Improving Speech Communication in High Noise Environments

Jacob B. Munger, Scott L. Thomson

Department of Mechanical Engineering
Brigham Young University
Provo, UT, USA

Abstract. During speech the vocal folds vibrate resulting in audible sounds that are transmitted through the vocal tract as well as vibrations that are transmitted through the body tissue to the skin surface. These skin surface vibrations can be detected by contact microphones and used to transmit speech. However, the skin attenuates high frequency content and in some locations muffles the signal resulting in poor speech quality. To reconstruct a signal that better matches the microphone signal a finite impulse response filter is fit to an average transfer function of the accelerometer signal. When implemented this filter restores much of the lost frequency content and in the presence of background noise results in a signal with good intelligibility and less noise than the microphone signal.

1 Introduction

During speech the vocal folds vibrate, resulting in audible sounds. In addition to being transmitted through the vocal tract, these vibrations are also transmitted through several layers of various types of tissue throughout the head and neck, resulting in small, but measurable, skin surface vibration. Contact microphones sense these skin surface vibrations for speech transmission, as opposed to acoustic microphones that sense air vibrations that radiate from the mouth.

Contact microphones have one significant advantage over acoustic microphones in environments with elevated ambient noise levels in that they sense very little background noise. In comparing the use of throat contact microphones to acoustic microphones for use in rotary-wing aircraft, Acker-Mills et al. (2004) found that throat microphones had approximately a 10 dB higher signal-to-noise ratio. Commercially available contact microphones, however, suffer from poor speech quality and intelligibility (Acker-Mills et al., 2004; Shimamura and Tamiya, 2005). This is a result of the skin vibrations being influenced by the many tissue layers (e.g. skin, fat, muscles, bones) of the neck or face between the contact microphone location and the vocal tract.

This study has been conducted in two parts. This paper will present the results from the second part of the study. The first portion of the study investigated where the frequency response on the skin is most like the signal picked up from the air (Munger and Thomson, in review). In this portion of the study we found that while the skin attenuates high frequency content, some high frequency content can still be detected at some locations on the face. We determined that locations other than on the throat, where many currently used contact microphones are placed, can pick up good speech signals. We also found that different types of sounds were picked up better at different locations. The nasal sounds were picked up better on the nasal bone while the vowel sounds were picked up best above the upper lip. Based on power spectral density (PSD) comparisons it was found that the best locations overall for speech transmission via a contact microphone are the nasal bone, above the upper lip, the temple and the zygomatic bone.

Although some of the locations yielded speech signals that were generally understandable, many locations on the face and neck produced signals that were muffled and hard to understand. Many of these locations that produced poorer speech signals were at locations that are more convenient to place a contact

microphone, such as over the vocal folds or in front of the ear. The objective of this portion of the research was to find a simple filtering method to reconstruct from the accelerometer data a clear signal that sounds more like the microphone speech signal. This filter was then be applied to accelerometer signals of speech recorded in the presence of elevated background noise to compared with the microphone signals of the same speech recorded in the presence of elevated background noise

2 Methods

2.1 Data Collection

To collect the skin vibration data small accelerometers were attached to 15 locations on the face and neck of 14 male and 10 female subjects using medical-grade double-sided adhesive tape. In this paper we will present the results for the accelerometers placed in front of the ear and over the vocal folds for one male subject. These accelerometers measure the magnitude and frequency of the skin vibration at each location while the subject speaks. An acoustic microphone was used to simultaneously acquire the audible speech.

The subjects sustained the vowels /a/ (b**a**t), /oo/ (b**oo**t), /ah/ (c**augh**t), /ee/ (f**ee**t), the nasals /m/ and /n/, and the fricative /f/ for 4 to 5 seconds each. The subjects also said the phonetically balanced phrases:

- *Rice* is often served in round bowls.

