TABLE OF CONTENTS

Page

INTRODUCTION 1
0.1 Context 1
0.2 Problem Statement 2
0.3 Background 3
0.4 Objectives 5
0.5 Structure 8

CHAPTER 1 LITERATURE REVIEW 11
1.1 Hearing Protection Devices 11
1.2 Communication in Noise: existing tools and techniques 13
1.3 Bandwidth Extension Techniques of Bone Conducted Speech 16
1.3.1 Speech Bandwidth Extension using Equalization Approach 16
1.3.2 Speech Bandwidth Extension using Analysis and Synthesis Approach 17
1.3.3 Speech Bandwidth Extension using Probabilistic Approach 17
1.4 Vocal Effort 18
1.4.1 Open Ear 19
1.4.2 Occluded Ear 20
1.4.3 Vocal Effort in Noise 21
1.4.4 In Noise, Open Ears 21
1.4.5 In Noise, Occluded Ears 22
1.4.6 Vocal Effort With Varying Distance 22

CHAPTER 2 IMPROVING THE QUALITY OF IN-EAR MICROPHONE SPEECH VIA ADAPTIVE FILTERING AND ARTIFICIAL BANDWIDTH EXTENSION 25
2.1 Introduction 25
2.2 Methods and Materials 29
2.2.1 Speech Corpus 29
2.2.2 Predicted Quality 30
2.2.3 IEM Noise Reduction 32
2.2.4 IEM Bandwidth Extension 37
2.2.5 Performance Evaluation 39
2.3 Results 40
2.3.1 Pre-Enhancement Objective Quality Assessment 40
2.3.2 IEM Speech Enhancement 42
2.3.3 Performance Evaluation 46
2.4 Discussion 47
2.5 Conclusions 52

CHAPTER 3 VARIATIONS IN VOICE LEVEL AND FUNDAMENTAL FREQUENCY WITH CHANGING BACKGROUND NOISE LEVEL AND TALKER-TO-LISTENER DISTANCE WHILE WEARING HEARING PROTECTORS: A PILOT STUDY 55
3.1 Introduction 56
3.2 Method 59
3.2.1 Apparatus 59
3.2.2 Participants 61
3.2.3 Task 62
3.2.4 Conditions 62
3.2.5 Procedure 63
3.2.5.1 Measurement of individual earplug transfer function 63
3.2.5.2 Assessment of well-fitted earplug 64
3.2.5.3 Adjustment of the background noise level 64
3.2.5.4 Analysis 65
3.3 Results 66
3.4 Discussion 70
3.5 Conclusions 71

CHAPTER 4 MODELING SPEECH LEVEL AS A FUNCTION OF BACKGROUND NOISE LEVEL AND TALKER-TO-LISTENER DISTANCE FOR TALKERS WEARING HEARING PROTECTION DEVICES 73
4.1 Introduction 74
4.2 Methods and Materials 77
4.2.1 Experimental Setup 77
4.2.2 Model Fitting 79
4.3 Results 80
4.4 Discussions 81
4.5 Conclusions 84

CHAPTER 5 CONCLUSIONS AND FUTURE WORK 85
5.1 Conclusions 85
5.2 Future Work 85
5.2.1 Intelligibility of IEM speech 85
5.3 Distance model as a predictive tool 86
5.4 Consideration of hearing-impaired listeners 87
5.5 Implementation and validation 88
5.6 Contributions 88

APPENDIX I INTEGRATION OF A DISTANCE SENSITIVE WIRELESS COMMUNICATION PROTOCOL TO HEARING PROTECTORS EQUIPPED WITH IN-EAR MICROPHONES 89

APPENDIX II PROTECTING MINERS' HEARING WHILE FACILITATING COMMUNICATION 99

APPENDIX III ON THE POTENTIAL FOR ARTIFICIAL BANDWIDTH EXTENSION OF BONE AND TISSUE CONDUCTED SPEECH: A MUTUAL INFORMATION STUDY 111

APPENDIX IV MODELING SPEECH PRODUCTION IN NOISE FOR THE ASSESSMENT OF VOCAL EFFORT FOR USE WITH COMMUNICATION HEADSETS 117

BIBLIOGRAPHY 123

LIST OF TABLES

Page

Table 1.1 Major advantages and disadvantages of different types of HPDs (Berger and Voix, 2016). 12

Table 2.1 Average POLQA MOS-LQO scores for the noisy IEM signal (N), the denoised IEM (NS), the bandwidth extended IEM (BWE) and the clean IEM signal (C). 48

Table 2.2 Statistical significance results based on a 95% confidence interval between the objective evaluation of different stages of enhancement. 49

Table 2.3 The average MUSHRA scores for the noisy IEM signal (N), the denoised IEM (NS), the clean IEM signal (C) the bandwidth extended IEM (BWE) and the hidden reference (REF). 51

Table 2.4 Statistical significance results based on a 95% confidence interval between the subjective evaluation of different stages of enhancement. 51

Table 2.5 Comparison of conventional BC enhancement approaches to the proposed approach. 51

Table 3.1 Experimental conditions with changing talker-to-listener distance for the quiet un-occluded ear and occluded ear in noise and in quiet. 63

Table 3.2 The mean (μ) and standard deviation (σ) of absolute level values across speakers for all conditions and distances in dB(A). 66

Table 3.3 The mean (μ) and standard deviation (σ) absolute F0 values across speakers for all conditions and distances in Hz. 67

Table 3.4 Overall change in linear level from 1 m to 30 m for different noise conditions. 69

Table 3.5 OOverall change in F0 as well as the overall rate of change of F0 per dB increase for each condition. 69

Table 4.1 Experimental conditions with changing talker-to-listener distance for the quiet un-occluded ear and occluded ear in noise and in quiet. 78

Table 4.2 Final parameter values optimized for all three noise conditions with the corresponding R^2 value. 81

Table 4.3 Mean (μ) and standard deviation (σ) in Δ_L for each noise level and distance. 83

LIST OF FIGURES

Page

Figure 0.1 Auditory research platform (a), its electroacoustic components (b), and equivalent schematic (c). 7

Figure 0.2 A graphic representation of the RAVE showing a talker and the potential distance of the transmitted signal based on the level of background noise. 7

Figure 1.1 An example of the attenuation of an active, level-dependent HPD with 82 dB(A) maximum level inside the ear and a passive attenuation of 20 dB as a function of the background noise level. The user may select one of three options: a 5 dB damping, a unity gain "pass-through" (L-D Unity), and a 10 dB gain (with permission from 3M^{TM}). 14

Figure 1.2 Illustration of the three paths affecting the perception of one's own voice: (a) direct air-conduction, (b) bone-conduction, and (c) indirect air-conduction. 19

Figure 1.3 Speech power level (L_w) plotted as a function of talker-to-listener distance in a reverberant environment..... 23

Figure 2.1 Auditory research platform (a), its electroacoustic components (b), and equivalent schematic (c). Placed inside the ear the ARP captures speech produced by a talker using either the IEM or the OEM and transmits communication to other users through a wireless link. 30

Figure 2.2 The linear predictive coding spectral envelope of the phoneme /i/ recorded with the REF, the OEM and the IEM simultaneously..... 31

Figure 2.3 Block diagram representing the NLMS filtering stage: (a) when the adaptation is ON and (b) when it is OFF..... 34

Figure 2.4 Offline identification stage of the earplug transfer function in the user's ear. White noise is played on a loudspeaker outside of the ear and recorded using both the IEM and the OEM. The transfer function of the earplug is calculated by assessing the noise outside the ear, recorded by the OEM, and the residual noise inside the ear, recorded by the IEM. 35

Figure 2.5 Test signal for the IEM to optimize speech detection criteria..... 36

Figure 2.6 Flow chart representing the adaptation process. 38

Figure 2.7 Block diagram illustrating the bandwidth extension process. 39

Figure 2.8 POLQA MOS-LQO results of clean IEM and OEM signals using the REF signal as reference, with sentences sorted by ascending order of IEM MOS-LQO scores. 41

Figure 2.9 POLQA MOS-LQO results of noisy IEM and OEM using the REF signal as reference, with sentences sorted by ascending order of IEM MOS-LQO scores. 42

Figure 2.10 Cumulative distribution plot of the difference in POLQA MOS-LQO results between the clean and noisy IEM and OEM signals. 43

Figure 2.11 A denoised IEM signal (IEM NS(SO)) using suboptimal criteria for speech detection during the adaptation process plotted against the clean IEM signal (IEM C). 44

Figure 2.12 Average POLQA MOS-LQO scores after denoising (NS) and after bandwidth extension (BWE) over different triggering percentages, showing a peak at $T_g = 1.06 - 1.07$. 45

Figure 2.13 The denoised IEM signal (IEM NS) as compared to the noisy IEM signal (IEM N). Zoomed portions of the denoised signal when only noise is present (a), when speech inside the ear is present (b) and when external speech is present (c) 45

Figure 2.14 The spectrograms of the sentence 'It is easy to tell the depth of a well' of the clean reference signal (REF), the noisy IEM signal (N), the denoised IEM signal (NS), and bandwidth extended denoised IEM signal (BWE). 46

Figure 2.15 Cumulative distribution of the difference in POLQA MOS-LQO scores between the denoised and noisy IEM ($\Delta_{NS/N}$), as well as the denoised and clean IEM ($\Delta_{NS/C}$). 48

Figure 2.16 Cumulative distribution of the difference in POLQA MOS-LQO scores between the bandwidth extended and noisy IEM ($\Delta_{BWE/N}$), the bandwidth extended and denoised IEM ($\Delta_{BWE/NS}$), and the bandwidth extended and clean IEM ($\Delta_{BWE/C}$). 49

Figure 2.17 Log-spectral distance between the clean IEM signals and the denoised IEM signals, with sentences sorted in ascending order. 50

Figure 2.18 Box and whisker plot comparing the MUSHRA results of the noisy IEM signal (IEM N),the denoised IEM signal (IEM NS), the clean IEM signal (IEM C), the bandwidth extended denoised IEM signal (IEM BWE) and the hidden reference. 50

Figure 3.1 Auditory research platform (a), its electroacoustic components (b), and equivalent schematic (c). 60

Figure 3.2 An example of the experimental setup with a participant (a), a close-up of the apparatus (b), includingFireface® UCX soundcard (i), the computer loudspeakers (ii), and the Windows™ computer running MATLAB™ (iii). The earpiece with and without the Comply™ tips (c), and the hallway where the experiments were held (d). 61

Figure 3.3 An example of a transfer function of a well-fitted earplug. 64

Figure 3.4 The spectral and temporal differences between the simulated residual noise inside the ear (IEM noise) and the noise as it would have been outside the ear (OEM noise). 65

Figure 3.5 Average increase in speech levels, Δ_l from the occluded quiet condition over increasing distance and noise levels. The level at 1 m distance for the quiet occluded condition is used as reference for all the curves. The standard deviation, σ_l, in Δ_l across talkers over different noise conditions and distance. 68

Figure 3.6 Average increase in F0 level, Δ_{F0}, from the occluded quiet condition over increasing distance and noise levels. F0 at 1 m distance for the quiet occluded condition is used as reference for all the curves. The standard deviation, σ_{F0}, in Δ_{F0} across talkers over different noise conditions and increasing distance. 70

Figure 4.1 Illustration of the three paths affecting the perception of one's own voice: (a) direct air-conduction, (b) bone-conduction and (c) indirect air-conduction. 75

Figure 4.2 Schematic of procedure ensuring a good acoustical seal in the ear canal. 78

Figure 4.3 An example attenuation curve of a well-fitted earplug. 79

Figure 4.4 The mean curves for the 70, 80 and 90 dB(SPL) conditions compared to their respective model curves with $a = -1.659$ $c =$

0.18 and $\varepsilon = -0.0675$ (a), and the respective standard deviation at each noise level (b). 82

Figure 4.5 Model curves vs. individual participant curves. 83

LIST OF ABREVIATIONS

AC	Air Conduction
ANC	Active Noise Control
ARP	Auditory Research Platform
BC	Bone and tissue Conduction
BWE	Bandwidth Extension
FLANN	Function Link Artificial Neural Networks
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HPD	Hearing Protection Device
ICA	International Congress on Acoustics
ICASSP	International Conference on Acoustics, Speech and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
IEM	In-Ear Microphone
LPC	Linear Predictive Coding
LSF	Line Spectral Frequencies
NIHL	Noise Induced Hearing Loss
NIOSH	National Institute for Occupational Safety and Health
OEM	Outer-Ear Microphone
OSHA	Occupational Safety and Health Administration

PPE	Personal Protection Equipment
RAVE	Radio Acoustical Virtual Environment
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level

INTRODUCTION

This dissertation is comprised of three journal and five conference papers and one patent which deal with speech bandwidth extension and vocal effort coding for implementation with a smart radio system for use in noisy industrial environments. The following introduction serves to present the context, problem statement, objectives of this PhD work as well as the overall organization of this document.

0.1 Context

To ensure a safe and efficient workplace, workers in noisy industrial environments must be able to protect their hearing health while still be able to communicate effectively. In the United States alone, over 22 milion workers are exposed to hazardous noise levels each year, which puts them at risk of Noise Induced Hearing Loss (NIHL) (NIOSH, 2015). Noise exposure in the workplace is not only responsible for occupational NIHL, but also for work related injuries and other diseases such as hypertension and sleep deprivation (Girard *et al.*, 2015; Concha-Barrientos *et al.*, 2004). Therefore, the Occupational Safety and Health Administration (OSHA) recommends three ways of reducing noise exposure for the workers (Katz *et al.*, 1994): engineering reduction of the noise, limiting exposure time, and enforcing the use of personal Hearing Protection Devices (HPD).

Noise reduction at the source and limiting exposure time are both effective ways of reducing noise exposure, however, for many occupations such as mining, construction and first responders, neither is always possible (Berger, 2003). For this reason, the use of HPDs plays a significant role in the reduction of noise exposure in the workplace. However, the effectiveness of the use of HPDs has been widely criticized. Concerns about the quality of the fit (Voix and Laville, 2009) as well as the frequency of the use (Neitzel and Seixas, 2005; Hughson *et al.*, 2002) are major concerns when considering the effectiveness of HPDs. Currently, however, objective individual fit checking of HPDs, such as field attenuation estimation systems (FAESs) (Voix *et al.*, 2014), are encouraged, recommended and more frequently implemented (Hager *et al.*, 2011) making HPDs a more reliable tool of noise reduction. Consequently, this work focuses

on the use of hearing protection devices. The National Institute for Occupational Safety and Health (NIOSH) recommends the use of hearing protection devices for workers exposed to levels equivalent to at least 85 dB(A) of noise for 8 hours (NIOSH, 1998). When worn correctly, HPDs can be very effective in reducing the risk of NIHL (Berger, 2003). Still, the use of HPDs among workers exposed to hazardous levels of noise can be as low as 10% (Hughson *et al.*, 2002). Workers attribute their reluctance to wear HPDs to difficulties in communication (Reddy *et al.*, 2012; NIOSH, 2005; Hughson *et al.*, 2002). Therefore, there is a need for a device that can provide workers with sufficient hearing protection without hindering their ability to communicate.

0.2 Problem Statement

Workers should not have to choose between adequately protecting their hearing health and the ability to communicate properly. Investigating and understanding the state of the art in HPDs, paves the way to improving the attitude of workers in noisy environments by highlighting the strengths and indicating the current weaknesses of communication in noise. There are three main weaknesses in the current ways of communicating in noise:

Weakness 1: Compromising hearing health

To enhance communication, workers may choose not to wear their assigned HPDs or remove them during communication. This degrades the effectiveness of the HPD and compromises the worker's hearing health.

Weakness 2: Degraded Speech Signal

There are many different types of communication headsets that allow for radio communication between users wearing HPDs. However, currently the communication speech signal either suffers degraded quality, as it is masked by the background noise, or a limited bandwidth, as a consequence of unconventional methods of capturing the speech signal.

Weakness 3: Lack of designated listeners while using radio communication

Radio communication is an effective, continually more affordable, and practical way allowing for verbal communication between users with communication headsets. However, its weakness lies in the fact that there are no designated receivers; the communication signal is transmitted to everyone on the same radio channel regardless of whether or not they are the intended receivers. Considering that the average preferred SNR for speech when wearing communication headsets has been shown to be 13.8 dB constantly receiving noisy signals and adjusting for the desired SNR is annoying and contributes to the noise dose received by each worker (Giguère *et al.*, 2012a).

Providing workers with an HPD that allows for a high quality communication signal between intended talkers and listeners would address the aforementioned weaknesses and could remedy the low usage rates of HPDs by workers in industrial noisy environments. This is the driving motivation of this Doctoral study.

0.3 Background

Using passive HPDs in the workplace is inexpensive and when worn properly, can be very effective in terms of protecting the user's hearing. However, conventional passive HPDs can hinder communication by not discriminating between relevant signals, i.e. communication and warning signals, and noise consequently attenuating all signals alike. Also, attenuation of HPDs is frequency dependent and increases with frequency resulting, in some cases, in an attenuation of speech content in the high frequencies below the audible threshold. This, as well as, an upward masking from the low-frequency noise content causes speech to sound muffled and decreases intelligibility (Berger and Voix, 2016). For this reason, there are currently many headsets that combine hearing protection and communication abilities. These headsets can consist of passive or active protection from the noise, combined with a microphone and radio capabilities for communication (Berger, 2003). These types of headsets decrease the need to remove the HPD for communication and therefore tackle *Weakness 1* described in Section 0.2.

Conventionally, many communication headsets utilize a boom microphone, placed in front of the mouth to pick up the user's speech during verbal communication. Although boom microphones are often highly directional by design (i.e. the use of so-called noise reduction microphones), they are still susceptible to background noise and capture a speech signal degraded by noise. One of the ways to remedy this issue is the use of adaptive filtering to denoise the speech signal before it reaches the listener (Gan and Kuo, 2003). Another way is to capture speech unconventionally, using bone and tissue conduction (BC) microphones. When it comes to providing a speech signal with relatively high signal-to-noise ratio (SNR), BC microphones are better than air conduction microphones because they are less susceptible to background noise and can be placed in various positions, which allows for simultaneous use with other Personal Protection Equipment (PPE). Bone and tissue conduction microphones can be placed inside the ear canal, on the forehead, on the temple, or on the throat (Tran *et al.*, 2008). Although less susceptible to noise, the bandwidth of speech captured using BC microphones is limited and is constricted to 2 kHz, thus reducing its quality. Many Bandwidth Extension (BWE) techniques aimed at enhancing the quality of BC speech have been developed (Shin *et al.*, 2012). However, these techniques are often computationally exhaustive, or require extensive training by the user, thus limiting their widespread use in practical settings.

An effective compromise between the two extremes of noisy air conducted speech and bandlimited BC speech captured by bone conduction sensors is speech captured from inside an occluded ear using an in-ear microphone. Despite occluded speech being also bandlimited to 2 kHz, it is captured acoustically thus can share a significant amount of information with clean speech captured in front of the mouth in the 0 to 2 kHz range (Bouserhal *et al.*, 2015a). This suggests that the enhancement of occluded speech may be reached using a BWE technique that is simple and practical.

As highlighted in Section 0.2, a third weakness of existing HPD solutions relies on the fact that, despite having a high-quality speech signal, radio communication is flawed. More specifically, radio communication protocols do not distinguish between receivers. This means that users on the same radio channel receive transmitted communication signals regardless of whether or not

they are the intended receivers. This contributes to the daily accumulated noise dose (Giguère *et al.*, 2012a) and makes for an unnatural communication environment. In a natural acoustical setting, only people within a specific spatial range are exposed to the communication signals between a listener and a talker. This spatial range is defined by the talker's vocal effort and the level of background noise. Mimicking a natural acoustical environment would enhance the experience of people wearing communication headsets. Studies have already shown a clear relationship between vocal effort and background noise level (Byrne, 2014). A relationship has also been established between vocal effort and talker-to-listener distance for people not wearing HPDs (Pelegrín-García *et al.*, 2011). Occluding the ear canal with an HPD causes deviations from these models. By studying the relationship between background noise, vocal effort and intended communication distance for the occluded ear a model can be created and implemented with radio systems to mimic communication in a natural acoustical environment while wearing HPDs.

The current state-of-the-art of communication headsets are lacking a high quality speech signal for communication, and existing BWE techniques are too computationally exhaustive for current DSP platforms or require a great deal of training from the user to be practically used. Although a few companies have looked into this application for military purposes, no headsets, currently on the market, take into consideration the intended communication distance in an effort to mimic a natural acoustical environment. This doctoral study aimed to fill this gap.

0.4 Objectives

This work focuses on the enhancement of communication when using a communication headset with similar configuration to that of the Auditory Research Platform (ARP) shown in Figure 0.1. All the studies in this PhD project were conducted using the ARP hardware. As can be seen from Figure 0.1, the ARP is an intra-aural HPD containing a miniature loudspeaker as well as a microphone inside the ear, a microphone outside the ear, a digital signal processor, and is equipped with a wireless link for radio communication. This makes the ARP a communication headset with passive hearing protection that captures speech with in-ear microphones

placed inside occluded ears. Although the ARP's passive attenuation and communication abilities tackle *Weakness 1*, there is still a need to improve the quality of the BC speech recorded inside the ear and to integrate a radio communication protocol that mimics a natural acoustical environment.

This PhD has therefore three specific objectives: (1) to remove any residual noise from the captured occluded speech, which is referred to as in-ear microphone (IEM) speech for the remainder of the document; (2) to find a simple yet efficient method to extend the bandwidth of the IEM speech signal, thus improving its perceived quality; (3) to model the relationship between vocal effort, background noise level, and intended communication distance for occluded speakers.

This model will serve as the fundamental knowledge needed to create a 'Radio Acoustical Virtual Environment' (RAVE) aimed at mimicking a natural acoustical environment. Under RAVE, only listeners that are within a specific spatial range of the talker receive the transmitted communication signal. This spatial range is a function of the talker's vocal effort and the background noise level. A graphic demonstrating the functionality of RAVE is shown in Figure 0.2. The next section details the structure of this thesis and identifies the contributions of this work.

