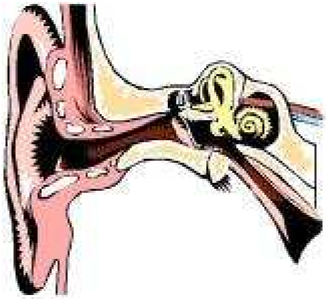
Speech Recognition Laboratory



CleveLabs Laboratory Course System Version 6.0

Introduction

The human body's ability to recognize speech is an amazing function in light of all the variables that exist in how speech is delivered. For example, we can recognize a word if it is said by two different people with two different types of accents or speeds at which they say the word. We can also recognize that same word regardless of the background noise that is present in the environment at the time. We can also recognize the word in unique sentences that place other sounds directly before and after the specific word. The brain has a complex system that is capable of processing sound waves into useful information. By studying that system we can gain insight into several important clinical areas including brain development, speech pathology, and human pattern recognition. For example, have you ever wondered why it is easier to learn a new language when you are a small child compared to an adult? Is it possible that it is easier to "steer" particular neural circuitry responsible for speech during infancy than during the later adult years? If so, can we take advantage of those mechanisms to help counter learning disabilities and other clinical speech disorders?

Speech recognition systems are currently available for many applications. Individuals who are paralyzed and in wheelchairs may use voice recognition to turn on and off various appliances or to light their homes. For example, simple speech commands of "on" and "off" may be used to control a light in their room. Speech recognition is also used in some security applications. For example, a person may have to speak into a microphone to gain entry into a building or room. Once the voice is recognized as the person with access, the door will open. Finally, computer programs are now using voice recognition for word processing applications and to improve access for disabled users. These voice recognition programs are far from perfect however. Many of them require much training with the users saying words and letters over and over so that the computer program can target the particular patterns of a user's speech. If a new user wants to use the program it must be retrained. On the contrary, when our brains meet a new person we can easily talk to them and recognize the words they are saying without having to be trained. Consider the following sentence: "It's hard to wreck a nice beach." This is a classic sentence that illustrates the difficulty in teaching a computer to understand speech. Say the sentence out loud faster and faster to understand why. Why is it so difficult to create a computer program to accurately detect speech that our brains seem to do so effortlessly?

Equipment required:

- CleveLabs Kit
- CleveLabs Course Software
- A microphone to input sound to the computer
- Microsoft® Excel, MATLAB®, or LabVIEW™

Background

What is Sound?

Sound consists of a series of pressure fluctuations moving through the air that are produced by a vibrating object. Sound can be created by many sources. For example, when your vocal cords vibrate, movement in one direction forces the air in that direction to be compressed together. This is called condensation. When the vocal cords move away from that direction, an almost empty space is created that contains few molecules. This is called rarefaction. Meanwhile, the molecules that were compressed have passed on some of their energy further on to other molecules and are refilling the empty space. The combination of a condensation and a rarefaction results in a sound wave. Because sound waves require the condensation and rarefaction of molecules, sound needs a medium such as air, a solid, and a liquid to propagate. Thus, sound cannot be produced in a vacuum.

Healthy humans are able to hear sounds ranging from about 20 Hz to 20000 kHz. The frequency of these sounds is determined by the frequency of the vibration. If a sound has a high pitch, it has a high frequency, and if the pitch is low, the sound has a low frequency. The intensity of the sound wave determines how loud the sound is. Intensity is determined by the amplitude of the wave and is measured in decibels.

Anatomy

The sensory organ of hearing, the ear, consists of three parts: the outer ear, middle ear, and inner ear.