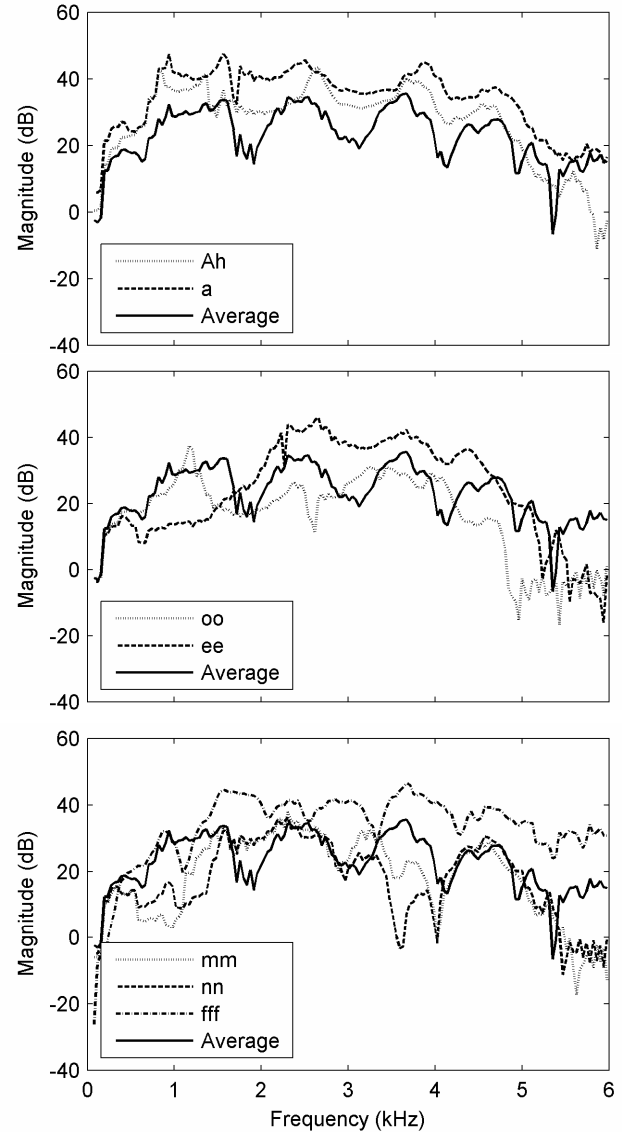

Figure 1. Transfer functions for the sounds with their average for one male subject over the vocal folds.