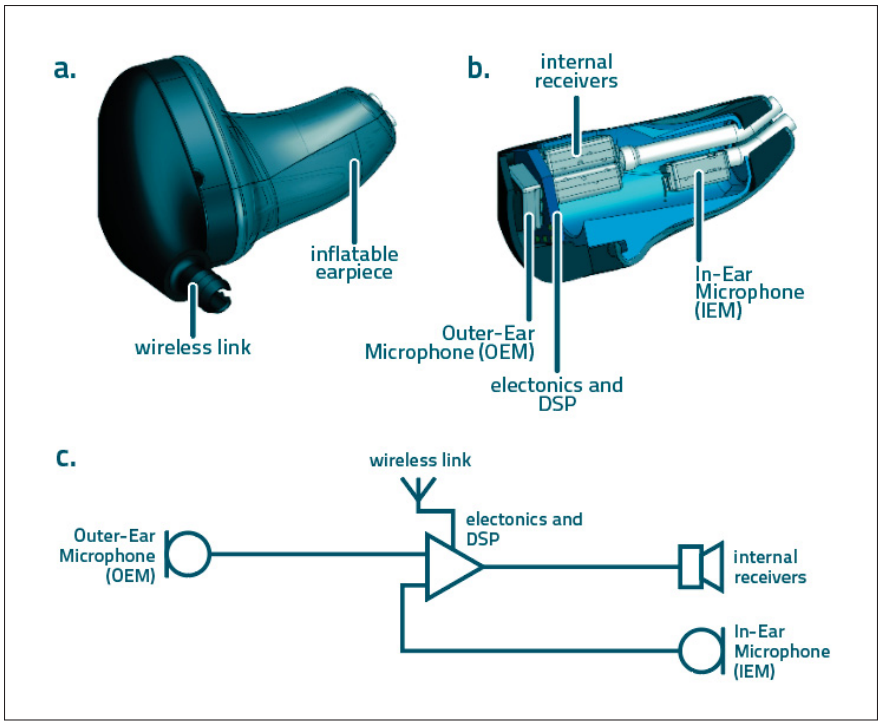

Figure 0.1 Auditory research platform (a), its electroacoustic components (b), and equivalent schematic (c).

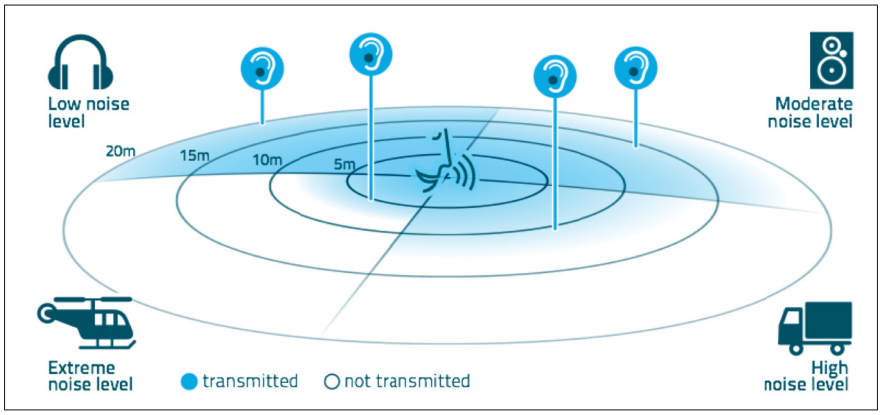

Figure 0.2 A graphic representation of the RAVE showing a talker and the potential distance of the transmitted signal based on the level of background noise.

0.5 Structure

This dissertation is comprised of a literature review, three journal articles, conclusions and five conference proceedings in the Annex. Chapter 1, is a literature review of HPDs and communication headsets. This is to clarify the choice of the ARP as the selected communication headset and HPD. Chapter 1 also reviews the literature on vocal effort, including: the variations that occur because of changes in the background noise level, i.e. the Lombard effect, the variations caused by the intended communication distance, and models that exist for speech production in noise with and without the use of HPDs. Note that the variations in vocal effort caused by emotion are not reviewed in this work because the model is only meant to describe changes caused by noise and by changing talker-to-listener distance. This review provides important background information that clarifies the choices made in the rest of the document. Chapter 2 is an article submitted to the Journal of Acoustical Society of America and is currently under review (Bouserhal *et al.*, 2016b). In Bouserhal *et al.* (2016b), the methodology of denoising and BWE used to improve the quality of the speech captured by the IEM is presented and builds on insights presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15) (Bouserhal *et al.*, 2015a). With an improved quality speech signal available, coding and vocal effort modeling is enabled to be used with RAVE. Next, Chapter 3 is an article detailing a study that proves that the vocal effort can be modeled as a function of the background noise level and the intended communication distance for occluded listeners. This article was published in the International Journal of Audiology (Bouserhal *et al.*, 2016a). Subsequently, the model relating the vocal effort to background noise level and intended communication distance is presented in Chapter 4. The article in Chapter 4, was submitted to JASA Express Letters and is currently under review (Bouserhal *et al.*, 2016c). This dissertation concludes with conclusions and future research directions presented in Chapter 5. At the end of this dissertation are 4 appendices as follows:

- **Appendix I** contains an article from the proceedings for meetings in acoustics that was presented at the *International Congress in Acoustics (ICA)* 2013, in Montreal between June

2-7, 2013. This article is a general proposal of the RAVE concept that contains a literature review and initial methodology.

- **Appendix II** contains an article presented at the *World Mining Congress (WMC)* in Montréal, Canada between August 11-13, 2013. This article is similar to the one presented to ICA, including a proposal and initial methodology, however, it was more tailored towards needs in the mining industry.

- **Appendix III** contains an article presented at the *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, in Brisbane, Australia between April 19-25, 2015. This article contained the results of a Gaussian Mixture Model (GMM) based mutual information study that showed the relationship between IEM speech, OEM speech and speech picked up in front of the mouth, referred to as REF speech. This study showed the possibility of artificial BWE to enhance the quality of IEM speech.

- **Appendix IV** contains an article presented at *Euronoise*, in Maastricht, Netherlands between May 31 to June 3, 2015. This article discussed the intended experimental protocol to arrive to a model between talker-to-listener distance, background noise level and speech level. In this article, the model presented by Pelegrín-García *et al.* (2011) was manipulated based on previous work in speech production in noise while wearing HPDs and speech production with varying talker-to-listener distance.

CHAPTER 1

LITERATURE REVIEW

The ARP, shown in Figure 0.1, is a good tool to tackle the weaknesses of current ways of communication in noise. The passive attenuation of the earplug and its radio capabilities reduce the need to compromise hearing health for verbal communication with good quality. This leaves three objectives for this work. The quality of the bandlimited speech captured with the IEM of the ARP must be enhanced by first, removing any residual noise and second, by artificially extending its bandwidth. The third objective is to find a model relating the vocal effort of occluded talkers to background noise level and intended communication distance. Therefore, this literature review is organized as follows. Section 1.1 reviews different types of HPDs followed by Section 1.2 which reviews current ways of communication in noise including a review of communication headset technologies. Sections 1.1 and 1.2 combined justify the choice of the ARP. A review of current bandwidth extension techniques on bone conducted speech is presented in Section 1.3. Finally, Section 1.4 presents vocal effort, the changes caused by speaking in noise, varying talker to listener distance and occluding the ear.

1.1 Hearing Protection Devices

Currently, many different forms of HPDs are available. There are two main types: intra-aural i.e. earplugs, and circumaural i.e. earmuffs. Circumaural HPDs are usually one size fits all and need only to be slightly adjusted. They are desirable because they fit over most people's heads and do not require a complicated fitting technique. They are convenient when used in intermittent noise because they are easy to remove and put on, this is especially convenient when working in unclean environments (Berger, 2003). Wearing them for long periods of time, however, could cause discomfort. The headband causes pressure on the top of the cranium as well as on the ears, underlying tissue and bone (Zannin and Gerges, 2006). Also, the cushion lining around the ears causes perspiration which may cause discomfort. Earmuffs are usually

bulky and are not appropriate for use in tight spaces or in conjunction with other PPE (safety glasses, hard hats, respiratory masks, etc...).

Intra-aural HPDs come in many shapes and sizes and could be made of various materials. They could be as generic as a yellow foam plug or as unique as a pair of custom molded ear plugs. Unlike earmuffs, they can be used with different PPE, such as helmets or safety glasses, without hindering their own performance or that of the PPE used alongside them. To alleviate this issue some circumaural HPDs are desgined to be used in particular with other PPEs, such as helmets and hard hats (Berger, 2003). Foam plugs are affordable and easy to attain, however, it is easy to wear them incorrectly. If worn incorrectly, plugs do not offer the correct attenuation and leave the wearer exposed to potentially hazardous noise levels. Custom molded earplugs are advantageous because they are unique to the user and are difficult to wear incorrectly, thus, providing better effective attenuation. However, molded ear plugs are expensive, and their creation requires several different materials and time (Berger, 2003). Furthermore, eventhough custom molded earplugs can provide a more consisten attenuation, they can still be worn incorrectly (Tufts *et al.*, 2012). Table 1.1 lists the advantages and disadvantages of different types of HPDs.

Table 1.1 Major advantages and disadvantages of different types of HPDs (Berger and Voix, 2016).

Type	Advantages	Disadvantages
Circumaural (Earmuffs)	fits most people, no complicated fitting technique	painful after long use, perspiration, limits use of other PPEs, bulky
Generic intra-aural (Generic Earplugs)	affordable, easily attained, convenient with use of other PPEs	difficult to fit properly, uncomfortable
Custom molded intra-aural (Custom Earplugs)	good attenuation, comfortable, convenient with use of other PPEs	expensive, takes a long time to produce

1.2 Communication in Noise: existing tools and techniques

Communication in noise is difficult regardless of whether or not hearing protection is worn. Both speech quality and intelligibility are degraded in the presence of noise (Abel *et al.*, 1982; Giguère *et al.*, 2011). Currently, there are several methods of communication in noise while wearing hearing protection. Often and unfortunately, workers remove conventional passive HPDs (passive earmuffs or earplugs) in order to communicate (Hughson *et al.*, 2002). This is understandable because most passive conventional HPDs offer a fixed attenuation regardless of the background noise level (Berger, 2003). This causes either over-protection or under-protection of the worker and thus reduces speech intelligibility (Giguere *et al.*, 2009; Giguère *et al.*, 2011). This is especially difficult for hearing-impaired wearers because HPDs may decrease speech levels to under the impaired wearer's hearing threshold greatly decreasing speech intelligibility (Berger and Voix, 2016). This may lead wearers to remove the HPD, however, removal greatly reduces the effectiveness of the protection of the user from hazardous noise levels (Berger, 2003). For normal hearing wearers, earplugs that attenuate all frequencies equally, i.e. uniform attenuation earplugs, have better speech intelligibility scores as they do not greatly attenuate the higher frequencies, which are important for speech quality and intelligibility. However, they also have a lower attenuation and cannot protect workers in extremely loud conditions (Casali, 2010).

Another method of communication in noise is through the use of communication headsets that combine hearing protection with radio capabilities. The hearing protection part of the headsets could be either passive or active, with fixed or level-dependent attenuation (Berger, 2003). Some communication headsets utilize Active Noise Control (ANC) to further reduce any residual noise under the HPD. However, due to physical and mechanical constraints, ANC is most effective below 3 kHz (Casali, 2010; Brammer *et al.*, 2008).

Between passive and active HPDs, the preferred way of communication in noise is with the use of active level-dependent HPDs (Tufts *et al.*, 2011). Active, level dependent HPDs can be intra-aural or circumaural and have a varying attenuation as a function of the ambient noise.

This variable attenuation can be done passively through a nonlinear mechanical component such as a valve or diaphragm or actively by monitoring the ambient noise level using an outer-ear microphone (OEM). The OEM allows for a "pass-through" feature of the background noise with a varying gain based on the level of the noise. Typically, level dependent HPDs have a maximum level for the noise under the HPD. Often, the user can select one of three features: unity gain, attenuation or amplification. Figure 1.1 shows an example of the attenuation of a level-dependent HPD with a maximum level of 82 dB(A) under the earpiece and a passive attenuation of 20 dB as a function of background noise level. Level dependent HPDs have shown to increase speech intelligibility in quiet and in noise for both normal hearing and hearing-impaired users (Giguère *et al.*, 2012b). They are currently among the best options for workers needing to verbally communicate in noise through radio communication.

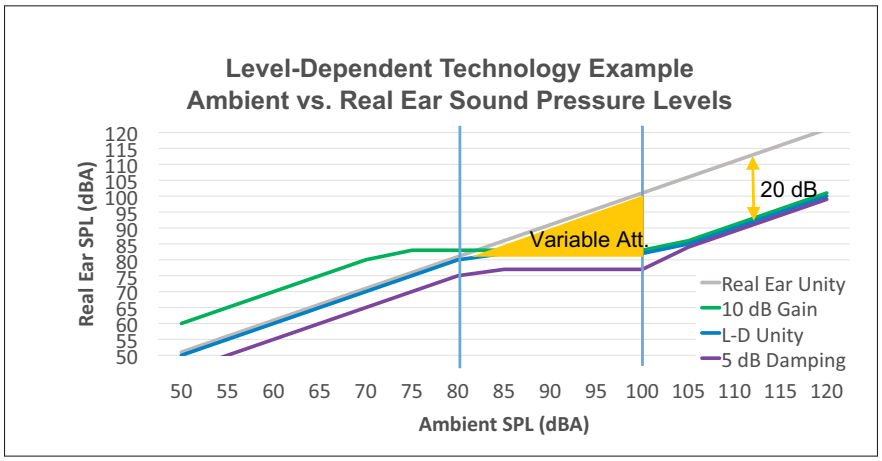

Figure 1.1 An example of the attenuation of an active, level-dependent HPD with 82 dB(A) maximum level inside the ear and a passive attenuation of 20 dB as a function of the background noise level. The user may select one of three options: a 5 dB damping, a unity gain "pass-through" (L-D Unity), and a 10 dB gain (with permission from 3M^{TM}).

Thus far, the ways of communication in noise described have used air conduction for speech transmission and capturing in front of the mouth. In highly noisy environments, unconventional methods of capturing the speech are sometimes utilized. Bone-conduction and throat

microphones are capable of capturing a high SNR speech signal despite elevated levels of background noise (Shin *et al.*, 2012; Turan and Erzin, 2013a). However, speech signals captured using bone and throat microphone suffer in quality as they have a limited bandwidth (Shin *et al.*, 2012; Turan and Erzin, 2013a). Many BWE techniques have been developed to enhance the quality of these bandlimited signals and a selected few are discussed in Section 1.3.

As previously mentioned, another effective way of capturing a relatively high SNR speech signal is through the use of IEMs. Speech captured from an occluded ear using an IEM maintains a relatively high SNR in noisy environments but suffers from a limited bandwidth. The smartPlug™ by Sensear (Sensear, 2016), the Honeywell QUIETPRO series, (Honeywell International, 2016), as well as the 3M™ Peltor™ ORA TAC (3M™, 2012) are among the existing commercial communication headsets that use IEMs to capture speech.

Given these insights, the choice of the ARP as a research tool was based on the fact that i) it is intra-aural, therefore it could be used with other PPE, ii) it can be equipped with radio capabilities, allowing for hearing protection and communication to exist simultaneously, iii) the ARP allows for different tips to be added, including roll-down foam tips, malleable silicone, flange and custom molds, thus allowing for a variety of research conditions that can be studied, and finally, iv) it has an OEM and IEM, allowing it to capture speech from inside the occluded ear using the IEM while monitoring the level of background using the OEM. Speech captured inside an occluded ear is less susceptible to background noise yet it suffers from a limited bandwidth. However, speech captured through air conduction using an IEM shares a significant amount of mutual information with wideband speech captured in front of the mouth in the 0 to 2 kHz range (Bouserhal *et al.*, 2015a). Frequencies above 2 kHz are important for both speech quality and intelligibility. In the frequency domain, speech is characterized by the fundamental frequency, F0, and the peaks and valleys of the spectral envelopes. The peaks of the envelope, i.e. the formants, are needed to distinguish between different phonemes. For some consonants, such as */f/* and */s/*, with formants greater than 2 kHz, a wide bandwidth is

important for discrimination between phonemes. Therefore, artificial BWE is only required in the 2 to 4 kHz range. The next section reviews different BWE techniques.

1.3 Bandwidth Extension Techniques of Bone Conducted Speech

Originating from bone and tissue conduction, the quality of BC speech must be enhanced by artificially extending its frequency bandwidth int the higher frequencies (2-4 kHz). There are many different techniques that could be used to extend the bandwidth of BC speech. A survey of these techniques is presented by Shin *et al.* (2012). In general there are three main approaches to the enhancement of BC speech: equalization, analysis and synthesis, and probabilistic. The main contributions for each of these approaches are discussed in the following three sections.

1.3.1 Speech Bandwidth Extension using Equalization Approach

The equalization approach is a simple way of extending the bandwidth of BC speech. First presented by Shimamura and Tamiya (2005), this approach involved obtaining the long-term spectra of both the BC and Air Conduction (AC) speech and finding an equalization filter based on the ratio of the two long-term spectra. Next, the BC speech is filtered using the reconstruction filter and enhanced using a reinforced spectral subtraction technique (Shimamura and Tamiya, 2005; Ogata and Shimamura, 2001). Results of this technique showed an overall improvement of the BC speech, however, these results were not consistent and varied with each speaker as well as the filter length. Kondo *et al.* (2006) built up on this technique and enhanced it by using a speaker-dependent short-term Fast Fourier Transform (FFT) for equalization. This technique is speaker dependent and required extensive training and smoothing. Although the results of this work were an improvement on the work by Shimamura and Tamiya (2005), the enhanced BC speech low-energy speech regions and silent regions were overly emphasized which affected the perceived quality of the speech (Kondo *et al.*, 2006). Finally, Shimamura *et al.* (2006) improved this approach by proposing a speaker dependent neural network based approach involving a normalized least-mean square adaptive filter. In summary, equalization

approaches are simple, however, they are not robust to any leakage noise in the BC microphone, resulting from flanking pathways, and their speaker dependent nature is not practical.

1.3.2 Speech Bandwidth Extension using Analysis and Synthesis Approach

In this approach an inverse speech transfer function between the AC and BC speech is obtained using either both the AC and BC speech (Yu *et al.*, 2005; Tat Vu *et al.*, 2008) or only the BC speech (Rahman and Shimamura, 2011) to reconstruct the BC speech. Yu *et al.* (2005) first introduced this approach by using the linear-predictive coding (LPC) coefficients to filter and extend the bandwidth of BC speech. However, LPC coefficients are susceptible to quantization noise and were thus replaced with line spectral frequencies (LSF) by Tat Vu *et al.* (2008). Thus far, the BWE techniques to improve the quality of BC speech all utilized an AC speech source as well. Rahman and Shimamura (2011) introduced a blind restoration technique that depended only on the BC speech. Nonetheless, speech distortion is introduced in this technique when the LPC filter is designed from mismatched LSF coefficients caused by BC channel noise or physiological noise such as teeth clack. In general, the analysis and synthesis approach is not very useful in practical application as it is not robust to BC channel noise or physiological noise (Shin *et al.*, 2012).

1.3.3 Speech Bandwidth Extension using Probabilistic Approach

To address the issues caused by noise, probabilistic approaches were introduced. These approaches estimated the transfer function between the BC and AC speech by utilizing a maximum likelihood estimation (Liu *et al.*, 2004). Liu *et al.* (2005) enhanced this technique by estimating the BC leakage noise, the background noise, the AC speech, the BC speech and any physiological noise as Gaussian distributions. The enhanced speech was a weighted sum of the AC speech and the noise reduced BC speech. This approach is advantageous because it does not require any pre-training and is speaker independent yet it still requires an AC microphone and does not use any speech model. A new probabilistic approach that utilized Gaussian Mixture Models (GMM) to model the speech was presented by Subramanya *et al.* (2008). Al-

though this technique showed improvements from past techniques it requires access to an AC microphone, training from multiple speakers, and significantly large databases.

More recent techniques have been proposed with promising results. Huang *et al.* (2014) used function link artificial neural networks (FLANN) to denoise and extend the bandwidth of BC speech. However, it requires training the neural network with clean AC speech data. Li *et al.* (2014) proposed a technique that uses geometric harmonics along with a Laplacian pyramid to denoise and enhance the BC speech. This technique introduces distortion and is computationally complex, making it unsuitable when considering constraints of real-time processing on an embedded hardware with limited resources.

Current forms of BWE techniques for BC speech either require large amount of training, require the use of an AC microphone, are speaker dependent or are computationally exhaustive. A simple speaker independent technique that requires no AC microphone nor training would be practical and applicable in a real life setting.

1.4 Vocal Effort

Talkers adjust their speech level in the presence of noise (Lane and Tranel, 1971), with varying talker-to-listener distance (Fux *et al.*, 2011), and to express emotion (Schröder, 2001). This work focuses on changes in vocal effort as a function of noise and talker-to-listener distance. These changes in vocal effort are governed by talkers' perception of their own voice (Tufts and Frank, 2003). There are 3 main auditory feedback paths, illustrated in Figure 1.2, that affect one's perception of his/her own voice (Pörschmann, 2000; Lehnert and Giron, 1995):

(a) Direct air conduction: sound travels from the talker's mouth to the ear through propagation in the open air.

(b) Bone conduction: sound transmitted through bone and tissue conduction inside the skull. Direct stimulation of the cochlea can occur through vibrations of the skull vibrating the

cochlear fluid or indirect stimulation can occur through excitation of the air entrapped in the ear canal vibrating the eardrum resulting in a direct stimulation the cochlea.

(c) Indirect air conduction: sound travels from the talker's mouth then reflects off of surfaces around the talker traveling back to the talker's ear.

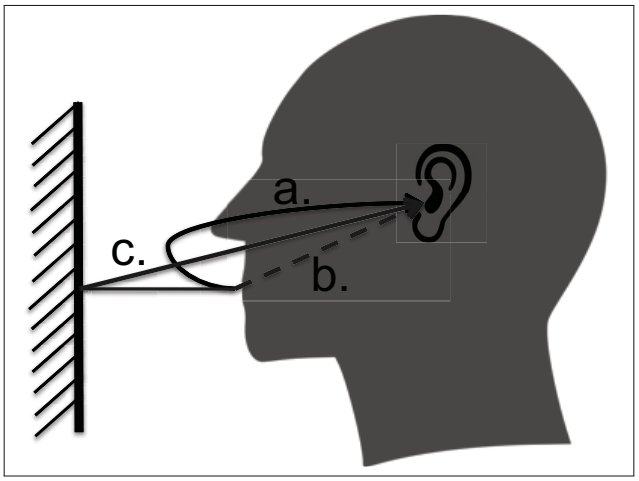

Figure 1.2 Illustration of the three paths affecting the perception of one's own voice: (a) direct air-conduction, (b) bone-conduction, and (c) indirect air-conduction.

