Alan Liu
CPSC 490
May 3, 2018
Professor Steven Zucker

Sound Enhancement and Source Localization with Bone Conduction

Abstract

Audio speech enhancement has been an active area of development with applications in areas including internet of things devices, hearing aids, virtual and augmented reality. Nearly every application that uses a microphone and processes speech employs some form of speech enhancement and noise filtering.
Objective
Improve on those speech enhancement and noise filtering capabilities in the bone conduction setting. Combine noise filtering with level differences in a bone conduction application to deliver high fidelity sound.
Method
Use an audio speech recognition(ASR) dataset for audio samples and a testing framework to test different speech enhancement techniques including beamforming and GMM models. Create a bone conduction Arduino application that listens to input sounds and filters out noise, playing the noise filtered sound through the bone conduction transducer. Employ a bone conduction headphone application on a high frequency audio signal with various ILDs to test various different head related transfer functions for bone conduction.
Results
I was able to improve on the Word Error Rate (WER) on the speech recognition dataset from 18.38% to 15.13% by applying a band pass filter at frequency cutoffs of 50 and 3000 Hz testing with a Gaussian mixture model and expectation maximization algorithm. I created a web application that uses various high frequency sounds with staggered level delays to empirically test a bone conduction transfer function using commercial bone conduction headphones. The main limitation after applying some of the techniques to the Arduino application was in the speed of the digital audio processing. Without specialized hardware, it was difficult to obtain real-time results that would be necessary in many applications such as hearing aids. This work may be useful in bone conduction applications such as Google Glass in augmented reality to realistically simulate binaural sound in virtual environments.

Motivation

In everyday settings, ordinary conversations occur in a multitude of environments. For these conversations to be correctly understood, its important to distinguish voices and sources from each other just as the human brain can with the cocktail party effect. Localization of the sound takes this a step further. Source localization using interaural time differences (ITD) and interaural level differences (ILD) in binaural processing for air conduction is a fairly developed field[1], but recent advances in bone conduction technology as a new method of sound delivery has raised the question if location from binaural processing is possible in those bone conduction settings as well[2]. Using the results from source localization, there are also ways to directly transmit the sound to

listeners in applications such as VR binaural audio using head-related transfer functions (HRTF). These functions take a position in space and map them to sounds transmitted through air to the left and right ears using different ITD and ILD to provide the same localization effect. However, there has not been extensive research into similar functions in bone conduction, which instead, transmit sound using vibration through the bone needing to be optimized for different levels of ITD and ILD due to a faster speed of transmission.

1 Audio Processing

I used an audio speech recognition(ASR) competition dataset CHiME3 [3] that contained over 23Gb of audio samples in clean and noisy city environments. It contained audio files that were recorded at 16kHz from multiple microphones. Each audio file was a couple of seconds long containing phrases read from the wall street journal corpus. I chose this dataset because it provided a large amount of training data for modeling purposes and also had base code to modify the existing data.

I will now review some of the techniques used in the CHiME3 experimental code that I built off of and how they would be useful to apply in the overall scope of the project.

1.1 Beamforming

Because the dataset was composed of multiple microphones, it allowed the use of a signal processing technique which combines multiple signal inputs called beamforming. It utilizes the audio delay between the various microphones, knowing their relative positions to align the signal and reduce noise. In particular, this implementation of beamforming is called MVDR beamforming and has been shown to be effective at analyzing speech input [4].

Since we are attempting to simulate binaural sound, we would use at least two microphones to capture the incoming sound. Since we would know the width of someone's head in this rigid structure, we would also know the relative position of the two microphones, allowing us to use beamforming to further enhance our audio data. Although beamforming applications have been evaluated in the hearing context before [5], advances in beamforming over the last two decades have suggested that recent results may be more promising.

1.2 Frequency Filters

Because our goal is to focus on the human voice, I tested a band pass filter with a series of frequency cutoffs that would allow us to enhance the frequencies of speaker, and reduce the amplitude of noise, which may lie outside of the frequency range.

After empirically testing a series of frequency cutoffs, I found that on the dataset provided, the best cutoffs were around the range of 50 Hz to 3200 Hz, which is fairly standard for the human voice. One drawback is that noise may also fall within this range, and the quality of the voice may decrease even if the speech recognition system is still able to classify the voice correctly as text.

