ON THE POTENTIAL FOR ARTIFICIAL BANDWIDTH EXTENSION OF BONE AND TISSUE CONDUCTED SPEECH: A MUTUAL INFORMATION STUDY

Rachel E. Bouserhal ⋆ *Tiago H. Falk*† *Jérémie Voix*⋆

⋆ École de Technologie Supérieure, Université du Québec, Montréal, Canada
†Institut National de la Recherche Scientifique, Université du Québec, Montréal, Canada

ABSTRACT

To enhance the communication experience of workers equipped with hearing protection devices and radio communication in noisy environments, alternative methods of speech capture have been utilized. One such approach uses speech captured by a microphone in an occluded ear canal. Although high in signal-to-noise ratio, bone and tissue conducted speech has a limited bandwidth with a high frequency roll-off at 2 kHz. In this paper, the potential of using various bandwidth extension techniques is investigated by studying the mutual information between the signals of three uniquely placed microphones: inside an occluded ear, outside the ear and in front of the mouth. Using a Gaussian mixture model approach, the mutual information of the low and high-band frequency ranges of the three microphone signals at varied levels of signal-to-noise ratio is measured. Results show that a speech signal with extended bandwidth and high signal-to-noise ratio may be achieved using the available microphone signals.

Index Terms— Mutual Information, Gaussian Mixture Models, Bandwidth Extension, Bone Conducted Speech, In-ear microphone

1. INTRODUCTION

Communication is a vital part of any workplace. Providing good communication becomes a difficult task in environments with excessive noise exposure where workers must be equipped with Hearing Protection Devices (HPD). Depending on the type of HPD used, the spectrum of the noise and the wearer's hearing ability, the use of HPDs can greatly limit speech intelligibility [1]. To compensate for these conflicting needs, radio communication headsets that aim at providing both good communication and good hearing protection have been developed. Their performance, however, is often suboptimal, especially in terms of communication. Currently available headsets either pick up a speech signal that is masked by noise or has a limited bandwidth. In either case, both the intelligibility as well as the quality of the signal are degraded. Ideally, a communication signal must have a high Signal-to-Noise Ratio (SNR) as well as a wide bandwidth. However, current communication headsets fail to provide both simultaneously. Most commonly, these headsets involve circumaural HPDs equipped with a boom microphone placed in front of the mouth. Although so-called "noise reduction" boom microphones are directional, they still pick up speech that is often degraded by background noise, resulting in low SNR. One way to alleviate this problem is the use of active noise reduction techniques on the recorded speech signal [1, 2, 3]. Active noise reduction techniques still remain a step in the right direction, however, their performance is unreliable in high frequency noise [4].

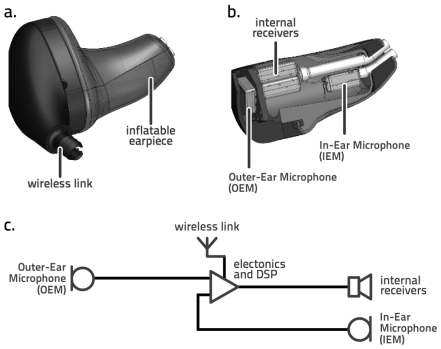

Fig. 1. Overview of communication headset (a), its electro-acoustical components (b), and equivalent schematic (c).

In an effort to solve the problem of low SNR, non-conventional ways of capturing speech that rely on bone and tissue conduction have been employed. Namely, throat microphones [5] and more recently occluded-ear speech capturing [6] have been used simultaneously with hearing protection. Signals originating from bone and tissue conduction have better SNRs than those recorded conventionally, but they have their own limitations such as a lower bandwidth, decreased quality and intelligibility.

Various bandwidth extension techniques have been employed for the enhancement of bone and tissue conducted speech [7, 8, 9]. Recently, a new communication headset was developed [6] comprised on an instantly custom molded HPD equipped with an Outer-Ear Microphone (OEM), an In-Ear

Microphone (IEM) and a Digital Signal Processor (DSP) (see Fig. 1), thus opening doors to new bandwidth extension capabilities.