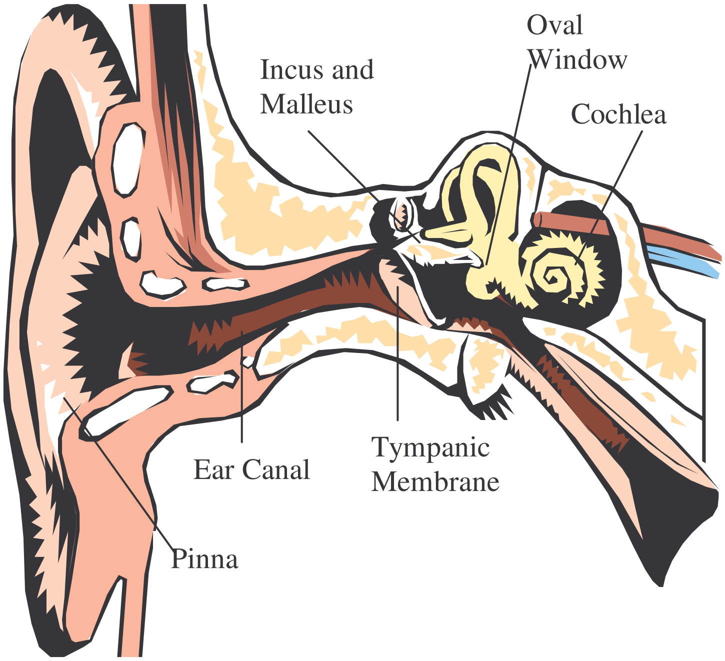


Figure 1. Anatomy of the human ear.

Outer Ear

The outer ear consists of the pinna and the ear canal. The purpose of the outer ear is to channel sound waves toward the tympanic membrane, or eardrum. This focuses sound on the anatomical structure used for deciphering it.

Middle Ear

The middle part of the ear consists of the tympanic membrane and the ossicular system. The tympanic membrane oscillates when struck by air pressure or sound vibrations. The middle ear is comprised of three bones including the incus, the malleus, and the stapes. The center of the tympanic membrane is attached to one end of the malleus. The other end of the malleus is connected to the incus, which is bound to the stapes at its opposite end. The faceplate of the stapes lies against the membrane at the oval window. The malleus and incus serve as a lever that moves when the tympanic membrane vibrates. When the tympanic membrane moves inwards, the lever pushes forward on the stapes and the cochlear fluid. When the tympanic membrane moves outwards, the lever pulls back on the stapes and the cochlear fluid.

Inner Ear

The inner ear consists of the cochlea and the nerves that transmit sound information to the brain for processing. The cochlea consists of three tubes coiled side by side: the scala vestibuli, scala media, and scala tympani. The Reissner’s membrane lies between the scala vestibuli and scala media; however, it is very thin so these two coils can be considered one compartment. Between the scala media and scala tympani is the basilar membrane. Sitting on the surface of the basilar membrane is the organ of corti, which produces nerve impulses in response to sound vibrations. These sound vibrations enter the cochlea when the stapes faceplate moves at the oval window. This movement causes the cochlear fluid to be moved in and out of the scala media and scala vestibuli.

The basilar membrane contains 20,000 to 30,000 fibers that extend from the bony center of the cochlea to an unconnected end at the outer wall. These stiff fibers vibrate much like musical reeds. Thus, when sound vibrations enter at the oval window, the basilar membrane will vibrate. In response to these vibrations, the organ of corti will generate nerve impulses. The organ of corti consists of inner and outer hair cells that act as electromechanical sensors. The upper portions of the hair cells are attached to the reticular lamina, which is connected to the basilar fiber by the rods of Corti. At the end of each hair cell are stereocilia that touch or extend into the tectorial membrane. The stereocilia on each hair cell are longer the farther they are from the modulus. A thin filament connects the tops of these stereocilia. When the basilar fiber moves upward, the reticular fiber moves in the direction of the longer stereocilia, which causes the shorter ones to be pulled outward from the hair cells. This results in the opening of ion channels that allow potassium ions to enter the hair cell, and the cell depolarizes. When the basilar membrane moves downward, the ion channels are closed, and the result is hyperpolarization of the cell. Thus, with vibration of the basilar fiber, the hair cells will alternate between depolarization and hyperpolarization. This stimulates the cochlear nerve, which sends information about the sound to the brain.

How are sounds of different frequencies distinguished?