This feedback mechanism is referred to as the *audio-phonation loop* (Garnier *et al.*, 2010).

1.4.1 Open Ear

Blauert *et al.* (1980) identified that direct transmission of sound from skull vibrations to the cochlea is 40 dB and 70 dB less effective in the high frequencies and the low frequencies respectively, making sound transmission through air conduction superior to bone conduction. This also implies that in the open ear condition the bone conduction pathway may be neglected. When it comes to the perception of one's own voice, however, the significance of the contribution from each path is debatable. Békésy (1949), concluded that the air and bone conduction paths equally contribute to one's hearing of one's own voice. However, Pörschmann (2000),

observed that except for the mid frequencies (700 to 1200 Hz) where bone conduction had a slightly superior contribution, air conduction was the dominant contributor to the hearing of one's own voice. It is important to note that there are large deviations between talkers in terms of the contribution of bone conduction to self perceived speech (Maurer and Landis, 1990). In summary, for different frequency ranges, both the bone and air conduction paths are significant contributors to the perception of one's own voice. They act as the main feedback paths that aid the talker in correcting his/her speech to become more intelligible.

1.4.2 Occluded Ear

Occluding the ear canal with an HPD reduces the effect of two of the three feedback paths, the direct and indirect air conduction paths, but amplifies the bone conduction path. While speaking the skull vibrates causing the soft tissue of the ear canal to vibrate as well. When the ear canal is open these vibrations are small and negligible. However, when the ear canal is blocked the energy from the soft tissue vibrations in the ear canal build up resulting in an amplification of the bone conduction sounds in the ear canal. This phenomenon is called the *occlusion effect*. The location at which the ear canal is blocked determines the strength of the occlusion effect. Blocking the ear canal right at its opening causes larger occlusion effect than when it is blocked closer to the eardrum. This can be explained by modeling the open and occluded ear canals as open and closed pipes of different lengths. The occlusion effect can also be modeled with electronic circuits where the open ear canal resembles a high pass filter and the occluded ear canal is the removal of this high pass filter (Brummund *et al.*, 2014). Either way, occluding the ear canal changes the perception of one's own voice and a 'boomy' version is perceived. Therefore, occluding the ear canal must affect the way talkers adjust their vocal effort as a function of noise and talker-to-listener distance.

These changes in vocal effort caused by occluding the ear must be studied and modeled for the application of RAVE. Speech production as a function of changing background noise level has been well studied for talkers with occluded and open ears. A review of this work is presented in Section 1.4.3. Variations in vocal effort as a function of talker-to-listener distance for talkers

with open-ears has been studied and modeled but not for the occluded ear. Section 1.4.6 reviews current work on changes in vocal effort with varying talker-to-listener distance.

1.4.3 Vocal Effort in Noise

Increase in speech level as function of the background noise level is known as the *Lombard effect* (Zollinger and Brumm, 2011). The Lombard effect, as well as increasing speech level with changing talker-to-listener distance, are done both involuntarily and voluntarily by a talker to enhance speech intelligibility by the listener. Studies have shown that the Lombard effect is manifested differently when talkers are trying to communicate in noise compared to performing a reading task (Junqua *et al.*, 1999). Garnier *et al.* (2006) showed that changes in speech acoustics for speech produced under the Lombard effect are not purely physiological in nature but are rather a controlled enhancement of speech intelligibility. For open ears, the air conduction pathways are the primary feedback paths for a talker (Henry and Letowski, 2007). Byrne (2014) extensively reviews the findings on changes in speech levels for the open ear and closed ear condition as a function of background noise and type of HPD. Here, a summary of the relevant findings is presented.

1.4.4 In Noise, Open Ears

Lombard speech refers to the significant changes in speech production when speech is produced in noise (Junqua *et al.*, 1999; Zollinger and Brumm, 2011; Brumm and Zollinger, 2011). Some of these changes include an increase in speech level of 1–6 dB for every 10 dB of noise increase (Lane and Tranel, 1971). Shifts in produced fundamental frequency, F0, as well as first formant, F1, have also been observed. Studies show an increase in the fundamental frequency (Junqua, 1993; Garnier and Henrich, 2014) of anywhere between 0.6–2.5 semitones (Lu and Cooke, 2008). Summers *et al.* (1988) report a decrease in spectral tilt, while more recent studies report a shift to the right in the spectral center of gravity (Tufts and Frank, 2003; Garnier and Henrich, 2014). Both of these findings indicate an increase in the high frequency content, which can improve speech intelligibility in noise.

1.4.5 In Noise, Occluded Ears

When the ear canal is occluded studies have shown that talkers do not react to an increase in noise levels as much as talkers not wearing HPDs. Tufts and Frank (2003) report that talkers wearing earplugs in noise decreased their speech level by 4–11 dB compared to their speech level in noise without HPDs. Also, they observed that the overall speech level increased by only 5 dB (from 66.6 dB (SPL) to 71.9 dB (SPL)) when wearing foam HPDs, even when the noise was increased by 40 dB (Tufts and Frank, 2003). In other words, while wearing HPDs, talkers adjust their vocal effort by only 1.25 dB for every 10 dB increase in noise. In quiet, however, talkers wearing earplugs did not significantly alter their overall speech level (Navarro, 1996; Tufts and Frank, 2003) from their open-ear level, with a slight decrease of 0.6 dB. These results contradict older studies (Kryter, 1946; Casali *et al.*, 1987) reporting that talkers increase their speech level by 4 dB while occluded in quiet. Tufts and Frank (2003) attribute this contradiction to the placement of the plug in the ear and its contribution to the occlusion effect, emphasizing again the role of perception of one's own voice on speech production.

1.4.6 Vocal Effort With Varying Distance

In quiet conditions, talkers raise their vocal effort to reach farther distances. A doubling in the talker-to-listener distance increases the vocal level between 1.3–6 dB (Traunmüller and Eriksson, 2000; Zahorik and Kelly, 2007; Pelegrín-García *et al.*, 2011). A study done by Zahorik and Kelly (2007) showed that talkers adjust their vocal effort according to their acoustical environment as well as the communication distance. The talkers' F0 as well as the first formant, F1, also increase as a function of distance. As the vocal level increases, F0 increases by 5 Hz/dB while F1 increases by 3.5 Hz/dB (Liénard and Di Benedetto, 1999). The change in F0 caused by an increase in the communication distance, and thus vocal level, was determined to be unique and distinguishable from changes that occurred in Lombard speech or other factors that may raise the vocal effort (Fux *et al.*, 2011).

Pelegrín-García *et al.* (2011) proposed a model for speech levels as a function of the room acoustics. The proposed model by Pelegrín-García *et al.* (2011) was as follows:

$$L_w = a_k + \alpha_i + (b_k + \beta_i) \times \log_2(d/1.5) + \varepsilon_{ijk}, \tag{1.1}$$

where L_w is the speech power level, a_k and b_k are fixed factors, α_i, ε_{ijk}, and β_i are random effects and d is the distance in meters. For purposes of this work, the model of speech power level in a reverberant acoustical environment is used. This is done to best reflect the acoustical conditions in a factory. Therefore equation 1.1 becomes:

$$L_w = 56.2 + 2.74 + (1.3 + 0.76) \times \log_2(d/1.5) + 1.33. \tag{1.2}$$

Figure 1.3 illustrates the relationship between speech power levels and talker-to-listener distance as presented by Pelegrín-García *et al.* (2011).

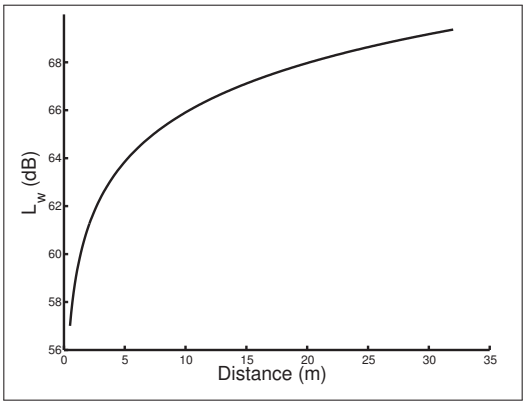

Figure 1.3 Speech power level (L_w) plotted as a function of talker-to-listener distance in a reverberant environment.

These findings relating the speech levels to talker-to-listener distance are only for the open-ear condition, thus these effects need to be well understood for the occluded ear.

This concludes the literature review relevant to this doctoral project. The choice of the ARP is justified by reviewing current HPDs available in the market. BWE techniques for BC speech are reviewed and their weaknesses are highlighted, demonstrating a need for a simple BWE technique that is robust to noise and is speaker independent. Finally, the current literature and findings on changes in vocal effort in noise and with talker to listener distance are presented. The reviewed literature comprises the necessary information to carry out the objectives of this doctoral study. It also highlights the need to denoise the IEM speech signal, to enhance it by extending its bandwidth and finally to find a model of speech levels in noise as a function of the talker-to-listener distance and background noise level for occluded talkers.

CHAPTER 2

IMPROVING THE QUALITY OF IN-EAR MICROPHONE SPEECH VIA ADAPTIVE FILTERING AND ARTIFICIAL BANDWIDTH EXTENSION

Rachel E. Bouserhal[1,3], Tiago H. Falk[2,3], Jérémie Voix[1,3]

[1]École de technologie supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Article submitted to the Journal of Acoustical Society of America

Abstract

Bone and tissue conducted speech has been used in noisy environments to provide a relatively high signal-to-noise ratio signal. However, the limited bandwidth of bone and tissue conducted speech degrades the quality of the speech signal. Moreover in very noisy conditions, bandwidth extension of the bone and tissue conducted speech becomes problematic. In this paper, speech generated from bone and tissue conduction captured using an in-ear microphone is enhanced using adaptive filtering and a non-linear bandwidth extension method. Objective and subjective tests are used to evaluate the performance of the proposed techniques. Both evaluations show a statistically significant improvement of the noisy in-ear microphone speech after enhancement.

2.1 Introduction

Traditionally, communication headsets use a boom microphone, placed in front of the mouth, to capture speech in noisy settings. Although directional, these microphones often suffer from a low signal-to-noise ratio (SNR) in excessively noisy environments and require active noise control for enhancement (Gan and Kuo, 2003). Alternatively, speech captured through bone and tissue vibrations has been used to provide a signal with a higher SNR (Casali and Berger, 1996). Bone conduction speech can be captured either by microphones placed inside an occluded ear (Bou Serhal *et al.*, 2013; Kondo *et al.*, 2006) or through bone conduction sensors

placed somewhere on the cranium (Zheng *et al.*, 2003). Although speech generated from bone and tissue conduction can have a relatively high SNR, it suffers from a limited bandwidth (less than 2 kHz), thus reducing signal quality and intelligibility (Turan and Erzin, 2013b). For applications in which quality and intelligibility are important (e.g. command and control), bone and tissue conduction speech can be a limiting factor. Therefore, to this day, communicating in noise is a difficult task to achieve as the communication signal either suffers from noise interference, in case of airborne speech, or from limited bandwidth, in case of bone and tissue conducted (BC) speech.

Moreover, in excessively noisy environments where workers are exposed to noise levels greater than 90 dB(A) for 8 hours, the Occupational Safety and Health Administration enforces the use of Hearing Protection Devices (HPD) (OSHA, 1983). When worn correctly, HPDs can be very effective in preventing noise induced hearing loss (Berger, 2003). However, limited communication remains the number one complaint of workers equipped with HPDs (NIOSH, 2005).

Communication headsets are a great way of combining good hearing protection and communication. Most commonly, headsets made up of circumaural HPDs equipped with a directional boom microphone placed in front of the mouth are used. Circumaural HPDs can generally provide better attenuation than intra-aural HPDs, because they are easier to wear properly (Berger, 2003). The disadvantages of these types of communication headsets is two-fold. First, the boom microphone is exposed to the background noise and can still capture unwanted noise that can mask the speech signal. Second, cirucumaural HPDs with boom microphones are not compatible with most other personal protection equipment. The use of other personal protection equipment alongside HPDs is common in noisy environments. For example, the use of helmets is required for construction workers as are gas masks for fire-fighters. Using bone and tissue conduction microphones to capture speech is a convenient way to eliminate both of those problems. Bone conduction sensors can be placed in various locations and can provide a relatively high SNR speech signal (McBride *et al.*, 2011). As mentioned previously, however, the elevated SNR comes at a price of very limited bandwidth, typically less than 2 kHz

(Shin *et al.*, 2012). As a consequence, the enhancement of bone and tissue conducted speech is a topic of great interest. Many different techniques have been developed for the bandwidth extension of BC speech (Turan and Erzin, 2013b; Li *et al.*, 2014; Dekens and Verhelst, 2013; Rahman and Shimamura, 2011). Even though theses techniques can enhance the quality of bone and tissue conducted speech, they are either computationally complex or require a substantial amount of training from the user (Shin *et al.*, 2012), thus limiting their widespread use in practical settings.

An effective compromise between the two extremes of noisy air conducted speech and bandlimited BC speech captured by bone conduction sensors is speech captured from inside an occluded ear using an in-ear microphone. Occluding the ear canal with an HPD causes bone conducted vibrations originating from speech to resonate inside the ear canal leading the wearer to hear an amplified version of their voice, this is called the occlusion effect (Bernier and Voix, 2013). By way of the occlusion effect, as a consequence of wearing an earplug, a speech signal is available inside the ear and can be captured using an in-ear microphone. Therefore, occluding the ear canal with a good acoustic seal via an earplug equipped with an in-ear microphone allows for the capturing of a speech signal that is not greatly affected by the background noise because of the passive attenuation of the earplug. Another advantage of using an in-ear microphone instead of a bone conduction microphone is that the speech is still captured acoustically and can share a significant amount of information with clean speech captured in front of the mouth in the 0 to 2 kHz range (Bouserhal *et al.*, 2015a). However, in extremely noisy situations, some residual noise can exist inside the ear and capturing speech through air-conduction can result in a reduced SNR. Additionally, the speech captured inside the ear depends on resonance from bone and tissue and thus also suffers from a limited bandwidth. Because of the shared mutual information between the in-ear microphone speech signal and the air-conducted speech signal captured in front of the mouth, extending the bandwidth of in-ear microphone speech, in quiet conditions, is possible. A bandwidth extension technique that utilizes nonlinear characteristics should extend the bandwidth of the in-ear microphone signal and add the high frequency harmonics (Iser and Schmidt, 2008).

In noisy conditions, however, extending the bandwidth of the bandlimited in-ear microphone speech becomes a difficult task because depending on the spectrum of the noise, simple bandwidth extension techniques may actually amplify the noise in the signal and decrease the SNR. Bandwidth extension techniques for noisy speech are rare and are typically computationally complex (Li *et al.*, 2014; Seltzer *et al.*, 2005). Since the SNR of the in-ear microphone speech is relatively high, denoising the speech signal becomes an easier task if the noise information inside the ear is known. In such extremely noisy conditions that the in-ear microphone signal becomes noisy, speech captured through air-conduction outside the ear has a very low SNR and is almost completely masked by the noise. Here, we propose the use of a microphone placed outside of the ear, on the outside of the earplug, such that the relationship between the sound outside the ear and inside the ear (i.e the transfer function of the earplug) is known. This provides insight about the "in-ear" noise and enabling denoising through adaptive filtering. Once the in-ear microphone speech signal is denoised, bandwidth extension can then be performed to further improve quality. Using combined techniques as such requires little training from the user and is computationally simple. Experimental objective and subjective results show that the proposed solution significantly improves the quality of the in-ear microphone speech. Increases of 44 points on the 100-point MUSHRA (MUlti Stimulus Test with Hidden Reference and Anchor) scale and 1.2 points on the 4.5-point POLQA (Perceptual Objective Listening Quality Assessment) scale were observed.

The remainder of this paper is organized as follows. Section 2.2 describes the methods and material used to perform and evaluate the proposed enhancement technique. The results are presented in Section 2.3 followed by a discussion and conclusion in Sections 2.4 and 2.5, respectively.

2.2 Methods and Materials

2.2.1 Speech Corpus

We propose the use of the Auditory Research Platform (ARP) shown in Fig. 2.1, as a communication headset. The ARP uses an intra-aural custom molded earpiece for passive attenuation of ambient noise. Within the earpiece there is an In-Ear Microphone (IEM) and a miniature loudspeaker. Located flush on the outer face of the earpiece is an Outer-Ear Microphone (OEM). It is of interest to see if in noisy conditions, with a configuration such as that of the ARP, a communication signal similar to that captured in front of the mouth in quiet conditions can be reached. Therefore, a speech corpus was recorded in an audiometric booth with the ARP as well as with a digital audio recorder (Zoom® H4n) placed in front of the speaker's mouth (i.e REF signal). A female speaker read out the first ten lists, totaling 100 sentences, of the Harvard phonetically balanced sentences (IEEE , 1969) and speech was recorded at 8 kHz sampling rate and 16-bit resolution across the three microphones, simultaneously. The recordings were made with an 8 kHz sampling rate to stay true to realistic conditions with radio communications.

A noisy speech corpus was then created from the clean corpus. Noise was injected to the OEM signals post recording to avoid any uncontrolled deviations in the speech between different recordings. To remain as close as possible to realistic conditions, the noise inside the ear, the IEM noise, was simulated using the OEM noise and the transfer function of the earplug. This was achieved by playing white noise over loudspeakers in the audiometric booth while the speaker was still equipped with the ARP (Nadon *et al.*, 2015). The noise signals collected by the IEM and OEM were 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 (Varga and Steeneken, 1993) was then added to the OEM signal at an SNR of -5 dB. The noise was then filtered using the previously calculated transfer function of the earpiece and added to the IEM speech signal. The REF signal was kept clean in order to provide an upper bound on the achievable performance. An SNR of -5 dB was chosen to simulate a typical industrial

factory workplace setting. At this level, the signal captured by the OEM contains inaudible speech information, as it is buried in the noise (Bouserhal *et al.*, 2015a).

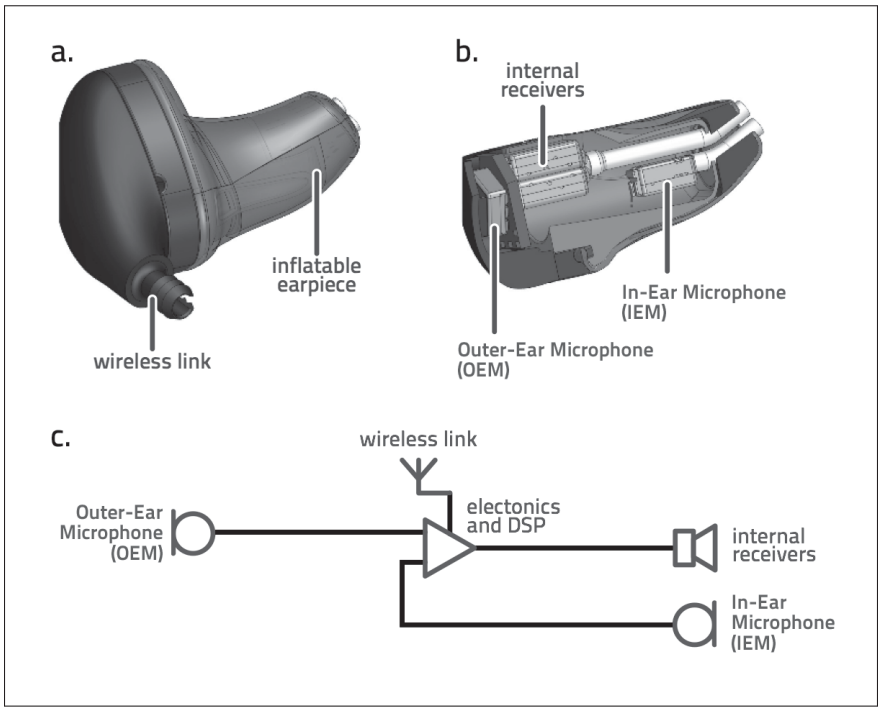

Figure 2.1 Auditory research platform (a), its electroacoustic components (b), and equivalent schematic (c). Placed inside the ear the ARP captures speech produced by a talker using either the IEM or the OEM and transmits communication to other users through a wireless link.

2.2.2 Predicted Quality

As shown in previous work, clear spectral differences between the IEM, OEM, and REF captured speech can be observed. The IEM signal has a boost in the low frequency range but has a high frequency roll-off at about 1.8 kHz. The OEM and REF signals share the same bandwidth but have slight shifts in the formants (Bou Serhal *et al.*, 2013). These formant shifts are minimal and should not affect the quality of the clean OEM signal. To illustrate these spectral differences once more, the linear predictive coding (LPC) spectral envelope of the phoneme /i/ recorded with the REF, OEM and IEM simultaneously is presented in Fig. 2.2.

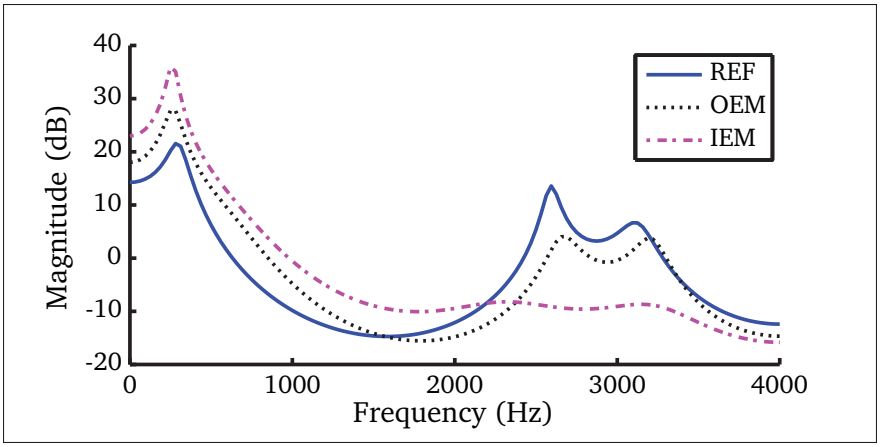

Figure 2.2 The linear predictive coding spectral envelope of the phoneme /i/ recorded with the REF, the OEM and the IEM simultaneously.