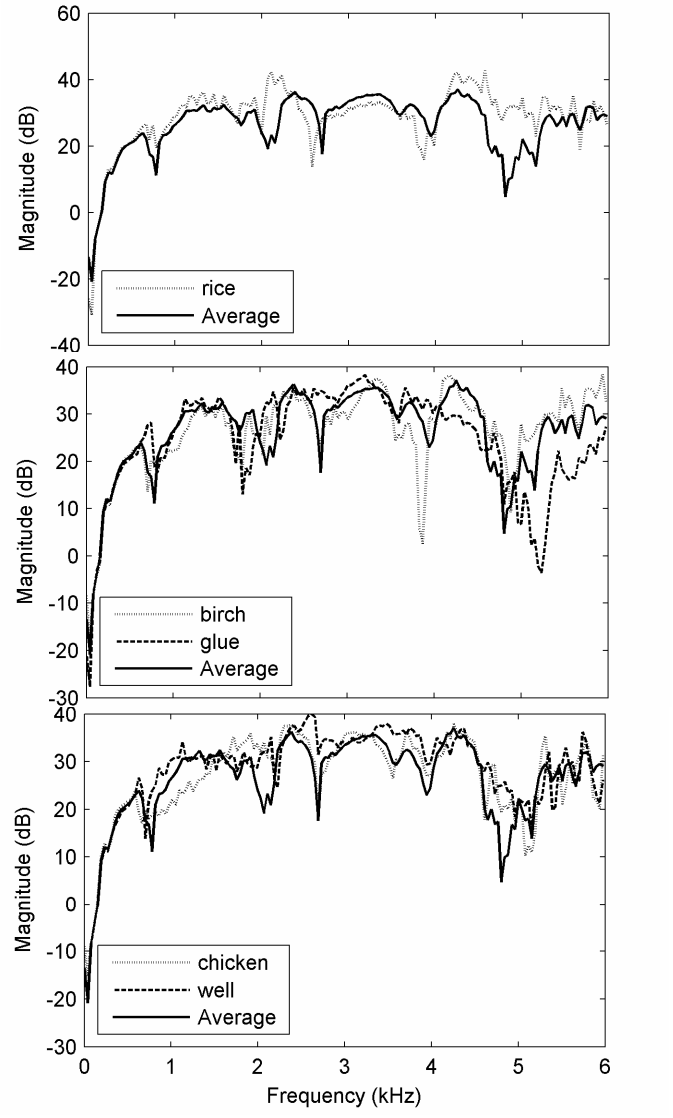

Figure 2. Transfer functions for the phrases with their average for one male subject over the vocal folds.

- The *birch* canoe slid on the smooth planks
- *Glue* the sheet to the dark blue background.
- These days a *chicken* leg is a rare dish.
- It's easy to tell the depth of a *well*.

(Italicized words are used to reference the phrases in the figures.) The sounds and phrases where recorded in a quiet environment as well as with 95 dB background white noise.

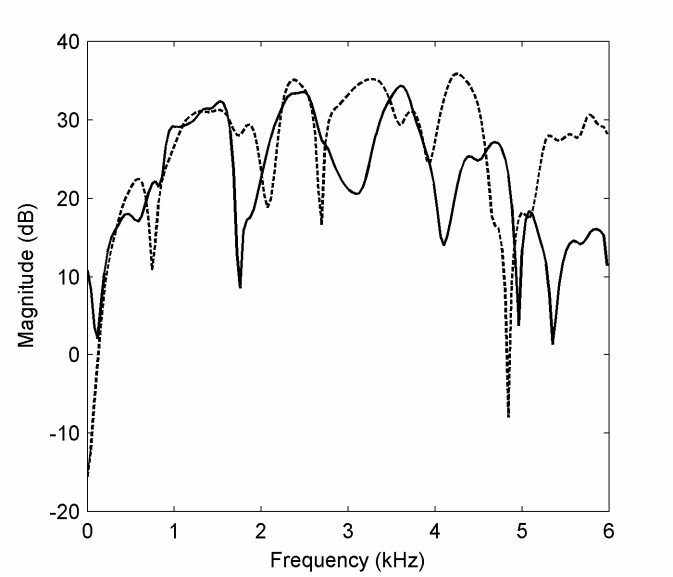

Figure 3. . Comparison of the average transfer functions for the sounds (——) and phrases (·····) for one male subject over the vocal folds.

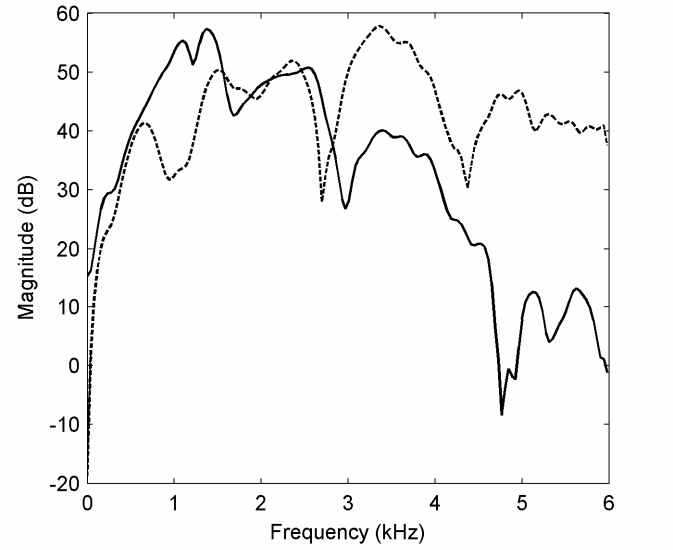

Figure 4. Comparison of the average transfer functions for the sounds (——) and phrases (·····) for one male subject in front of the ear.

