** respond to only a narrow band of frequencies. Bushy cells are nonlinear, hard to depolarize but linear when hyperpolarized. This means that for long depolarization they only produce one or two action potentials. They do not integrate over time, thereby preserving the temporal characteristics of the auditory nerve input. The temporal information can be used for sound localization. **Octopus cells** are situated to integrate information across frequency (detect synchrony). They are activated by synchronous activity of many auditory fibers, and are able to fire very rapidly (1000 spikes/s). They are also nonlinear, hard to depolarize and also hard to hyperpolarize. So it responds to a tone with *onset* (very short duration) PST histogram. Finally, **multipolar cells** projects to the DCN (Dorsal Cochlear Nucleus), the AVCN (Anterior Ventral Cochlear Nucleus), and cochlear nucleus on the other side of the brainstem.

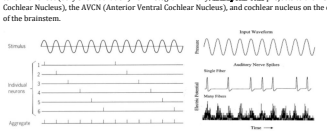


Fig. 17 Tone frequency coding (left: simplified case right: more general case)

	Spherical Bushy cells	Globular Bushy cells	Stellate cells	Multipolar cells	Octopus cells
Tuning	Narrow	Narrow	Narrow	Broad	Broad
Intrinsic Membrane Properties	Nonlinear	Nonlinear	Linear	Linear	Nonlinear
PST Histogram	Primary-like	Primary-like with notch	Chopper	Onset - Chopper	Onset
Function	Preserve timing information	Preserve timing information	Loudness - integrate in critical band	Loudness -- overall (AGC?)	Synchrony - integrate across freq.
Project to - send info	Superior Olivary Complex	Superior Olivary Complex *	Inferior Colliculus	Cochlear nucleus on other side	Lateral Lemniscus

Fig. 18 Summarization of characteristics of Ventral Cochlear Nucleus Neurons

Masking

The JND is sometimes called self-masking. We also have a concept called simultaneous masking (opposite to non-simultaneous masking). Both concepts will be discussed in this section.

The simultaneous dynamic-masking experiments consist of upward spread of masking and neutrally measured two-tone suppression. They are closely related dynamic-masking phenomena. Mayer was the first one to find the asymmetry of masking. Then in 1923 and 1924, Fletcher and Wegel and Lane published quantitative results of tonal masking. I really love Wegel and Lane's paper in 1924. They explain everything so clearly that even for a person out of the area like me can understand the experiment quickly. They first gave the definitions of masking. First, the magnitude of a tone is defined as the ratio of its pressure to that of its minimum audible value. If a minimum audible pressure of a tone, p_1, increases to p_2 because of the introduction of a second tone. Then p_2/p_1 is defined as the magnitude of masking of the first tone by the second tone. In all the experiments, the masking tone is designated by F_1 and the masked tone is F_2. With these information, we can analyze the results in Fig. 19 [8] directly, where each sub-figure has a different masking tone, with frequency F_1 (highlighted), and in a specific sub-figure the masked tones frequencies are indicated on each curve. From the figure, we can get some evident facts. If the masked tone has much lower freq. than the masking tone, then it is very hard to be masked, even when the magnitude of the masking tone is raise to very high. On the other hand, if the masked tone has much higher freq. than the masking tone, although it is still not affected too much when masking tone is in low intensity, its masking magnitude (reflect how much it is masked) raise rapidly when the masking tone is increased in high-intensity range. Generally, when two tones lie closer, the masking curve is straight and has 45 degree slope, intercepting Masking tone = 10 in x axis. From these results we can see the upward spread of masking: a loud low freq. signal can mask a high freq. signal significantly, but a loud high freq. signal cannot affect a low freq. signal too much. One thing the authors point out is that in these curves, when the two tones are too close to beat (the curves that have 45-degree slope and pass the origin), the masking curves turn to have different meanings. Such two tones, separately inaudible but not lower than 1/2 of the audible threshold, will beat obviously and become alternately audible and inaudible.

The two-tone suppression (2TS), I would say, is closely related to the upward spread of masking. Its experiment involves neural tuning curve measurement and neural firing rate measurement so it goes

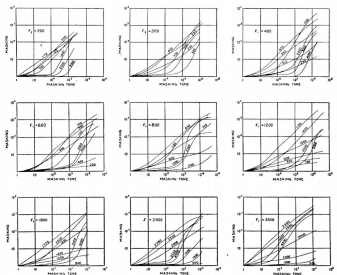


Fig. 19 Masking for tones in same ear

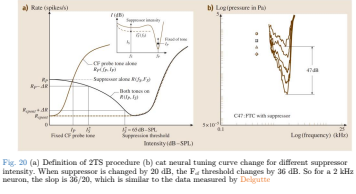


Fig. 20 (a) Definition of 2TS procedure (b) cat neural tuning curve change for different suppressor intensity. When suppressor is changed by 20 dB, the F_2 threshold changes by 36 dB. So for a 2 kHz neuron, the slop is 36/20, which is similar to the data measured by Delgutte

deeper into the neural scale and has great significance. As shown in Fig. 20a [3], the CF probe tone is fixed so that its rate is fixed. The intensity of the suppressor is gradually increased so that the firing rate of the neuron gradually goes down. Another very clear example of 2TS is the neural tuning curves in Fig. 20b [3]. We see that as the suppressor tone is turned higher and higher, the threshold of the tuning curve increases quickly. This means that the audible threshold of the masked tone increases quickly.

The masking reflects the nonlinearity nature of the cochlea. Another example for nonlinearity is the resonant tectorial membrane (RTM) model, described by Allen and Neely. When the excitation tone is increased from 14 to 124 dB, the BM stiffness changes and thus the resonant frequency shift to the base, as shown in Fig. 21a [4]. In Fig. 21b [4], it is shown that when a low-freq. suppressor is turned on, the high-freq. probe tone is not only suppressed but also shifts towards the base. It is because the BM stiffness is decreased with increased input level (because of the add-in of the suppressor).

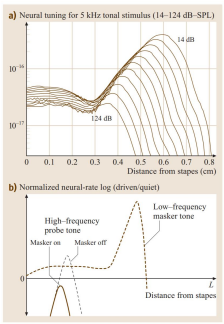


Fig 21. (a) Compression in NL RTM model, as the BM stiffness decreases, the peak is reduced and shift toward the base. (b) After the low-freq. suppressor is turned on, the high-freq. probe tone is suppressed and move towards the base because of the reduction of BM stiffness.

Next, I will talk about JND (Just Noticeable Difference). It is called self-masking as an indication that it is determined by the internal noise of the auditory system. As we discussed before in the neural cell section, the "loudness" we perceive is due to the temporal summation of the spikes of the neural cells. Therefore, each time we hear a tone pulse, the loudness we perceive is not fixed but follows some probability distribution. In 1846, Weber first hypothesized that the higher the sound (physical intensity), larger the JND (intensity JND), i.e. Weber's Law:

$$\frac{\Delta I}{I} = \text{constant}\ .$$

The significance in this law is that it reflects that the ear is a floating point convertor. The ear has dynamic-range convertors inside, ex. Outer hair cell, so that it can deal with very loud sound. However, when dynamic-range is large, the resolution is lower because the "number of significant digit" is fixed. Therefore, as the intensity goes higher, the intensity JND increases. The problem in Weber's Law is formulated only in physical domain, but actually the noise is internal. As found by Riesz in his experiment (near-miss to Weber's Law), intensity JND over intensity is not a constant.

In 1860, Fechner dealt with the internal psychophysical noise, on the basis of Weber's Law. He assume that Weber's Law is correct, and also assumed that the loudness JND is a constant, independent of the loudness. In this way, he proposed the relationship between loudness and intensity, i.e. Fechner's Law by Counting JNDs.

$$N_{JND} = \int \frac{dL}{\Delta L(L)} = \int \frac{dI}{\Delta I(I)} \quad \text{(Counting of JND)} \quad \rightarrow \quad L(I) \propto \log(I) \quad \text{(Fechner's Law)}$$

Fechner's Law was also wrong because both of his assumptions were wrong. The loudness JND is not independent of the loudness. Instead, it is

$$\Delta L(L) = \sqrt{L}.$$

As loudness goes up, the loudness JND also goes up.

In 1933, Fletcher and Munson obtained loudness growth data based on loudness additivity, as shown in Fig. 22 [8]. Based on the data, Steven's Law was formulated:

$$L(I) = I^{\upsilon}, \text{where } \upsilon \approx 1/3.$$

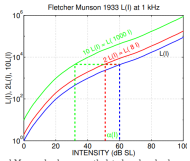


Fig. 22 Fletcher and Munson loudness growth data based on loudness additivity.

This is an amazing result. For a pure tone to reach 10 times loudness in human's perception, it needs to have 1000 times the physical intensity! Also it is shocking that for a pure tone with its intensity enhanced by 1000 times, its loudness perceived by human will only enhance by 10 times. So human ear is born with a μ-law compressor! Our auditory systems are very efficient and protective.

Human Speech Recognition

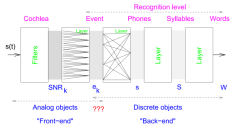


Fig. 23 A very nice illustration of human speech recognition.

As shown in Fig. 23 [5], when a sound signal enters our cochlea, it will pass through a bunch of the filters. For example, one major filter is the ear canal, as we discussed in Part I. After this stage the signal is characterized by SNR (signal-to-noise ratio).

Then, the signal passes through filter banks along the BM, as we just discussed. Then the outer hair cells at each frequency band will act as a dynamic-range convert, i.e. floating point converter. Then the potentials will induce spikes and other activities of the neuron cells in ventral cochlear nucleus and other cochlear nucleus (this happens in each band). After these processes, each band will contribute some information for our recognition of phones. When the information which come from different frequency band adds up and complement each other, we have more and more information for correct recognition of a phone. Therefore, if we divide the incoming nonsense phone into k frequency bands, the probability that we can correctly recognize the phone will decrease if signal in some of the bands are missing. The more bands we are provided, the more likely we can correctly recognize the phone. This accords with our intuition. But the great thing Fletcher did was that he successfully put this into a practical formula. If the phone is divided into k frequency bands, also the error rate (of recognition) when a certain band k is individually provided is e_k, then the probability of correct recognition of nonsense phone when all bands are provided, s, will be

$$1 - s = e_1 e_2 \ldots e_k$$

$$= (1 - s_1)(1 - s_2) \ldots (1 - s_k)\ .$$

It is also found that when the SNR varies, the single-band error rate can be expressed as

$$e_k = e_{min}^{\ SNR_k/k}\ .$$

Therefore the total error

$$e = 1 - s = e_{min}^{(SNR_1 + SNR_2 + \cdots + SNR_k)/k}$$

We can also find the probability of correct recognition of nonsense syllable, S. For a CVC syllable, since there are three phones contained, the correct recognition of nonsense syllable requires that all three phones are correctly recognized. Therefore

$$S = s^3$$

III. Information Theory

Shannon's channel capacity

Assuming that for a given channel, N(t) is the number of allowed signals of duration T, Shannon defines the capacity of a discrete channel to be

$$C = \lim_{T \to \infty} \frac{log N(T)}{T}$$

We notice that in information theory, log operation appears a lot. In this formula, the base of the log is not specified. I think if we use bit (0 or 1) as the unit of C, the base of the log should be 2. Then, for a certain channel, we should have some finite number of allowed sequences, $S_1, S_2 \ldots S_n$, whose durations are $t_1, t_2 \ldots t_n$, for n types of signals. Then we can do the **recursion** (this word was not mention in Shannon's paper but I think it fits very well). To count the total number of allow signals, since the last signal t can be any of the allowed sequence, with durations of $t_1, t_2 \ldots t_n$, we do recursion of one signal back to express the total number of allowed signal N(t) as

$$N(t) = N(t - t_1) + N(t - t_2) + \cdots + N(t - t_n)$$

This is a finite difference equation (a filter, in some sense). According to a well know property, when t is very large,

$$C = \lim_{t \to \infty} \frac{log N(T)}{T} = log X_0$$

where X_0 is the largest real solution of the characteristic equation:

$$X^{-t_1} + X^{-t_2} + \cdots + X^{-t_n} = 1$$

As an example, in telegraphy case, the allowed sequences are Dot (2 units of time: line close to 1, open for 1), Dash (4 units of time: close for 3, open for 1), Letter space (3 units of time: line open for 3), and Word space (line open for 6). Therefore, it seems like

$$N(t) = N(t - 2) + N(t - 4) + N(t - 3) + N(t - 6)$$

However, this is not correct, because there is another restriction in this channel: there cannot be a word/letter space followed by a word/letter space. Before a word/letter space, it can be either a dot (2 units of time) or a dash (4 units of time). This rule is not contained in the difference equation above. To incorporate this rule, we examine one more signal back for the case when a word/letter space is at the end.

$$N(t - 3) \rightarrow N(t - 3 - 2) + N(t - 3 - 4)$$

$$N(t - 6) \rightarrow N(t - 6 - 2) + N(t - 6 - 4)$$

Therefore, the difference equation becomes:

$$N(t) = N(t - 2) + N(t - 4) + N(t - 5) + N(t - 7) + N(t - 8) + N(t - 10)$$

By using solving the characteristic equation, we can find that

$$C = 0.539\ .$$

Another formula proposed by Shannon for channel capacity applies to the case when the signal average power P is in additive to white Gaussian noise with power N. For a bandwidth W (when talking about white spectrum, it is important to specify the bandwidth), the channel capacity is

$$C = W log_2(1 + P/N)\ \ bits/s\ .$$

This is the signaling rate for the best possible encoding.