Predicted Quality in Quiet Conditions

Considering the shared mutual information between the OEM and REF signals (Bouserhal *et al.*, 2015a) as well as their spectral differences (see Fig. 2.2), it is expected that the OEM speech signal is perceptually very similar to that of the REF speech in quiet conditions. Moreover, the "boomy" effect of the IEM, its limited bandwidth, and reduced shared mutual information with the REF signal should reduce its perceptual quality considerably compared to both the OEM and REF signals. To validate these predictions, as an objective quality measure, the International Telecommunication Union ITU-T Standard P.863, Perceptual Objective Listening Quality Assessment, POLQA, (ITU-T, 2011) with the REF signals as the reference was calculated for both the OEM and IEM signals. POLQA is the current recommendation for benchmarking and is used to evaluate speech for new and upcoming networks. The results of these measurements are shown in Section 2.3.

Predicted Quality in Noisy Conditions

In noisy conditions, because of the passive attenuation of the earplug, the quality of the IEM signal should not be greatly degraded by the presence of noise. The quality of the signal

picked up by the exposed OEM, however, should be substantially reduced as the OEM speech is masked by the high level of noise. This prediction is supported by the maintenance of the amount of mutual information between the IEM and the REF signals in noisy conditions and the significant degradation observed between the OEM and the REF in noisy conditions (Bouserhal *et al.*, 2015a). To validate the predicted changes in quality, POLQA was calculated for the noisy condition (SNR=-5 dB) of the IEM and the OEM speech as compared to the clean REF speech. The results of these measurements are shown in Section 2.3.

2.2.3 IEM Noise Reduction

NLMS Filtering

Once the noise level is high enough that the OEM speech is almost completely masked (SNR < -5 dB), the IEM speech signal can be denoised using normalized least mean squared (NLMS) adaptive filtering. The choice of adaptive filtering comes from a need to create an algorithm that assumes no properties about the noise and is, thus, robust to various types of noise. Therefore, using adaptive filtering is beneficial for the user by enhancing the received communication signal.

To properly denoise the IEM speech signal produced by the user without affecting the speech content, the adaptation process must be frozen (OFF) when the user is speaking and active (ON) when the user is not speaking. This ensures that the adaptive filter cancels only the noise and does not interfere with any speech produced by the user. The two states of the adaptive filter are shown in Fig. 2.3. When the adaptation is ON the structure of the proposed adaptive filter follows the well-known structure commonly described in the literature (e.g Manolakis *et al.* (2005)); the only exception being that the signal of interest is the error signal, $e(n)$. Here, $H(z)$ is the true transfer function of the earplug while $\hat{H}(z)$ is the estimated earplug transfer function. The method of estimating $\hat{H}(z)$ is further explained in the next section. When the adaptation is ON, the user is not speaking. The OEM captures the noise outside the ear, $n_o(n)$, while the IEM captures the residual noise inside the ear $n_r(n)$, colored by $H(z)$. The signal

captured by the IEM is defined as the desired signal, $d(n)$. The input, $x(n)$, to the adaptive filter is the signal captured by the OEM filtered with the adaptive filter which is initialized by the estimated transfer function of the earplug $\hat{H}(z)$. The output of the adaptive filter, $y(n)$, is thus a close estimate of the residual noise inside the ear and the difference between $d(n)$ and $y(n)$ should approach 0. The adaptive filter of order 160 is defined as follows:

$$\begin{aligned} y(n) &= w^T(n-1)x(n), \\ e(n) &= d(n) - y(n), \\ w(n) &= w(n-1) + \frac{\mu e(n)x(n)}{\varepsilon + x(n)^T x(n)}, \end{aligned} \tag{2.1}$$

where n is the current time index, μ is the adaptation step size, $w(n)$ is the vector of filter weights at time index n , and ε is a very small number to avoid division by zero.

As presented in Fig. 2.3, when the adaptation is OFF, let $s_o(n)$ and $n_o(n)$ be the speech signal produced by the user and noise signal outside the ear, respectively. Therefore, the OEM picks up the sum of these two signals, $x(n)$. Meanwhile, the IEM picks up the residual noise signal after the attenuation of the earplug, $n_r(n)$, and the residual speech signal $s_r(n)$. The speech signal originating from bone and tissue conduction, $s_i(n)$, is also picked up by the IEM. The sum of all the three signals picked up by the IEM is the desired signal $d(n)$. The signal $x(n)$ picked up by the OEM is then filtered using the $\hat{H}(z)$ and the output, $\hat{x}(n)$, is fed to the input of the NLMS adaptive filter. The output of the adaptive filter, $y(n)$ is then subtracted from $d(n)$. The adaptive filter brings the difference between the residual noise, $n_i(n)$, and the estimated residual noise, $\hat{n}_i(n)$ to zero. Since the OEM speech is almost entirely masked by the noise, the effect of $s_r(n)$ and $\hat{s}_r(n)$ is negligible. Therefore, the resulting difference between the output of the adaptive filter and the signal captured by the IEM is the speech signal originating from bone and tissue conduction, $s_i(n)$, with minimal effects of noise.

To properly denoise the IEM speech signal produced by the user without affecting the speech content, adaptation must only be performed when the user's speech is not present inside the ear. This ensures that the adaptive filter cancels only the noise and does not interfere with

any speech produced by the user. Therefore, the adaptive filtering algorithm must also include a robust speech detection procedure that switches the adaptation process ON and OFF as a function of the speech inside the ear. This adaptation process, including the speech detection method, is described in the following sections.

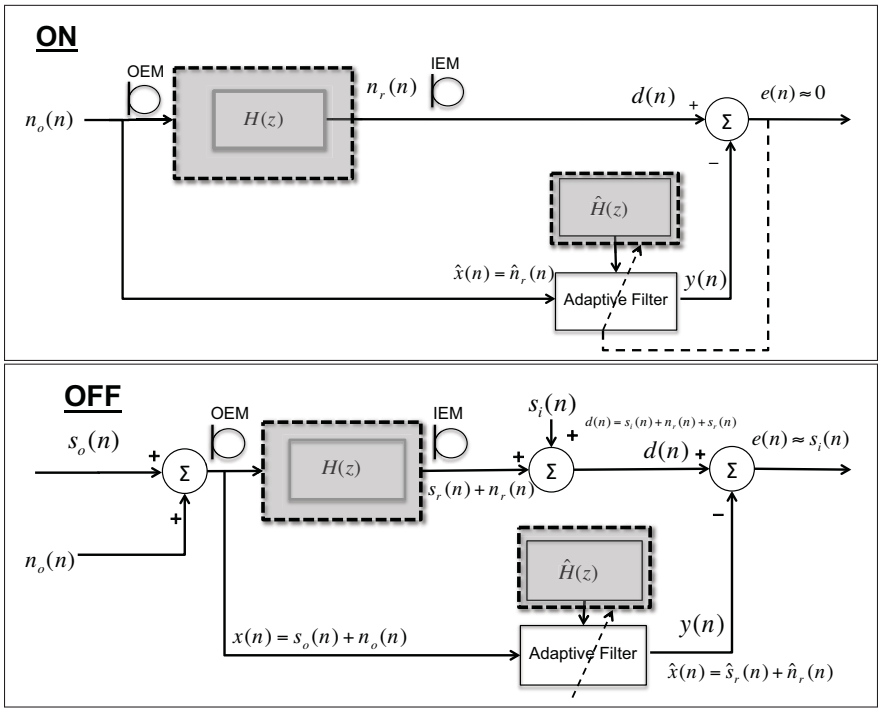

Figure 2.3 Block diagram representing the NLMS filtering stage: (a) when the adaptation is ON and (b) when it is OFF.

Offline Transfer Function Identification

First, the earplug transfer function must be estimated, as it varies from user to user. This is done in an offline identification stage. As shown in Fig. 2.4, the ARP is worn and the user is exposed to white noise at 85 dB (SPL) using a loudspeaker outside the ear for at least 2 seconds. The OEM and IEM simultaneously capture the signals outside and inside the ear respectively. After the OEM and IEM signals are collected the transfer function of the earplug, $H(z)$, is estimated as $\hat{H}(z)$.

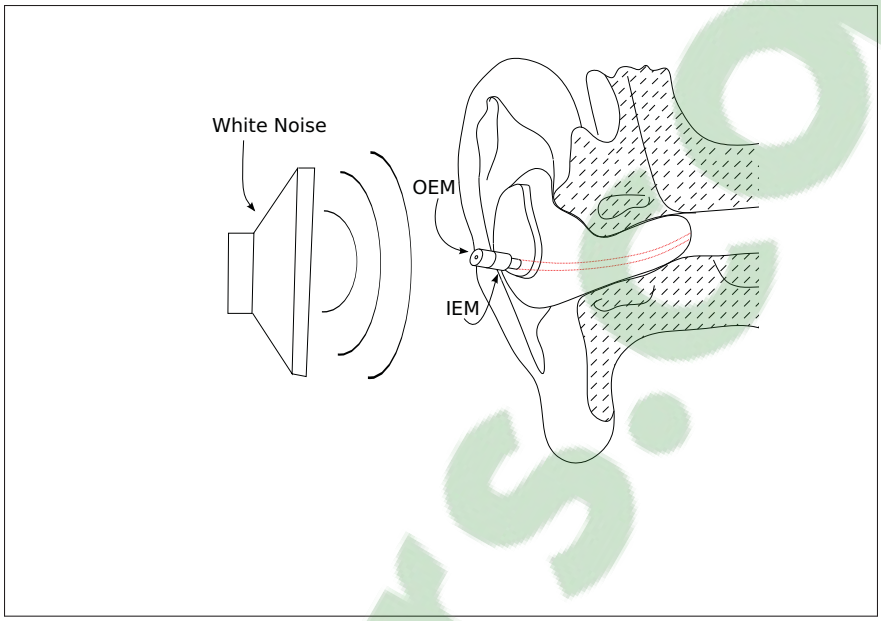

Figure 2.4 Offline identification stage of the earplug transfer function in the user's ear. White noise is played on a loudspeaker outside of the ear and recorded using both the IEM and the OEM. The transfer function of the earplug is calculated by assessing the noise outside the ear, recorded by the OEM, and the residual noise inside the ear, recorded by the IEM.

The Adaptation Process

To achieve denoising without affecting the speech content, the adaptation process is a function of whether or not the user is speaking. To denoise the user's speech, the adaptive filter must only adapt when the user is not speaking. This ensures that the filter is adapting to the earplug transfer function and thus the noise and only the noise is subtracted from the signal and not any relevant speech information. To guarantee robustness of the speech detection process, voice activity detection inside the ear is achieved in the current project by monitoring the value of the coefficients of the adaptive filter. After completion of the two second identification stage the vector of filter weights over the entire index of time, w, is used to detect if the user is speaking. To decide what criteria can be used to detect speech inside the ear using filter weights, test signals were developed using the first 10 lists of the recorded Harvard phonetically balanced

sentences discussed in Section 2.2.1, for both the OEM and the IEM. The test signals always started with at least 2 seconds of noise followed by 8 to 10 seconds of speech either by the user or by an external competing speaker. Exterior speech was added to simulate a case where the user is not speaking but an external speaker is loud enough that some residual speech exists after the passive attenuation of the plug. The residual speech should not trigger the speech activity of the adaptation process. For the IEM signal, the residual speech was simulated by passing the speech through $\hat{H}(z)$. The location of the user's speech and the residual speech was randomized to avoid any trends in the adaptation process. Fig. 2.5 is an example of a randomly chosen IEM test signal with both user speech and external speech segments.

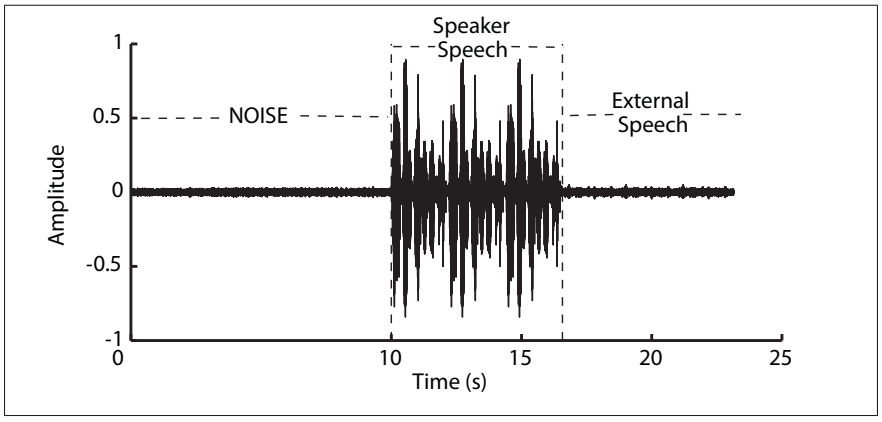

Figure 2.5 Test signal for the IEM to optimize speech detection criteria.

Through analysis of the changes in the filter weights for the test signals recorded by the female speaker as described in Section 2.2.1, it was concluded that the maximum valued filter weight can be chosen as a good triggering criteria. Once the maximum filter weight increases more than a triggering threshold, T_g, from one time-index to the other, it is predicted that the user is speaking. Therefore once

$$\frac{\max(w(n))}{\max(w(n-1))} \geq T_g,$$

speech by the user is detected and the adaptation is turned OFF. The choice of value for T_g, had to be done in a way that was not particular to one female speaker. Recorded conversation

using the IEM and the OEM from 4 different speakers (2 female, 2 male) was used to analyze the effect of using different triggering thresholds. Noise was inserted using the same procedure as discussed in Section 2.2.1. A sweep of the voice activity detection triggering threshold, T_g, from 1.01 to 1.2 was performed during the adaptation process. The bandwidth of the denoised signals for the 4 speakers resulting from the sweep was extended using the BWE process described in Section 2.2.4. The quality of these signals was measured before and after the bandwidth extension to see the effect of the different values for the triggering criteria. The choice of T_g was made as the triggering percentage value that produced the optimal objective quality over the 4 speakers as is shown in Section 2.3.

The change in filter weights is triggered at the onset of speech but not the end. To ensure that the adaptive process starts back once speech inside the ear is no longer present the overall change in energy, Δ_ε, at the onset of speech is also measured and monitored, per sample, i.e. $\Delta_\varepsilon(n)$. Once triggered by the user's speech, the adaptation is disabled for at least one second and as long as Δ_ε is maintained. When the adaptation is OFF the filter weights of the adaptive filter are updated with those from the previous second, $w(n-fs)$. This is to ensure that the filter weights are those from when no speech is produced by the user. Once the change in energy is less than the onset change, $\Delta_\varepsilon(n) < \Delta_\varepsilon$, the adaptation starts again. The process of monitoring the change in Δ_ε gives a non ad-hoc way to turn ON the adaptation once the user is no longer speaking. The adaptation process is demonstrated by the flow chart in Fig. 2.6.

The adaptive filtering denoises the IEM signal by utilizing the information about the noise captured by the OEM. Once the IEM is denoised its quality can be enhanced by extending its bandwidth in the high frequencies as described in the next section.

2.2.4 IEM Bandwidth Extension

Artificially extending the bandwidth of a clean bandlimited signal has been very well studied. Since the IEM signal shares mutual information with the REF signal between 0-2 kHz (Bouserhal *et al.*, 2015a), it is only necessary to extend the bandwidth in the high frequency

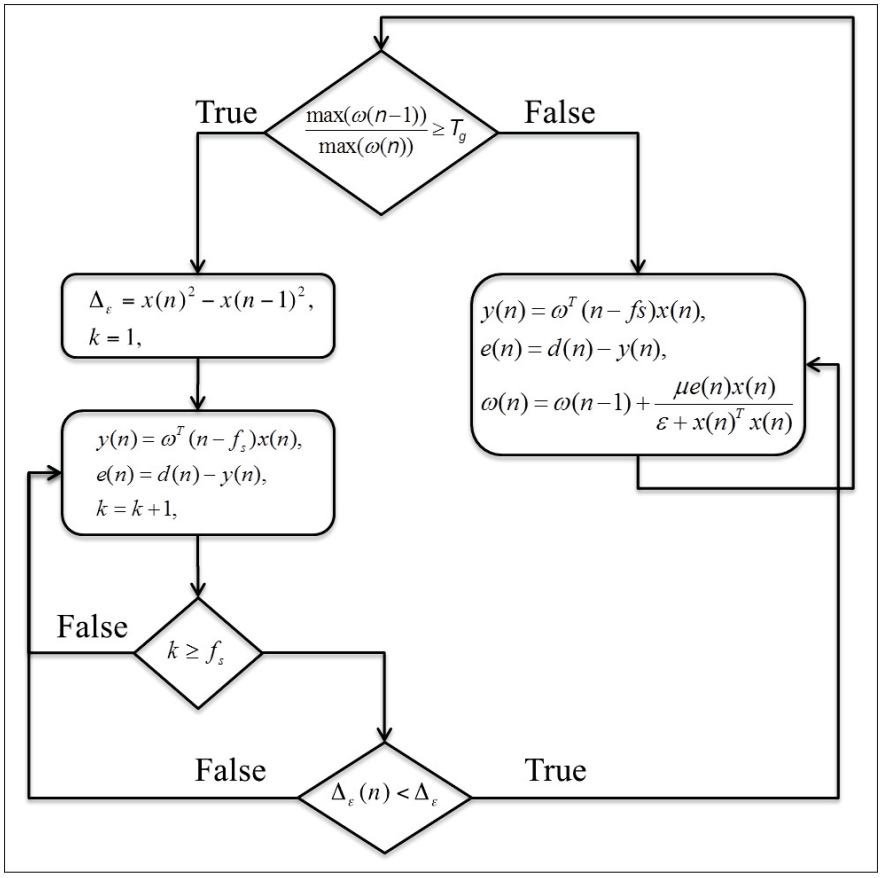

Figure 2.6 Flow chart representing the adaptation process.

range, 2-4 kHz. As described by Iser et. *al* (2008), a simple yet effective way of extending the bandwidth is through the application of the signal's nonlinear characteristics. A block diagram of the bandwidth extension process is shown in Fig. 2.7. First, the signal is upsampled by a factor of 2 to avoid aliasing. The excitation signal is extracted using a whitening filter and then cubed (Iser and Schmidt, 2008). The whitening filter is a finite infinite response filter whose coefficients are those of an 18th order LPC filter at that time frame. Cubing the excitation reproduces the odd harmonics along the entire bandwidth including the high band, in this scenario from 1.8 kHz to 4 kHz. Since the high frequencies are the only region of interest and to eliminate any overlap, the excitation signal is bandpassed between 1.8 kHz and 3.5 kHz using a second order Butterworth filter. The IEM signal is also bandpassed with a second order Butterworth filter between 160 Hz and 1.8 kHz to eliminate the boomy effect coming from the bone and tissue conduction and because it contains no relevant frequency information above

1.8 kHz (see Fig. 2.2). The sum of the two bandpassed signals is then low passed with a fourth order Butterworth filter at 3.5 kHz. This is done to eliminate any residual ringing caused by the odd harmonics of the cubed excitation signal. The overall output is then donwnsampled by a factor of 2 to go back to an 8 kHz sampling frequency. It is important to note that this BWE technique adds missing harmonics in the high frequencies. However, missing formants and frication noise are not recovered.

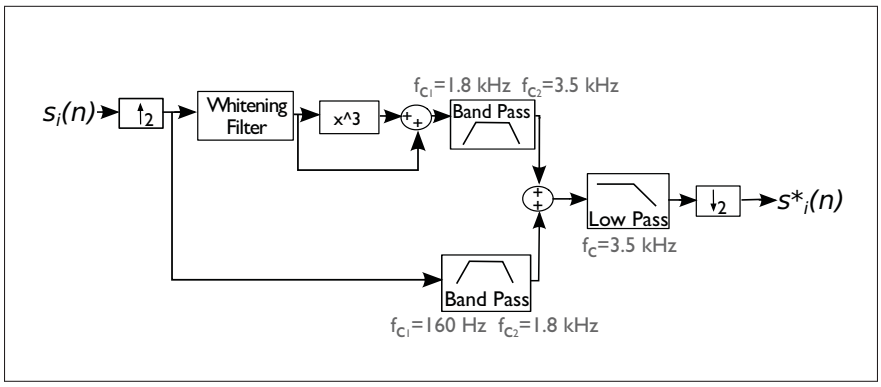

Figure 2.7 Block diagram illustrating the bandwidth extension process.

2.2.5 Performance Evaluation

The performance of the denoising and bandwidth extension processes were evaluated using both objective and subjective measures. Objectively, the quality of the signals was measured using POLQA. To confirm the results from the objective measures the quality of the denoised and bandwidth extended speech was also measured subjectively. The MUlti Stimulus Test with Hidden Reference and Anchor (MUSHRA) (ITU-R, 2001) was used for this latter test. As part of the MUSHRA evaluation, four test signals were compared to the reference: the clean IEM signal, the noisy IEM signal, the denoised IEM signal, and the bandwidth extended denoised IEM signal. The reference signals chosen are those recorded in front of the mouth (REF). The noisy IEM served as the anchor, since it is a bandlimited noisy version of the reference signal. Ten randomly selected REF speech signals and their corresponding test signals were chosen.

The test was performed online and participants were invited to take part of the test through an email that was approved by the internal review board at École de technologie supérieure. A description of the nature of the test, as well as detailed instructions, were described in the email invitation. No assumptions were made on the participants' hearing abilities. To measure the statistical significance of improvements in the quality both objectively and subjectively, the Analysis of Variance (ANOVA) was tested on the gathered data.

2.3 Results

2.3.1 Pre-Enhancement Objective Quality Assessment

Quiet Condition

The POLQA MOS-LQO (mean opinion score - listening quality objective) results comparing the IEM and OEM signals to the REF singals are shown in Fig. 2.8. It can be seen that the quality of the OEM signal in quiet is high and in some cases was measured to have the maximum POLQA score of 4.5, thus indistinguishable from the REF signal. The inferior quality of the IEM can also be seen. The great variability in the POLQA scores for the IEM signals could be attributed to the fact that POLQA was not designed to measure the quality of speech originating from bone and tissue conduction.