2.2 Data Analysis

2.2.1 Transfer Function Estimate

The sound files were first truncated so that only the portion of the data where the subject was speaking is analyzed. For a given sound or phrase the transfer function estimate is first calculated for a particular location using the following equation

$$T_{xy}(f) = \frac{P_{xy}(f)}{P_{xx}(f)},$$

where T_{xy} is the transfer function estimate, P_{xy} is the cross power spectral density of the accelerometer to the microphone, and P_{xx} is the PSD of the accelerometer. The MatLab command tfestimate was used to perform this calculation.

The transfer functions for all the sounds were averaged at each frequency to obtain an average transfer function for all sounds. An average transfer function was also found for the phrases. These average transfer functions $T_{xy,avg}$ were then smoothed using a triangular smoothing weighted average

$$T_{xy,avg,i,Smooth} = \frac{T_{xy,i-2} + 2T_{xy,i-1} + 3T_{xy,i} + 2T_{xy,i+1} + T_{xy,i+2}}{9}.$$

This smoothing average includes the two points on either side of the current value but weights them less than the current value. Smoothing was performed in order to better fit the filter coefficients.

2.2.2 Filter Coefficients

A finite impulse response (FIR) filter was then fit to these average transfer functions. These filter coefficients were found using the fir2 function in MatLab. The fir2 function returns the n^{th} order filter numerator coefficients given the frequency and magnitude information of the transfer function estimate. The filter has the form

$$B(z) = b(1) + b(2)z^{-1} + \cdots + b(n+1)z^{-n}$$

The filter coefficients were only calculated for data up to a selected cutoff frequency of 6 kHz. This was done in order to have a better fit for the portion of data that is important for speech transmission. For the results presented here a FIR filter order of 300 was used and was found to match the average transfer functions with very little error. However, a lower order