Information and entropy

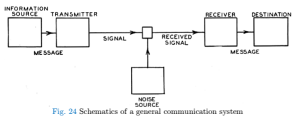


Fig. 24 Schematics of a general communication system

Shannon's information theory begins with a vector of probabilities $[p_k]$, k=1…K. When the transmitter, as shown in Fig. 24 [8], sends a message signal k to the receiver through the channel, due to the noise source, the message has a probability to be sent correctly (send k, receive k), or in error (send k, receive one of the other k-1 signals). There is a probability confusion matrix with dimension of k*k that describes all these probabilities. The probability vector obeys

$$\sum_k p_k = 1.$$

And the information density is the reciprocal of the probabilities:

$$i_k = log_2 \frac{1}{p_k}.$$

As an example, in a channel, if a message such as a phone is often transmitted, it contains very little information because there is a tons of syllables/words that contains this phone. Even we receive this phone correctly, we need to receive a lot more phones to have the full recognition. However, imagine that now there is a phone that is very seldomly transmitted, and only a few words contain it. Then after receiving this phone, we have much more information. Therefore, lower probability of occurrence means more information.

Entropy measures the extent of uncertainty inside a random variable or the outcome of a random process. In Shannon's information theory, entropy is defined as the expected value of the information:

$$\mathcal{H} = \mathcal{E}[I_k] = \sum_k p_k * log_2 \frac{1}{p_k}.$$

As an example, if there is only one type of data in the dataset, we have the least uncertainty, and thus the least entropy, $\mathcal{H} = 0$.

Zipf's Law

Zipf's law is an empirical law which states that many types of data in physical and social sciences can be approximated with a Zipfian distributions. As an example in English language, the frequency/probability of any word is inversely proportional to its rank in the frequency table. This means that the most frequent word will occur approximately twice as often as the second most frequent word. In one of our homework, we made word frequency table in Shakespeare's "As You Like It". The word appearance probability vs. its rank in the word frequency table is plotted in Fig. 25. An approximately reverse-proportional relationship can be observed.

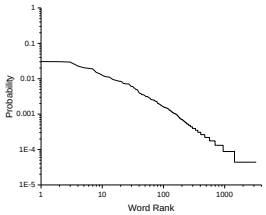


Fig. 25 Zipf's Law type plot

EM Algorism

Expectation-maximization (EM) algorism is an iterative method to obtain the maximum likelihood estimates of parameters in statistical models. For example, in our homework we are provided with a data file, and we are told that all the data in it belong to two unknown Gaussian distributions, but we do not know which data belongs to which Gaussian distribution. Our mission is to use EM algorism to find, with most likelihood, what are the two Gaussian distributions (their mean and variance) that these data come from.

How does this work? At first, two random Gaussian PDFs, G1 and G2, are generated. Then for all the data X, the relative probability that the data X belongs to G1 (Probability of G1 pile * G1's probability density at X) and it belongs to G2 (Probability of G2 pile * G2's probability density at X) are calculated. Then this data X is partially assigned into G1 and partially assigned to G2. The proportion depends on the relative probability that X belongs to G1 or G2. After doing this to all data in the dataset, new mean, variance and pile weight sets are calculated for G1 and G2 based the assignment we just performed. G1 and G2 will very likely move after this one iteration. Then another iteration start, until it converges (does not move any more). In my case, the MatLab program shows that 91 iterations are performed to obtain the final results, which converges.

In Fig. 26, the two Gaussian distributions are shown. Note that the pile probabilities have been multiplied to the two gaussians respectively. So the sum of the integrals of two Gaussian PDF will be 1. The existence of pile probability means that the two piles have different weights, i.e. the two piles have different probability of being taken data. The dataset is also plotted (as circles). But since it is too dense in the middle, it is hard to visualize it.

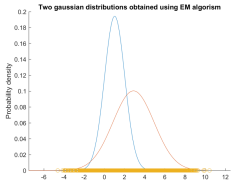


Fig. 26 Two Gaussian distributions obtained using EM algorism

4 FINAL REPORT: B+ and below

Only the 9 pages (2-10) of the remaining six reports are presented, one page per report. Page 1 has been removed to restrict the identity of these six authors.

Basic Acoustics

The speed of sound is given by the formula:

$$c_0 = \sqrt{\frac{compressibility}{stiffness}} = \sqrt{\eta_0 \frac{P_0}{\rho_0}}$$

$\eta_0 = \sqrt{\frac{c_p}{c_v}} = 1.4[dimensionless]$ is a physical constant. c_p is the heat capacity of gas under the constant pressure condition. c_v is the heat capacity of gas under the constant volume condition. $P_0 = 1 \times 10^5[Pa]$ is the barometric pressure. It is independent of the temperature because it only depends on the total weight of the air above us. The weight of air does not depend on temperature. $\rho_0[kgm/m^3]$ is the density of air. It is dependent on the temperature and the pressure. This relationship is given by:

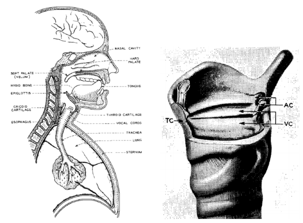
$$\rho(P_0, T) = 1.29 \frac{P_0}{1 \times 10^5} \frac{273}{T} [kgm/m^3]$$

where P_0 is in Pascal and T is in Kelvin. This implies that the speed of sound c_0 is dependent on temperature T and is proportional to the square root of temperature $\sqrt{T}$.
The intensity of sound is $I = \frac{|P|^2}{\rho c}$. The reference intensity is $I_{ref} = 1 \times 10^{-12}\mathrm{W\,m^{-2}}$. The corresponding reference pressus is $P_{ref} = 2 \times 10^{-5}$pascal. This can be verified by:

$$\frac{|P_{ref}|^2}{\rho_0 c_0} = \frac{|2 \times 10^{-5}|^2}{407} = 1 \times 10^{-12} = I_{ref}$$

Intensity indicates the power of the sound over unit area. Therefore, the total power of a sound source can be found by integrating the intensity over a surface area enclosing the source. The intensity in decibel can be calculated by $10\log_{10}(\frac{I}{I_{ref}})$. Similarly, the sound pressure in decibel can be calculated by $20\log_{10}(\frac{P}{P_{ref}})$. For example, in homework 2 problem 1 we were given that each person is talking at talking at 20 dB-SPL measured at 1 meter. Here, SPL stands for sound pressure level. Therefore, the sound pressure created by each person is $P = P_{ref} \times 10^{20/20} = 2 \times 10^{-5} \times 10 = 2 \times 10^{-4}$[Pa]. The intensity is $I = \frac{|P|^2}{\rho_0 c_0} = \frac{|2 \times 10^{-4}|^2}{407} = 1 \times 10^{-10}[\mathrm{W\,m^{-2}}]$. Assuming that instensity is uniform, the power of each person talking is $P = Area \times I = 4\pi \times 1^2 \times 1 \times 10^{-10} = 1.26 \times 10^{-9}$[W]. In order to light up a 60W light bulb we need $\frac{60}{1.26e-9} = 5 \times 10^{10}$ people talking.

Human Vocal System and Speech



(a) Human Vocal System Adapted from [1, fig. 2.1] (b) Human Larynx Adapted from [1, fig. 2.2]

Figure 1: Anatomy of Human Vocal System

Figure 1a shows the various parts of the human vocal system. Larynx is the voice box holding the vocal cords. The cavity between larynx and the nasal cavity is the pharynx. 1b shows a cut-away view of the human larynx. On the graph, AC stands for arytenoid cartilages, TC stands for thyroid cartilage, and VC stand for vocal cords. The slit between the vocal cords is called the glottis.

	Continuant [+continuant]	Discontinuant [−continuant]
Sonorant [+sonorant]	Glides (/w,j/) and Semivowels (/l,r/)	Nasals (/m,n,ŋ/)
Obstruent [−sonorant]	Fricatives (/f,v,θ,ð,s,z,ʃ,ʒ,h/)	Stops (/p,b,t,d,k,g/) and Affricates (/tʃ,dʒ/)

Table 1: Four Groups of Consonants Characterized by Sonorant and Continuant Adapted from [1, Table 2.2]

Speech sounds can be categorized in different ways. Voiced sounds are the sounds created by the vibration of the vocal cords. In contrast, unvoiced sounds are created without the vibration of the vocal cords. Vowels are the sounds produced without narrow constriction on the vocal tract. They are usually voiced. They can be further characterized by the degree of constriction, the tongue hump position, and whether they are nasalized. Consonants are the sounds produced with some constriction in the vocal tract. Consonants can be divided into four groups depending on whether they are sonorant and whether they are continuant. Sonorant means the consonant is produced without increased air pressure in the vocal tract. Continuant means the consonant is produced without complete closure in the vocal tract. The opposite of sonorant is called obstruent. The opposite of continuant is called discontinuant. Continuant and sonorant.Table 1 shows the four groups of consonants characterized by the two features. Stop consonants are also called plosives. Consonants are also described by their positions of articulation. For example, bilabial sounds are created by the two lips, palatal sounds are created at the tip of the tongue and the roof of the mouth, and alveolar sounds are produced at the tip of the tongue and the back ridge along the teeth.
A phone is a sound of speech. A phoneme is a spoken sound that is meaningful in the context of a specific language. For example, a nasalized vowel and its non-nasalized version are two distinct sound. Therefore they are two different phones. In the context of English, nasalizing a vowel does not change the meaning of the word. Therefore, the two phones are represented by the same phoneme. The International phonetic alphabet (IPA) is a standard way to represent different phones.

Wave Equation and Transmission Lines

Newton's second law givens that: $-\nabla\rho(\boldsymbol{x}, t) = \rho_0 \frac{\partial}{\partial t} u(\boldsymbol{x}, t)$. Hooke's law gives that: $-\nabla \cdot u(\boldsymbol{x}, t) = \frac{1}{\eta_0 P_0} \frac{\partial}{\partial t} \rho(\boldsymbol{x}, t)$. Combining the two equations we get the scalar wave equation:

$$\nabla^2 \rho(\boldsymbol{x}, t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} \rho(\boldsymbol{x}, t)$$

where $c_0 = \sqrt{\frac{\eta_0 P_0}{\rho_0}}$ is the velocity of the sound.
In 1747, D'Alembert found that the solution to the one-dimensional wave equation is: $f(t - x/c) + g(t + x/c)$. This is simply a superposition of a forward-going wave and a backward-going wave.
The acoustic equations

$$-\frac{d}{dr}\begin{bmatrix} P(r,\omega) \\ V(r,\omega) \end{bmatrix} = \begin{bmatrix} 0 & sM(r) \\ sC(r) & 0 \end{bmatrix} \begin{bmatrix} P(r,\omega) \\ V(r,\omega) \end{bmatrix}$$

describes the relations between pressure and velocity of the acoustic wave. They are analogous to voltage and current in the electric transmission line.

$M(r) = \frac{\rho_0}{A(r)}$ is the unit=length mass. $C(r) = \frac{A(r)}{\eta_0 P_0}$ is the unit-length compliance.
These equation can be used to derive the Webster horn equation:

$$\frac{1}{A(r)} \frac{\partial}{\partial r} A(r) \frac{\partial}{\partial r} p(r, t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} p(r, t)$$

It allows the area of the horn to vary along the direction of wave propagation instead of being uniform. It is based on the quasi-static approximation, which requires the frequency of the wave to be small enough. The critical frequency is $f_0 = \frac{c_0}{2d}$, where d is the diameter of the horn.
The reflectance is defined as $\Gamma = \frac{P_-}{P_+} = \frac{V_-}{V_+}$. In a transmission line with characteristic impedance z_0 terminated with a load impedance Z_L, the reflectance is: $\Gamma = \frac{Z_L - z_0}{Z_L + z_0}$. For a tube with cross-section area A, the characteristic impedance is $z_0 = \frac{\rho c}{A}$. The reflectance equation can be derived using the fact that the reflected wave adds to the total pressure but subtracts from the total velocity. To satisfy the boundary condition, we need $\frac{P_+ + P_-}{U_+ - U_-} = z_0 \frac{1+\Gamma}{1-\Gamma} = Z_L$.

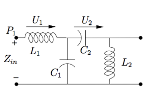


Figure 2: Transmission Line Circuit for ABCD Matrix, Adapted from Homework 2

The ABCD transmission matrix method is a simple way to calculate the input-output relation in a transmission line. We can divide the circuit into simple components. The circuit consisting of several components in cascade can be described by multiplying the ABCD matrices of each components. Figure 2 shows a simple circuit with four components. Its transmission matrix can be obtained by multiplying the transmission matrices of all the components:

$$\begin{bmatrix} P_1 \\ U_1 \end{bmatrix} = \begin{bmatrix} 1 & sL_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{sC_2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{sL_2} & 1 \end{bmatrix} \begin{bmatrix} P_2 \\ U_2 \end{bmatrix}$$

This method is much simpler than the traditional circuit analysis method.