For our bone conduction application, it would be fairly straightforward to pre-process the signal using a frequency filter. This would allow us to focus on speech, but it may mean a lack of support in other frequencies in niche use cases such as for musical purposes.

1.3 Kaldi Speech Recognition

I used the open-source Kaldi speech recognition toolkit [6] to evaluate the performance of my improvements. Published in 2011, it is the standard for evaluating audio speech enhancement techniques. It uses acoustic modeling with subspace Gaussian mixture models along with standard Gaussian mixture models to evaluate potential word matches in the enhanced audio data.

1.4 Results

Best results after applying a band pass filter on a frequency range from 50-3200 Hz and evaluating with a GMM:

```
best overall dt05 WER 32.08% (language model weight = 11)
--------------------
dt05_simu WER: 22.29% (Average), 16.14% (BUS), 29.00% (CAFE), 17.23% (PEDESTRIAN), 26.81% (
    STREET)
--------------------
dt05_real WER: 41.87% (Average), 45.40% (BUS), 43.11% (CAFE), 34.05% (PEDESTRIAN), 44.90% (
    STREET)
--------------------
et05_simu WER: 27.78% (Average), 18.83% (BUS), 31.99% (CAFE), 31.57% (PEDESTRIAN), 28.74% (
    STREET)
--------------------
et05_real WER: 76.38% (Average), 86.60% (BUS), 83.68% (CAFE), 78.57% (PEDESTRIAN), 56.69% (
    STREET)
--------------------

local/chime3_calc_wers.sh exp/tri3b_tr05_multi_enhanced1 enhanced1
compute dt05 WER for each location

--------------------
best overall dt05 WER 15.13% (language model weight = 11)
--------------------
dt05_simu WER: 10.19% (Average), 8.63% (BUS), 12.21% (CAFE), 8.82% (PEDESTRIAN), 11.09% (STREET
    )
--------------------
dt05_real WER: 20.07% (Average), 23.01% (BUS), 18.73% (CAFE), 17.72% (PEDESTRIAN), 20.84% (
    STREET)
--------------------
et05_simu WER: 11.11% (Average), 8.33% (BUS), 12.22% (CAFE), 11.34% (PEDESTRIAN), 12.57% (
    STREET)
--------------------
et05_real WER: 37.39% (Average), 48.07% (BUS), 40.87% (CAFE), 36.17% (PEDESTRIAN), 24.45% (
    STREET)
--------------------
```

Control GMM analysis using beamforming without frequency filter:

```

--------------------
best overall dt05 WER 18.38% (language model weight = 10)
--------------------
dt05_simu WER: 18.06% (Average), 18.97% (BUS), 22.21% (CAFE), 14.07% (PEDESTRIAN), 17.01% (
    STREET)
--------------------
dt05_real WER: 18.70% (Average), 26.20% (BUS), 17.64% (CAFE), 13.06% (PEDESTRIAN), 17.90% (
    STREET)
--------------------
et05_simu WER: 21.30% (Average), 19.16% (BUS), 23.76% (CAFE), 21.52% (PEDESTRIAN), 20.77% (
    STREET)
--------------------
et05_real WER: 33.09% (Average), 49.52% (BUS), 33.26% (CAFE), 28.12% (PEDESTRIAN), 21.46% (
    STREET)
--------------------

local/chime3_calc_wers.sh exp/tri3b_tr05_orig_clean noisy
compute dt05 WER for each location

--------------------
best overall dt05 WER 52.62% (language model weight = 10)
--------------------
dt05_simu WER: 49.45% (Average), 44.88% (BUS), 60.40% (CAFE), 42.85% (PEDESTRIAN), 49.69% (
    STREET)
--------------------
dt05_real WER: 55.78% (Average), 70.33% (BUS), 61.71% (CAFE), 41.98% (PEDESTRIAN), 49.09% (
    STREET)
--------------------
et05_simu WER: 63.13% (Average), 60.85% (BUS), 66.53% (CAFE), 67.61% (PEDESTRIAN), 57.53% (
    STREET)
--------------------
```

```
et05_real WER: 80.00% (Average), 95.83% (BUS), 82.52% (CAFE), 81.41% (PEDESTRIAN), 60.25% (
    STREET)
```

As we can see, we significantly improved the best WER from an average of 18.38% to 15.13%. After reviewing the analysis, it performed significantly better on the simulated environments, which may have skewed some of the results.