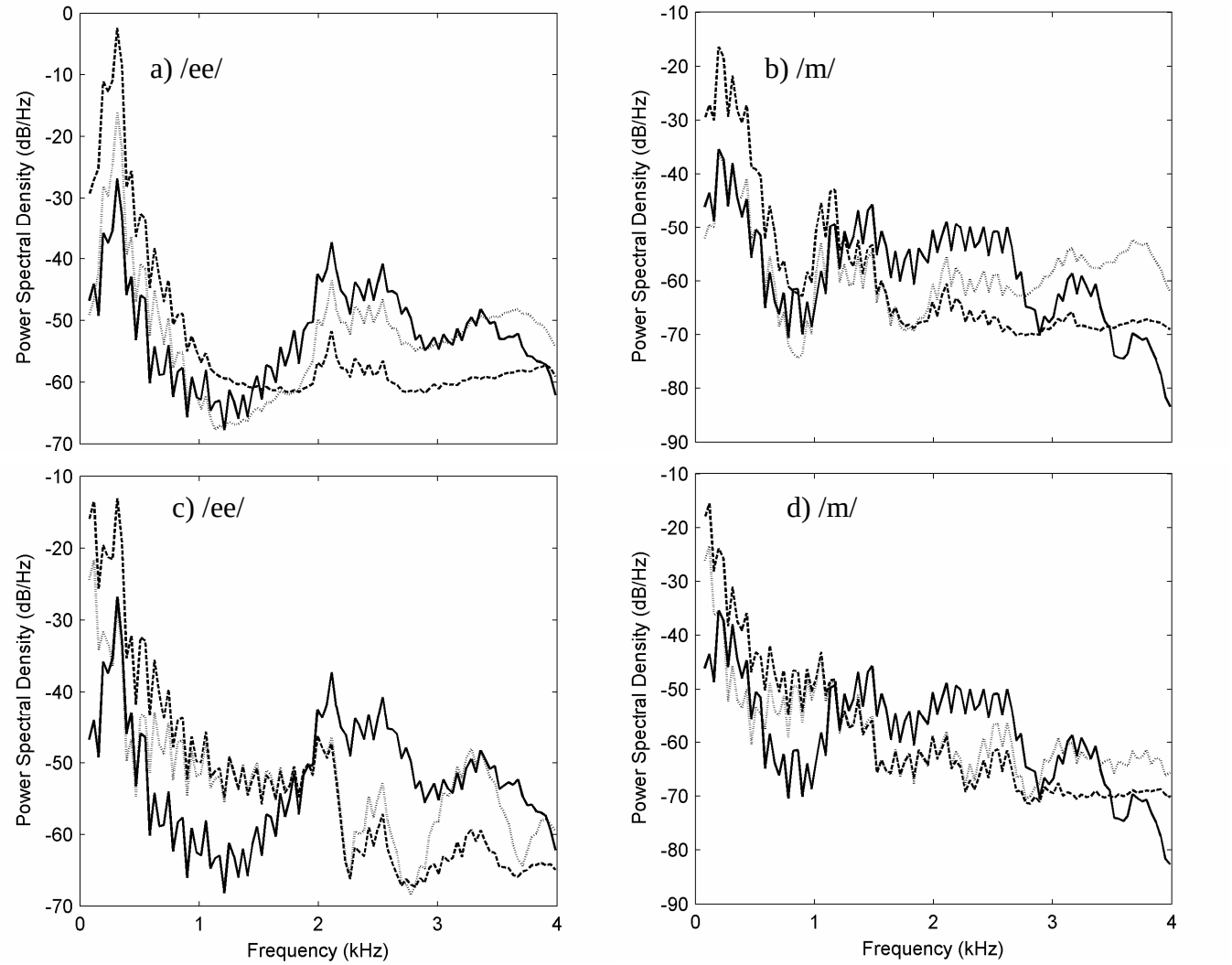

Figure 5. Power spectral density for the ── Microphone; ····· Filtered accelerometer signal; - - - Unfiltered accelerometer signal. a) Front of ear sound /ee/. b) Front of ear /m/. c) Over vocal folds sound /ee/. d) Over vocal folds ear /m/.

filter would likely work just as well since the data being fit is an average so slight deviation would likely not effect the results.

The accelerometer signal was then passed through a low pass filter to remove all frequency content above that for which the transfer function coefficients were calculated. This was done using a butterworth filter with a cutoff frequency of 4 kHz.

The low pass filtered accelerometer signal was then passed through a filter with the calculated FIR coefficients to reconstruct a signal that sounds more like the microphone. This was done using the 'filter' function in MatLab.

2.2.3 Power spectral Density

The PSD of the microphone, accelerometer, and the filtered accelerometer were found using Welch's method (Welch, 1967) via the "pwelch" function in MatLab. The accelerometer signals were then normalized to yield the same area under the PSD curve as the microphone signal between 0 and 4 kHz.

$$PSD_{i,norm} = PSD_i + \frac{\int_0^{f_c} PSD_{mic}(f)\,df - \int_0^{f_c} PSD_i(f)\,df}{f_c},$$

where $PSD_{i,norm}$ is the normalized PSD for location i, PSD_{mic} is the PSD of the microphone, PSD_i is the PSD of the accelerometer at location i, f is the frequency and f_c is the upper frequency (4 kHz). The integrals where numerically calculated using the trapezoidal method.

2.2.4 Spectrogram

A spectrogram for the phrase, "These days a *chicken* leg is a rare dish," was calculated for the location in front of the ear. The spectrogram was generated using the 'spectrogram' function in MATLAB, which calculates the PSD estimate over select time intervals using Welch's method.