Simulation of the Middle Ear

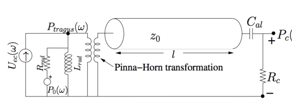


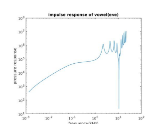
Figure 3: Model of the Middle Ear, Adapted from Homework 3

Figure 3 shows the model we used to simulate the middle ear. The ear canal is model by a transmission line with characteristic impedance $z_0 = \frac{\rho c}{A_{canal}}$. At the cochlea side, the transmission line is terminated by a capacitor C_{al} and a resistor R_c. At the tragus side, the transmission line is terminated by the radiation resistance $R_{rad} = \frac{\rho c}{A_{rad}}$ and the radiation inductance $L_{rad} = \frac{r_c \rho}{A_{rad}}$ in parallel. A_{rad} is the effective radiation area. The pinna-horn effect is modelled by a transformer.
At the cochlea side, the reflectance $R(L, s) = \frac{R_c + \frac{1}{sC_{al} - z_0}}{R_c + \frac{1}{sC_{al}} + z_0}$. In order to simulate this, we can first use bilinear transform to convert the from Laplace domain to Z domain. Then, we can use the resulting coefficients to define a filter.

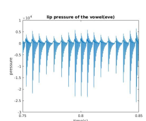
Simulation of Vowel Sounds

In homework 4, we synthesized four vowels sounds by modelling human speech production. Specifically, we model the glottis as a velocity source in parallel with an admittance, the lips as a mass in parallel with a resistance, and the vocal tract as two tubes in cascade. The length of vocal tract is approximately 17 cm. The speed of sound at body temperature is 367 m/sec. We set the sample rate at 44.1 kHz. The distance sound travels during one sample period is $\frac{367\mathrm{m/s}}{44.1\mathrm{kHz}} = 8.3 \times 10^{-3}$m. Thus, we need the vocal tract to be 21 samples long and have a length of $21 * 8.3 \times 10^{-3}\mathrm{m} = 17.43$cm. The density of air is $\rho_0 = 1.18\mathrm{kg/m^2}$.
In this model, we assume that the area of the glottis is constant. The boundary condition at the glottis is modelled by a reflection coefficient $\Gamma_g = 0.95$. This describes the admittance part of the model.
To illustrate the effect of the velocity source, we first define a pitch signal f_0 with duration 1.5 seconds:
$f_0(t) = 150 + 100 \times t/1.5 + 25 \times sin(2 \times \pi \times t/1.5) + 0.2 \times \sin(2 \times \pi \times 20 \times t)$.
The lip is modelled as a mass in parallel with a resistance. Taking the vowel /e/ as an example. The area of lip is $A = 1\mathrm{cm}^2 = 1 \times 10^{-4}\mathrm{m}^2$. The

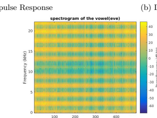
corresponding radius $a = \sqrt{\frac{A}{\pi}} = 5.6 \times 10^{-3}$m. The mass $m = \frac{0.27 \times \rho_0}{a} = 56.47$H. The resistance $r = \frac{0.459 \times \rho_0 \times c}{a^2} = 6.24 \times 10^6$ohm. Because at 1 kHz $\frac{r}{s \times m} = \frac{6.24 \times 10^6}{2 \times \pi \times 1000 \times 56.5} = 17.6 >> 1$, the impedance at the lips is $z = \frac{s \times r \times m}{r + s \times m} \approx s \times m$. The characteristic impedance of the vocal tract just before the lips is $z_0 = \frac{\rho_0 \times c}{A} = 4.33 \times 10^6$ohm. The reflectance at the lips is $\Gamma(s) = \frac{z - z_0}{z + z_0} = \frac{56.47s - 4.33 \times 10^6}{56.47s + 4.33 \times 10^6}$. Taking the bilinear transform, we get $\Gamma(z) = \frac{0.0698 - z^{-1}}{1 - 0.0698z^{-1}}$. In the time domain, we have $u_m(t) = 0.0698 \times u_p(t) - u_p(t - T) + 0.0698 \times u_m(t - T)$. If the area is time-varying, we will have time-varying mass and resistance. Assuming the change of lip area is slow, we can compute the reflectance at each time step and use it to find the time domain relationship at each step.
The vocal tract is modelled by two tubes with areas A_1 and A_2. Their characteristic impedance are $z_1 = \frac{\rho_0 \times c}{A_1}$ and $z_2 = \frac{\rho_0 \times c}{A_2}$.
In tube 1: $u_{1m} = u_{1p} \times \frac{z_2 - z_1}{z_2 + z_1} + u_{2m} - u_{2p}$.
In tube 2: $u_{2p} = u_{2m} \times \frac{z_1 - z_2}{z_1 + z_2} + u_{1p} - u_{1m}$.



(a) Impulse Response



(b) Lip Pressure



(c) Spectrogram

Figure 4: Simulated Responses of the Vowel /e/

The pressure response is computed by applying bilinear transform on the impedance $Z(s) = \frac{srm}{r + sm} = \frac{2.40 \times 10^7 s}{1.04 \times 10^6 + 23.06s}$. This gives $Z(z) = \frac{6.8841 - 6.8841 \times z^{-1}}{1 - 0.323 \times z^{-1}}$. Therefore, the time domain relation is:
$p(t) = 6.8841 \times u(t) - 6.8841 \times u(t - T) + 0.323 \times p(t - T)$.
For each of the vowels /e/, /æ/, /a/ and /ə/ we plot the frequency impulse response. Then, we drive the glottis with u_g and plot the pressure waveform at the lips as well as the spectrogram of this waveform. The results for the vowel /e/ is shown is figure 4. The formants F_1 and F_2, of the generated data, are much larger compared to those measured by Peterson and Barney (1952) [3].

Loudness and JND

JND stands for just noticeable difference. The loudness JND and the intensity JND are measures of the internal perceptual noise. They are the amount of increment in loudness or intensity that makes the difference noticeable. The loudness JND ΔL is proportional to the internal noise $\sigma_L(L)$. Intensity is a physical property of the sound. In contrast, loudness is subjective. It describes the perception of the sound.
Weber's Law states that the intensity JND ΔI is proportional to the intensity I. The ratio $\frac{\Delta I}{I}$ is defined to be the Weber Fraction. Weber's Law does not always hold. Riesz used two tones 3 Hz apart to show that Weber's Law does not hold for pure tones. This is called the near-miss to Weber's Law.
Fechner's Hypothesis states that the loudness JND is constant. Fechner's idea is that the number of loudness JND's should equal the number of intensity JND's. Assuming that Weber's law holds:

$$N_{JND} = \int_{L_1}^{L_2} \frac{dL}{\Delta L(L)} = \frac{L_2 - L_2}{\Delta L} = \int_{I_1}^{I_2} \frac{dI}{\Delta I(I)} \propto \int_{I_1}^{I_2} \frac{dI}{I} = \log(I_2 - I_1)$$

This is Fechner's Law: $L(I) \propto \log(I)$.
Fletcher and Munson demonstrated the additivity of loudness. For a tone f_1 with intensity I_1 they adjusted the intensity I_2 of the tone f_2 such that the two tones have equal loudness: $L(I_1, f_1) = L(I_2, f_2)$. Then, they played the two tones together to get the loudness twice of each tone played alone: $L(I_1, f_1, I_2, f_2) = L(I_1, f_1) = +(I_2, f_2)$. Increasing the intensity of f_1 to reach this loudness they found that the intensity was increased by 9 dB. Their data gives Stevens' Law:

$$L(I) = I^v$$

where $v \approx \frac{1}{3}$.

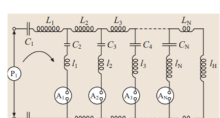


Figure 5: Equivalent Electrical Circuit of Cochlea. Adapted from [2, fig. 3.2b]

Cochlea Physiology

Figure 5 illustrates the electrical model of cochlea Wegel and Lane proposed in 1924 [4]. The inductors represent the mass of the fluid in cochlea. The capacitors represent the stiffness of the basilar membrane.
There are two types of auditory masking: neural masking and dynamic masking. Neural masking is caused by the internal noise that is measured by loudness and intensity JND's. Dynamic masking is caused mechanically by the nonlinear outer-hair-cells in the cochlea. There are two types of dynamic masking: non-simultaneous and simultaneous dynamic masking.
Non-simultaneous dynamic masking is also called forward masking or post-masking. It is the masking of a sound by the sound that immediately precedes it. The delay can be as long as 200 ms.

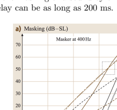


Figure 6: Upward Spread of Masking. Adapted from [2, fig. 3.6a]

Simultaneous dynamic masking is observed in upward spread of masking and two-tone suppression. Upward spread of masking is the effect when a low

frequency sound masks a high frequency sound. This effect is documented by Wegel and Lane in 1924 [4]. In their experiment, they used a masker tone with frequency $f_m = 400$Hz. For each masker intensity I_m and probe frequency f_p, they measured the threshold probe intensity I_p that made the probe tone audible. As illustrated in figure 6, the threshold probe intensity is plotted against the masker intensity for each probe frequency. When $f_p > f_m$, the thresholds and slopes of the cuves are larger than those when $f_p < f_m$ or when $f_p \approx f_m$. This shows that while the masking of higher frequency tones is small when the masker intensity I_m is small, it grows quickly after a threshold I_m is reached. The masking of higher frequency tones can be even greater than the masking of the tones in the critical band of the masker tone.

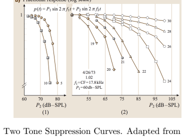


Figure 7: Two Tone Suppression Curves. Adapted from [2, fig. 3.6b]

Two tone suppression is similarly by Abbas and Sachs in 1976 [5]. They fixed a probe at the characteristic frequency of the neuron at an intensity several decibels above the tuning curve. They then created suppressor tones with different frequencies and varied their intensities. Finally, they computed the ration of the neuron firing rate when both the probe and a suppressor is present to the rate when only the probe is present. As illustrated in figure 7, the suppression is greater when the suppressor frequency is lower than the probe frequency. This is consistent with upward spread of masking if we consider the suppressor as a masker.

Auditory Pathway

The cochlear nucleus is divided into dorsal cochlear nucleus (DCN) and ventral cochlear nucleus (VCN). The VCN is further divided into anterior (AVCN) and posterior (PVCN) parts.
There are five types of neurons in the VCN: stellate cells, spherical bushy cells, globular bushy cells, multipolar cells and octopus cells. Each stellate cell is tuned to a narrow band. Its dendrites are parallel to the auditory nerve fibers. It receives inputs from only a few nerve fibers. It responds linearly to

=10 cm long. The barrel can be treated as a short piece of transmission line with a compliance $C = V_{barrel}/\eta P_0$. The neck will look like a mass $M = \rho h a_0 l/A_{neck}$. The mass and compliance constitute a resonance frequency which is

$$f_0 = \frac{c}{2\pi}\sqrt{A/Vl}$$

Using the given parameters, the resonance frequency is equal to 347.3 [Hz].

The acoustic equations can be represented in matrix form

$$\frac{d}{dx}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & Z(x,s) \\ Y(x,s) & 0 \end{bmatrix}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix}$$

From the matrix equation, we can write the partial differential equation as

$$P' + ZU = 0$$

and

$$U' + YP = 0$$

And the wave equation is

$$\frac{\partial^2 Q}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P}{\partial x^2}$$

The intensity of sound wave can be represented as

$$I_0 = \frac{|P^2|}{\rho_0 c}$$

The unit of intensity is usually W/m^2. This unit can be converted into log scale (dB-SPL) by using 20uPa as a reference pressure level. 20uPa is also the same as 10^{-12} $[W/m^2]$ according to the formula $I = |P|^2/\rho c$. By this definition, A sound intensity of $3.93uW/m^2$ is equivalent to around 66dB-SPL. 50 Billion people speaking at the same time at 20dB-SPL at a distance of 1m can light a 60W light bulb. Only 5 million people are required if the sound pressure is 60 dB-SPL instead.

A pure tone plane wave with a pressure of 0dB-SPL at 1kHz has a displacement magnitude $X_i = p/\omega z_0 = 25/\pi$ [pm]. This distance is much smaller than the radius of a single Hydrogen atom. However, since it is a collective average over a large amount of particles, the number is actually meaningful.

The previous discussions are about sound propagation in tubes with constant diameter. For horns, the area is changing in the direction of the propagation. The d'Alembert solution represents all waves as a linear combination forward going and backward going wave.

$$p(r,t) = f(ct-r) + g(ct+r)$$

The wave solution in horns can be acquired by solving the spherical wave equation.

$$\frac{\partial^2 p}{\partial x^2} + 2m\frac{\partial^1 p}{\partial x^1} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}$$

where A(x) depends on the propagation direction x. And the forward going wave is

$$f(ct-r)/r = \frac{p_0}{r}\delta(t-(r-r_0)/c)$$

Using a Fourier transform, the waves in the frequency domain is

$$F(w,r) = \frac{p_0}{r}e^{-j(r-r_0)w/c}$$

For example, a transmission line with an exponential change of area given as

$$A(x) = A_0 E^{2mx}$$

will have a solution in the form of

$$P(x,s) = e^{-mx}e^{\sqrt{m^2c^2+s^2}x/c}$$

For a rectrograde wave having a pressure of 1Pa at r_0m when t=0. The solution to the rectrograde wave as a function of time is

$$g(ct+r)/r = \frac{r_0}{r}\delta(t+(r-r_0)/c)$$

And the time domain plot at $r = r_0$ is shown in Figure 1.