Noisy Condition

To show the decrease in the POLQA MOS-LQO results between the clean and the noisy condition (SNR=-5 dB), the MOS-LQO of the noisy IEM and the noisy OEM is shown in Fig. 2.9. Again the large variability in the POLQA scores for the IEM signals could be due to the fact that POLQA was not intended for use with bone and tissue conducted speech. Descriptive statistics are used to evaluate the degradation caused by noise. The cumulative distribution of the difference between the clean and the noisy POLQA MOS-LQO scores of the IEM signals

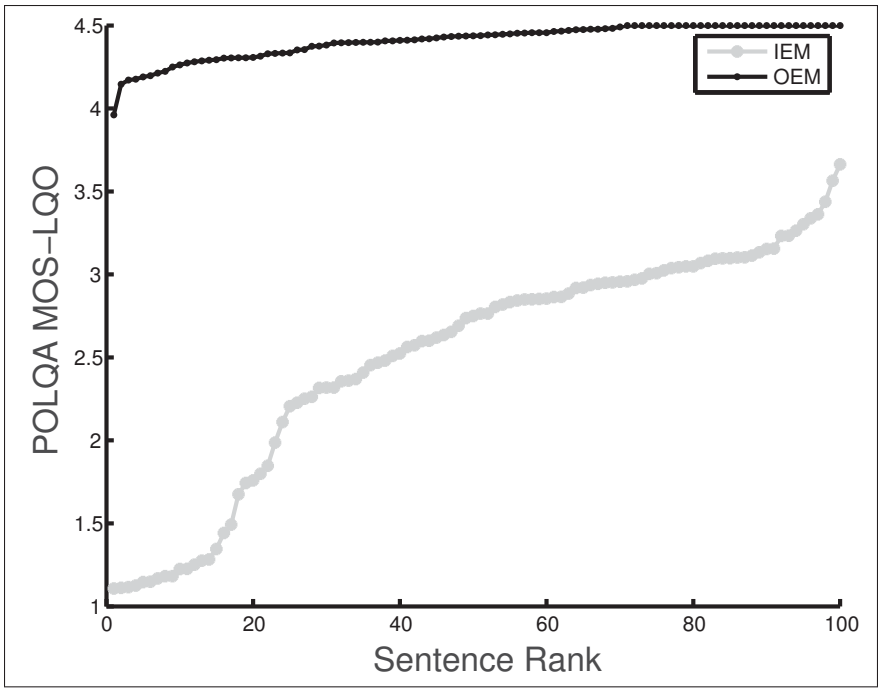

Figure 2.8 POLQA MOS-LQO results of clean IEM and OEM signals using the REF signal as reference, with sentences sorted by ascending order of IEM MOS-LQO scores.

as well as the difference between the clean and the noisy POLQA MOS-LQO scores of the OEM signals are shown in Fig. 2.10.

It can be seen that the decrease in the OEM quality is much greater than the decrease in the IEM quality. As a consequence of noise, half of the OEM sentences were degraded by at least 3.39 points on the 5-point POLQA MOS-LQO scale, while half of the IEM sentences were at most degraded by 1.01 points. This confirms that the passive attenuation of the earplug prevents major degradations from noise on the IEM speech. Therefore, in noise, the IEM signals have superior quality relative to the OEM signals. The completely degraded signal captured by the OEM in noisy conditions can be utilized to denoise the relatively superior quality speech signal captured by the IEM, as described in Section 2.2.3.

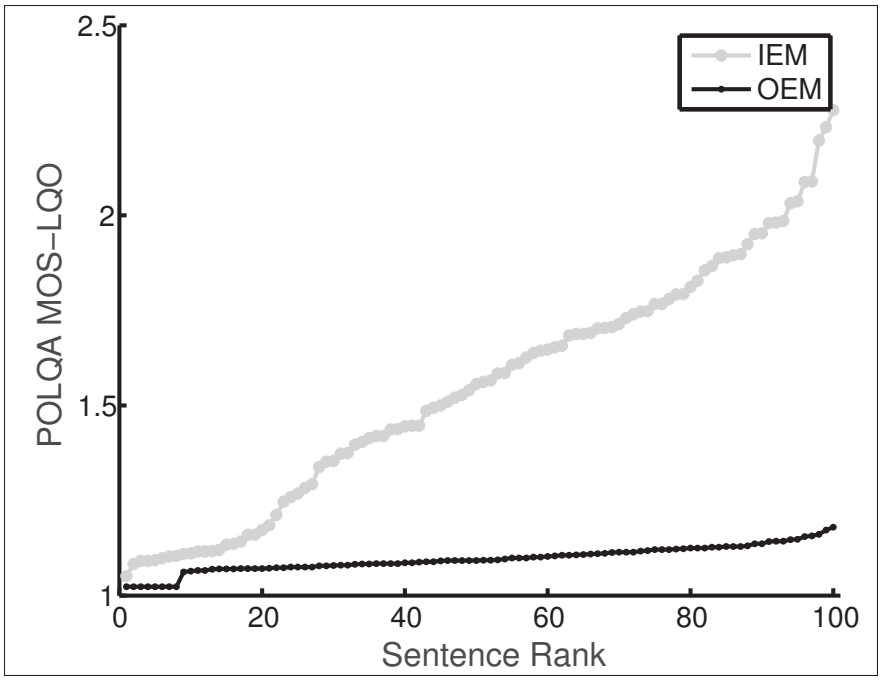

Figure 2.9 POLQA MOS-LQO results of noisy IEM and OEM using the REF signal as reference, with sentences sorted by ascending order of IEM MOS-LQO scores.

2.3.2 IEM Speech Enhancement

Adaptive Process Triggering Threshold

The adaptation must be ON only when the user is not speaking. This way the IEM speech is denoised without affecting the speech content. To show the importance of choosing an optimal triggering threshold, a denoised speech signal using a triggering criterion of $T_g = 1.15$ is plotted in Fig. 2.11. Since the adaptation is ON in this case even after the onset of speech by the user, the IEM speech content is affected by the denoising because the filter coefficients are not adapted only to $H(z)$. It is therefore important to have an optimal voice activity detection criteria for the adaptive filtering process. The choice of triggering threshold, T_g, was chosen based on the results of one female speaker but validated for 4 other subjects. A sweep from $T_g = 1.01$ to $T_g = 1.2$ with a step size of 0.01 was done on 4 different speakers, as described in Section 2.2.3. The average POLQA MOS-LQO scores with only noise reduction and with

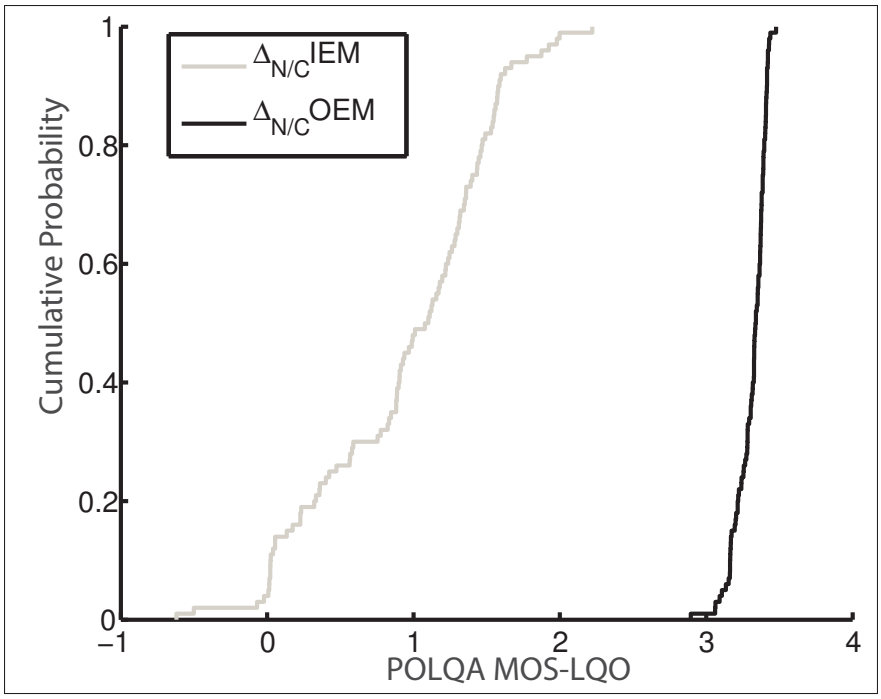

Figure 2.10 Cumulative distribution plot of the difference in POLQA MOS-LQO results between the clean and noisy IEM and OEM signals.

bandwidth extension are shown in Fig. 2.12. The results show a clear peak around 1.06-1.07, suggesting that triggering threshold of $T_g = 1.06$ to detect speech activity inside the ear is best and can be extended to more than just one speaker.

IEM Noise Reduction

With a triggering threshold chosen as, $T_g = 1.06$, the denoising and bandwidth extension techniques described in Section 2.2.3 were performed on the test speech signals. For the denoising phase $\mu = 0.7$ and $\varepsilon = 0.001$ were chosen empirically. To show the performance of the denoising using adaptive filtering, a randomly selected denoised IEM signal, IEM NS, is plotted against its corresponding noisy IEM test signal, IEM N, in Fig. 2.13. As can be seen, the adaptive filtering process denoises the entire signal, when only noise is present (a), when the user is speaking (b) and when external speech is present (c). The adaptation process stops adapting once the user is speaking and relevant IEM speech content is preserved.

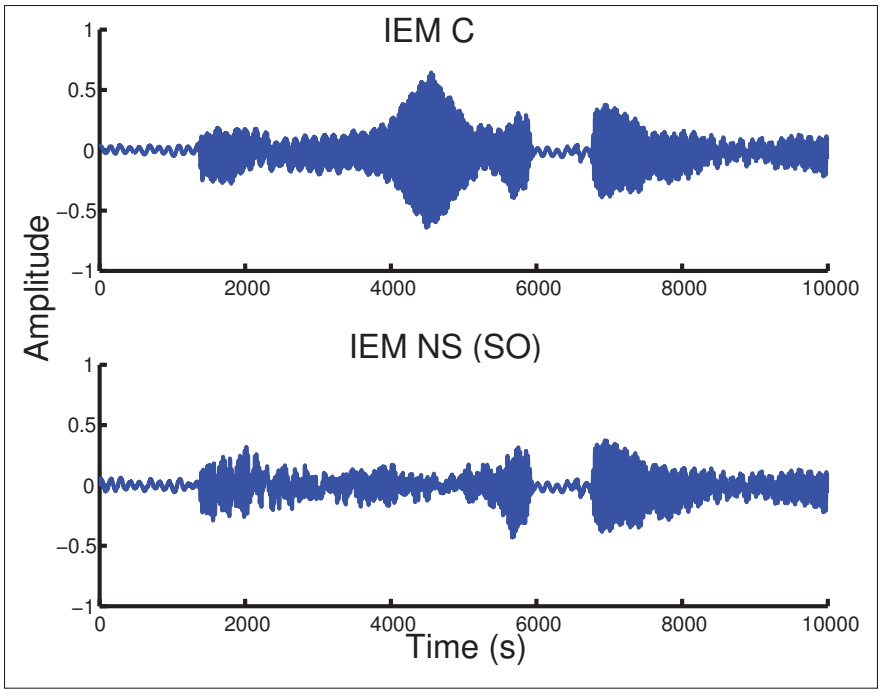

Figure 2.11 A denoised IEM signal (IEM NS(SO)) using suboptimal criteria for speech detection during the adaptation process plotted against the clean IEM signal (IEM C).

Bandwidth Extension

Artificial bandwidth extension is then applied to the denoised signals. To show the regeneration of the the odd harmonics and to compare the spectral content, the spectrograms of the REF signal, the noisy IEM, the denoised IEM and the bandwidth extended IEM signal are shown in Fig. 2.14. The noise reduction can be seen, as well as the 'noise-like' effects of the bandwidth extension. Overall, however, the missing mid and high frequency harmonics lost in the IEM signal are regenerated after the bandwidth extension.

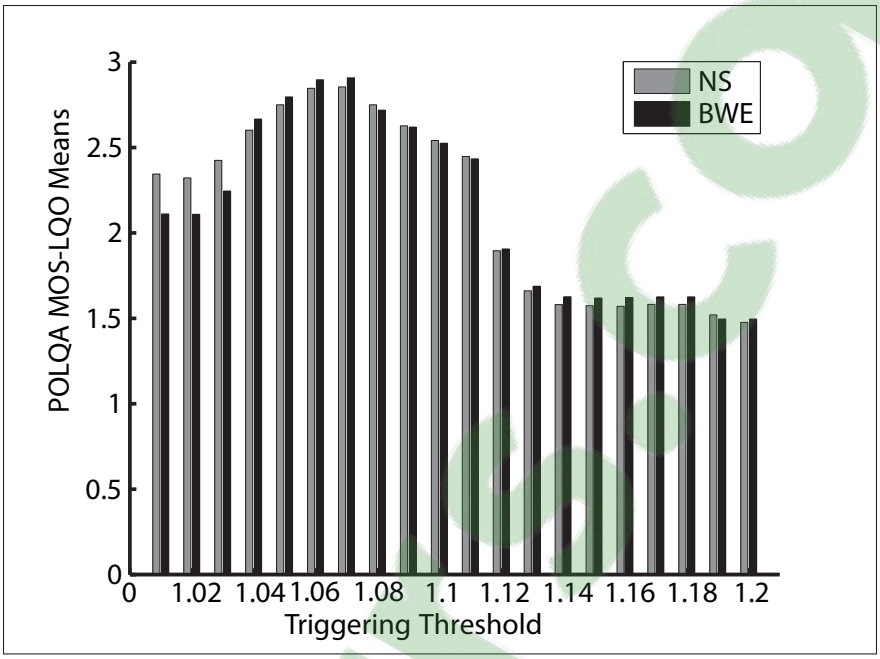

Figure 2.12 Average POLQA MOS-LQO scores after denoising (NS) and after bandwidth extension (BWE) over different triggering percentages, showing a peak at $T_g = 1.06 - 1.07$.

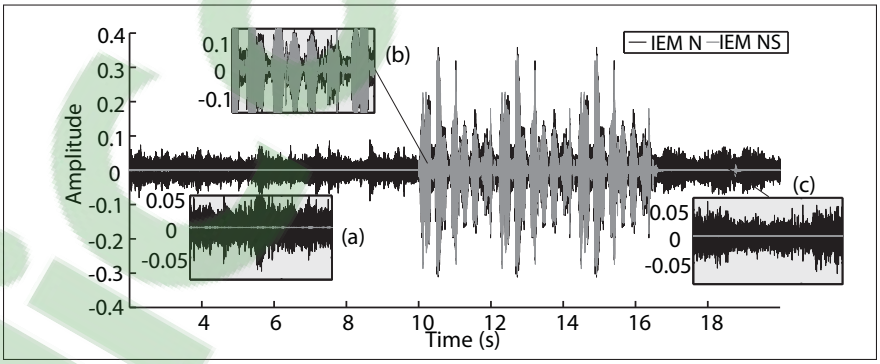

Figure 2.13 The denoised IEM signal (IEM NS) as compared to the noisy IEM signal (IEM N). Zoomed portions of the denoised signal when only noise is present (a), when speech inside the ear is present (b) and when external speech is present (c)

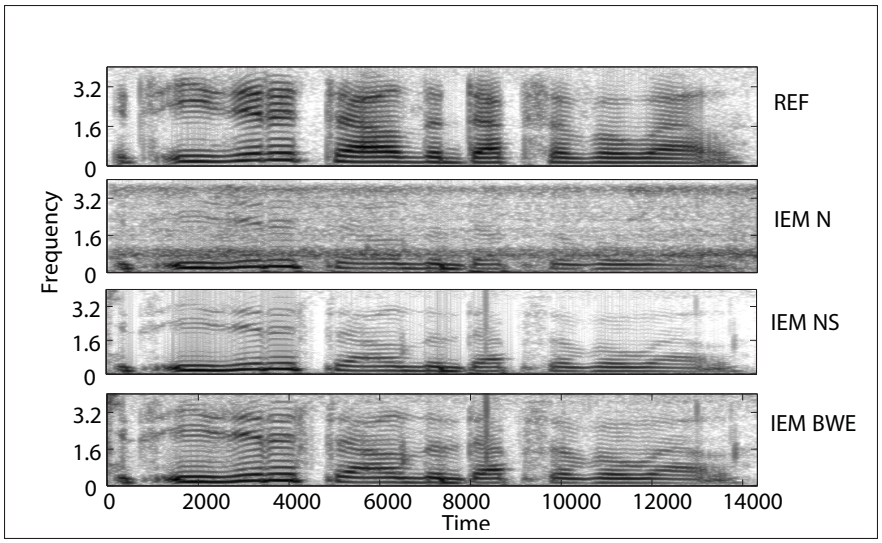

Figure 2.14 The spectrograms of the sentence 'It is easy to tell the depth of a well' of the clean reference signal (REF), the noisy IEM signal (N), the denoised IEM signal (NS), and bandwidth extended denoised IEM signal (BWE).

2.3.3 Performance Evaluation

Objective Evaluation

To compare POLQA MOS-LQO results, the cumulative distributions of the difference in POLQA MOS-LQO scores between the denoised IEM signal (NS), the noisy IEM signal (N) and the clean IEM signal are plotted in Fig. 2.15. The same comparison made with the bandwidth extended signals are plotted in Fig. 2.16. This is done to show if, objectively, the bandwidth extension does increase perceived quality. The results show that the denoising enhanced at least half the sentences by 1 point on the POLQA MOS-LQ0 scale and that at least half the sentences measured no differently than the clean signals after denoising. Objectively, the results comparing the effects of bandwidth extension show that bandwidth extension increases the quality from the noisy signal by 1.2 points for at least half of the sentences. However, very small improvements of 0.02 and 0.14 points caused by bandwidth extension between the denoised signals and the clean signals respectively for at least half the sentences is seen. Therefore, results point that both the denoising as well as the bandwidth extension enhance the

quality of the IEM noisy signal. The mean POLQA scores and p-values from one dimensional ANOVA tests are shown in Tables 2.1 and 2.2, respectively. Results show a statistically significant enhancement from the noisy signal caused by the denoising and the bandwidth extension techniques. There is also a significant improvement from the bandwidth extension technique and the clean IEM signal, thus signaling the importance of the higher frequency components for quality perception.

Subjective Evaluation

The subjective results from the MUSHRA listening test confirm the objective trends found using POLQA. The results averaged over 42 participants are shown in Fig. 2.18. The mean MUSHRA scores and p-values from one dimensional ANOVA tests are shown in Tables 2.3 and 2.4, respectively. A statistically significant increase in quality can be seen as a consequence of the denoising and the bandwidth extension. Objectively and subjectively, there is no statistical significance between the quality of the clean IEM signals and the denoised IEM signals, thus suggesting indistinguishable differences. To confirm this, the log-spectral distance (LSD) between the de-noised IEM and the clean IEM signals was measured. The LSD is defined as follows (Falk *et al.*, 2010):

$$LSD = \sqrt{\frac{1}{2\pi}\int_{-w}^{w}\left[10\log_{10}\frac{s_i(w)}{s_i^*(w)}\right]^2 dw} \tag{2.2}$$

As shown in Fig. 2.17, all LSD values were under 1 dB. In speech coding, two signals with LSD < 1 dB are considered to be perceptually indistinguishable (Paliwal and Kleijn, 1995).

2.4 Discussion

Experimental results with POLQA showed a statistically significant improvement in the speech quality between the noisy IEM speech and the enhanced (denoised and bandwidth extended) speech. Looking only at POLQA scores, however, it is not as apparent that the bandwidth extension enhances the quality much more than the denoising does. This could be because

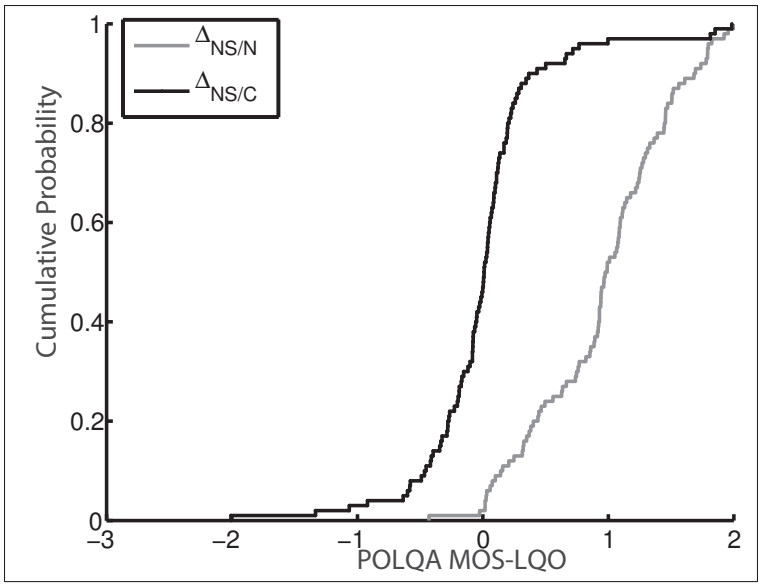

Figure 2.15 Cumulative distribution of the difference in POLQA MOS-LQO scores between the denoised and noisy IEM ($\Delta_{NS/N}$), as well as the denoised and clean IEM ($\Delta_{NS/C}$).

Table 2.1 Average POLQA MOS-LQO scores for the noisy IEM signal (N), the denoised IEM (NS), the bandwidth extended IEM (BWE) and the clean IEM signal (C).