The OEM can capture a wideband speech signal transmitted through air conduction. OEM signal quality and intelligibility are directly related to the background noise levels and types. By contrast, the IEM, placed inside the ear canal is less affected by background noise due to the attenuation offered by the custom-molded earpiece. The IEM also takes advantage of the occluded ear canal [10], thus enabling the recording of bone and tissue conducted speech from inside the ear. While the IEM is less sensitive to environmental noise, it does suffer from other limitations, such as a narrow bandwidth around 2 kHz. Such limited bandwidth poses a challenge for the HPD, particularly in extremely noisy environments where residual noise "leaks" to the IEM hindering its intelligibility. In this paper, we explore the potential benefits of having an IEM and an OEM for bandwidth extension purposes. For comparison, we also utilize an ideal reference microphone (REF) placed in front of the mouth, thus capturing a high SNR, wide bandwidth speech signal.

As mentioned previously, the IEM signal has a limited bandwidth, typically around 2 kHz. The Linear Predictive Coding (LPC) spectral envelopes of the phoneme $/i/$ captured using the REF, IEM and the OEM simultaneously, are shown in Fig. 2. It can be seen that the OEM and the REF signals are similar in the high frequencies. The IEM, however, has a high frequency roll-off around 2 kHz, and has more energy in the low frequencies. The similarity between the OEM speech and the REF speech suggests that the OEM signal could potentially be used to extend the bandwidth of the IEM signal and make it sound closer to the REF signal.

In this paper, we explore the potential of enhancing (i.e., bandwidth expanding) the IEM signal via information captured from the OEM. We measure this potential by means of the mutual information shared between different frequency bands of the three microphone signals captured simultaneously. The remainder of this paper is organized as follows. In Section 2, the Gaussian Mixture Model (GMM) based mutual information approach used to evaluate the similarities between the three signals is described. The experimental setup as well as the simulations are presented in Section 3. The results are presented and discussed in Section 4, followed by the conclusions drawn in Section 5.

2. MUTUAL INFORMATION COMPUTATION

In this section, we briefly describe the methodology as it relates to the context of this work. To measure the mutual information between the different frequency bands of all three microphone signals, the GMM based mutual information approach described in [11] was used. The speech spectrum was modeled using the Mel-Frequency Cepstral Coefficients (MFCC) as they provide a good representation of human

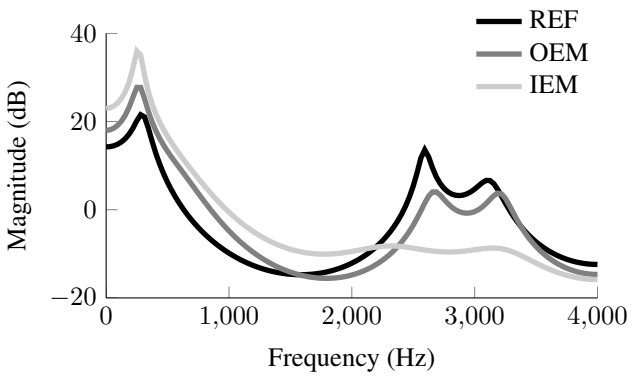

Fig. 2. The LPC spectral envelope of the phoneme /i/ recorded with the REF, the OEM and the IEM simultaneously.

speech perception in the low frequencies. Since the signals used in this study were recorded at a sampling frequency of 8 kHz, we use 16 triangular filters to stay in accordance with the amount of critical bands in that frequency range [12]. Because the IEM signal is bandlimited to about 2 kHz, we are particularly interested in the mutual information of the 0-2 kHz and 2-4 kHz sub-bands of the different microphone signals. We use the first 11 filters to derive the low-band MFCC's covering the range between 0-2 kHz, and the last 4 to derive the high-band MFCCs covering the 2-4 kHz range. The 12th filter, spanning both ranges, is ignored to avoid any overlap between the two frequency bands. For each of the signals and ranges of interest, we use a GMM to model their joint density functions, as defined in [11]:

$$f_{GMM}(x,y) = \sum_{m=1}^{M} \alpha_m f_G(x,y|\theta_m), \tag{1}$$

where x and y represent the different microphone signals at different ranges of interest, M is the number of mixture components, α_m is the mixture weight of the mixture component m, and $f_G(.)$ is the multivariate Gaussian distribution defined by $\theta_m = \{\mu_m, C_m\}$, where μ_m is the mean vector and C_m is the diagonal covariance matrix calculated using the standard expectation-maximization (EM) algorithm. Once the probability density functions of the signals are determined, the mutual information measure can then be calculated as follows:

$$I(\widehat{X;Y}) = \frac{1}{N}\sum_{n=1}^{N}\left(\log_2\left(\frac{f_{GMM}(x_n,y_n)}{f_{GMM}(x_n)f_{GMM}(y_n)}\right)\right). \tag{2}$$

This mutual information measure is used in the next section to understand the relationship between the REF, OEM and IEM signals and their respective low and high frequency subbands.

3. EXPERIMENTAL SETUP

3.1. Speech Corpus

A speech corpus was recorded in an audiometric booth with the communication headset shown in Fig. 1 as well as a digital audio recorder (Zoom® 4Hn) placed in front of the speaker's mouth (i.e REF signal). A female speaker read out the first ten lists of the Harvard phonetically balanced sentences and speech was recorded at 8 kHz sampling rate and 16-bit resolution across the three microphones, simultaneously.

3.2. Measuing the Transfer Function of the Earpiece

It is of interest to see the change in mutual information at varied levels of SNR. To avoid any uncontrolled deviations in the speech between different recordings, the noise is injected post recording. The transfer function between the OEM and IEM is calculated to remain as close as possible to realistic conditions. This is achieved by playing white noise over loudspeakers in the audiometric booth while the speaker is still equipped with the in-ear HPD. The noise signals collected by the IEM and OEM are then used to calculate the transfer function between the two microphones, i.e. the transfer function of the earpiece. Factory noise from the NOISEX-92 database [13] was then added to the OEM signal for a range of SNRs from -5 dB to +30 dB in 5 dB increments. The same procedure was done with the IEM signal, but the noise was first filtered using the previously-calculated earpiece transfer function. The REF signal was kept clean in order to provide an upper bound on the achievable performance.

3.3. Computation of Mutual Information

MFCC features are extracted for both the low-band and the high-band for each of the three microphones for the entire range of SNRs. Therefore, 6 different features are generated for each SNR and are represented as REF_k, OEM_k, IEM_k, where the subscript k indicates either the 0-2 kHz or 2-4 kHz speech subbands. For example, REF_{0-2} and REF_{2-4} would represent the MFCC features extracted for the low-band and the high-band from the REF signals, respectively. For every SNR, we investigate the mutual information between the signal pairs as shown in Fig. 3, for both the 0-2 kHz and 2-4 kHz sub-bands.

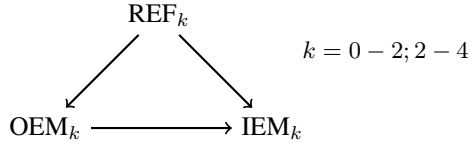

Fig. 3. Schematic showing the signal pairs used in the mutual information calculation, for each tested SNR value.

This calculation yields the shared information between the three microphone signals. Most notably, it indicates whether the OEM shares enough information with the REF in the high band, thus allowing for artificial bandwidth extension from it. As a secondary analysis, we also investigate the relationship between the low-band of the OEM and the IEM with the high-band of the REF as shown in the schematic of Fig. 4.

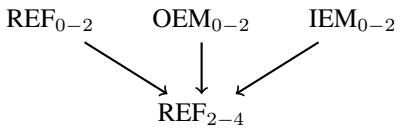

Fig. 4. Schematic showing the cross-band signal pairs used in the mutual information calculation for each tested SNR value.

This relationship indicates if enough information is shared that the high-band of the REF could be predicted using the low-band of the IEM or the OEM. The results are discussed in the next section.

