Grace Sacco, Cosette Newcomb, Zion Johnson

Project Proposal

Overview:

The topic of our project is transforming someone's voice or speaking style so that it can be more easily understood. This is a topic with far-reaching implications and importance because over 5% of the population suffers from hearing loss. There are also over 700 million non-Native English speakers in the world. It would be incredibly powerful to create a program which would make it easier for those who suffer from hearing loss and those who are non-Native speakers to better understand what is being spoken to them. Our initial hypothesis is that the most important speech features to look at will be clarity (making sure that the recording is not muffled by noise), the volume, the pacing, and the frequency. We find these components to be the most important for the following reasons. Starting with clarity, we believe that too much background noise will result in a lower understandability from the listener. We also consider volume to be a very important component because it can be much harder to understand someone when they are speaking very low. We believe that pacing will also prove to be an important factor because when someone speaks at a very fast pace or a very slow pace, understandability is often diminished. Finally, we believe that frequency will be an important factor because, as we learned in class, people at different ages can hear sounds at different levels of frequency. Thus, by editing the frequency of the sound, we are able to make the sample more understandable to different age groups.

Data:

We will be generating our own corpus of semantically unpredictable sentences by recording ourselves reading this list of sentences. We may also generate additional sentences using this program. We chose to generate this corpus because we think that using semantically unpredictable sentences will be the most useful in our human trials because there will be no context clues available to aid in the understanding the sentence, and thus, there will be less bias and a great accuracy. We will also be using Common Voice by Kaggle for additional sentences and testing, if necessary, because this corpus holds over 500 hours of recorded speech and its respective demographics.

Methods:

With the speech recordings from Common Voice, we will modify them using Praat transformation tools. Most specifically we will be using the Praat Manipulation feature that will allow us to manipulate the pitch and duration contours of a sound, configuring the intensity contour, and removing the noise feature. We hope that combining these tools, as well as any others that we find, would be helpful to make the sound recordings more clear and facilitate understanding of what was said. We will investigate our hypothesis using automated speech

recognition programs both before and during our modification process to see which speech transformations make the greatest impact on understandability.

Evaluation:

Currently, we plan to evaluate our results by doing both of the following: running our original and modified sound files through an automated speech recognition program (e.g. Speech-to-Text: Automatic Speech Recognition) and running a study in which we ask participants to type what they hear in certain recordings (both modified and original recordings) as well as comparing the original and modified recordings.

If our hypothesis holds true, we expect that our modified recordings will be more readily understandable in the human trial and will translate to text with a lower WER than the original recordings.

References:

Hawkins, Sarah. "Roles and representations of systematic fine phonetic detail in speech understanding." *Journal of phonetics* 31.3-4 (2003): 373-405.

Huckvale, Mark. "The new accent technologies: recognition, measurement and manipulation of accented speech." Beijing: Language and Culture Press, 2006.

Kain, Alexander, Akiko Amano-Kusumoto, and John-Paul Hosom. "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility." *The Journal of the Acoustical Society of America* 124.4 (2008): 2308-2319.

Picheny, Michael A., Nathaniel I. Durlach, and Louis D. Braida. "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech." *Journal of Speech, Language, and Hearing Research* 29.4 (1986): 434-446.

Nuesse, Theresa, et al. "Measuring speech recognition with a matrix test using synthetic speech." *Trends in hearing* 23 (2019): 2331216519862982.

Skuratovsky, Ilya. "Dynamically changing voice attributes during speech synthesis based upon parameter differentiation for dialog contexts." U.S. Patent No. 8,326,629. 4 Dec. 2012.

Vemuri, Sunil, et al. "Improving speech playback using time-compression and speech recognition." *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004.