Signal	Mean POLQA	Signal	Mean POLQA
IEM N	1.559	IEM BWE	2.790
IEM NS	2.655	IEM C	2.757

extending the bandwidth can introduce noise-like features in the high frequencies that could be misinterpreted by the objective measure as noise. The subjective results support this hypothesis. From the MUSHRA results it is evident that the denoised bandwidth extended signal is perceived to have significantly better quality than the denoised IEM without bandwidth extension. Extending the bandwidth after denoising results in even better perceived quality than the clean IEM signal. The mean and p-values between the MUSHRA scores of the clean IEM and the denoised IEM show that there was no statistical significance between the two. This suggests that the perceived quality between the denoised IEM and the clean IEM is undistin-

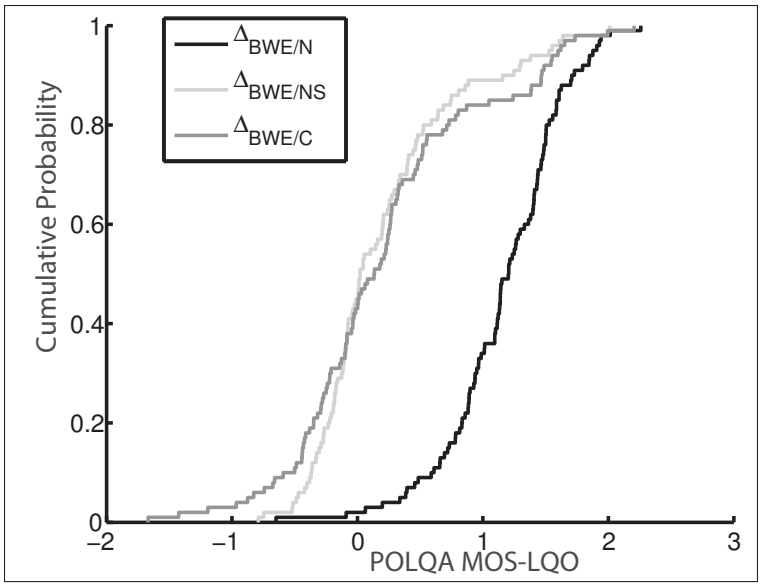

Figure 2.16 Cumulative distribution of the difference in POLQA MOS-LQO scores between the bandwidth extended and noisy IEM ($\Delta_{BWE/N}$), the bandwidth extended and denoised IEM ($\Delta_{BWE/NS}$), and the bandwidth extended and clean IEM ($\Delta_{BWE/C}$).

Table 2.2 Statistical significance results based on a 95% confidence interval between the objective evaluation of different stages of enhancement.

Signals	p-value	Significant?
N vs. NS	$p < 0.0001$	Yes
N vs. BWE	$p < 0.0001$	Yes
NS vs. BWE	$p < 0.01$	Yes
C vs. BWE	$p < 0.01$	Yes
C vs. NS	$p = 0.9413$	No

guishable. In fact, in about 30% of the time participants gave the clean IEM speech and the denoised IEM speech identical MUSHRA scores.

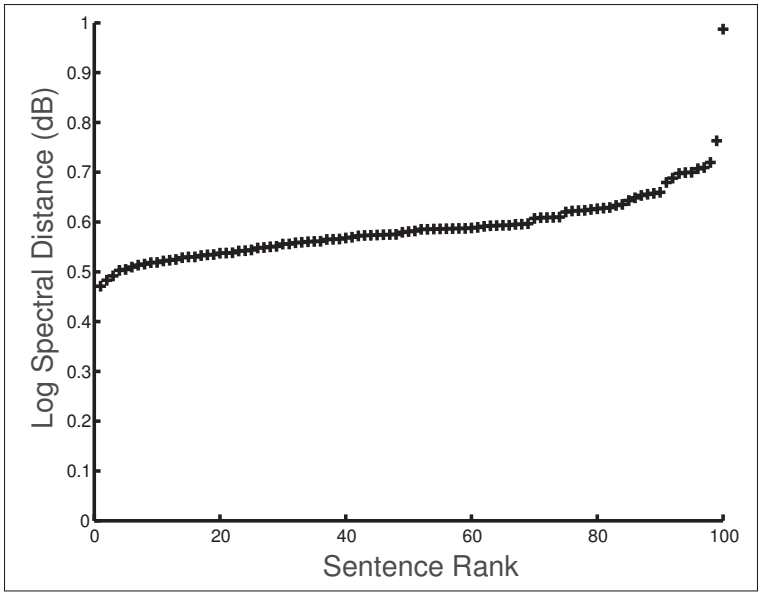

Figure 2.17 Log-spectral distance between the clean IEM signals and the denoised IEM signals, with sentences sorted in ascending order.

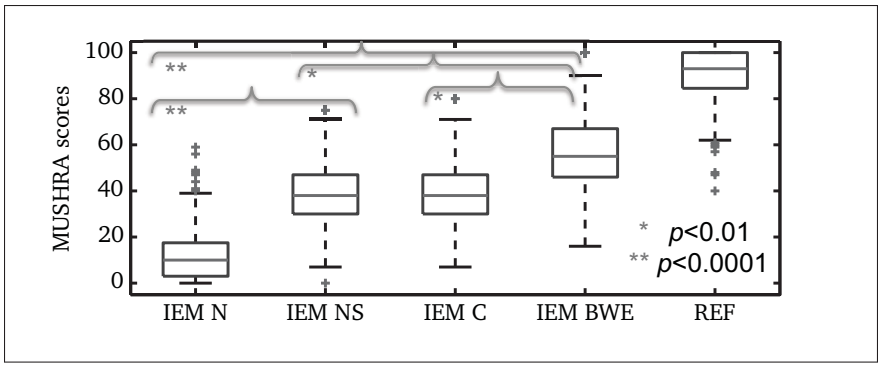

Figure 2.18 Box and whisker plot comparing the MUSHRA results of the noisy IEM signal (IEM N),the denoised IEM signal (IEM NS), the clean IEM signal (IEM C), the bandwidth extended denoised IEM signal (IEM BWE) and the hidden reference.

The proposed approach is speaker independent, computationally simple, robust to noise and requires no speech training by the user. Utilizing the review of conventional BC speech done by (Shin *et al.*, 2012) a comparison with the proposed solution is shown in Table 2.5.

Table 2.3 The average MUSHRA scores for the noisy IEM signal (N), the denoised IEM (NS), the clean IEM signal (C) the bandwidth extended IEM (BWE) and the hidden reference (REF).

Signal	Mean MUSHRA
IEM N	10
IEM NS	38
IEM C	38
IEM BWE	55
REF	93

Table 2.4 Statistical significance results based on a 95% confidence interval between the subjective evaluation of different stages of enhancement.

Signals	p-value	Significant?
N vs. NS	$p < 0.0001$	Yes
N vs. BWE	$p < 0.0001$	Yes
NS vs. BWE	$p < 0.0001$	Yes
C vs. BWE	$p < 0.0001$	Yes
C vs. NS	$p = 0.9782$	No

A possible limitation of this work is that it is done with speech data from only one female speaker. However, in close observation the only criteria that is potentially speaker dependent is the choice of triggering threshold, T_g, for voice activity detection during the adaptation process.

Table 2.5 Comparison of conventional BC enhancement approaches to the proposed approach.

Approach	**Requires Training?**	**Complex?**
Equalization (Tamiya and Shimamura, 2004; Kondo *et al.*, 2006)	Yes	No
Analysis-and-Synthesis (Tat Vu *et al.*, 2008, 2006)	No	Yes
Probabilistic (Liu *et al.*, 2004)	No	Yes
Proposed	No	No

The proposed concept of adaptive filtering for noise reduction has been shown to work in the past and is not speaker dependent (Davis, 2002; Martinek and Zidek, 2010). Once denoising is achieved, the bandwidth extension process has also been well studied and proven to work regardless of the speaker (Iser and Schmidt, 2008). Therefore, the only speaker dependent factor that may arise is the value of the speech detection triggering criteria used in the adaptation process. Since the choice of T_g was made based on results from 1 speaker but validated on 4 speakers, it is assumed that this threshold can be extended for use with N number of speakers and should not greatly affect the enhancement process. In this work the main priority was to evaluate and enhance the quality of the IEM speech. In future work, it is relevant as well to measure and evaluate the intelligibility of the IEM speech and how the proposed enhancement process affects it.

2.5 Conclusions

Using bone and tissue conducted speech in noisy environments is a reliable way of providing a high SNR speech signal to the listener. The downfall usually lies in the limited bandwidth of the bone and tissue conducted speech. This paper focuses on the enhancement of speech generated from bone and tissue conduction picked up using a communication device equipped with an in-ear microphone and an outer-ear microphone. An adaptive filtering approach is used to denoise the in-ear microphone signal using the outer-ear microphone. A novel voice activity detection criteria using the filter coefficients of the adaptive filter is used to ensure that only noise is reduced while the speech content remains unaffected. Once the signal is denoised the bandwidth of the signal is extended by exploiting the nonlinear characteristics of a cubic operator. Both objective and subjective evaluations show that the bandwidth extension of the denoised in-ear microphone signal significantly enhances its quality. For factory noise, the techniques shown in this paper provide a simple, speaker independent, non computationally exhaustive method to enhance the quality of speech picked up using an in-ear microphone. Overall, gains of 1.23 (out of 4.5) in POLQA MOS-LQO scores and 45 (out of 100) in MUSHRA scores show the benefits of the proposed speech enhancement solution.

Acknowledgment

This work was made possible via funding from the Centre for Interdisciplinary Research in Music Media and Technology, the Natural Sciences and Engineering Research Council of Canada, and the Sonomax-ETS Industrial Research Chair in In-Ear Technologies. The authors would also like to acknowledge the help of João Felipe Santos for his help with the online subjective evaluation.

CHAPTER 3

VARIATIONS IN VOICE LEVEL AND FUNDAMENTAL FREQUENCY WITH CHANGING BACKGROUND NOISE LEVEL AND TALKER-TO-LISTENER DISTANCE WHILE WEARING HEARING PROTECTORS: A PILOT STUDY

Rachel E. Bouserhal[1,3], Ewen N. MacDonald[4], Tiago H. Falk[2,3], Jérémie Voix[1,3]

[1]École de Technologie Supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Centre EMT, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

[4]Technical University of Denmark, Lyngby, Denmark

Article published in the International Journal of Audiology

Abstract

OBJECTIVE: Speech production in noise with varying talker-to-listener distance has been well studied for the open ear condition. However, occluding the ear canal can affect the auditory feedback and cause deviations from the models presented for the open-ear condition. Communication is a main concern for people wearing Hearing Protection Devices (HPD). Although practical, radio communication is cumbersome, as it does not distinguish designated receivers. A smarter radio communication protocol must be developed to alleviate this problem. Thus, it is necessary to model speech production in noise while wearing HPDs. Such a model opens the door to radio communication systems that distinguish receivers and offer more efficient communication between persons wearing HPDs. DESIGN: This paper presents the results of a pilot study aimed to investigate the effects of occluding the ear on changes in voice level and fundamental frequency in noise and with varying talker-to-listener distance. STUDY SAMPLE: Twelve participants with a mean age of 28 participated in this study. RESULTS: Compared to existing data, results show a trend similar to the open ear condition with the exception of the occluded quiet condition. CONCLUSIONS: This implies that a model can be developed to better understand speech production for the occluded ear.

3.1 Introduction

Finding the balance between good hearing protection and communication in noisy environments has been a difficult task. It is no question that workers in noisy environments must be protected to avoid noise induced hearing loss (Berger, 2003). However, communication remains a major concern for those equipped with Hearing Protection Devices (HPD) (NIOSH, 2005). Understanding the changes in speech production by talkers with occluded ears in noise can provide the groundwork for better communication in noisy environments.

Using radio communication in noisy environments is a practical and affordable solution allowing communication between people with HPDs. Traditionally, one of its weaknesses lies in the lack of designating receivers: all those carrying a personal radio (e.g. walkie-talkie) are subjected to the broadcast signal regardless of whether or not they are the intended listeners. Receiving irrelevant communication is annoying and contributes to the daily accumulated noise dose (Mazur and Voix, 2013). A new concept of a "Radio-Acoustical Virtual Environment" (RAVE) is being developed (Bou Serhal *et al.*, 2013). RAVE intends to mimic a natural acoustical environment by transmitting a radio communication signal only to people within a specific spatial range. This range is defined as the intended communication distance of the talker.

To predict the talker's intended communication distance, speech production in the presence of noise while wearing HPDs must first be understood. Talkers with normal hearing adjust their vocal effort in the presence of noise (Junqua *et al.*, 1999), when trying to communicate at a distance (Fux *et al.*, 2011) and to express emotion (Schröder, 2001). These adjustments still occur when wearing HPDs, however, they are altered as a function of the effects of the HPD on the wearer's perception of his/her own voice (Tufts and Frank, 2003; Casali *et al.*, 1987). The type of HPD influences the residual noise level inside the ear and the level of occlusion, which affects the perception of the wearer's own voice.

For the open ear condition, variations in the vocal effort have been well studied in the presence of noise and as a function of communication distance. Lombard speech refers to the signif-

icant changes in speech production when speech is produced in noise (Junqua *et al.*, 1999; Zollinger and Brumm, 2011). Some of these changes include an increase in vocal level of 1-6 dB for every 10 dB of noise increase (Lane and Tranel, 1971). Shifts in fundamental frequency, F0, as well as first formant, F1, have also been observed. Studies show an increase in the fundamental frequency (Junqua, 1993; Garnier and Henrich, 2014) of anywhere between 0.6-2.5 semitones (Lu and Cooke, 2008). Summers *et al.* (1988) report a decrease in spectral tilt, while more recent studies report a shift in the spectral center of gravity (Tufts and Frank, 2003; Garnier and Henrich, 2014). Both of these findings indicate an increase in the high frequency content, which can improve speech intelligibility in noise.

In quiet conditions, in turn, talkers raise their vocal effort to reach farther distances. A doubling in the talker-to-listener distance increases the vocal level between 1.3-6 dB (Traunmüller and Eriksson, 2000; Zahorik and Kelly, 2007; Pelegrín-García *et al.*, 2011). A study done by Zahorik and Kelly (2007) showed that talkers adjust their vocal effort according to their acoustical environment as well as the communication distance. The talkers' F0 as well as first formant, F1, also increase as a function of distance. As the vocal level increases, F0 increases by 5 Hz/dB while F1 increases by 3.5 Hz/dB (Liénard and Di Benedetto, 1999). The change in F0 caused by an increase in the communication distance, and thus vocal level, was determined to be unique and distinguishable from changes that occurred in Lombard speech or other factors that may raise the vocal effort (Fux *et al.*, 2011). It is clear from previous studies that adjustments in the vocal effort as a consequence of either increase in communication distance or the presence of noise varies from talker to talker, but follows the same trend across talkers. Vocal level and changes in the talker's F0 are good indicators of increased vocal effort as a consequence of either larger communication distance or presence of background noise. It is also relevant to consider the importance and effects of auditory feedback received by a talker on speech production (Hansen and Varadarajan, 2009). On one hand, when auditory feedback is lost, talkers produce "disorganized" speech in noise. On the other hand, with maskers that did not affect the auditory feedback, speech intelligibility increased (Dreher and O'Neill, 1957;

Ladefoged, 1972). Therefore, one's perception of one's own voice can have significant effects on changes in speech production.

Auditory feedback is received through two paths: air conduction and bone-conduction (Pörschmann, 2000). Occluding the ear canal with an HPD creates a resonance of the bone conducted vibrations originating from speech, causing talkers to hear an amplified 'boomy' version of their voice as they speak. This phenomenon is called the "occlusion effect" (Bernier and Voix, 2013). The occlusion effect changes the balance between the air-conduction and the bone-conduction paths, thus causing a change in speech production. A talker's perception of his/her own voice level compared to the level of noise is the driving factor in the speech production process (Tufts and Frank, 2003). Studies have shown that talkers wearing HPDs do not react to an increase in noise levels as much as talkers not wearing HPDs. Tufts and Frank (2003) report that talkers wearing earplugs in noise decreased their speech levels by 4-11 dB compared to their speech levels in noise without HPDs. Also, overall speech levels increased by only 5 dB (from 66.6 dB (SPL) to 71.9 dB (SPL)) when wearing foam HPDs, even when the noise was increased by 40 dB Tufts and Frank (2003). In other words, while wearing HPDs, talkers adjust their vocal effort by only 1.25 dB for every 10 dB increase in noise. In quiet, however, talkers wearing earplugs did not significantly alter their overall speech levels Tufts and Frank (2003); Navarro (1996) from their open-ear level with a slight decrease of 0.6 dB. These results contradict older studies (Casali *et al.*, 1987; Kryter, 1946) reporting that talkers increase their speech levels by 4 dB while occluded in quiet. Tufts and Frank (2003) attribute this contradiction to the placement of the plug in the ear and its contribution to the occlusion effect, emphasizing again the role of perception of one's own voice on speech production.

Although the effects of occluding the ear on speech production in noise have been studied, to the authors' knowledge no studies examine the effects of both background noise and changes in talker-to-listener distance. Based on the literature available, however, predictions can be made on the effects of occluding the ear while in noise for varied distances. The research hypothesis is that production changes with different distances for occluded ears should resemble those for open ears but should be smaller in magnitude.

This paper presents the results of a pilot study aimed to validate this hypothesis about the effects of occluding the ear on variations in level and F0 as a function of varied background noise and talker-to-listener distance. Conversational speech was recorded from users wearing HPDs in varying noise levels and talker-to-listener distances. Preliminary results show that variations in the vocal effort when occluded in noise follow a similar trend as the un-occluded condition. However, interesting changes are observed for the occluded quiet condition.

3.2 Method

Speech was recorded from 12 different talkers at 5 different distances in 3 different noise conditions and 2 quiet conditions.

3.2.1 Apparatus

Each participant was equipped binaurally with the intra-aural communication earpiece shown in Fig. 3.1. This communication earpiece was chosen for several reasons:

a. It is intra-aural, so it can be fitted into a participant's ear using different tips (roll-down foam plug, rounded flanged tips, malleable silicon wax, custom molded earpiece) causing different levels of the occlusion effect. In this way different tips could be used to better understand how the level of occlusion can affect speech production in noise. For the purposes of this study, foam tips (Comply^{TM} Tx 200) were used to provide the best acoustical seal without a custom fit.

b. It contains a microphone and miniature loudspeaker (internal receiver) inside the ear, as well as a microphone outside the ear. This allows direct assessment of how well the earpiece is worn and is further explained below.

c. It is the earpiece used for the RAVE application described in Section 3.1.

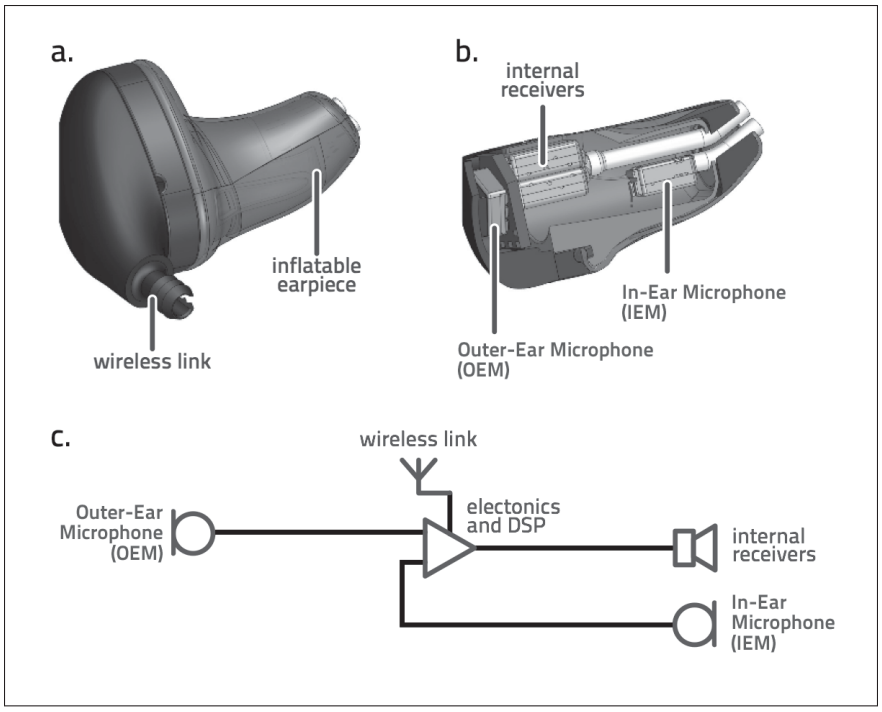

Figure 3.1 Auditory research platform (a), its electroacoustic components (b), and equivalent schematic (c).

An omnidirectional studio microphone (Sennheiser® MD 211 N) was placed 0.3 m in front of each talker's mouth. The choice of a microphone placed in front of the mouth instead of a headset microphone that is fixed in front of the mouth was made to ensure that the fit of the earpiece is not altered. Although an omnidirectional microphone placed at 0.3 m from the mouth would capture some of the room acoustics, it would minimize the proximity effect and capture a more genuine speech signal. Speech was recorded using the in-ear microphones, the outer-ear microphones, and the microphone placed in front of the mouth. Recordings were made at a sampling frequency of 48 kHz using a Fireface® UCX soundcard and a Windows™ computer running MATLAB™ (MathWorks, Natick, Massachusetts, United States). Two computer loudspeakers were used to send white noise at 85 dB (SPL) to assess the acoustic seal and attenuation achieved by the earplug as well as to give talkers timing cues. The experimental set-up portraying the apparatus, the room and the earpiece is presented in Fig. 3.2.

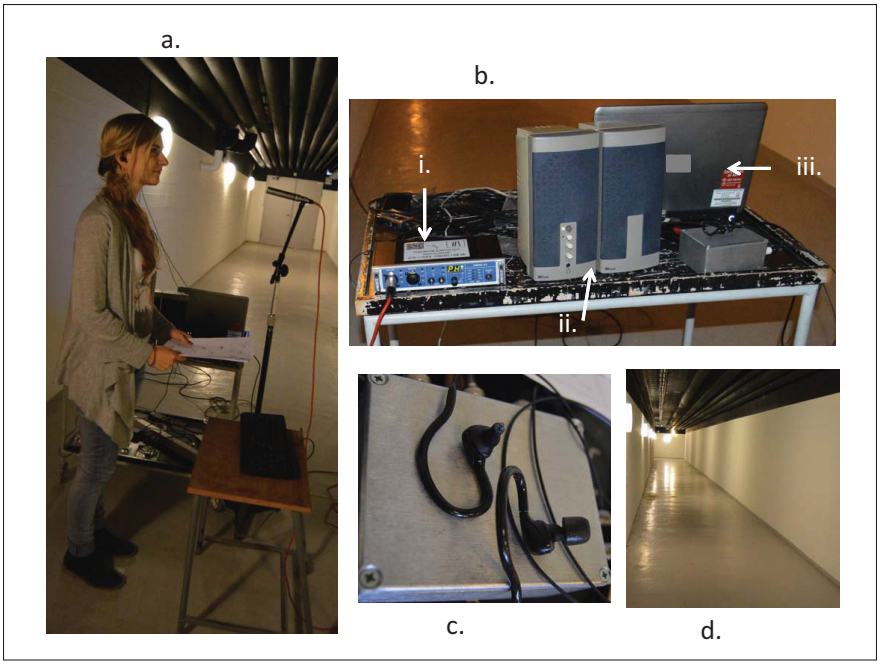

Figure 3.2 An example of the experimental setup with a participant (a), a close-up of the apparatus (b), includingFireface® UCX soundcard (i), the computer loudspeakers (ii), and the Windows™ computer running MATLAB™ (iii). The earpiece with and without the Comply™ tips (c), and the hallway where the experiments were held (d).