The neural fibers located in the ear are organized tonotopically. Sounds of different frequencies are determined by where the basilar membrane has the most neural activation. The length of the basilar fibers increase and the diameter decreases from the base of the cochlea to the tip. This results in the stiffness of the fibers decreasing from the base to the tip. The transmission of sound waves is weak at the beginning of the basilar membrane but becomes stronger as the natural resonant frequency of the basilar fibers equals the sound frequency. When the sound wave reaches the fibers with the necessary resonant frequency, the fiber will vibrate so easily that the energy of the sound wave will be dissipated and will not travel much further. Because of this principle, the shorter fibers will vibrate best at higher frequencies while the longer fibers will vibrate best at shorter frequencies. As a result, the site at which there is maximum activation determines the frequency of the sound. The nerve fibers of the cochlear nerve are situated spatially so that these sites can be determined.

Broca’s area, which is located in the frontal lobe of the brain, is responsible for many speech functions. This area is mainly responsible for producing speech. On the other hand Wernicke's area, located near the auditory cortex, is responsible for understanding speech or associating words with object identification. These two regions of the brain work together to allow us to both produce and understand speech.

Sound Restoration

There are many underlying causes as to why a person may be deaf. In some cases, the connection between the auditory nerve and the brain is destroyed and it may be difficult to restore the ability to hear. However, if the nerve fibers remain intact and functional, there is a chance that the ability to hear can be restored. For example, the bones of the middle ear responsible for transmitting sound may be destroyed. Therefore, sound has no way of reaching the inner ear and generating impulses in the auditory nerves. Therefore, a new mechanical device may be used to transduce sound waves and detect the frequency at which they are occurring. Once that is known, functional electrical stimulation may be used to electrically stimulate the corresponding nerve fibers for that frequency and restore sound. In this way, an artificial system is taking advantage of the body’s tonotopic organization of the auditory nerve fibers to restore sound.

Experimental Methods

Experimental Setup

You should make sure that your BioRadio receiver is connected to your computer before starting this laboratory session. If the receiver is not connected, most of the functionality in the laboratory session will be disabled.

Procedure and Data Collection

1. Run the CleveLabs Course software. Log in and select the "Speech Recognition" laboratory session under the Advanced Physiology subheading and click on the "Begin Lab" button.

2. Click on the tab labeled "Collect and View Voice Data". At the top left you will see settings for the microphone. These settings will default to the typical settings for most microphones on a computer. However, you may need to adjust them based on your particular microphone settings.

3. Due to the default buffer size in this laboratory session, you should not record files longer than approximately 30 seconds or there is a chance that the software will lock up.

4. To begin recording and viewing sound, click on the "Record" button. You should see amplitude fluctuations in the temporal pattern of your speech as you speak into the microphone. Save a screen capture of this. If you want to save a data file, simply click on record, speak, click on stop, then click on save and give the file a name.

5. The purpose of this laboratory session is to develop an algorithm that can successfully determine the number that a person is speaking from 0 through 9. Each number should have a unique spectral content. Click on the tab labeled "Spectral Analysis". Then, say each of these numbers a few times and note the temporal feature of each number that you say. You may need to adjust the scale of the plot. You also may want to examine what the effects of filtering have on the signal.

6. Stop any data recording that is occurring and then click on the "Processing and Application" tab. For the following voice recordings you should make sure that you use good annunciation. Click on record, speak the word "zero" into the microphone, and then click on stop. The temporal and spectral features of the word you just said will be displayed in the plots.

7. Now we are going to attempt to develop an algorithm that will detect which particular number that you are saying. For the purposes of this first experiment, you should try to be in a location that minimizes the amount of background noise that is occurring. Each word from "zero" to "nine" should have a distinct spectral pattern. While some words

may be close to each other such as “four” and “five” or “six” and “seven” which start with the same letters, there should be some differences. After you say the word, stop the recording, and when the spectral and temporal features appear, click on the button labeled “Add to Zero”. This will add this trial to the spectral model of the word “zero”. When you click on this button, the amplitude spectral features of the word are normalized, plotted as a function of frequency, and averaged with other trials that you added to the model for the word “zero”. The column labeled “samples” next to each button shows how many trials you added to each model. After you add this to the “zero” model, then click on the save button (disk) at the top of the screen and save the sound file as “zero1”.