3 Results

3.1 Transfer Function Comparison

Figure 1 shows the transfer functions over the vocal folds of each sound as well as the average transfer function over all sounds for one male subject. The sounds /Ah/ and /a/ have similar behavior and generally have magnitudes greater than the average. The sound /oo/ has magnitudes generally at or below the average, while the sound /ee/ starts below the average and then at 1.5 kHz the values become greater than the average. Sounds /m/ and /n/ also have very similar transfer functions, with some variations around 1 kHz and 3.5 Khz. Their magnitudes are generally a little below the average magnitude. Sound /f/ generally has magnitudes greater than the average. All sounds, except /f/, decline in magnitude after 5 kHz.

The transfer functions of each phrase were also calculated and averaged over all the phrases and are shown in figure 2. Figure 2 shows that all of the transfer functions for the phrases are very similar and generally follow the average with only slight variations.

The average transfer function of the sounds and phrases are compared in figures 3 and 4 for the locations over the vocal folds and in front of the ear. For the location over the vocal folds figure 3 shows that the two average transfer functions are fairly similar out to 2.5 kHz, but differ quite a bit after that. For the location in front of the ear figure 4 also shows that the transfer functions of the sounds and phrases are similar out to 2.5 kHz. After 2.5 kHz both transfer functions follow the same general trends but the transfer function of sounds is attenuated at frequencies greater than 3 kHz, but the transfer function of the phrases is not. This could be due to the presence of more fricative sounds and high frequency content in the phrases than are in the sounds analyzed.

3.2 Power Spectral Density

Figure 5 shows the PSD for sounds /ee/ and /m/ for the locations in front of the ear and over the vocal folds. Figure 5a shows that for the sound /ee/ in front of the ear the filter was able to reconstruct a signal which matches the frequency content of the microphone much better than the unfiltered accelerometer. Figure 5b shows that the filter worked well out to about 2 kHz. Figures 5c,d however, show that the filter over the vocal folds had minimal improvement for the sounds /ee/ and /m/.

Figure 6 shows the PSD for the phrase "These days a chicken leg is a rare dish" for the locations in front of the ear and over the vocal folds. For both locations the filtered accelerometer signal matches the microphone signal much better than the unfiltered accelerometer signal.