4. RESULTS AND DISCUSSION

Figures 5 and 6 show the mutual information of the low-band of the three microphone signals and the high-band, respectively as a function of SNR. It can be seen that the OEM and REF share some mutual information in both the low-band and high-band which decreases proportionally with the decrease in SNR. As expected, at high levels of SNR the OEM and the REF share more mutual information in the high-band than the IEM and the REF. Interestingly, however, the IEM and REF share more in the low-band than the OEM and REF. We expect that this is due to high frequency components within the 0.5-2 kHz range that are missing in the OEM due to its placement [14], away from the mouth, yet still conducted in the ear canal. Interestingly, the very little information that is present in the high-band of the IEM still contains shared information with the REF. At low SNRs the mutual information between the IEM and REF surpasses that of the OEM and the REF. Due to the attenuation of the earpiece, the mutual information between the IEM and the REF does not drastically decrease as the noise increases. It is beneficial that the REF and the IEM share information in the low frequencies even at low SNRs. If the high-band of the REF can be predicted from its low-band then the low-band of the IEM could be used to predict the high frequencies of the REF. In turn, Fig. 7 shows relationships between the low-band of IEM and OEM signals with the high-band of the REF signal. The average mutual information between the low-band and high-band within the REF signal is also plotted (dashed line) for comparison. As can be seen, the mutual information between the low-band of the IEM and the high-band of the REF is very close to the mutual information between the two frequency bands within the REF. Again, the shared information is not greatly affected

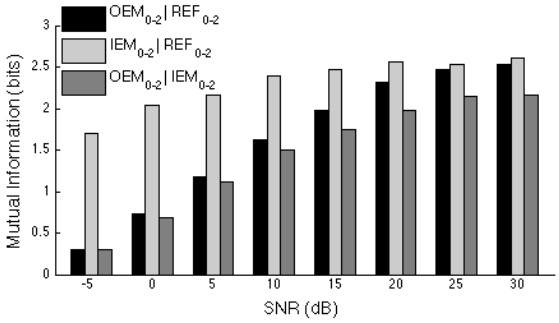

Fig. 5. Mutual information of the low-band between the REF, OEM and IEM signals.

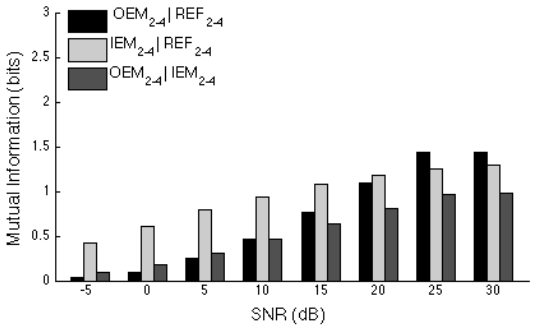

Fig. 6. Mutual information of the high-band between the REF, OEM and IEM signals.

by the increase in noise. The OEM shares information with the REF but is significantly affected by noise and is not very reliable in low SNRs.

These results aid in discovering ways to extend the bandwidth of the IEM as a function of SNR. In high SNRs (above 20 dB) the IEM can be mixed with the OEM using power complementary filtering to achieve a signal that is closer to the REF signal. Since the IEM is restricted to a bandwidth of 2 kHz, the IEM signal can be low passed at that frequency to reject any unwanted overlap with the OEM signal above 2 kHz. The OEM signal can then be high-passed at the same frequency and added to the low-passed IEM signal. This way the extended signal will contain a low-band and a high-band that are more closely related to the REF signal. Although at those levels of SNR the OEM may be used on its own as an intelligible signal, preliminary trials show that the enhanced IEM signal contains less noise and has higher objective quality values. Simple filtering is not computationally exhaustive and this method of extension would be worth its subtle enhancements.