The total energy of the wave over the time period is the integral of the power.The power of the wave corresponding to 1Pa is $20log_{10}(\frac{1Pa}{threshold})$ = $93.98dB-SPL$. Since the wave is a delta function, the total energy of this wave approaches 0. The total energy can be represent mathematically as $9.84\pi r_0^2\delta(t)$.The intensity of the wave is $(r_0/r)^2/407 = \frac{2.46r_0^2}{r^2}W/m^2$.

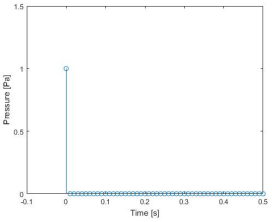


Figure 1: The plot of the retrograde wave as a function of time

2 Middle Ear & Modeling

The middle ear can be treated as a stub of transmission line with length of 2.5cm. The characteristic impedance of the ear canal can be calculated as

$$Z_0 = \frac{\rho c}{A} = 9.211M\Omega$$

The ear canal is terminated in a series combination of a stiffness . The experimental details of cat ear have been explored by Guinan and Peake. The model of the middle ear is shown in the figure below.

The diameter of the ear canal is around 0.75cm. The source is a combination of resistance $R_{rad} = \rho c/A_{rad}$ and mass $L_{rad} = r_e\rho/A_{rad}$. These two components are connected in parallel which gives a total impedance of

$$Z_{rad} = (\frac{A_{rad}}{\rho c} + \frac{A_{rad}}{sr_e\rho})^{-1}$$

The zero is located at s = 0 and the pole is located at -92000.

The load impedance is composed of the cochlear load resistor and capacitor. The load resistor is twice the value of the characteristic impedance which is

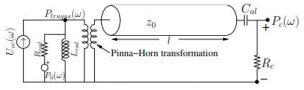


Figure 2: The model of the ear canal. It is terminated by a radiation impedance $Z_{rad}(s)$ at the tragus.

18.4MOhm. The impedance of the cochlear capacitance is the same as the load resistance at 0.8kHz. The capacitance value is 10.8pF.

With the load and source impedance, the reflection coefficient is

$$\Gamma = \frac{Z - Z_0}{Z + Z_0}$$

Through a bilinear transform, the reflection coefficients can be transformed from s-domain to z-domain where it can be implemented as a network of delay and gain. The reflection coefficients at the source end and the load end are respectively

$$\Gamma(L, z^{-1}) = 0.0347\frac{1 - 0.9011z^{-1}}{1 - 0.9689z^{-1}}$$

$$\Gamma(0, z^{-1}) = -0.4940\frac{1 - 0.0244z^{-1}}{1 - 0.5181z^{-1}}$$

The input impedance at the source end of the transmission line is

$$Z_L = Z_0\frac{Z_L + jZ_0 tan(wL/c)}{Z_0 + jZ_L tan(wL/c)}$$

where $Z_L = R_c + 1/(jwC_{ad})$

The transfer function is the ratio between the pressure at cochelea nd the pressure at the tragus.

$$H(w) = \frac{P_c(w)}{P_{tragus}(w)} = \frac{Z_{in}}{Z_{in} + Z_s}\frac{R_c}{R_c + 1/jwC}$$

To implement the wave propagation scenario in MATLAB, we can use two chains of registers. For every clock cycle, the signal from each register is pass on to the next one, representing wave propagating forward and backward. At

Figure 3: The implementation of wave propagating forward and backward using a series of registers.

the boundary, the transmission and reflection is determined by the reflection coefficient calculated based on the load/source impedance.

Based on the simulation, the input impedance and the transfer function of the system is shown in the figure below.

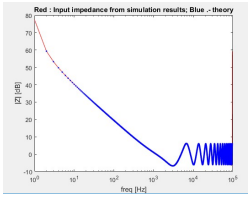


Figure 4: The input impedance of the middle ear with respect to frequency. The red line is from the simulation and the blue curve is from the theroy (equation)

If a 16 electrical ohm earphone delivers 120 dB SPL into ear canal with 1V rms input at 1kHz. The electrical power is $|V^2|/2R_e = 1/32W$. And the acoustic power is $0.5|P|^2/R$ where the cochlear resistance is $R_c = 2\rho c/A$ ohms. And the acoustic power can be evaluated as 22nW. The power ratio is 0.06%. That means this ear phone is very inefficient.

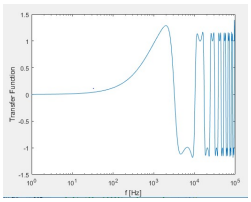


Figure 5: The transfer function of the middle ear system

3 Vocal Tract & Vowel Sound Simulation.

The source of the sound for the vocal tract is a Larynx. It is just a few cm across and it produces periodic signals. A picture of the Larynx is shown below.The pharynx is encompassed by the larynx and the nasal cavities

The system to generate audio sound consists of three part: glottis, vocal tract and lips. The glottis gives a pitch signal which serves as an put signal to the system. The vocal tract can be approximated by two tubes. In this case, it is simulated into two acoustic transmission lines connected in series. The two transmission lines have different length and area which are listed in the figure below

The lips can be modeled as a radiation impedance load which consists of a mass and a resistor. All the parts of the system will be discussed further in the later sections. The acoustic constant used in this model are measured at 37 degree celsius which is the temperature of human body. The density of the air $\rho = 1.18kg/m^3$. The speed of sound c = 367m/s.

The glottis is connected to the first section of the transmission line. The glottis serves as the source of sound. A high frequency signal is generated at the source as input of the network.This is the pitch signal. We model the pitch signal to have a frequency which is slowly rising from 150 to 250Hz.

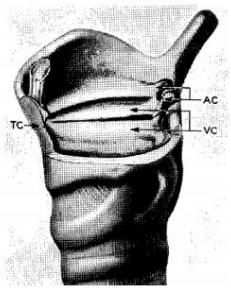


Figure 6: The cut-way view of a human larynx

Vowel	Section	Length [cm]	Area [cm^2]
/i:/ve	1	9	8
	2	8	1
/æ/t	1	4	1
	2	13	8
ɪ/ə/ther	1	9	1
	2	8	7
b/ə/rd	1	17	6
	2	0	6

Figure 7: The vocal tract can be simulated as two transmission lines. The picture shows approximate shape of the vocal tract when producing sound like /a/ in ɪ/ə/ther.

$$f_0(t) = 150 + 100 * t/\tau_0 + 25 * sin(2\pi * t)/\tau_0 + 0.2sin(2\pi * 20 * t)$$

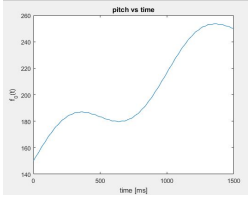


Figure 8: The pitch is the frequency of the input glottis signal. It is increased from 150Hz to 250Hz as shown in the figure.

The total duration τ_0 of the signal is selected to be 1.5sec. And the figure below shows the first 100ms of the signal as well as its FFT.

The forward going velocity wave in the first transmission line will be resulted from the input signal (pitch) plus the reflection of the backward going velocity wave at the glottis boundary. The reflection coefficient is selected to be 0.95 for simplicity

The lip boundary condition is more complex than the glottis boundary condition. The radiation load at the tragus is assumed to be consisted of a mass and a resistor in parallel. The area of the lip is assumed to be of the same size as the area of the tube which constitute the second section of the transmission line. If the lip is rounded shaped, the radius of the lip is found to be $r_2 = \sqrt{A_2/\pi}$. Let $\tau = 0.459\ \rho * c/a^2$ and $m = 0.27\rho/a$. The impedance of this system is

$$Z_L = \frac{s m\tau}{s m + r}$$

The Reflection coefficient will be

$$R = \frac{Z_L - Z_2}{Z_L + Z_2}$$

With a bilinear transform it can be modified to be a time domain delay system. For example, the radiation load when pronouncing the /a/ of ɪ/ə/ther can be

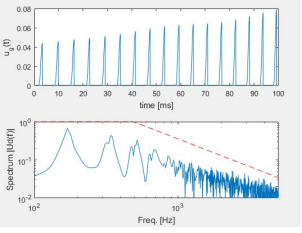


Figure 9: Time domain plot and frequency spectrum of 100ms pulse signals generated by the glottis approximated by a signal with frequency rising from 150Hz and 250Hz.

represented with a bilinear z transform of

$$-0.2197\frac{1 - 2.0889z^{-1}}{1 + 0.3215z^{-1}}$$

The recursion relation related to this bilinear transform is

$$Y[n] = -0.2197X[n] + 0.2197 * 2.0889X[n-1] - 0.3215Y[n-1]$$

This relation can be easily implemented in MATLAB by designating two variables to store the previous state($X[n-1]$) and output($Y[n-1]$).

If lip have a time-varying shape, the load impedance will be time-dependent. In that case, it is necessary that the lip impedance be simulated at each time instant in the simulation and stored as a matrix. During the simulation when the wave is propagating, the Load impedance and bilinear system has to be updated constantly.

There are a total of 21 spatial samples which divides up the 17cm vocal tract. The spatial samples are corresponding to each small sections of the vocal tract. According to the relative length between the two large sections, the 21 small

1 The History of Acoustics

A fantastic summary of the history of acoustics can be found in Pierce's book on acoustics, pages 3-6. A list of many important contributers to acoustics, psychoacoustics, and speech processing can be found in Table 1 in the appendix. This list is certainly not all-inclusive, but it is a good sampling of names to know.

2 Relevant Topics in Physics and Acoustics

2.1 Basic Acoustics

In acoustics, we study mechanical waves in a medium. Some important quantities associated with these waves are pressure and volume. Typically, we study a pressure disturbance p in air, where $P_{total} = P_0 + p(P_0$ is the equilibrium pressure). We are also interested in particle velocity u, and the volume velocity of the medium $U = uA$, where A is some cross-sectional area through which the fluid flows(the units of u are $\frac{length}{second}$ and the units of U are $\frac{length^3}{second}$).

Perhaps the most important equation in acoustics is the wave equation. It is derived from two important relations between pressure and particle velocity. The first is the linearized version of Euler's equation describing forces in the fluid

$$\bigtriangledown p = -\rho_0 \frac{\partial u}{\partial t} \tag{1}$$

and the second is the linear continuity equation

$$\bigtriangledown \cdot u = -\frac{1}{\eta_0 P_0}\frac{\partial p}{\partial t} \tag{2}$$

where $\eta_0 = c_P/c_V$ is the adiabatic index and ρ_0 is the equilibrium density of the medium. The linearized versions of these equations assume that the condensation $s = (\rho - \rho_0)/\rho_0$ is very small. Combining these equations, we get the 3D wave equation.

$$\bigtriangledown^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{3}$$

Another way to represent the equations for acoustics is in matrix form, as follows(in 1D):

$$\frac{\mathrm{d}}{\mathrm{d}x}\begin{bmatrix}P\\U\end{bmatrix} = -\begin{bmatrix}0 & \mathbb{Z}(s)\\ Y(s) & 0\end{bmatrix}\begin{bmatrix}P\\U\end{bmatrix} \tag{4}$$

where $\mathbb{Z}(x,s) = \frac{s\rho_0}{A(x)}$ and $Y(x,s) = \frac{sA(x)}{\eta_0 P_0}$.

We find that the speed of sound, c is as follows:

$$c = \sqrt{\frac{\eta_0 P_0}{\rho_0}} \tag{5}$$

For air at 20 degrees celcius and 1 atm, $\rho_0 = 1.21kg/m^3$, $\eta_0 = 1.402$, and c=343m/s.

Examining the speed of sound, it is possible to substitute known relations and solve for c in terms of other variables. For instance, the temperature dependence of the speed of sound can be shown using the ideal gas law.

$$P_0 V = NkT \tag{6}$$

$$M \cdot \frac{P_0}{NkT} = \frac{1}{V} \cdot M \tag{7}$$

$$\frac{mP_0}{kT} = \rho_0 \tag{8}$$

where k is the Boltzman constant, V is the volume, T is the temperature in Kelvin, M is the total mass, N is the total number of molecules, and m is the mass of one molecule of the gas. This gives

$$c = \sqrt{\frac{\eta_0 kT}{m}} \tag{9}$$

For plane waves and spherical waves, the intensity of the wave is

$$I = \frac{|pu|}{2} = \frac{p^2}{2\rho_0 c} \tag{10}$$

From the intensity, power can be easily calculated as $P = IA$, where A is some area through which the power is radiating.

2.2 d'Alembert Solution to the Wave Equation(1D and 3D)

The general solution to the wave equation was developed by Jean le Rond d'Alembert. In one dimension, the solution is

$$p(x,t) = f(t - x/c) + g(t + x/c) \tag{11}$$

Examing this equation, we see that the solution to the wave equations is a set of travelling waves, one in the opposite direction of the other. The shape of each individual function, f or g, does not change, but their superposition will change as they travel. The general solution in 3D(spherical coordinates) is

$$p(r,t) = \frac{1}{r}f(t - r/c) + \frac{1}{r}g(t + r/c) \tag{12}$$

The general solution to the wave equation has the same form for both p and u. It is possible to show that these general solutions are true by plugging them into the 1D and 3D wave equations.