3.2.2 Participants

For this pilot study, 12 graduate students (10 males, 2 females) were asked to participate as talkers in the study. They ranged in age from 23 to 34 with a mean age of 28. No formal audiogram was performed to measure their hearing; however, none of the participants reported any known hearing loss. One of the authors participated as the listener for all the experiments. All but one participant were involved in hearing research at the time of the study and most had basic knowledge of the Lombard effect.

3.2.3 Task

Each talker was given a set of geographical maps that contained landmarks, including a path that marks a start and a finish. The listener was provided the same maps with no path, but corresponding landmarks. The use of these maps has been used in the past to establish conversational speech (Pelegrín-García *et al.*, 2011; Anderson *et al.*, 1991). Talkers were instructed to direct the listener from start to finish in one minute. Talkers were encouraged to use eye contact with the listener and keep a conversational flow, avoiding long pauses and maintaining continuous speech. They were asked to notice the position of the listener and speak in a manner that would be intelligible. The listener was instructed to give no auditory or visual clues of intelligibility.

3.2.4 Conditions

Talkers repeated the task in 25 different conditions as shown in Table 3.1 an un-occluded, quiet condition where talker-to-listener distance varied from 1 m to 30 m (1, 5, 10, 20, and 30), four occluded conditions, in quiet and in 3 different simulated levels of noise (70, 80, 90 dB (SPL)). Factory noise from the NOISEX-92 database (Varga and Steeneken, 1993) was only played inside the ear through the internal receiver depicted in Fig. 3.1, leaving the outer-ear microphones, as well as the microphone placed in front of the mouth, free of noise. Noise played inside the ear depended on the transfer function of each participant's earpiece. The measurement of the individual earpiece is further explained in section 3.2.5. The experiments were conducted in a 45 m long corridor, in a basement connecting two buildings. Talkers were positioned on a reflective surface (concrete) at a distance of 2 m from the closest wall. Even though it is believed that the room gain is considerably low, since all the experiments were conducted in the same room, with talkers at the same position, any effects from the room gain affected all conditions similarly. Furthermore, since most of the conditions were occluded and the auditory feedback restricted to bone conduction, the effect of the room acoustics was assumed to be of minor concern. Any effects on speech production caused by reverberation

may be mainly attributed to the characteristics of the simulated residual noise and the visual feedback of the hallway.

Table 3.1 Experimental conditions with changing talker-to-listener distance for the quiet un-occluded ear and occluded ear in noise and in quiet.

Ear Condition	Noise (dB(SPL))	Distances (m)
Un-occluded	Quiet	1, 5, 10, 20, 30
Occluded	Quiet	1, 5, 10, 20, 30
Occluded	70, 80, 90	1, 5, 10, 20, 30

3.2.5 Procedure

After instructions on the nature of the experiment were given, each talker was equipped with the communication headset from Fig. 3.1. After the earpiece was inserted, three main steps, explained below, were taken to ensure a well-fitted earplug and proper residual noise under the HPD. Prior to recording, the microphone in front of the mouth was adjusted and measured to be 0.3 m away from the center of the mouth. The talker was instructed not to touch the earpiece and to remain in one position at one end of the room. The listener gradually changed distances from closest (1 m) to farthest (30 m) at each condition. The choice of increasing the distance consecutively was made to mimic the experimental conditions of Pelegrín-García *et al.* (2011) and to be able to later compare their open-ear speech production model to the occluded-ear model. Once the occluded conditions were completed, the talker was then asked to remove the earpiece and perform the quiet open-ear task.

3.2.5.1 Measurement of individual earplug transfer function

To ensure a good acoustical seal, the transfer function of the earpiece was measured for each talker. This was done by playing white noise over the loudspeakers while the talker's head

is placed in front of them for two seconds and recording simultaneously using the in-ear and outer-ear microphones and calculating the transfer function using MATLAB™.

3.2.5.2 Assessment of well-fitted earplug

A good acoustical seal was defined as a transfer function with no amplifications in the low frequencies; an example is given in Fig. 3.3. A leakage in a closed volume behaves like a vent in an acoustical volume, acting like a Helmholtz resonator, and would show up as an amplification of the low frequencies in the transfer function (Voix and Laville, 2004). The fit was adjusted by giving more detailed direction on proper earplug insertion, followed by asking the participant to reinsert the earpiece until a good acoustical seal was reached.

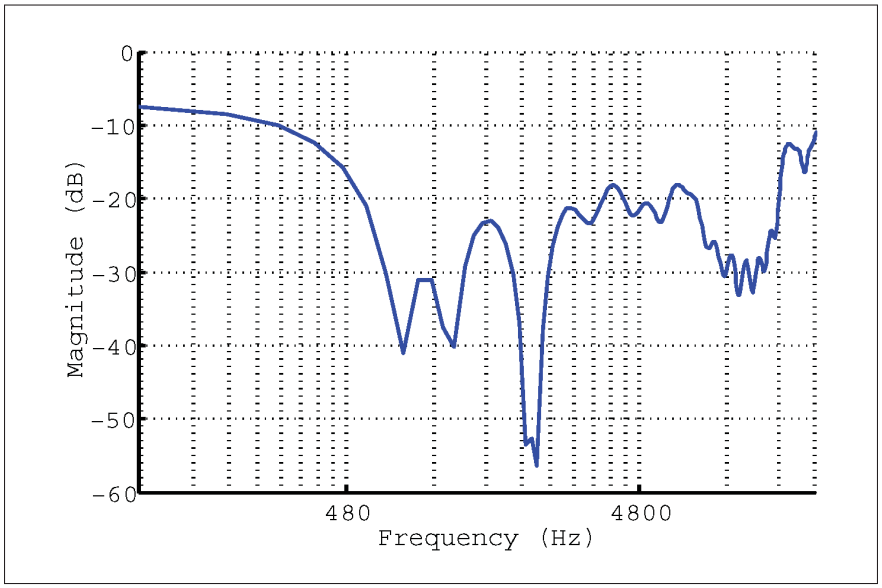

Figure 3.3 An example of a transfer function of a well-fitted earplug.

3.2.5.3 Adjustment of the background noise level

If the fit was acceptable, the transfer function of the fit for each ear was recorded and stored. The transfer function of each of the participant's ears was used to calculate the residual noise

in each ear for each noise condition. Prior to the experiments, 70, 80 and 90 dB (SPL) noise was recorded in an audiometric booth using the outer-ear microphones equipped by one of the authors. For each talker and for each ear, the three levels of noise were then passed through the individual's earplug transfer function before they were played directly inside the respective ear canal using the internal receivers, simulating the residual noise. A randomly selected example of the simulated noise in each ear showing the spectral and temporal differences from the OEM noise and the binaural differences as a function of each earplug's transfer function is shown in Fig. 3.4.

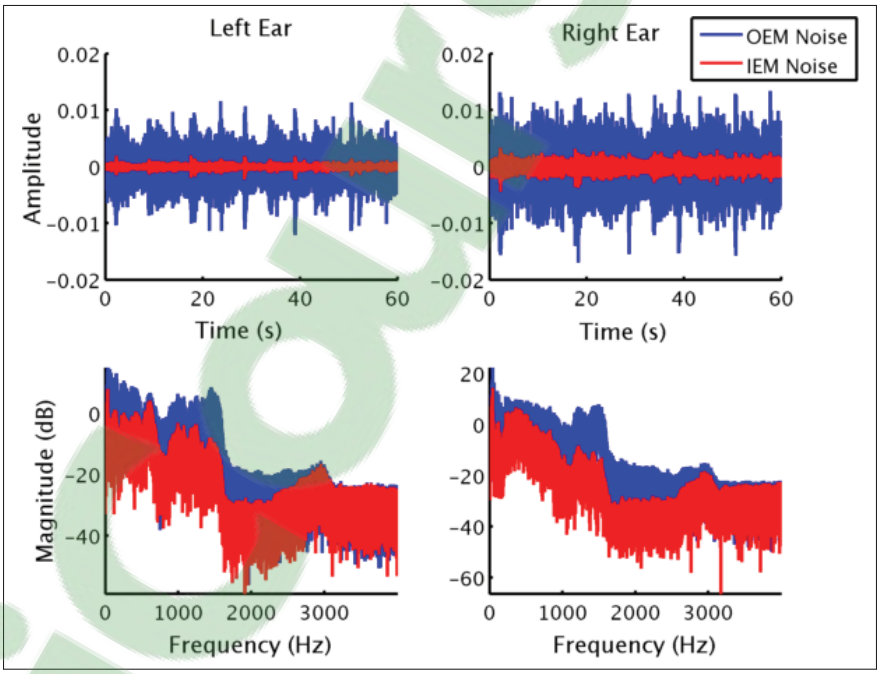

Figure 3.4 The spectral and temporal differences between the simulated residual noise inside the ear (IEM noise) and the noise as it would have been outside the ear (OEM noise).

3.2.5.4 Analysis

All speech recordings were run through an A-weighted filter. A-weighted SPL value was chosen over an overall SPL value, to better match the analysis bandwidth to the speech communication bandwidth and to apply less weight to any extraneous low-frequency parasitic noise that

could have been picked up by the microphones, given the ambient background noise (HVAC). While some of the energy present in the voice at F0 frequencies, presented in Table 3.3, may be affected by the roll-off of the A-weighting filter and while this effect may be slightly changed as F0s shift towards higher frequencies, equivalent patterns were also observed with overall unweighted sound pressure levels. For these reasons, and for convenience, the RMS value of the A-weighted signal was then calculated throughout the whole analysis. The fundamental frequency, F0, was extracted using the speech processing toolbox, (Brookes *et al.*, 1997).

3.3 Results

Excluding the quiet occluded condition, the changes in vocal levels and F0 were as expected: as the noise level and distance increased, so did the speech level and the fundamental frequency, F0. The average speech level and F0 for all conditions across speakers are presented in Table 3.2 and Table 3.3, respectively.

Table 3.2 The mean (μ) and standard deviation (σ) of absolute level values across speakers for all conditions and distances in dB(A).

Distance (m)		1	5	10	20	30
Conditions		Level (dB(A))				
Un-occluded (Quiet)	μ	55.93	58.04	58.86	60.45	61.85
	σ	3.86	3.57	3.34	3.18	2.91
Occluded (Quiet)	μ	63.06	65.65	66.99	68.86	70.36
	σ	3.69	3.53	3.62	3.13	3.19
Occluded (70 dB (SPL))	μ	65.79	67.33	68.17	69.41	70.23
	σ	3.38	2.71	2.77	2.66	2.81
Occluded (80 dB (SPL))	μ	66.20	67.99	69.35	70.45	71.86
	σ	3.39	2.48	2.62	2.41	2.40
Occluded (90 dB (SPL))	μ	68.43	70.11	71.27	72.36	73.69
	σ	3.40	2.43	3.03	2.62	2.66

Since the occluded quiet condition would be used as a baseline for applications such as RAVE, changes in the speech levels and F0 were thus normalized to the occluded quiet condition. The trend in vocal level changes, as well as the standard deviation across talkers, as the distance

Table 3.3 The mean (μ) and standard deviation (σ) absolute F0 values across speakers for all conditions and distances in Hz.

Distance (m)		1	5	10	20	30
Conditions		Level (dB(A))				
Un-occluded (Quiet)	μ	136.72	136.03	137.75	141.79	146.41
	σ	25.80	26.74	28.69	28.93	28.46
Occluded (Quiet)	μ	136.94	139.07	142.41	146.42	152.89
	σ	26.60	30.28	29.71	30.32	32.56
Occluded (70 dB (SPL))	μ	140.94	142.29	144.45	148.53	153.49
	σ	31.15	30.70	34.73	32.74	33.38
Occluded (80 dB (SPL))	μ	139.68	144.65	149.28	153.78	158.60
	σ	28.77	31.79	33.92	35.06	34.00
Occluded (90 dB (SPL))	μ	146.61	149.95	156.41	162.72	169.34
	σ	31.08	33.47	32.47	36.04	34.69

and noise increase are presented in Fig. 3.5 . Contradictory to the findings of Tufts and Frank (2003); Navarro (1996) a decrease of 6 dB in level is observed in the un-occluded quiet condition compared to the occluded quiet condition. It is also interesting to note that as the distance increased, particularly from 20 m to 30 m, talkers adjusted more in the occluded quiet condition than in the 70 dB noise condition. On average, for every 10 dB (SPL) increase in noise, the voice level increased by 1.8 dBA. A maximum standard deviation of 3.4 dBA is observed for the 90 dB (SPL) noise condition at the farthest distance of 30 m. Table 3.4 shows the overall change in the level from 1 m to 30 m for different noise conditions. Using Greenhouse-Geisser correction, a significant main effect was found for both noise condition ($F(2.38,44) = 223.127$, $\rho<0.001$) and distance ($F(1.28,44) = 87.902$, $\rho<0.001$) as well as a significant interaction ($F(16,176) = 4.519$,$\rho<0.001$). Multiple pairwise t-test comparisons with Bonferroni correction confirmed that all noise conditions and distances were significantly different from each other (all $\rho<0.03$). Two follow-up repeated measures ANOVAs were conducted. In the first, only the data from the two quiet conditions were examined. Again, the main effects of condition ($F(1,11) = 192.297$, $\rho<0.001$) and distance ($F(1.67,44) = 97.758$, $\rho<0.001$) as well as the interaction ($F(1.69,44) = 3.920$, $\rho<0.05$) were all significant. In the second, only the data from the three simulated noise conditions were examined. While the main effects of noise

condition (F(1.22,44) = 67.091, ρ<0.001) and distance (F(1.39,22) = 81.841, ρ<0.001) were significant, the interaction was not (F(8,88) = 1.245, ρ=0.28).

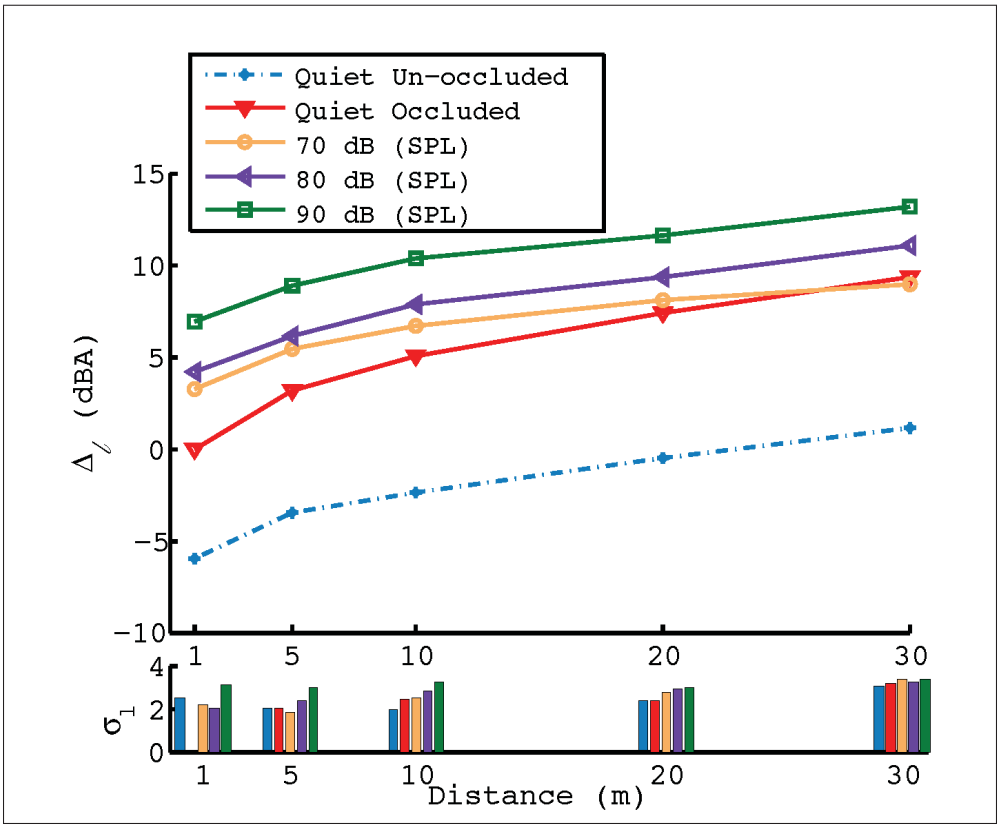

Figure 3.5 Average increase in speech levels, Δ_l from the occluded quiet condition over increasing distance and noise levels. The level at 1 m distance for the quiet occluded condition is used as reference for all the curves. The standard deviation, σ_l, in Δ_l across talkers over different noise conditions and distance.

Many studies have shown that, when un-occluded, a talker's F0 increases as the vocal level increases (Titze and Sundberg, 1992; Sundberg and Nordenberg, 2006; Garnier and Henrich, 2014). Fig. 3.6 shows the average changes in F0, as well as the standard deviation across talkers, as talker-to-listener distance increases for the varying noise conditions. As can be seen, F0 increases at the onset of noise and as the distance increases. However, at 1 m, changes in F0 between the quiet, 70 dB (SPL), and 80 dB (SPL) are relatively small; a maximum of 30.3 cents (note 100 cents = 1 semitone) is observed between the quiet occluded and the 70 dB noise

Table 3.4 Overall change in linear level from 1 m to 30 m for different noise conditions.

Condition	Overall Change (dBA)
Un-Occluded Quiet	7.5
Occluded Quiet	9.3
70 dB (SPL)	5.7
80 dB (SPL)	6.9
90 dB (SPL)	6.2

condition. At 90 dB of noise, on average, F0 increased by 79 cents at the 1 m position. Again, the largest variability is observed at the 90 dB (SPL) noise condition at the farthest distance of 30 m. Table 3.5 shows the overall change in F0 from the closest to the farthest position for each noise condition as well as the rate of change in F0 with the increase in level. A maximum overall change in F0 of 173 cents is observed for the 90 dB (SPL) condition. Using Greenhouse- Geisser correction, a significant main effect was found for both noise condition ($F(4,44) = 48.041$, $\rho<0.001$) and distance ($F(1.70,44) = 39.990$, $\rho<0.001$) as well as a significant interaction ($F(16,176) = 2.413$, $\rho=0.003$). Multiple pairwise t-test comparisons with Bonferroni correction confirmed that, with the exception of unoccluded quiet vs. occluded quiet and occluded quiet vs. 70 dB, all conditions were significantly different from each other (all $\rho<0.02$). Similarly, with the exception of 1 vs. 10 m and 1 vs. 5 m, all distances were significantly different from each other (all $\rho<0.03$).

Table 3.5 OOverall change in F0 as well as the overall rate of change of F0 per dB increase for each condition.

Condition	Overall Change (cents)	Rate (cents/dB)
Un-Occluded Quiet	81.1	10.8
Occluded Quiet	128.9	13.9
70 dB (SPL)	102.2	17.9
80 dB (SPL)	150.1	21.8
90 dB (SPL)	173.2	27.8

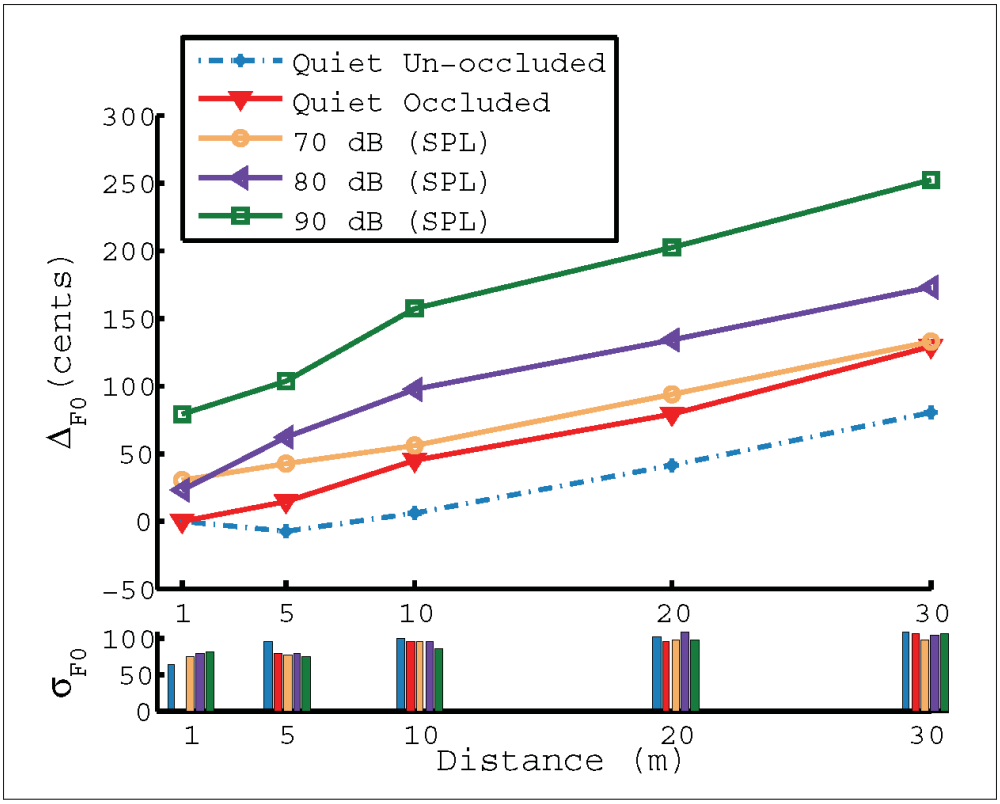

Figure 3.6 Average increase in F0 level, Δ_{F0}, from the occluded quiet condition over increasing distance and noise levels. F0 at 1 m distance for the quiet occluded condition is used as reference for all the curves. The standard deviation, σ_{F0}, in Δ_{F0} across talkers over different noise conditions and increasing distance.

