Abstract

The abstraction of phenomena of everyday things can be seen in modern technology. Language can be abstracted down to letters and characters that can be typed into computers that will then interpret that information and process it. Even the human can be abstracted into general shapes (eyes, nose, mouth) which machine learning algorithms can recognize and attribute to a specific person's face. This abstraction is a process of breaking down such phenomena such as recognizing someone's face into simpler and simpler phenomena (recognizing face → recognizing shape of eye → recognizing eyelid → etc) until it can be represented by something simple enough such that a machine can handle it. So in order to recognize speech, we need to be able to abstract the human voice into such features. My research involves finding such a method to abstract the human voice and to do that, I explored two existing methodologies: NSIM and iVector, to see if it is able to discern one speaker from another to a degree that it performs better than humans who try to discern the same speakers.

1. Background

Speaker recognition has been researched for decades already, and it's really making its becoming very big right now as machine learning has really risen in the past decade as well. The impact of speaker recognition is very wide as it can be used in many applications. For example, recognition a specific person's speech can be used as a security measure. Being able to abstract someone's speech into data the machine efficiently can also allow things like a translator mobile app come true. Other knowledge that was acquired while trying to crack the problem of speaker recognition such as understanding how speech works can also be used in applications such as slowing down the speed at which words come in a sound file. My research really comes down to whether or not there is a way to find a discerning factor between people's voices.

Two things to understand before going into detail on my research are how the speech chain works and what features are.

1.1 Human Speech and Spectrograms

Human speech starts with the vibration of the vocal folds which creates an excitation signal. This signal then travels through the vocal tract, from the vocal folds to the lips, where the sound is shaped even more. A person can pucker their lips to create an "oo" sound or open their mouths and draw back their tongue to create an "aah" sound.

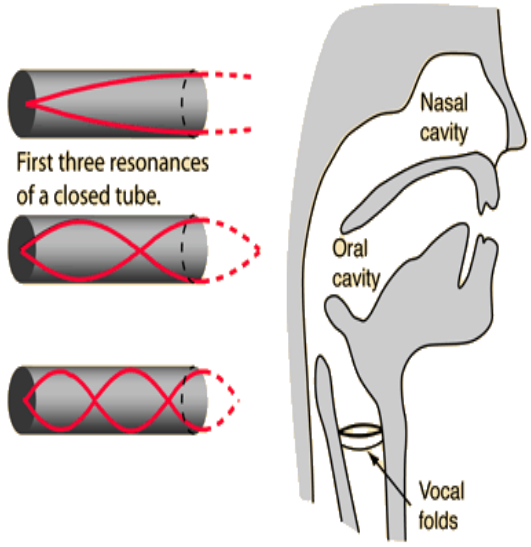

Figure 1. Human speech consists of an excitation signal from the vocal folds then speech forming through the vocal tract.

Speech is a combination of the signal from the vocal folds and the filter from the vocal tract, and that signal can be abstracted into an amalgamation of different frequencies [1]. It is also possible to abstract speech using digital signal processing techniques such as the Fast Fourier Transform to into key frequencies called formants which help define someone's vocal tract when they try to produce a certain sound. A way of representing these frequencies is with a spectrogram, which is a graph of a speech signal that has gone through the Fast Fourier Transform.

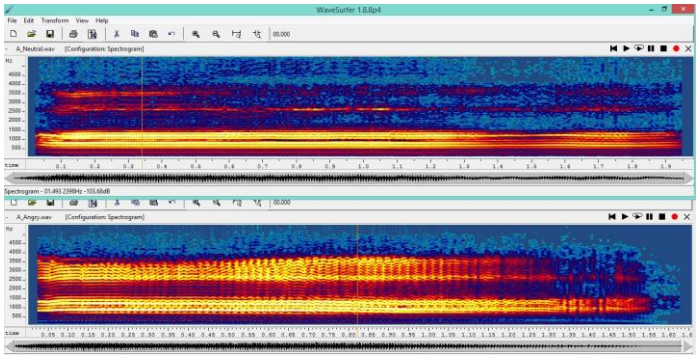

Figure 2. A man saying "aah" with neutral emotion (top) vs angry emotion (bottom)

The y-axis of this graph represents frequencies, the color represents the intensity of the frequencies, and the x-axis is time. The formants can be seen on this spectrogram as the thin white lines. Formants are a good way of discerning one speaker from another, but as shown here, emotions can also heavily vary what values the formants can get. This is one of the many problems in speaker recognition. Regardless, spectrograms and key aspects of how human speech is formed is used repeatedly in speaker recognition studies.