2.3 Acoustic Impedance

Acoustic impedance is defined (relative to the volume velocity) as $Z = \frac{p}{U} = \frac{P}{vS}$ Specific acoustic impedance is defined (relative to the particle velocity) as $z = \frac{p}{u} = ZA$ The characteristic acoustic of air is $\rho_0 c$.

2.3.1 Acoustic Impedance of a Tube

The acoustic impedance of a tube is easily calculated as follows:

$$Z_{tube} = \frac{z_{tube}}{A_{tube}} \tag{13}$$

Assuming the tube is filled with air, $z_{tube} = \rho_0 c$, and

$$Z_{tube} = \frac{\rho_0 c}{A_{tube}} \tag{14}$$

2.3.2 Radiation Impedance of a Sphere

Given the form for a diverging spherical wave $p = \frac{C_0}{r}e^{j(wt-kr)}$ where C_0 is some amplitude coefficient, and the velocity potential $\phi = -\frac{p}{jw\rho_0}$ we find that

$$u = \bigtriangledown\phi = \frac{1}{\rho_0 c}(1 - \frac{j}{kr})p = (\frac{1}{\rho_0 c} + \frac{1}{jw\rho_0 r})p \tag{15}$$

Hence, the specific acoustic impedance of the sphere is

$$\frac{p}{u} = \frac{1}{\frac{1}{\rho_0 c} + \frac{1}{jw\rho_0 r}} \tag{16}$$

and the acoustic impedance is

$$\frac{P}{U} = \frac{1}{A_{rad}(\frac{1}{\rho_0 c} + \frac{1}{jw\rho_0 r})} \tag{17}$$

Where A_{rad} is some radiation area. Hence, for a sphere,

$$\frac{1}{Z_{rad}} = \frac{A_{rad}}{\rho_0 c} + \frac{A_{rad}}{jw\rho_0 r} \tag{18}$$

In the circuit analog, this looks like a mass in parallel with a resistance as shown in figure 1.

Often, we are interested in the radiation impedance of a half sphere(i.e. at the opening of a tube). For this we set $A_{rad} = \frac{4\pi r^2}{2} = 2\pi r^2$.

2.4 Decibels

A Bel is the logarithm of the ratio of a power or intensity to a reference level. Bels are quite small for sound level ratings, so we typically work with decibels. Decibels are defined as follows $dB = 10log_{10}\frac{\Pi}{\Pi_{ref}}$ where Π is the power. We also use the factor of 10 to get a dB rating for intensity, because

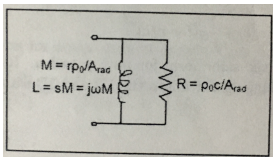


Figure 1: Electrical equivalent circuit for the radiation impedance of a sphere.

$$dB = 10log_{10}\frac{\Pi}{\Pi_{ref}} = 10log_{10}\frac{\Pi/A}{\Pi_{ref}/A} = 10log_{10}\frac{I}{I_{ref}} \tag{19}$$

Because power and intensity are proportional to the pressure squared, we can also use pressure to find dB. However, in this case the logarithm is preceded by a factor of 20.

$$dB = 10log_{10}\frac{\Pi}{\Pi_{ref}} = 10log_{10}\frac{p^2}{p_{ref}^2} = 10log_{10}(\frac{p}{p_{ref}})^2 = 20log_{10}\frac{p}{p_{ref}} \tag{20}$$

The accepted reference sound pressure level is $p_{ref} = 20\mu Pa$ or $I_{ref} = 10^{-12}W/m^2$ (these are characteristic of the same sound level, even though they are different quantities). The units of a number calculated with this reference would be dB-SPL.

We can consider p/p_{ref} to be the gain G, though we use this quantity less in acoustic than we do in electronics(usually we think of the voltage gain, $G = V/V_{ref}$), where V is analogous to p because $\Pi \propto V^2$). 0dB is the same thing as $G = 1$. Negative dB indicates that the signal in question is smaller than the reference, and positive indicates that it is larger.

2.5 Sound Propagation in Tubes

For the purposes of speech processing and acoustics, it is very useful to describe sound propagation in tubes with tube transmission line models, in many cases composed of multiple tubes of varying areas, and various kinds of inputs and loads. The 1D transmission line model will be developed in section 2.6.1.

Another relevant concept is sound propagation in a tube of varying area, a 'horn' (i.e. the bell of a trumpet). This behaviour is described by the Webster equation for a

horn. In deriving the Webster horn equation, for a tube with a cross-sectional area $A(x)$ that varies with length x, the 1D version of equation 1 becomes

$$-\frac{\partial p_x}{\partial x} = \frac{\rho_0}{A(x)}\frac{\partial(u_x A(x))}{\partial t} \tag{21}$$

Combining this with equation 2, we get the Webster equation for a horn:

$$\frac{1}{A(x)}\frac{\partial}{\partial x}(A(x)\frac{\partial p_x}{\partial x}) = \frac{1}{c^2}\frac{\partial^2 p_x}{\partial t^2} \tag{22}$$

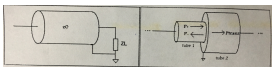


Figure 2: (a) Tube transmission line terminated in a load Z_L. (b) Two tubes (of different areas A_1 and A_2) in series. Transmission Line examples. Note that Z_L can be complex. For instance, in some cases Z_L will be the radiation impedance of a half sphere, which is complex.

2.6 Transmission Lines and Networks

2.6.1 1D Acoustic Transmission Line

For an acoustic transmission line(as shown in figure 2) at any point x, the acoustic impedance $Z(x,w)$(in the frequency domain) is defined by

$$Z(x,w) = \frac{P(x,w)}{U(x,w)} = \frac{P^+(x,w)e^{-jk(x-L)} + P^-(x,w)e^{jk(x-L)}}{U^+(x,w)e^{-jk(x-L)} - U^-(x,w)e^{jk(x-L)}} \tag{23}$$

where $k = w/c$. The characteristic impedance of a tube is defined as follows(see Section 2.3.1 for acoustic impedance of a tube)

$$z_0 = \frac{P^+}{U^+} = \frac{P^-}{U^-} = \frac{\rho_0 c}{A} = \sqrt{\frac{Z}{Y}} \tag{24}$$

At the end of the line, $x = L$, we define Z_L as the impedance of the 'load', and we define a reflection coefficient for the backward going waves (U^-andP^-) as

$$\Gamma = \frac{P^-}{P^+} = \frac{U^-}{U^+} \tag{25}$$

substituting these quantities into equation 18, we find that

$$Z_L = Z(L,w) = \frac{P^+(L,w)(1 + P^-(L,w)/P^+(L,w))}{U^+(L,w)(1 + U^-(L,w)/U^+(L,w))} = z_0\frac{1+\Gamma}{1-\Gamma} \tag{26}$$

solving equation 21 for Γ, we find

$$\Gamma = \frac{Z_L - z_0}{Z_L + z_0} \tag{27}$$

It is also possible to calculate the input impedance of the line. For the setup in figure2a, the input impedance will be

$$Z(0,w) = z_0\frac{e^{jkL} + \Gamma(L,w)e^{-jkL}}{e^{jkL} - \Gamma(L,w)e^{-jkL}} \tag{28}$$

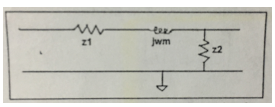


Figure 3: Karal correction for discontinuity in circular cross section.

because if there are no other impedance mismatches in the line, we can take $\Gamma(L,w) = \frac{P^-(L,w)}{P^+(L,w)} = \frac{P^-(0,w)}{P^+(0,w)}$ (and the same for U^+andU^-), because P^+, P^-, U^+ and U^- are amplitudes, so the only thing that affects them is Γ (they are amplitudes, so they don't vary with oscillation). From this relation, we can solve for $\Gamma(0,w)$, finding that

$$\Gamma(0,w) = \Gamma(L,w)e^{-2jkL} \tag{29}$$

There is a phase difference of -2kL between $\Gamma(L,w)$ and $\Gamma(0,w)$, but their amplitudes are the same. The transmission coefficient through the boundary in figure 2b is $T = 1 - \Gamma$ (sometimes we call $\Gamma'R'$), and it is equal to the ratio P^{trans}/P^+ (where P^{trans*} is analogous to P^+, but for the next section of the transmission line.)

The equations for Γ were been developed with the setup of a tube transmission line terminated by a load(as in figure 2a) in mind, but they can be generalized to apply to any boundary where there is an impedance mismatch. For example, if two tubes are connected as in figure 2b, with neither of them necessarily being the beginning or the end of the line, the reflection coefficient for that particular boundary can be calculated as

$$\Gamma\frac{z_2 - z_1}{z_2 + z_1} = \frac{A_1 - A_2}{A_1 + A_2} \tag{30}$$

where z_1 and z_2 are the characteristic impedances of tubes 1 and 2. Notably, Karal found that this is not entirely correct. When two tubes are in series(as in figure2b), there is an effective mass component that we must consider in the 'load'. This is drawn in figure 3, and the reflection coefficient is calculated as follows

$$Z_L = z_2 + jwm \tag{31}$$

where m is the effective mass, such that

$$\Gamma = \frac{Z_L - z_1}{Z_L + z_1} = \frac{z_2 + jwm - z_1}{z_2 + jwm + z_1} \tag{32}$$

2.6.2 ABCD Transmission Matrix(2 Port Networks)

The ABCD chain matrix method of transmission line analysis can be very useful for simplifying a problem and saving time. Generalized transmission line components are given in figure 4a,b. A transmission matrix can be written for each of the two types of components such that

$$\begin{bmatrix}V_i\\I_i\end{bmatrix} = T\begin{bmatrix}V_2\\-I_2\end{bmatrix} \tag{33}$$

where the transmission matrix T is given by

$$T = \begin{bmatrix}A & B\\C & D\end{bmatrix}. \tag{34}$$

For the configuration in figure 4a, the transmission matrix is

$$T = \begin{bmatrix}1 & 0\\Y & 1\end{bmatrix} \tag{35}$$

and for the configuration in figure 4b, the matrix is

$$T = \begin{bmatrix}1 & Z\\0 & 1\end{bmatrix} \tag{36}$$

ABCD transmission matrices are so useful because we define the inputs and outputs such that they can be conveniently multiplied together to find the transmission matrix for the entire line. In figure 4c, the equation for the transmission line is given by

$$\begin{bmatrix}V_{in}\\I_{in}\end{bmatrix} = T_{total}\begin{bmatrix}V_{out}\\-I_{out}\end{bmatrix} = T_1T_2T_3\begin{bmatrix}V_{out}\\-I_{out}\end{bmatrix} \tag{37}$$

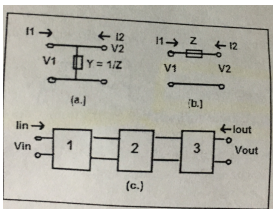


Figure 4: (a,b) Smallest two-port network components. (c) Transmission line with 'black box' components 1, 2 and 3.

The transmission matrix T is related to the impedance matrix Z, defined as

$$\begin{bmatrix}V_1\\V_2\end{bmatrix} = Z\begin{bmatrix}I_1\\I_2\end{bmatrix} = \begin{bmatrix}z_{11} & z_{12}\\z_{21} & z_{22}\end{bmatrix}\begin{bmatrix}I_1\\I_2\end{bmatrix} \tag{38}$$

by

$$T = \frac{1}{z_{21}}\begin{bmatrix}z_{11} & det(Z)\\1 & z_{22}\end{bmatrix} \tag{39}$$

This can be determined by solving the system of equations given by equations 27 and 31.

2.6.3 N-Port Networks

For analysis of the N-port matrix, we consider the acoustic quantities

$$P_i = P_i^+ + P_i^- \tag{40}$$

$$U_i = U_i^+ + U_i^- \tag{41}$$

by KCL, we require $\Sigma U_i = 0$ which yields $\Sigma U_i^+ = \Sigma U_i^-$ Using the relations

$$\frac{U_i^+}{P_i^+} = \frac{A_i}{\rho_0 c} = \frac{U_i^-}{P_i^-} \tag{42}$$

$$\frac{d}{dx}\begin{bmatrix}P(x,\omega)\\V(x,\omega)\end{bmatrix} = -\begin{bmatrix}0 & SM\\SC & 0\end{bmatrix}\begin{bmatrix}P(x,\omega)\\V(x,\omega)\end{bmatrix}$$

Where $M = \frac{m}{l}$ and $C = \frac{cl}{S^2}$.The M can be thought of as the series inductance per unit length and the C can be thought of as the shunt capacitor per unit length.Figure.1 shows a simple transmission line model in acoustic wave that is commonly used to model the acoustic wave transmission. In order to be symmetric, we can split the M into two inductance element of M/2.

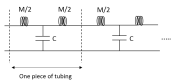


Figure 1: Simple transmission line model for acoustic wave with series inductance and shunt capacitance.