Other techniques I attempted that were not as successful included applying a Gaussian smoothing filter before reducing the noise, zeroing out noise between spikes in volume, dynamically ignoring or isolating various inputs to reduce introduced noise, using various different frequency cutoffs for the band pass filter.

1.5 Code

Code for this section can be found under the CHiME3 folder. The modified matlab script used to enhance the data can be found under `CHiME3/tools/enhancement`.

2 Microprocessor Bone Conduction Prototype

As a part of the analysis, I attempted to create a prototype of a bone conduction device using an Arduino and a bone conduction transducer. The goal was to create a device that would allow for prototyping of the same audio processing techniques that were explored earlier.

I attached a Electret microphone to the analog input of a Sparkfun Redboard device. This allowed me to then read the audio signal and play it out from the digital pins of the microprocessor.

A challenge is that the Ardunio is not designed for audio processing applications, and thus is not suited for real-time polling or output. As a result, it was necessary to use multiple ports to obtain the speed necessary for high fidelity audio output.

After attempting to use digital filtering on the Arduino, it was clear that the filtering delay was causing stuttering in the resulting output, suggesting that it would not be powerful enough to test real-time audio applications for bone conduction. Further work may be explored in using a specialized analog to digital converter board to optimize the input.

Arduino Code can be located under the `arduino` folder.

3 Evaluating ILD of Bone Conduction

To evaluate the head-related transfer function of bone conduction, I used a commercial bone conduction device with a frequency response of up to 20kHz. The high frequency response was important due to the higher attenuation between the ears in bone conduction. This is a problem for bone conduction, since bone, which is a denser medium than air, allows faster and quicker conduction, interfering with the effect of the head shadow, which is one of many binaural location cues. [7] In addition, higher frequencies also usually create a larger acoustic head shadow effect. By simulating a larger head shadow and using a higher frequency, we can artificially create an acoustic environment.

The binaural cues in particular we are interested in are the interaural time differences and the interaural level differences. However, because we have focused in particular on higher frequencies, we can focus on the interaural level differences, which dominates our judgment of location for higher frequency sounds.

To test these results, I used a high frequency audio sample, and created 31 different versions with various levels of audio gain equidistant from -15dB to 15dB which simulates

a 180 degree arc if we were using a normal air conduction head related transfer function. These gain levels mirror the arc of locations given by the observations done by SADIE in their binaural air conduction localization measurements (https://www.york.ac.uk/sadie-project/binaural.html).

I have uploaded a test website in the links below. The source code is available under the `audiotest.html` file in the root directory.

4 Future Work

Further work could be in developing a more rigorous auditory model with variable frequencies, time delays and variable gains. The tool I created may be useful in preliminary evaluations of location transmission in bone conduction. However, I made a series of simplifications including using a fixed audio sample rather than dynamically changing frequencies, and using a fixed environment, whereas real-world applications would need to handle both of these issues.

Additionally, we could apply beamforming and frequency filters digitally to a specialized analog to digital converter, which would allow us to use more complex techniques like a neural net model to determine the level of gain or filtering for each environment.

5 Acknowledgments

I would like to thank Professor Steven Zucker for his advice in finding a direction to tackling this problem. Dr. Joseph Santos-Sacchi for his valuable insights on the auditory system, Dr. Sunil Puria for his research on current bone conduction techniques as well as Professor Mary Lui and the TD Mellon Forum for gracious funding and connecting me with TD alumni and advisers. I appreciate all the help and support.

References

[1] Kohlrausch A., Braasch J., Kolossa D., Blauert J. (2013) An Introduction to Binaural Processing. In: Blauert J. (eds) The Technology of Binaural Listening. Modern Acoustics and Signal Processing. Springer, Berlin, Heidelberg

[2] Zeitooni, Mehrnaz, Elina Mki-Torkko, and Stefan Stenfelt. "Binaural hearing ability with bilateral bone conduction stimulation in subjects with normal hearing: Implications for bone conduction hearing aids." Ear and hearing 37.6 (2016): 690-702.

[3] Yoshioka, Takuya, et al. "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices." *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on.* IEEE, 2015.

[4] Higuchi, Takuya, et al. "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[5] Greenberg, Julie E., and Patrick M. Zurek. "Evaluation of an adaptive beamforming method for hearing aids." The Journal of the Acoustical Society of America 91.3 (1992): 1662-1676.

[6] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[7] Puria, Sunil, Richard R. Fay, and Arthur N. Popper. The middle ear. Springer,, 2013.