8. Repeat step 6 for four additional trials of “zero” and save these files as “zero2” and “zero3” and so on. When you are finished you should have 5 samples saved for the word “zero” in the model.

9. Repeat steps 6 and 7 for words “one” through “nine”.

10. When you are finished you should have 5 samples saved for each word. You now have a model built up for words “zero” through “nine” with five trials averaged for each word. Click on “Save Model” and save this model to a file named “model1”.

11. The model is now saved and you can recall it later if you want by selecting it from the drop down list and clicking on load model. However, since the model is still loaded into memory at this point, we will now use it to predict which word you are saying. The algorithm uses a least squared fit to determine which word you are saying. Click on record, say the word “zero”, and then click stop. Make sure that the sound file switch is set to “current”, and then click on “Predict”. The number that the algorithm predicted will appear in the box below.

12. The algorithm normalized the current input word values and then calculated the sum of the mean squared error for each point between the current input and each word model that you had built up. These values are shown in the boxes labeled “Prediction Errors” to the right of the numbers. The algorithm then selects the number with the lowest error.

13. Try saying each number “zero” through “nine” and note what the algorithm predicts that you said. If the algorithm did not predict the correct number, note how many and which numbers produced a lower error than the correct number.

14. Now change the sound file switch to “Saved”. Select one of the save sound files that you used to create the model and then click on “Predict”. You should do this for each saved sound file. Note what the algorithm predicts that you said. If the algorithm did not predict the correct number, note how many and which numbers produced a lower error than the correct number.

15. Now switch the sound file switch back to “Current”. Have another subject say the words “zero” through “nine”. Note what the algorithm predicts the subject said. If the algorithm did not predict the correct number, note how many and which numbers produced a lower error than the correct number.

16. Finally, add some background noise such as people talking and repeat step 15 with the original subject that was used to create the model.

Data Analysis

Review the results of your model predictions. You should have four separate prediction trials. Once when the original subject said new words to the model for predictions, once when the original sound files used to create the model were used for predictions, once when a different subject used the model for predictions, and once when background noise was added to the original subject saying words for model predictions.

Review the results of each of these four data sets. What percent of the time did the model accurately predict the number for each of these data sets? If it did not correctly predict the number, how close did it come? In other words, did the correct number produce an error that was close to being the lowest error compared to the other numbers?

Discussion Questions

1. Which areas of the brain are important for speech processing?

2. What structures of the ear are important for transducing sound waves into auditory stimulation to the brain?

3. How well did the algorithm work to predict what number a person was saying? Which sets of trials (original words from the original subject, new words from original subject, new words from a new subject, or new words from original subject with background noise) did the algorithm predict most accurately? Why do you think this was?

4. Were there any words that produced similar spectral patterns and hence the model had a difficult time distinguishing between them? If so, which ones were they and why?

5. Can you suggest different algorithms which may be more effective at recognizing speech from a wide array of variables such as different people, background noise, and speeds at which people speak?

6. Why is it so difficult to develop a computer program to detect speech from humans?

7. Dyslexia is a disorder in which people have phonological difficulties. People with this disorder have difficulty in sorting out the sounds within words. Ultimately they may have trouble with reading, writing, and spelling. Can you describe how understanding the neural circuitry of human speech recognition may also lead to effective treatments and techniques for treating dyslexia?

8. How would you go about developing a cochlear prosthetic to restore hearing to an individual with a damages inner ear? How can you take advantage of the tonotopic organization of the neural circuitry?

References

1. Guyton and Hall. Textbook of Medical Physiology, 9th Edition, Saunders, Philadelphia, 1996.

2. Rhoades, R and Pflanzer, R. Human Physiology. *Third Edition.* Saunders College Publishing, Fort Worth 1996.