For the cascaded tubes (one can think of them as cascaded transmission lines), the ABCD matrix method is essentially useful for the analysis of the acoustic wave transmission. From Fig.1, we can treat each repeating section as one cascaded unit, namely the pressure at port 1 and port two can be related with one ABCD matrix. So can port 3, port 4..etc as shown in Fig.2.

$$\begin{bmatrix}P1\\V1\end{bmatrix} = \begin{bmatrix}A & B\\C & D\end{bmatrix}\begin{bmatrix}P2\\V2\end{bmatrix}$$

$$\begin{bmatrix}P2\\V2\end{bmatrix} = \begin{bmatrix}A & B\\C & D\end{bmatrix}\begin{bmatrix}P3\\V3\end{bmatrix}$$

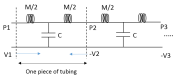


Figure 2: Cascaded transmission line models for acoustic wave with series ports definition.

Note that we define the -V2 to be "into the port". In each repeating unit, we can find the corresponding ABCD matrix for a series inductance and a shunt

capacitance, respectively. With only a series inductance present between tow ports, the ABCD matrix can be written as followed: (assume V2 is "moving out from the port 2" and -V2 is "moving into the port 2")

$$\begin{bmatrix}P1\\V1\end{bmatrix} = \begin{bmatrix}1 & Z\\0 & 1\end{bmatrix}\begin{bmatrix}P2\\V2\end{bmatrix}$$

With only a shunt capacitance present between tow ports, the ABCD matrix can be written as followed:

$$\begin{bmatrix}P1\\V1\end{bmatrix} = \begin{bmatrix}1 & 0\\Y & 1\end{bmatrix}\begin{bmatrix}P2\\V2\end{bmatrix}$$

Combine the two of the matrix by multiplying them together we can get the ABCD for a repeating cell shown in Fig.3:

$$\begin{bmatrix}P1\\V1\end{bmatrix} = \begin{bmatrix}1 & Z\\0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\Y & 1\end{bmatrix}\begin{bmatrix}P2\\-V2\end{bmatrix}$$

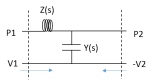


Figure 3: Single cell from the repeating transmission line models for acoustic wave transmission with port definition.

2 Acoustic Horns

If a tube has an area that is changing rather than has constant cross-section area along the tube, the direction of the wave propagation will be different than what we just discussed previously. When the area changes, the impedance also changes. The equation of a typical horn is absed on Newton's second law of conservation of momentum and the Hooke's law:

$$\nabla\cdot\mathrm{U} = \frac{-sP}{\gamma P_0} \text{ (Newton's law)}$$

$$\nabla P = -s\,\rho_0 U \text{ (Hooke's law)}$$

where P is the pressure and U is the vector particle velocity. The s is the Laplace frequency which is equal to $\sigma + j\omega$. The ratio of pressure to the particle velocity is defined as the specific acoustic impedance with the unit of [Rayls] and the pressure over a volume velocity is the acoustic impedance in [acoustic ohms].

The conical horn, which is one the special cases of the acoustic horns, has an exact solution in spherical coordinates:

$$\nabla_r P = \frac{\rho U}{4\pi} = -s\,\rho_0 U_r$$

$$\nabla\cdot\mathrm{U} = \frac{1}{S}\times\frac{\rho}{4\pi}\times r^2 U_r = \frac{-sP}{\gamma P_0}$$

The series acoustic impedance per unit length is defined as $\mathrm{Z} = \frac{A(r)\rho_0}{4\pi r^2}$ while the shunt acoustic admittance per unit length is $\mathrm{Y} = \frac{r A(r)}{\gamma P_0}$. The horn area now becomes $\mathrm{A(r)} = A_0 r^2$. We can then write the transformed spherical equations in Webster form in terms of pressure P and the volume velocity $\mathrm{V} = \mathrm{A(r)}U_r$ with respect to r in equation (1.1). The characteristic impedance (Z_0) depends on the radius: $Z_0 = \sqrt{\frac{m}{c}} = \frac{\rho_0}{A(r)}$. The propagation constant $\gamma = \sqrt{ZY} = \frac{s}{v}$. Therefore, we cam rewrite the Webster horn matrix in pressure as:

$$\frac{d^2P}{dr^2} = \gamma^2 r P$$

Following the solution derived from D'Alembert in 1747, the solution is relative to the spherical symmetry, which is the sum of the forward-traveling and backward-traveling wave ($P^+(r,\omega)$ and $P^-(r,\omega)$):

$$P(r,\omega) = P^+\frac{e^{-\gamma r/r}}{r} + P^-\frac{e^{+\gamma r/r}}{r}$$

which will be inversely proportional to the distance of the radiation source. The radiation impedance (Z_{rad}) is defined as $\frac{P}{U}$ and the equivalent circuit of the radiation impedance of a conical horn is just like two shunt resistance and inductance in parallel where they are both dependent on the radius of the horn area.

3 Reflectance

We are now going to talk about how to define the reflectance and the input or output impedance of a transmission line. The reflectance $\Gamma(s)$ can be expressed as $\frac{Z_{in}-Z_0}{Z_{in}+Z_0}$. The input impedance can be expressed in terms of $\Gamma(s)$: $Z = \frac{P}{U} = \frac{P^+ + P^-}{U^+ + U^-} = Z_0\frac{1+\Gamma}{1-\Gamma}$. For a typical transmission line terminated with the impedance of R_l shown in Fig.4, the ABCD matrix expression is :

$$\begin{bmatrix}P\\U\end{bmatrix} = \begin{bmatrix}1 & R_l\\0 & 1\end{bmatrix}\begin{bmatrix}cos(j\omega l/c) & jZ_0 sin(j\omega l/c)\\ jY_0 sin(j\omega l/c) & cos(j\omega l/c)\end{bmatrix}\begin{bmatrix}R_l\\1\end{bmatrix}$$

4 HW2 Tube Simulation

In HW2 we have simulated an impulse traveling along a transmission line with a sealed input end (glottis) and an output end (mouth). The output is terminated with a resistor $Z_m = 2Z_0$ where Z_0 is the characteristic impedance

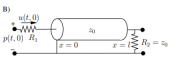


Figure 4: Figure captured from HW2 for an electrical circuit with transmission line terminated with R_1 at input and R_2 at output.The p and u in the figure are the same as the P and U respectively in the matrix.

of the tube. With the given sampling frequency F_s = 44100 Hz, we can calculate the vocal tract length L = 3450/44100 = 0.782m (since we are asked to make the vocal tract length 10 samples long). The reflectance at the mouth $Gamma_R$ = $(2z_0-z_0)/(2z_0+z_0) = 1/3$. The reflectance at the glottis is 1 due to the input impedance is assumed to be infinity. The simulation results show that the impulse response of initial input will become 1/3 of the initial value after it hits the last segments of the tube (the tenth one) and then this 1/3 of the initial impulse response will travel backwards to the glottis, being reflected by the infinity input impedance and become 1/3 of the initial value of the impulse response. The impulse response will therefore repeat the same process again and again.

The impulse response and the frequency response at the mouth are plotted in Fig.5 and Fig.6, respectively. The velocity at the mouth needs to consider in both positive and negative direction: $u_m(t,L) = u_m^+(t,L) - u_m^-(t,L) = u_m^+(t,L) * (1-Gamma_R) = u_m^+(t,L)*(1-1/3) = 2/3*u_m^+(t,L)$ After the impulse response travels back to the glottis, the incident velocity u_m^+ at this point will be 1/3 of the initial value due to the reflection coefficient at glottis is 1.Therefore, the total velocity at mouth in time domain can be written as:

$$u_m(t,L) = \frac{2}{3}[\delta(t-\frac{L}{c}) + \frac{1}{3}\delta(t-\frac{3L}{c}) + \frac{1}{3}\delta(t-\frac{5L}{c})....]$$

and if we perform FFT on this series we can obtain:

$$u_m(\omega,L) = \frac{2}{3}[e^{-j\omega\frac{L}{c}} + \frac{1}{3}e^{-j\omega\frac{3L}{c}} + \frac{1}{3}e^{-j\omega\frac{5L}{c}}....] = 2/3[(e^{-j\omega\frac{L}{c}})/(1-1/3e^{-j\omega\frac{2L}{c}})]$$

A similar analysis can be done to find the impulse and frequency response of the pressure at the mouth.

5 HW3 Model and Simulation of Middle Ear

The model of the middle ear was first simulated in the time domain by Kelly and Lochbaum in 1963. There is some more comprehensive discussion in Bilbao's thesis published in 2001. The middle ear can be treated as a stub of 2.5 cm long

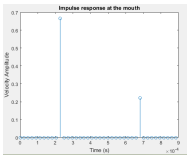


Figure 5: Impulse response at the mouth.The amplitude decreases to 1/3 of the initial value after traveling to the last segment of tube.

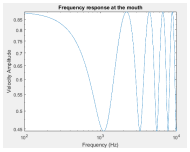


Figure 6: Frequency response at the mouth.

transmission line which is terminated in a series of combination of a stiffness of a capacitor ((C_{ad}), stiffness of the annular ligament) and a resistor ((R_c), cochlear impedance) as shown in Fig.7 when Rosowski et al. did some research on cats' middle ears. The experimental details of the cats' ears have been explored by Guinan and Peake in 1967 and it was first modeled by Zwislocki in 1962 and later by Lynch et al. in 1982. R_c is twice of the characteristic impedance Z_0 which is equal to $\frac{\rho_0}{A_0}$ = 9.21 × 10^6 ohms if r_c = 0.375 × 10^{-2}m. C_{ad} can be calculated with the assumption of the impedance of this element is equal to the cochlear impedance R_c at a frequency of 0.8 kHz.

The radiation impedance Z_{rad} is composed of a series resistance of R_{rad} =

Laplace transform is continuous in both time and frequency domain. It is causal, which means the time domain function is only defined at time greater than zero. The time domain signal is zero when $t < 0$. Laplace transform has poles and zeros in complex domains and it's a linear transform. The relation between the signal in time domain and that in s domain is:

$$F(s) = \int_0^\infty f(t)\times e^{-st}dt$$

Fourier transform, on the other hand, is defined from $-\infty < t < \infty$ and it's non-causal. Fourier transform is a linear transform in ω domain and the relation between the signal in time domain and that in the ω domain is :

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\times e^{-j\omega t}dt$$

Fourier transform is also continuous in time and frequency domain.

7 History of Acoustic

Figure.9 lists some important parts of the history in acoustic:

Name	Significant Contribution
Pythagoras	Musical relationship (550 B.C.)
Aristotle	One of the first scientists, but legally qualitative (383-322 B.C.)
Marsenne	Father of acoustics (1588-1648)
Galilae Galilei	$\Omega = C$, Resonance, Harmonics (1638)
Hooke	Law of a spring (1676)
Boyle	Sonic gas law, no sound in a vacuum (1660)
Newton	Principia — first model of speed of sound (1686)
	Wave slit experiment (1642-1727)
Euler	Speed of sound contribution (1744,1755)
D'Alembert	First solution of wave equation (1747)
Lord Rayleigh	Published *" The Theory of Sound"* (1876-1877)
Gauss	Conservation law (1811)
Laplace	Adiabetic gas law (1816)
Fourier	Heat transfer analysis method (1822)
Lagrange	Thermodynamic gas laws (1736-1813)
Nyquist	Formula for the thermal noise power (1928)

Figure 9: Table of some significant contribution from pioneers in acoustic field.

8 Cepstral Analysis

The cepstrum is defined as the inverse of the DFT of the log magnitude of the DFT of a signal:

$$c[n] = F^{-1}log|F(\omega)|$$

and the mass L_{rad} in parallel. We can use the effective area of the pinna $A_{rad} = 0.5\times\pi d_c^2$ where $d_c = 0.75$ cm is the diameter of canal.

$$Y_{rad} = \frac{1}{sL_{rad}} + \frac{1}{R_{rad}}$$

where $R_{rad} = \frac{\rho_0 c}{A_{rad}} = 4.6\times10^6$ ohms.

$$L_{rad} = \frac{\rho_0 a_d}{A_{rad}} = 50 \text{ H}.$$

The cutoff frequency is defined by $\frac{R}{2\pi L}$ = 14.6 kHz. The radiation load is well approximated by a mass in this model. Therefore the reflectance at the input Γ_{tragus} is ≈ -1 due to the resonant frequency is $\frac{\rho_0 c}{2\pi L}$ = 50 kHz.

The frequency response of the radiation impedance should be similar to the Fig.8 from the paper of Rosowski et al. The cat data indicates that the ear radiates above 3kHz and the Fig.8 shows a matched radiation impedance at above 3kHz as well.

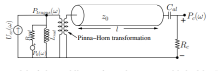


Figure 7: The model of the middle ear from the paper published by Rosowski, Carney and Peak in 1988 about the cats' middle ears.

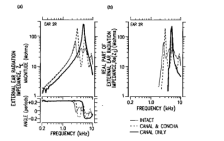


Figure 8: Radiation impedance for an actual cat ear with and without the pinna from from the paper of Rosowski et al. in 1988.The effect of the pinna flatterns out the frequency response of the ear's radiation impedance above 3 kHz. The frequency of the pole is also lowered a little bit from 5 kHz to 2.7 kHz.