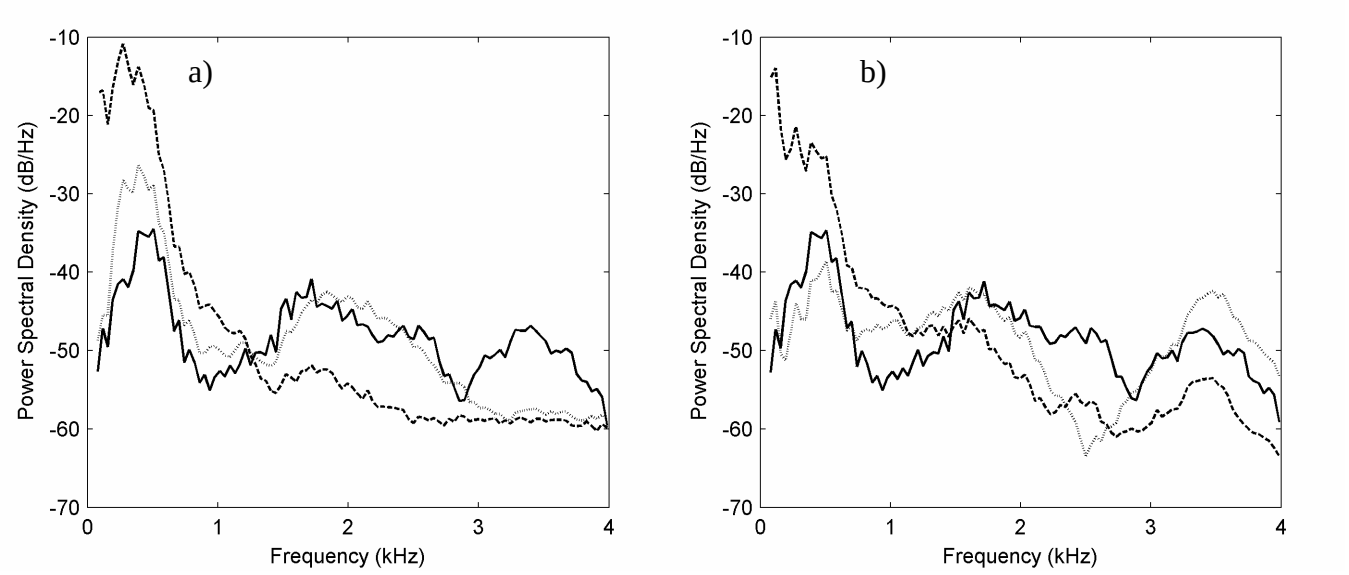

Figure 6. PSD of "These days a chicken leg is a rare dish". —— Microphone; ····· Filtered accelerometer signal; - - - Unfiltered accelerometer signal. a) In front of ear. b) Over vocal folds.

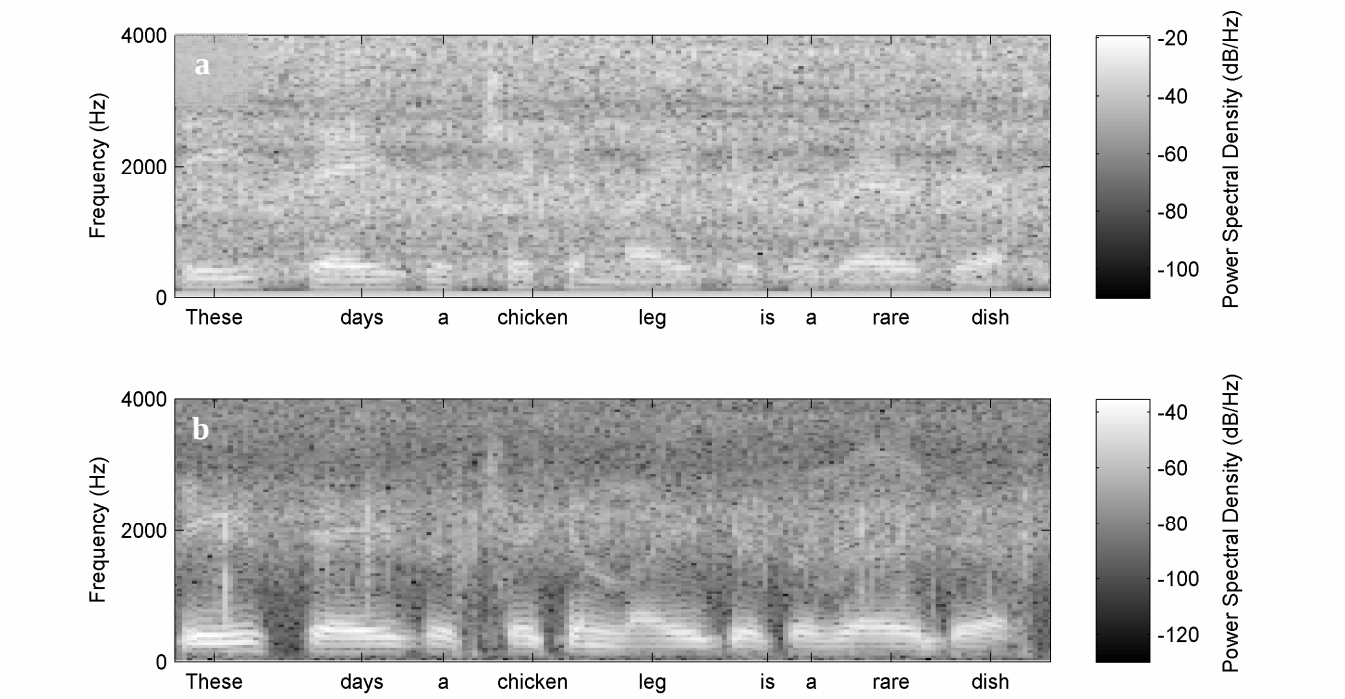

Figure 7. Spectrogram of the phrase 'Rice is often served in round bowls" recorded with 95 dB background noise. a) Microphone b) Filtered signal from in front of the ear.