1.2 Features

The second thing to understand are what features are. There has already been a lot of discussion of being able to abstract speech into some key quantities. These quantities are better known as features in the field of artificial intelligence [3]. Features are essentially measurable quantities of the phenomena. In an example simpler than human speech, consider bench pressing. If someone bench pressing is given a smartwatch to continuously gather acceleration data in the x, y, and z direction, features can be extracted from that continuous plot of accelerations. Something as simple as the three averages of the x, y, and z accelerations can be used as a data point to be put on a feature graph.

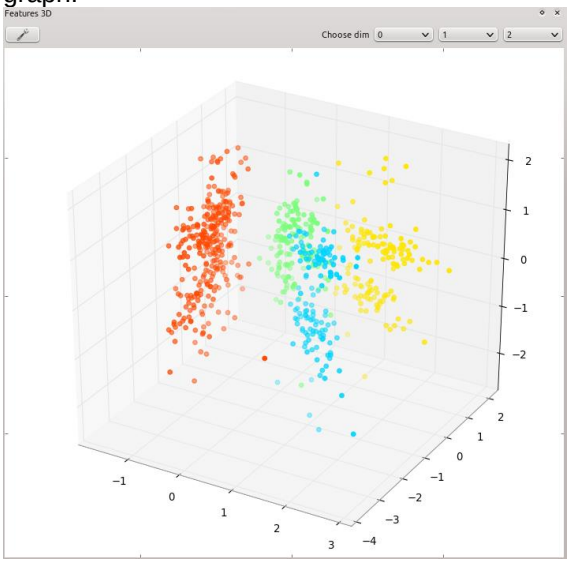

Figure 3. 3D Feature graph example

When a lot of data is collected for bench press, and perhaps other exercises such as deadlift and squats, you can see groups start to from the features that were collected from the raw data. Machine learning algorithms can be applied to differentiate one set of data points from the other and therefore differentiate the motion of one exercise from another. The same methodology can be applied to human speech, which is why features is a very key factor in speech recognition research.

2. Research Question

So the end goal of my research is to test some existing methodologies against human perceived results to see if these are potentially usable as features. I worked with two methodologies, and I needed to see if it could match or exceed the accuracy of human perception in telling the difference between two speakers. The first methodology was called NSIM, which utilizes an Auditory Neural Network and spectrograms to model the brain's response to hearing a sound, and consequently comparing two brain responses created from two speaker signals as images to spit out a similarity index [2].

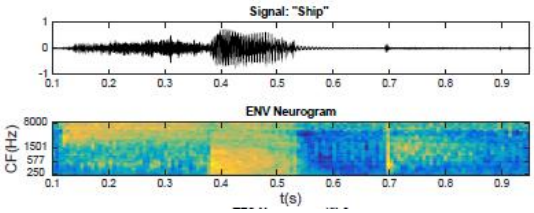

Figure 4. A speech signal compared with its neurogram using NSIM

The second approach was named iVector, which is a more machine learning based approach that trains the variability of a speech signal and derives from a universal model of human speech a long vector, named supervector that represents the qualities of a specific person's speech [4].

2.1 Challenges

Two big challenges in speaker recognition studies are finding good features and tackling the issue of intersession variability. When it comes to finding a good feature, a lot of testing is required to validate a certain methodology. Should my work with NSIM prove to match or exceed human perception, it could potentially be used as a means to measure the difference between people's speech. My work with iVector tries to tackle the second problem – people tend to sound different in separate speaking sessions even

though they are in a controlled environment saying the same words.

2.2 Specific Goals

My work with NSIM requires me to gather data on a set of 45 different speech signals (each signal is a 300 ms sample from the beginning of one of five females' saying the "aah" sound). These 45 speech signals were comprised of samples from 5 different female subjects. Each subject participated in three speaking sessions (held on separate days) and in each session they produced three speech samples. This created a total of 5 * 3 * 3 = 45 data. First I would need to tune the given NSIM code to compare two speakers – it was originally intended to test if a signal degrades through the telephone and therefore only compares a regular speech signal against a degraded version of that speech signal (there was some processing involved in creating that degraded speech signal). When I finish gathering data, there would be a 45 by 45 matrix when I finish gathering data, with each entry being the value that was outputted from plugging the two sound files in the NSIM code I created. I then needed to scale this matrix into a scaling that can be compared to the human perceptual results that were prepared prior by my SPAPL lab. Then I would be able to shoot for my first goal in determining whether or not NSIM is a methodology that can match or even exceed human perception.