As for the simulation of the D'Alembert's wave solution, we first need to define the F_s = 96.04 kHz if we assume there are 7 sections in the transmission line model. The input impedance $Z_{(0,s)} = Z_0\times\frac{1+\Gamma_{(0,s)}}{1-\Gamma_{(0,s)}} = Z_0\times\frac{1+e^{\frac{-2sL}{c}}}{1-e^{\frac{-2sL}{c}}}$ when the cochlear is blocked. With further simplification, $Z_{(0,s)} = \frac{2Z_0 cos(\frac{sL}{2c})}{sin(\frac{sL}{2c})} = Z_0 coth(\frac{sL}{2c})$. The reflectance at the output $\Gamma_{(0,s)} \approx \frac{Z_{load}-Z_0}{Z_{load}+Z_0}$ and the reflectance at the cochlear end $\Gamma_{(L,s)} = \frac{R_c+1/sC_{ad}-Z_0}{R_c+1/sC_{ad}+Z_0}$.

We can use the bilinear function from Matlab to find the poles and zeros in both s and Z^{-1} domain. The coefficients of the $Gamma_{L,s}$ in time domain are as followed: B = [0.3447,-0.3106] and A = [1,-0.9659], which means the $H(z) = (0.3447-0.3106z^{-1})/(1-0.9659z^{-1})$. The coefficients of the $Gamma_{0,s}$ in time domain are as followed: B = [-0.494,0.012] and A = [1,-0.5181], which means the $H(z) = (-0.494+0.012z^{-1})/(1-0.5181z^{-1})$.The pole and zero in $Gamma_{L,s}$ are 1/0.9659 = 1.0353 and 0.3447/0.3106 = 1.1098 in z^{-1} domain, respectively. In s domain, the pole is at s = 0 and s = -3351.1. The zeros are at s = 0 and s = -10053. The pole and zero of $Gamma_{0,s}$ are 1/0.5181 = 1.93 and 0.494/0.012 = 41.167 in z^{-1} domain, respectively. In s domain, the poles are at s = -92000 and s = -61333. The zeros are at s = 92000 and s = -184000.

The ratio of the cochlear pressure to the ear canal pressure $\frac{P_c(\omega)}{P_{ec}(\omega)}$ can be found by finding the impulse response to a current pulse in the ear canal. The cochlear velocity can be calculated as $u_c(L,t) = u^+(L,t) - u^-(L,t)$ and then we can find the cochlear pressure which is equal to $R_c\times u_c(L,t)$. The ear canal pressure can be calculated from the canal velocity: $u_{ec}(0,t) = u^+(0,t) - u^-(0,t)$ with the relation of $P_{ec}(0,t) = Z_{rad}\times u_{ec}(0,t)$. The transfer function of the ratio of the cochlear pressure to the ear canal pressure $\frac{P_c(\omega)}{P_{ec}(\omega)}$ can be found by performing FFT to the impulse response and it should look like a high pass filter with a 6-9 dB high pass slope and an in-band ripple of no more than few dB between 0.8 kHz to 0.5F_s.

6 Signal Processing

Here we discussed about some important transform in frequency and time domain that are important in acoustic research:

a) DTFT: Discrete Time Fourier Transform
This transform is discrete in time. It is periodic and continuous in frequency.
b) DFT: Discrete Fourier Transform
This transform is discrete in time and frequency. It is periodic in frequency.
c) STFT: Short Time Fourier Transform
This transform is continuous in time and frequency.

The speech signal can be modeled as the convolution of the glottal source and vocal tract in time domain, which means we can add the log magnitude of each component:

$$H(\omega) = H_1(\omega)H_2(\omega)$$

$$=> log(|H(\omega)|) = log(|H_1(\omega)|) + log(|H_2(\omega)|)$$

It is like we can add the source and filter together and then we can return to the time domain through the inverse DFT:

$$c[n] = c_1[n] + c_2[n]$$

The cepstrum is useful since we can separate the source and the filter. Figure.10 shows the block diagram of the computation of cepstrum.

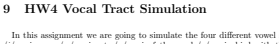


Figure 10: Block diagram of the computation of cepstrum.

9 HW4 Vocal Tract Simulation

In this assignment we are going to simulate the four different vowel sounds: /i/ as in eve, /ɑ/ as in at, /a/ as in father and /ɔ/ as in bird with the two-tube model (two transmission lines connected in series and the second one will be terminated with a mass and a radiation load resistance in parallel). This model will simulate each of the sound by varying the length of two transmission lines, much like the behavior of vocal tract when human is trying to pronounce the sound. The average length vocal tract is L = 17 cm. Therefore, for each sound, we need to use the corresponding length for each transmission line in the simulation model.

For the parameters and assumption that we will use are defined as followed:

a) Speed of sound (C): 367 m/s
b) Radiation mass terminated at the transmission line: $0.27\times1.18/a$ where a is the radius of lips.
c) Radiation load resistance terminated at the transmission line: $0.459\times1.18\times367/a^2$ where a is the radius of lips.
d) Sampling Rate (F_s): 44100 Hz
e) Total number of section of the transmission line (N): 21
f) The length of the vocal tract (L) : 21 × 367/44100 = 0.1748 m = 17.48 cm
g) Radius of lips (a): a = $\sqrt{A/\pi}$ where A is the lip area.

Contents

1 Generation of Sound

1.1 Sound propagation and wave equation

This world is made of sounds. Everyday, for many people, the first thing happens is waking up to the sound of an alarm clock. During the day, we listen to the vibrant kinds of sounds coming from the bustling streets, from the music played on a pair of headphones, and most importantly, from conversation with other people. Along with visual contexts, talking and listening is one of the main approaches people use to exchange information and communicate with one another. Naturally, it is essential to answer the question: How exactly do sounds travel from the source to the receiver?

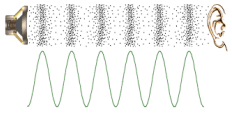


Figure 1: Illustration of how a sound travels

In Figure. 1, it demonstrates how a sound is delivered from a loudspeaker to a human's ear. Although the sound, the information, "travels" from one end to the other, nothing travels or moves physically from one end to the other. Rather, it is the air pressure change that is encoding the information. As shown in the figure, air particles expand and contract due to the vibration of the loudspeaker cone. It is a transmission line where the expansion and contraction happen along the line of all air particles from the loudspeaker to the human's ear. The air pressure changes due to the expansion and contraction, and this change is recorded by the human's ear, the eardrum to be specific. From this recording of the change of air pressure, the human perceives the sound.

Therefore, theoretically if we can have a parametric model of the air pressure at a location, we then can predict what sound will be delivered at that location. This idea is the core reason of the derivation of the wave equation. The 1-D wave equation was first derived by French scientist Jean-Baptiste le Rond d'Alembert as

followed

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{1}$$

where x corresponds to the location along a 1-D line, t corresponds to time and c is the speed of sound which has the relationship

$$c = \sqrt{\frac{\eta P_0}{\rho_0}} \tag{2}$$

where η is the adiabatic expansion coefficient of air, P_0 is the static (barometric) air pressure, and ρ_0 is the static density of air.

Derived from combining conservation of mass, Hooke's law and the Gas law, the wave equation captures the relationship between time and location for the air pressure. In addition to the derivation of the 1-D wave equation, d'Alembert derived a general solution $p(x,t)$ to it.

$$p(x,t) = f_1(t - \frac{x}{c}) + f_2(t + \frac{x}{c}) \tag{3}$$

where f_1 and f_2 are arbitrary functions.

The solution can be seen as having two parts. f_1 corresponds to a forward-traveling wave and f_2 corresponds to a backward-traveling wave. These two waves constitute the air pressure at location x.

1.2 Webster horn equation

From the 1-D wave equation stated as (1), it is easy to see that in higher dimensions, the wave equation has the form

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{4}$$

In 2-D cylindrical coordinate system, suppose that we have a radiating wave (in which case the air pressure does not depend on the angle θ), the wave equation becomes

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}p\right) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{5}$$

Closely related to (5) is the Webster horn equation that describes the propagating wave in a tube. The Webster horn equation is stated as followed

$$\frac{1}{A(r)}\frac{\partial}{\partial r}\left(A(r)\frac{\partial}{\partial r}p\right) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{6}$$

where $A(r)$ is the cross-sectional area that depends on the distance r. An illustration of the tube and the variables is shown below in Figure. 2

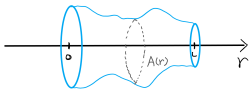


Figure 2: A horn and its variables for Webster horn equation

For example, if we have an exponential horn where $A(r) = A_0 e^{2ar}$, then we can find the solution $p(r,t)$ by using the Webster horn equation

$$\begin{aligned}\frac{1}{A_0 e^{2ar}}\frac{\partial}{\partial r}\left(A_0 e^{2ar}\frac{\partial}{\partial r}p(r,t)\right) &= \frac{1}{e^{2ar}}\left[2ae^{2ar}\frac{\partial}{\partial r} + e^{2ar}\frac{\partial^2}{\partial r^2}\right]p(r,t) \\ &= \left[2a\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2}\right]p(r,t) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}\end{aligned} \tag{7}$$

Now, in the frequency domain where $s = \sigma + j\omega$, we have the equation

$$\left[2a\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2}\right]p(r,t) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \longleftrightarrow P^{"} + 2aP^{'} = \frac{s^2}{c^2}P \tag{8}$$

Solving this differential equation will give us the air pressure $P(r,s)$ in the frequency domain

$$P(r,s) = e^{-ar}e^{\sqrt{a^2 c^2 + s^2}\frac{r}{c}} \tag{9}$$

1.3 Sound pressure, volume velocity and acoustic impedance

As mentioned earlier, a propagating sound wave determines the air pressure change along the path, and this contributes to the delivery of a sound. It turns out that this acoustical path can be modeled using circuit notations, such as the potential, the current and the impedance in the frequency domain. The corresponding terms are the sound pressure (air pressure), the volume velocity and the acoustic impedance in the frequency domain.

The sound pressure $p(r,t)$ in the time domain is often represented in dB-SPL. This can be calculated as

$$p(\text{in dB-SPL}) = 20 * log_{10}\frac{p}{p_0} \tag{10}$$

where p_0 is the reference pressure at 20 μPa

The volume velocity $v(r,t)$ in the time domain is defined as the air flow rate for a certain cross-section. It has a unit of m^3/sec

The acoustic impedance Z, which is defined in the frequency domain just like the impedance of a circuit, is defined as the amount of resistance to the propagation of sound wave through air.

Analogous to the Ohm's law for a circuit, the three terms obey the relationship

$$Z(r,s) = \frac{P(r,s)}{V(r,s)} \tag{11}$$

1.4 Helmholtz resonator

As a demonstration of how we can model an acoustic path as a circuit, a Helmholtz resonator is considered. Suppose that we have a Helmholtz resonator shown in Figure. 3, where we have a open-ended neck connected to a barrel with closed end. Now if we blow air into the neck, which is acting like a current source, we want to find out what the resonant frequency f_0 is in this case.

First of all, it is known that we can model the neck and the barrel as a mass and a compliance in series since we have the same volume velocity v. In addition, the mass will have an impedance of $j\omega M$, and the compliance will have an impedance of $\frac{1}{j\omega C}$, where

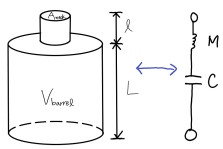


Figure 3: A Helmholtz resonator and its equivalent circuit

$$\begin{cases} M &= \frac{\rho_0 l}{A_{neck}} \\ C &= \frac{V_{barrel}}{\eta p_0} \end{cases} \tag{12}$$

Therefore, we have an expression for the impedance of the circuit

$$Z = j\omega\frac{\rho_0 l}{A_{neck}} + \frac{\eta p_0}{j\omega V_{barrel}} \tag{13}$$

Now, to solve for the resonant frequency, we set the impedance $Z = 0$ just like how we would in order to find the resonant frequency of a circuit. Then, solving for ω will give us the formula for finding the resonant frequency f_0

$$f_0 = \frac{\omega}{2\pi} = \frac{c}{2\pi}\sqrt{A_{neck}/(V_{barrel} l)} \qquad Hz \tag{14}$$

1.5 Transmission line model and ABCD matrix

As shown in the previous section, an acoustic path can be modeled as a circuit. Moreover, we can model it using a 2-port transmission line model. For example, suppose that an acoustic path is modeled as an equivalent transmission line shown in Figure. 4

There are two ways to solve for the input impedance Z_{in}. One is the traditional way of adding up the impedance of each element, and the other is using the ABCD

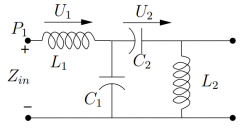


Figure 4: An equivalent transmission line model for an acoustic path

matrix method from 2-port transmission line model. Here, the ABCD matrix method is demonstrated. In this 2-port transmission line, we have four elements. For each of the elements either on one port only or across the two ports, a matrix can represent the transfer function from the ouput to the input for both the sound pressure P and the volume velocity V.

In this specific case, we will have the relationsip represneted in the matrix form shown as

$$\begin{bmatrix} P_1 \\ U_1 \end{bmatrix} = \begin{bmatrix} 1 & sL_1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ sC_1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1/sC_2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1/sL_2 & 1 \end{bmatrix}\begin{bmatrix} P_2 \\ U_2 \end{bmatrix} \tag{15}$$

Now, to calculate the input impedance Z_{in}, we know it is the same as looking into the circuit from the input side while holding the output current, volume velocity U_2 in this case, to be zero. Plugging in $U_2 = 0$ into (15) will give us the ratio P_1/U_1 which is equal to Z_{in}.