At low levels of SNR, more complex ways of bandwidth extension must be investigated. The GMM bandwidth extension technique used in [15] could be used to extend the bandwidth of the IEM signal. The GMM can be trained offline in

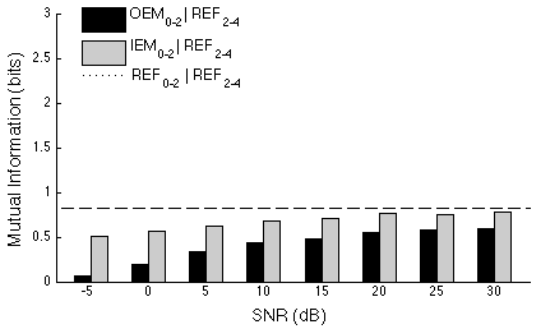

Fig. 7. Cross-band mutual information between the OEM, IEM and REF signals compared with the average cross-band mutual information within the REF signal.

a quiet environment using the IEM and OEM. In quiet, the OEM signal shares enough information in the high-band with the REF that it can be tuned to be used in its place. Once the training is complete, even in low levels of SNR, the low-band of the IEM signal can be used to predict the high-band of the OEM signal and ultimately the REF signal. Having a robust bandwidth extension technique, as such, in low levels of SNR could enhance the communication experience of those equipped with the earpiece.

Overall, we have found that, in quiet, the OEM and the REF signals share mutual information in the 2-4 kHz range while the IEM and the REF signals share information in the 0-2 kHz range for all SNRs. This suggests that it may be possible to use either the high-band of the OEM signal or the low-band of the IEM signal to artificially extend the bandwidth of the IEM signal thus creating a better quality/intelligibility signal that is less prone to environmental factors.

5. CONCLUSIONS

In this paper, we study of the GMM based mutual information between signals of three different microphones at different SNRs. We reveal the relationship between frequency bands of the three microphone signals, which opens up the door to various ways of bandwidth extension by capitalizing on the information present in the signals available. It brings up the potential of an enhanced communication experience using bone and tissue conducted speech with increased SNR that is bandwidth extended in its high frequencies.

6. ACKNOWLEDGMENTS

This work was made possible via funding from the Center for Interdisciplinary Research on Media, Music, and Technology, the Natural Sciences and Engineering Research Council of Canada, and the Sonomax-ETS Industrial Research Chair in In-Ear Technologies.

7. REFERENCES

[1] E.H. Berger, *The Noise Manual*, AIHA, 2003.

[2] W.S. Gan and S.M. Kuo, "Integrated active noise control communication headsets," *Proceedings of International Symposium on Circuits and Systems.*, vol. 4, pp. IV–353–IV–356, 2003.

[3] W.S. Gan, S. Mitra, and S.M. Kuo, "Adaptive feedback active noise control headset: implementation, evaluation and its extensions," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 3, pp. 975–982, 2005.

[4] S.M. Kuo and D.R. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–975, June 1999, 00625.

[5] J.G. Casali and E.H. Berger, "Technology advancements in hearing protection circa 1995: Active noise reduction, frequency/amplitude-sensitivity, and uniform attenuation," *American Industrial Hygiene Association*, vol. 57, no. 2, pp. 175–185, 1996.

[6] R.E. Bou Serhal, T.H. Falk, and J. Voix, "Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones.," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2013, vol. 19, p. 040013.

[7] T. Turan and E. Erzin, "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7049–7053.

[8] M.S. Rahman and T. Shimamura, "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1–4, Aug. 2011.

[9] B. Geiser and P. Vary, "Speech bandwidth extension based on in-band transmission of higher frequencies," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7507–7511.

[10] A. Bernier and J. Voix, "An active hearing protection device for musicians," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2013, vol. 19, p. 040015.

[11] M. Nilsson, H. Gustaftson, S.V. Andersen, and W.B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, vol. 1, pp. I–525.

[12] H. Fastl and E. Zwicker, *Psychoacoustics, facts and models*, Springer, 2001.

[13] A. Varga and H.JM. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[14] G.A. Studebaker, "Directivity of the human vocal source in the horizontal plane," *Ear and hearing*, vol. 6, no. 6, pp. 315–319, 1985.

[15] K. Park and H.S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2000, vol. 3, pp. 1843–1846.