My work with iVector is incomplete as I started working on it late. The main task I wanted to complete within this program was to assist Soo, my mentor in creating a Universal Background Model for females by manipulating existing code.

3. Methodology

As mentioned above, I modified the pre-existing NSIM code to be able to take in two speech signals and use the algorithms they provided to spit out a similarity index. I would gather a set of 45 by 45 data, and then scale the matrix from its current range (-1 to 1) to the range the human perceptual results were gathered in (1 to 10). Then I will be able to plot my scaled data against the perceptual data [5]. When fitting a linear regression on my plot, I can then determine based on the correlation whether or not NSIM is a methodology that can be useful in measuring speech difference.

When trying to create the female UBM, I simply needed to be able to trace through the current work and learn how to train the model with female data rather than male data.

4. Data and Results

Gathering data required a great deal of time, but I was able to grab the NSIM set of data by the time the perceptual results came in near the sixth week. When that happened, I was able to categorize the comparisons into three categories. The first category determined the comparison between the same speakers in the same session. The second category determined the comparison between the same speakers in different sessions. The third determined the comparison between different speakers. The following color coded plot was produced with I plotted my NSIM results against the human perceptual results.

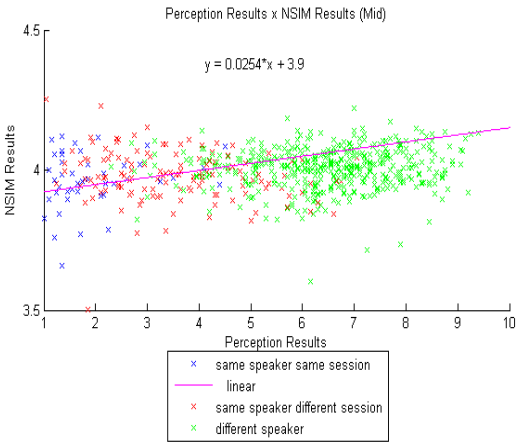

Figure 5. NSIM results plotted against human perceptual results.

As shown in the figure above, there is close to no correlation and NSIM is nowhere near being as good of a speech difference detector as human perception.

5. Discussion

It was slightly questionable when I read the description of how NSIM exactly works. The Auditory Neural Network portion of the computation seemed promising (and also quite time consuming), but in the end, what the algorithms really compared was the two neurograms that were produced as *images*. This

probably was not the best comparison method. Nevertheless, it worked for calculating speech degradation (of the same file) which explains why when I compared one token against the same token it yielded the highest similarity results by a great magnitude. This NSIM method must be very sensitive to differences in the neurograms. Additionally, the neurograms creation was not constant every time I called it. When I pass in a sound file in one trial and then the same sound file in another trial, the output is different! This probably added to the already existing variability between the data.

6. Further Work

The part that was the most uncertain was the comparison of the neurograms as images. Perhaps the Auditory Neural Network code can be used more effectively. Should people continue work with NSIM, it would most likely involve expanding on the neurogram itself rather than the comparison method. I would have continued working with iVector should my time here at research have been longer. It seems to be a much more credible approach as it incorporates modern computing capabilities.

7. References

[1] S. Möller et al., “Speech quality estimation: Models and trends, ” Signal Processing Magazine, IEEE 28.6 (2011): 18 - 28

[2] Hines, Andrew, and Naomi Harte. "Speech intelligibility prediction using a neurogram similarity index measure." *Speech Communication* 54.2 (2012): 306-320.

[3] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech Commun., vol. 52, no. 1, pp. 12–40, 2010.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio, Speech Lang. Process., vol. 19, no. 4, pp. 788–798, 2011.

[5] S. J. Wenndt and R. L. Mitchell, “Machine recognition vs human recognition of voices,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 4245–4248, 2012.

[6] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, “Automatic versus human speaker verification: The case of voice mimicry,” Speech Commun., vol. 72, pp. 13–31, 2015.