1.6 Free-ended tube, terminated tube and reflectance

Derived from the wave equation (1), the state equations are two differential equations decribing the relationship between the sound pressure $P(x,\omega)$ and the volume velocity $U(x,\omega)$ in a 1-D uniform tuve having cross-sectional area of A.

In matrix form, the state equations can be written as

$$\frac{\partial}{\partial x}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & Z \\ Y & 0 \end{bmatrix}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} \tag{16}$$

where $Z = s\rho_0/A$ and $Y = sA/\eta P_0$. A is the cross-sectional area of the tube.

Now, if we consider a tube that has a terminated end, then a backward traveling will be present due to reflection. Reflectance Γ is a term that quantifies how much of reflection is present at a specific location. It is defined as

$$\Gamma = \frac{P_-}{P_+} = \frac{U_-}{U_+} \tag{17}$$

Using eqn. (17), we can write the impedance of an acoustic circuit as the one shown in Figure. 5 with a terminated tube in terms of its reflectance Γ. First, before we derive that, we want to define the term characteristic impedance z_0.

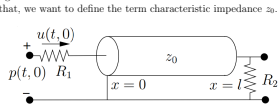


Figure 5: An acoustic circuit of a tube with a terminated end

Characteristic impedance z_0 is defined as the impedance of the transferring medium, in this case the tube, when no reflection is present. In other words, $z_0 = P_+/U_+$. Using the characteristic impedance, it can be demonstrated that the impedance $Z(x,\omega)$ can be rewritten as

$$Z(x,\omega) = \frac{P}{U} = \frac{P_+ + P_-}{U_+ - U_-} = z_0\frac{1+\Gamma}{1-\Gamma} \tag{18}$$

Furthermore, as we can see, the terminated end of the tube can actually be modeled as a load impedance placed at location $x = l$ as shown in Figure. 5. It is natural to seek an expression for the reflectance Γ in terms of this load impedance Z_l. After rearranging terms in eqn. (18), it can be seen that

$$\Gamma(x = l) = \frac{Z_l - z_0}{Z_l + z_0} \tag{19}$$

1.7 Middle ear modeling

One of the applications of using the terminated tube model would be the simulation of the middle ear. As we know that the sound signal is captured by the dish-like pinna, and then the sound wave propagates down through the ear canal to stapes and cochlea afterwards shown in Figure. 6.

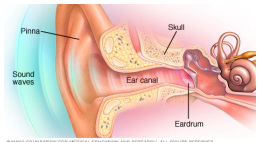


Figure 6: Illustration of how sound propagates through the ear canal

In fact, The ear canal is essentially a tube with terminated ends. One end is the inner ear, and the other is the radiational end that goes from the ear canal to the outer ea. Thus, we can do a crude simulation with a simple equivalent circuit as indicated in Figure. 7

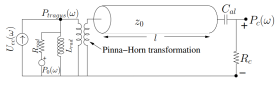


Figure 7: An equivalent circuit of a rough simulation of the middle ear

Contents

Chapter 1

Basic Acoustics

1.1 History of Acoustics

I start my paper with some paragraphs about the history of acoustics to provide context and motivation. I shall divide significant advances in the field of acoustics into different centuries.

1.1.1 17th Century

Marin Mersenne:

Figure 1.1: Marin Mersenne

Marin Mersenne is regarded as the father of acoustics for his work on the music theory. The first ever book on music theory was by Mersenne in 1836 that documented the sum of his musical knowledge that he collected throughout his lifetime. The book included Mersenne's Laws, which describe the frequency of oscillation of stretched strings. Although it would not be until later on that his experiments and theories were verified. Image credit: wikipedia.org (Fig 1.1)

Galileo Galilei:

Figure 1.2: Galileo Galilei

Galileo Galilei was the first person to study and model sound as waves. He was able to come to this conclusion after taking a chisel and scraping it against a brass plate. From there, he began to observe that the pitch of the screech varied directly with the spacing of the grooves, which are created from the contact between the chisel and the brass plate. Image credit: wikipedia.org (Fig 1.2)

Robert Hooke:

Figure 1.3: Robert Hooke

Robert Hooke as the first to discover and experiment on the law of elasticity. This discovery had so much significance in the understanding of materials. Using Hook's Law, Newton's Law and other mathematical results, the modern wave equation for sound was derived. Image credit: wikipedia.org (Fig 1.3)

Robert Boyle:

Figure 1.4: Robert Boyle

Robert Boyle was the first to discover that sound does not travel in vacuum. This was an important discovery that helped scientists realize that sound needed fluid medium to travel. The experiment that Boyle conducted consisted of a vacuum chamber with a ticking watch inside. As he kept decreasing the air inside the vacuum chamber, the ticking became softer and softer until it could no longer be heard. He was able to conduct this experiment alongside his other experiments that eventually led to discovery of Boyle's Law (relationship between volume,pressure and temperature of fluids.) Additionally, he also pioneered the modern scientific method, an important philosophical contribution that is still being used to this day. Image credit: wikipedia.org (Fig 1.4)

1.1.2 18th Century

Daniel Bernoulli:

Figure 1.5: Daniel Bernoulli:

Daniel Bernoulli was a swiss mathematician that discovered relationships between pressure and fluid velocity. This was a fundamental finding that helps explain how Pressure and wave velocity can be tied together. For instance For the same energy, the faster the speed of the wave, the lower its pressure. Image credit: wikipedia.org (Fig 1.5)

Jean le Rond d'Alembert:

Figure 1.6: Jean le Rond d'Alembert:

Jean le Rond d'Alembert was a French mathematician who made an immensely crucial contribution to acoustics through the discovery of the D'Alembert's formula for obtaining solutions to the wave equation. Till, today, we use d'Alembert's solutions to wave equations to model sound waves. Image credit: wikipedia.org (Fig 1.6)

1.1.3 19th Century

Pierre-Simon Laplace:

Figure 1.7: Pierre-Simon Laplace

Pierre-Simon Laplace was a French scholar who worked on a variety of different problems that would influence the world immensely. In the context of acoustics ,however, his most important contribution was the Laplace Transform (although this was initially just a mathematical technique that was used my Laplace to simply some solutions for differential equations). The Laplace transform was a variant of the many integral transforms that were being used/formulated in the 19th century (Euler and Lagrange both had other integral transforms for probability theory). It was not until the 1930s and later that Laplace transforms were used in the engineering domain for frequency analysis by Doetsch. Image credit: wikipedia.org (Fig 1.7)

Joseph Fourier:

Figure 1.8: Joseph Fourier

Joseph Fourier was a French Mathematician that discovered the Fourier series and the applications for heat transfer and vibrations. He first published his method of representing a wave as a superposition of multiple simsoids. Although initially formulated to solve heat transfer equations, it wasn't long before others used the Fourier series to model solutions in other areas. It wasn't until later when Gauss applying the fourier series formulation to solve wave equations that the fourier transform was formulated. Image credit: wikipedia.org (Fig 1.8)

John William Strutt, 3rd Baron Rayleigh:

Figure 1.9: John William Strutt, 3rd Baron Rayleigh

Rayleigh was an English scientist that worked on theoretical and experimental physics. His most important contribution ot acoustics was that he wrote the extremely famous textbook called theory of sound. For his contribution, the acoustic impedance unit Rayl is named after him. The textbook is still used today by acousticians and engineers. Image credit: wikipedia.org (Fig 1.9)

1.1.4 20th Century

George Ashley Campbell:

Figure 1.10: George Ashley Campbell

George Campbell was an American engineer that first invented wave filter. He researched the theory and implementation of loading coils and used them to create filter banks. Because of his work, for the first time signals were able to be sent on the same line modulated at different frequencies due to accurate passbands achieved from Campbell's filters. Image credit: wikipedia.org (Fig 1.10)

Harvey Fletcher:

Figure 1.11: Harvey Fletcher

Harvey Fletcher is credited to be the father of stereophonic sound. his contributions to acoustics are immensely important. Before Bell Labs, Fletcher was a PhD student under Robert Millikan (of Millikan oil drop experiment, or more accurately Fletcher oil drop experiment) and was the first ever student to graduate summa cum laude (with a PhD) from the University of Chicago. When he joined Bell Labs, he undertook the gargantuan task of trying to understand human hearing. Image credit: wikipedia.org (Fig 1.11)

Claude Shannon:

Figure 1.12: Claude Shannon

write about Shannon's Information theory . Image credit: wikipedia.org (Fig 1.12)

James L. Flanagan:

write about everything that Flanagan achieved

Homer Dudley:

write about VOCODER - eventually led to vocal systems like the one used by stephen hawking

Nyquist and Bode:

write about the math that these two invented to analyze waves in frequency domain.

1.2 Wave Equations

1.2.1 Elements of sound waves

It was discovered in 1660s by Galileo Galilei that sound propagates in waves. Since then many advanced have been made to mathematically describe sound waves. To start understanding sound waves, I first list out the most important variables that help us describe a sound wave.

Pressure P:
The most basic quantity that is used to understand sound waves is pressure. All our methods of sensing sound involve membranes that vibrate to pressure from sound waves. This is true for our ear drums, microphones and most loudspeakers. Pressure is measured in Pascals. However, for convenience and convention in the acoustics domain, te pressure is often reported in the decibel scale.

It was measured that the human threshold for hearing is around 20 μpa. The convention was therefore established to set this pressure value as a zero reference. The Decibel scale for pressure uses this zero reference. The decibel scale follows the formula:

$$pressure(db) = 20 \times \log_{10}\left(\frac{P_{measured}}{20 \times 10^{-6}}\right) \tag{1.1}$$

Volume Velocity V or u:
The next basic quantity that is used to understand sound waves is volume velocity. This is the aggregate velocity of a volume of particles that are traveling through space. Volume velocity is a result of basic Newtonian mechanics and is reported in m/s. The relationship between impedance and volume velocity is established by another quantity - Acoustic Impedance.

Acoustic Impedance $Z(\omega)$:
Acoustics impedance is basically the resistance to sound propagation in a medium. This is an important characterization to understand and model sound analogous to electrical circuits. The SI unit is $Pa \cdot s/m^3$ but often reported as $rayl/m^2$. Generally, the impedance is reported is the frequency domain. The relationship between Pressure, Volume Velocity and Acoustic impedance is as follows:

$$P(s) = V(s) \cdot Z(s) \tag{1.2}$$

This is very closely analogous to Ohms Law for electrical circuits. Expanding from the analogy, acoustic impedance is comprised of acoustic resistance, acoustic inductance and acoustic capacitance. This relationship with physical waves is explained further in section 1.3

Frequency and Wavelength:
Since sound is modeled as a wave, it will have to described in terms of frequency, wavelength and speed. This is analogous to light waves.

1.2.2 1D Wave Equation

1D wave equation was first conceived by D'Alembert in 1746. He derived the wave equation with existing knowledge of Hooke's Law, Newtons Laws of motion and experiments on strings The one dimensional wave equation for acoustics is written as:

$$\frac{\partial^2}{\partial x^2} p\,(x,t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} p\,(x,t) \tag{1.3}$$

Where $p(x,t)$ is distrubution of pressure along dimensions x over time, t is time and c_0 is the speed of sound in air. To expand the wave equation into any dimension and in any coordinate system frame and taken in the frequency (Laplace Transform) domain:

$$\nabla^2\, p\,(r,s) = \frac{s^2}{c_0^2}\, p\,(r,s) \tag{1.4}$$

Where $p(r,s)$ is Laplace transform of pressure along dimensions r (general coordinates).

The wave equation written in this was the first step to understand the mechanics of sound propagation. The solution to the wave equation is generally in the elegant form of a sum of a back propagating wave and a forward propagating wave.

$$p(r,s) = F(r - k_1 \cdot s) + G(r + k_2 \cdot s) \tag{1.5}$$

From speed of sound c_0:

It is possible to derive a relationship between pressure and velocity from the wave equation and the knowledge of speed of sound equation in air. The equation for the speed of sound in air is:

$$c_0 = \sqrt{\frac{\eta_0 \cdot P}{\rho_0}} \tag{1.6}$$

From the wave equation, formulation for pressure and velocity can be split using two equations:

$$-\frac{d}{dr}\, p(r,s) = \frac{\rho_0}{A(r)} V(r,s) \tag{1.7}$$

$$-\frac{d}{dr}\, V(r,s) = \frac{A(r)}{\eta_0 \cdot P_0} P(r,s) \tag{1.8}$$

When we differentiate equation 1.7 and 1.8 and substitute for pressure, we attain equation 1.3. Verifying the relationship between speed of sound formula and the wave equation. Finally, Equations 1.7 ad 1.8 can be written elegantly in a form of a matrix:

$$-\frac{d}{dr}\begin{bmatrix} P(r,s) \\ V(r,s) \end{bmatrix} = \begin{bmatrix} 0 & \frac{\rho_0}{A(r)} \\ \frac{A(r)}{\eta_0 P_0} & 0 \end{bmatrix} \cdot \begin{bmatrix} P(r,s) \\ V(r,s) \end{bmatrix} \tag{1.9}$$