3.3 Spectrogram

This filter was then applied to the phrase recorded with background noise. The spectrograms of the microphone and filtered accelerometer signals recorded with 95 dB background noise are shown in figure 7. Figure 7 shows that the when compared to the filtered accelerometer signal the noisy microphone signal has a much lower signal to noise ratio. Although there is still some noise in the filtered accelerometer signal at high frequencies the signal to noise ratio at lower frequencies is much better. When listening to the noisy microphone signal and the filtered accelerometer signal there is still some noise in the filtered signal but it is much quieter than in the microphone signal. The filtered signal is slightly muffled but the speech quality overall is good.

4 Discussion of Results

4.1 Filter Design

Figure 1 shows that each sound has a transfer function that follows a trend but each transfer function varies depending on the sound. However, figure 2 shows there is much less variation in the transfer functions of the phrases. In comparing figures 5 and 6 the PSDs of the filtered signal for the phrases matches the microphone much better than the PSDs of the filtered signals for the sounds. This indicates that generating a FIR filter using the average transfer function of the phrases results in a more accurate reconstruction of the microphone signal than does the filter generated from the average transfer function of the sounds. The phonetically balanced sentences are more representative of how often each phoneme is used in speech thus resulting in a filter that is more accurate. Using a filter that weights each phoneme the same would result in a filter that may disproportionably favor certain phonemes that are not used as often.

4.2 Signal Reconstruction

Figures 6 and 7 show that the FIR filters can reconstruct the signal from the accelerometer to more accurately match the frequency response of microphone signal. Figure 7 shows that in the presence of background noise the filtered accelerometer signal results in a signal which has much less background noise than the microphone while keeping good intelligibility.

4.3 Filter Type

In this paper a FIR filter was used to fit an average transfer function in order to reconstruct a more

intelligible signal from the skin vibrations. Using the phrases to develop an average transfer function has resulted in a filter that restores much of the lost frequency content and results in a good speech signal. However, this average transfer function only represents five sentences which may limit its overall effectiveness. Other advanced filter techniques, such as a least means squares (LMS) adaptive filter, may result in a filter with improved results and will be explored.

5 Conclusions

This paper has shown that even though the original accelerometer signal may not match the microphone signal the implementation of a FIR filter that corresponds to an average transfer function can reconstruct a signal that is a better representation of the microphone signal. When implemented in the presence of background noise the filtered signal has reduced noise and provided good intelligibility when compared to the noisy microphone signal.

Future research will involve expanding this work over more subjects and sounds to generate more generalized results. More advanced filtering techniques will be explored to improve intelligibility and noise reduction of the filtered signal. Jury listening tests will be needed to verify that the filtered signals are preferred over the noisy microphone signals.

References

Acker-Mills, B. E., Houtsma, A. J. M., and Ahroon. W. A. (2004). “Speech intelligibility in noise using throat and acoustic microphones,” USAARL Report No. 2004-13

Munger, J. B., Thomson, S. L., (in review), “Frequency response of the skin on the head and neck during speech”, J. Acoustical Society of America

Shimamura, T., Tamiya, T. (2005). “A reconstruction filter for bone conduction speech” *Circuits and Systems, 2005. 48th Midwest Symposium on* 1847-1850

Welch, P.D. (1967), "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," IEEE Trans. Audio Electroacoustics, **AU-15**:70-73.