3.4 Discussion

It is clear that the trend of increased vocal level and fundamental frequency as distance and noise increase is still present when wearing HPDs. Similarly to Tufts and Frank (2003) who found an average of 1.25 dBA increase in level for every 10 dB increase of noise when occluded, this study showed a 1.8 dBA increase. However, in contrast to Tufts and Frank (2003) and Navarro (1996), in this study, speech levels decreased in the un-occluded condition in quiet with a talker-to-listener distance of 1 m. One explanation for this is that the open ear condition was the last one to be performed and occurred immediately after the 90 dB (SPL) occluded condition. The drastic change between the two feedback conditions could have caused the talkers to dramatically decrease their vocal level from a natural level. Another explanation

could be the room acoustics of the hallway. Since the un-occluded condition was the only condition where the room acoustics could affect speech production the small changes in level as the distance increased and the overall decrease in levels could be attributed to the effects of the room on the talker's perception of his/her own voice (Pelegrín-García *et al.*, 2011). The elevated rate of change in the level as the distance increased for the occluded quiet condition is unclear. Once noise was introduced, the rate decreased, causing a crossover between the 70 dB (SPL) noise condition and the occluded quiet condition at 30 m. The variability across talkers is relatively large when considering the small variations that occurred as the noise and distance changed. However, other studies observed similar levels of variability across talkers (Garnier and Henrich, 2014; Lu and Cooke, 2008). An important thing to note about this study is that, since the noise introduced inside the ear canal under the HPD was based on each talker's personal earplug attenuation, the level of noise inside the ear was not the same across talkers. For example, a participant with a really good fit may have had less noise exposure than a participant with a fit that is not as good. It would be of interest to examine the relationship between the type of fit and the magnitude of differences produced by each talker. For the application of RAVE, it is crucial to look at the trends in the changes of level and F0 as recorded by the in-ear microphone, since in practice the in-ear microphone would be a more reliable source of information in high noise environments. So far, studies have only included normal hearing listeners. It would be of great relevance to conduct a similar study to include hearing-impaired talkers to create a speech production model better tailored to hearing-impaired users.

3.5 Conclusions

Overall, this pilot study demonstrated that tracking the differences in F0 and level for each talker could be used to determine a talker's intended communication distance. With access to level of background noise and the level of residual noise even the unexpected changes caused by occluding the ear in quiet can be accounted for. A study involving more participants and access to each participant's earplug transfer function could open up a door to a unified model of speech production in noise as a function of talker-to-listener distance and background noise

level for the occluded ear. Such a model could be used for applications such as RAVE to enhance the communication experience of occluded persons in noisy environments, thus promoting the use of hearing protection devices and reducing the risk of noise induced hearing loss.

Acknowledgment

This work was made possible via funding from the Centre for Interdisciplinary Research in Music Media and Technology, the Natural Sciences and Engineering Research Council of Canada, the Erasmus Mundus student exchange program in Auditory Cognitive Neuroscience and the Sonomax-ETS Industrial Research Chair in In-Ear Technologies.

CHAPTER 4

MODELING SPEECH LEVEL AS A FUNCTION OF BACKGROUND NOISE LEVEL AND TALKER-TO-LISTENER DISTANCE FOR TALKERS WEARING HEARING PROTECTION DEVICES

Rachel E. Bouserhal[1,3], Tiago H. Falk[2,3], Jérémie Voix[1,3]

[1]École de Technologie Supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Centre EMT, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Article submitted to JASA Express Letters

Abstract

Purpose: Studying the variations in speech levels with changing background noise level and talker-to-listener distance for talkers wearing hearing protection devices (HPDs) can aid in understanding communication in background noise. Methods: Speech was recorded using an intra-aural HPD from 12 different talkers at 5 different distances in 3 different noise conditions and 2 quiet conditions. Results: This paper proposes a model that can illustrate the difference in speech level as a function of background noise level and talker-to-listener distance. The proposed model complements the existing model presented by Pelegrín-García *et al.* (2011), and improves on it by taking into account the effects of occlusion and background noise level on changes in speech sound level. Conclusions: A model describing the relationship between speech level, talker-to-listener distance and background noise level for occluded talkers, can be incorporated with radio protocols to transmit verbal communication only to an intended set of listeners within a given spatial range, this range being dependent on the changes in speech level and background noise level.

4.1 Introduction

Talkers adjust their vocal effort with varying talker-to-listener distance (Fux *et al.*, 2011), in the presence of noise (Lane and Tranel, 1971) and to express emotion (Schröder, 2001). The increase in speech levels as a result of the onset of noise is known as the Lombard effect (Zollinger and Brumm, 2011). The Lombard effect as well as increasing speech levels with changing talker-to-listener distance are done both involuntarily and voluntarily by a talker to enhance speech intelligibility by the listener. Studies have shown that the Lombard effect is manifested differently when talkers are trying to communicate in noise compared to performing a reading task (Junqua *et al.*, 1999). Garnier *et al.* (2006) showed that changes in speech acoustics for Lombard speech are not purely physiological in nature but are rather a controlled enhancement of speech intelligibility.

Another implication of the Lombard effect is the presence of a feedback mechanism between vocal production and perception, working to adapt speech performance (Brumm and Zollinger, 2011). This feedback mechanism is referred to as the audio-phonation loop (Garnier *et al.*, 2010). As illustrated in Figure 4.1, there are three main components that affect the perception of one's own voice (Pörschmann, 2000; Lehnert and Giron, 1995):

(a) direct air conduction: sound travels from the talker's mouth to the ear through propagation in the open air.

(b) bone conduction: sound transmitted through bone and tissue conduction inside the skull. Direct stimulation of the cochlea can occur through vibrations of the skull vibrating the cochlear fluid or indirect stimulation can occur through the excitation of the air entrapped in the ear canal vibrating the eardrum resulting in a direct stimulation the cochlea.

(c) indirect air conduction: sound travels from the talker's mouth then reflects off of surfaces around the talker traveling back to the talker's ear.

For open ears, the air conduction pathways are the primary feedback paths for a talker (Henry and Letowski, 2007). Blauert *et al.* (1980) identified that direct transmission of sound from skull vibrations to

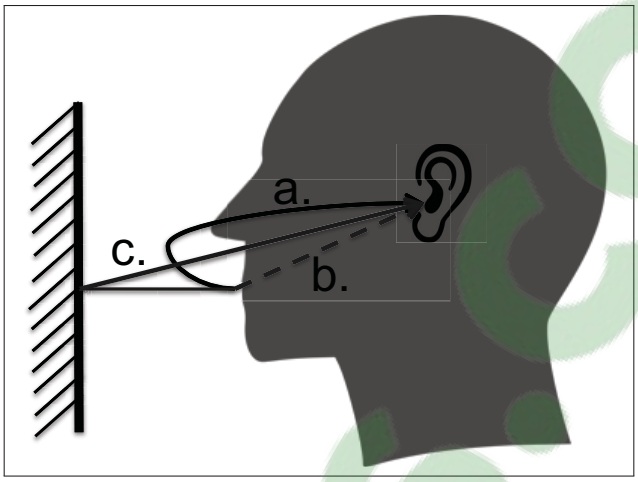

Figure 4.1 Illustration of the three paths affecting the perception of one's own voice: (a) direct air-conduction, (b) bone-conduction and (c) indirect air-conduction.

the cochlea are 40 dB and 70 dB less effective in the high frequencies and the low frequencies respectively, making sound transmission through air conduction superior to bone conduction. This also implies that in the open ear condition the bone conduction pathway may be neglected. When it comes to the perception of one's own voice, however, the significance of the contribution from each path is debatable. Békésy (1949), concluded that the air and bone conduction paths equally contribute to one's hearing of one's own voice. However, Pörschmann (2000), observed that except for the mid frequencies (700 to 1200 Hz) where bone conduction had a slightly superior contribution, air conduction was the dominant contributor to the hearing of one's own voice. It is important to note that there are large deviations between talkers in terms of the contribution of bone conduction to self perceived speech (Maurer and Landis, 1990). In summary, for different frequency ranges, both the bone and air conduction paths are significant contributors to the perception of one's own voice. They act as the main feedback paths that aid talkers in correcting their speech to become more intelligible. It is therefore reasonable to assume that wearing Hearing Protection Devices (HPDs) affects this feedback path and therefore causes deviations in speech production.

Occluding the ear canal with an HPD reduces the effect of two of the three feedback paths, the direct and indirect air conduction paths, but amplifies the bone conduction path. While

speaking the skull vibrates causing the soft tissue of the ear canal to vibrate as well. When the ear canal is open these vibrations are small and negligible. However, when the ear canal is blocked the energy from the soft tissue vibrations in the ear canal build up resulting in an amplification of the bone conduction sounds in the ear canal. This phenomenon is called the *occlusion effect*. The location at which the ear canal is blocked determines the strength of the occlusion effect. Blocking the ear canal right at its opening causes larger occlusion effect than when it is blocked closer to the eardrum. This can be explained by modeling the open and occluded ear canals as open and closed pipes of different lengths. The occlusion effect can also be modeled with electronic circuits where the open ear canal resembles a high pass filter and the occluded ear canal is the removal of this high pass filter (Brummund *et al.*, 2014). Since the ear canal is blocked and bone conduction is amplified, the bone-conduction path dominates the audio-phonation loop. Consequently, speech production is altered when wearing HPDs in quiet and in noise (Casali *et al.*, 1987; Tufts and Frank, 2003; Byrne, 2014). The changes in speech levels and fundamental frequency caused by occluding the ear in noise have been well studied. However, studies on the effect of occluding the ear canal on variations in speech levels caused by changing talker-to-listener distance are still limited.

However, for the open ear, changes in vocal effort as a function of the varying talker-to-listener distance have been well studied and modeled (Traunmüller and Eriksson, 2000; Zahorik and Kelly, 2007; Pelegrín-García *et al.*, 2011). A study done by Zahorik and Kelly (2007) showed that talkers adjust their vocal effort according to their acoustical environment as well as the communication distance, demonstrating again the effect of each path of the audio-phonation loop. Pelegrín-García *et al.* (2011) proposed a model of speech levels as a function of the talker-to-listener distance as well as the room acoustics.

In this paper, a model of the talker-to-listener distance as a function of the background noise level and the talker's speech levels for the occluded ear is presented. This model is a manipulation of the model presented by Pelegrín-García *et al.* (2011) and based on the results of a recent study by the authors (Bouserhal *et al.*, 2016b). The model is meant to better illustrate changes in speech levels for a talker wearing HPDs in noise with changing communication distance. It

could be integrated with HPDs equipped with radio capabilities to enhance the communication experience for users (Bou Serhal *et al.*, 2013). This could be done by instructing the radio to transmit verbal communication from the talker only to listeners within a specific spatial range. This range will be determined using the model, based on the talker's changes in speech levels and the level of background noise.

4.2 Methods and Materials

4.2.1 Experimental Setup

The model developed in this work is based on the data collected from a recent study by the authors. For details on the experimental setup and procedure, the reader is encouraged to refer to that work (Bouserhal *et al.*, 2016b). To summarize, the study involved 12 participants (10 males, 2 females) ranging in age from 23 to 34 with a mean age of 28. Participants were equipped with an intra-aural HPD containing outer-ear microphones, in-ear microphones, and miniature loud speakers in the ear, as depicted in Figure 4.2. They stood in a long corridor and were asked to lead a listener through a set of geographical maps from start to finish at varying levels of noise and varying talker-to-listener distances. Table 4.1 shows the 25 different experimental conditions performed. Speech was recorded using the outer-ear microphones placed on the outside surface of the HPD. Recordings were made at a sampling frequency of 48 kHz using a Fireface ® UCX soundcard and a Windows™ computer running MATLAB™ (MathWorks, Natick, Massachusetts, USA). Background noise levels were unweighted (SPL) values, however, all speech recordings were run through an A-weighting filter. A-weighted SPL value was chosen over an overall SPL value, to better match the analysis bandwidth to the speech communication bandwidth and to apply less weight to any extraneous low-frequency parasitic noise that could have been picked up by the microphones, given the ambient background noise. At the start of each experiment, the fit of the HPD was tested to ensure a good acoustic seal by undertaking the following procedure: a broadband white noise was played over external loudspeakers with the talker's head facing two external loudspeakers and recording simultane-

ously for two seconds using in-ear and outer-ear microphones on each ear, and calculating the attenuation of the earplug using MATLAB. The schematic of this procedure is illustrated in Figure 4.2. An example of the attenuation curve of a well-fitted earplug is shown in Figure 4.3.

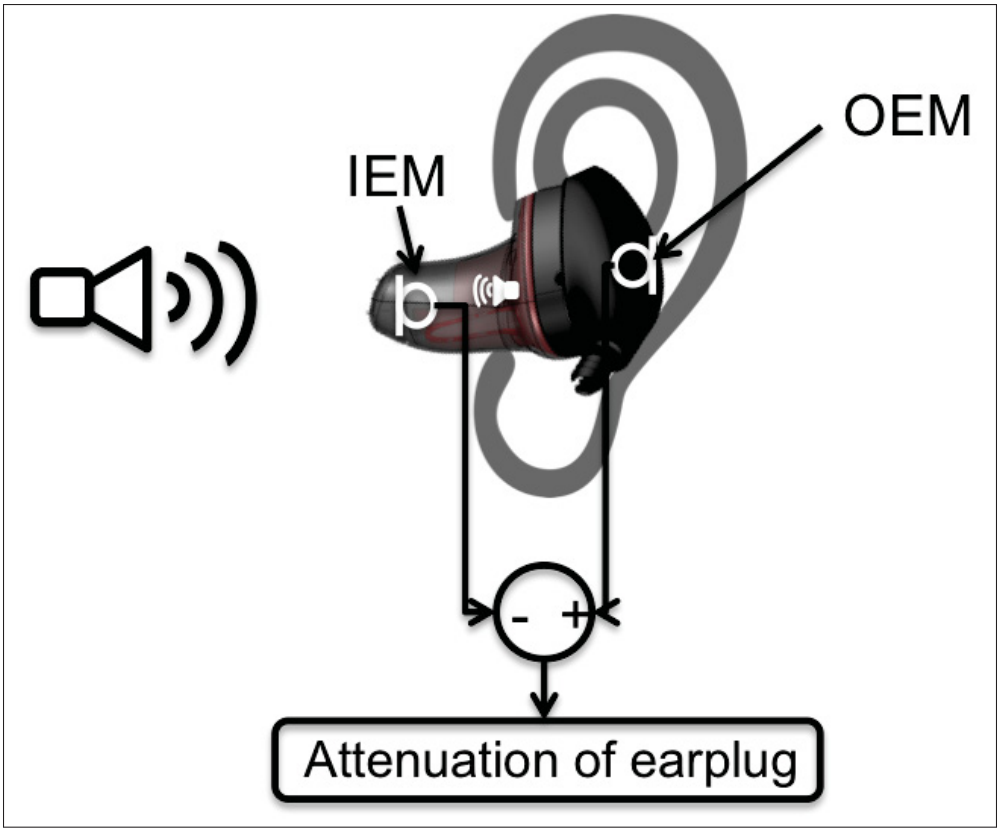

Figure 4.2 Schematic of procedure ensuring a good acoustical seal in the ear canal.

Table 4.1 Experimental conditions with changing talker-to-listener distance for the quiet un-occluded ear and occluded ear in noise and in quiet.

Ear Condition	Background Noise (dB(SPL))	Distances (m)
Un-occluded	Quiet (<50)	1, 5, 10, 20, and 30
Occluded	Quiet (<50), 70, 80, and 90	1, 5, 10, 20, and 30

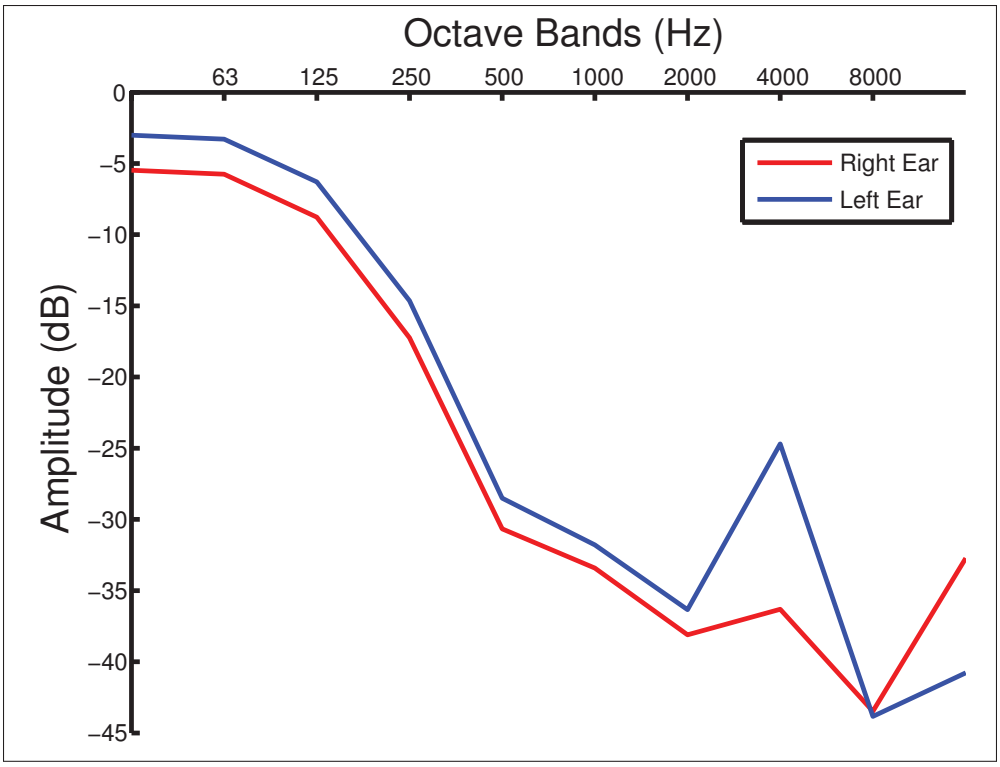

Figure 4.3 An example attenuation curve of a well-fitted earplug.

4.2.2 Model Fitting

To find the most appropriate model to fit the data, the model presented by Pelegrín-García *et al.* (2011) was used as a starting point. The proposed model by Pelegrín-García *et al.* (2011) was as follows:

$$L_w = a_k + \alpha_i + (b_k + \beta_i) \times \log_2(d/1.5) + \varepsilon_{ijk}, \tag{4.1}$$

where L_w is the speech power level, a_k and b_k are fixed factors, α_i, ε_{ijk}, and β_i are random effects and d is the talker-to-listener distance in meters. Pelegrín-García *et al.* (2011) use speech power levels, L_w, to represent the strength of speech sounds. However, in this work, unlike Pelegrín-García *et al.* (2011), this magnitude is represented using on-axis SPL which is a different yet valid way of representing the strength of speech sounds (Pelegrín-García *et al.*, 2011). For the purposes of this study, the effect of noise was added, and instead of speech power levels, the difference in speech sound level in dBA from the occluded quiet condition at

1 m was used. The equation therefore reduces to:

$$\Delta_L = a + c(N-60) + (b_k - \varepsilon) \times \log_2(d/1.5), \tag{4.2}$$

where Δ_L is the difference in speech sound level in dBA from the occluded quiet condition at 1 m, a is the change in speech sound level from the un-occluded to the occluded quiet condition, c is the slope at which Δ_L increases for every increase in noise from 60 dB(SPL), b_k is a fixed factor representing the slope as the talker-to-listener distance increases, d is the talker-to-listener distance in meters, and ε is any error presented from random factors. To optimize the value of the variables in this model, the curve fitting tool from MATLAB was used. Since the occlusion effect causes a type of amplified feedback of the talker's voice, it could be compared to speaking in a reverberant acoustical environment. The slope of increase due to talker-to-listener distance, b_k, could take one of four values based on the room acoustics: anechoic room, lecture hall, corridor and reverberant room. In a reverberant room the indirect feedback path is amplified and can thus be compared to being occluded where the bone conduction path is amplified. Also, results from Tufts and Frank (2003) and Bouserhal *et al.* (2016b), show that the rate of change in speech sound level when occluded is significanlty smaller than the open-ear condition. For both of these reasons the room dependent factor in a reverberant acoustical environment was chosen. The value $b_k = 1.3$ given by Pelegrín-García *et al.* (2011), is the slowest rate of increase due to talker-to-listener distance and would reflect an amplified indirect air-conduction path caused by the reverberation of the room, which is hypothesized to have a similar effect on variations in speech sound level as an amplified bone-conduction path caused by the occlusion effect (Bouserhal *et al.*, 2015b).

4.3 Results

The difference in speech sound level in dBA from the occluded quiet condition at 1 m, Δ_L, was averaged for all participants at each distance for each noise level and compared to its respective model. The optimal values for the variables a, c and ε were found using Equation 4.2.2 and the curve fitting tool *cftool* provided by MATLAB, for each noise level and averaged across the

three to find the values that can best represent all three noise conditions. The final values for the parameters a, c and ε are shown in Table 4.2. Figure 4.4 shows the deviation between the mean curves for each of the noise conditions and their respective model. With R-squared $= 0.965$, it can be seen that the models can well describe the relationship between speech level and talker-to-listener distance at different background noise levels. To show the large variability between talkers, the model curves at each of the three noise levels are plotted against the individual curves of each of the 12 participants. These comparisons for the 70 dB(SPL), 80 dB(SPL), and 90 dB(SPL) noise levels are shown in Figure 4.5. The mean (μ) and standard deviation (σ) in Δ_L for each noise level and distance is presented in Table 4.3. The standard deviations for each noise level are plotted in Figure 4.4. It can be seen that the standard deviation generally increases as the distance and level of noise increase. At 1 m the greatest standard deviation is at the 90 dB(SPL) conditions. However, the largest variability is observed at 70 dB(SPL) at the 30 m distance.

Table 4.2 Final parameter values optimized for all three noise conditions with the corresponding R^2 value.

Parameter	Value	Parameter	Value
a	-1.66	ε	-0.07
c	0.18	R^2	0.965

4.4 Discussions

The fixed factor $a = -1.695$ represents an initial change in speech level in quiet caused by wearing HPDs. This would imply that talkers increase speech level by about 1.7 dBA when wearing HPDs in quiet at a 1 m talker-to-listener distance. This is contradictory to Tufts and Frank (2003) and Navarro (1996) who found no significant increase in speech level when wearing HPDs in quiet at a 1 m distance. However, it is in accordance with older studies such as Casali *et al.* (1987) and Kryter (1946) that reported up to 4 dB increase in speech level after wearing HPDs. As Tufts and Frank (2003) have explained, this could be a consequence of different

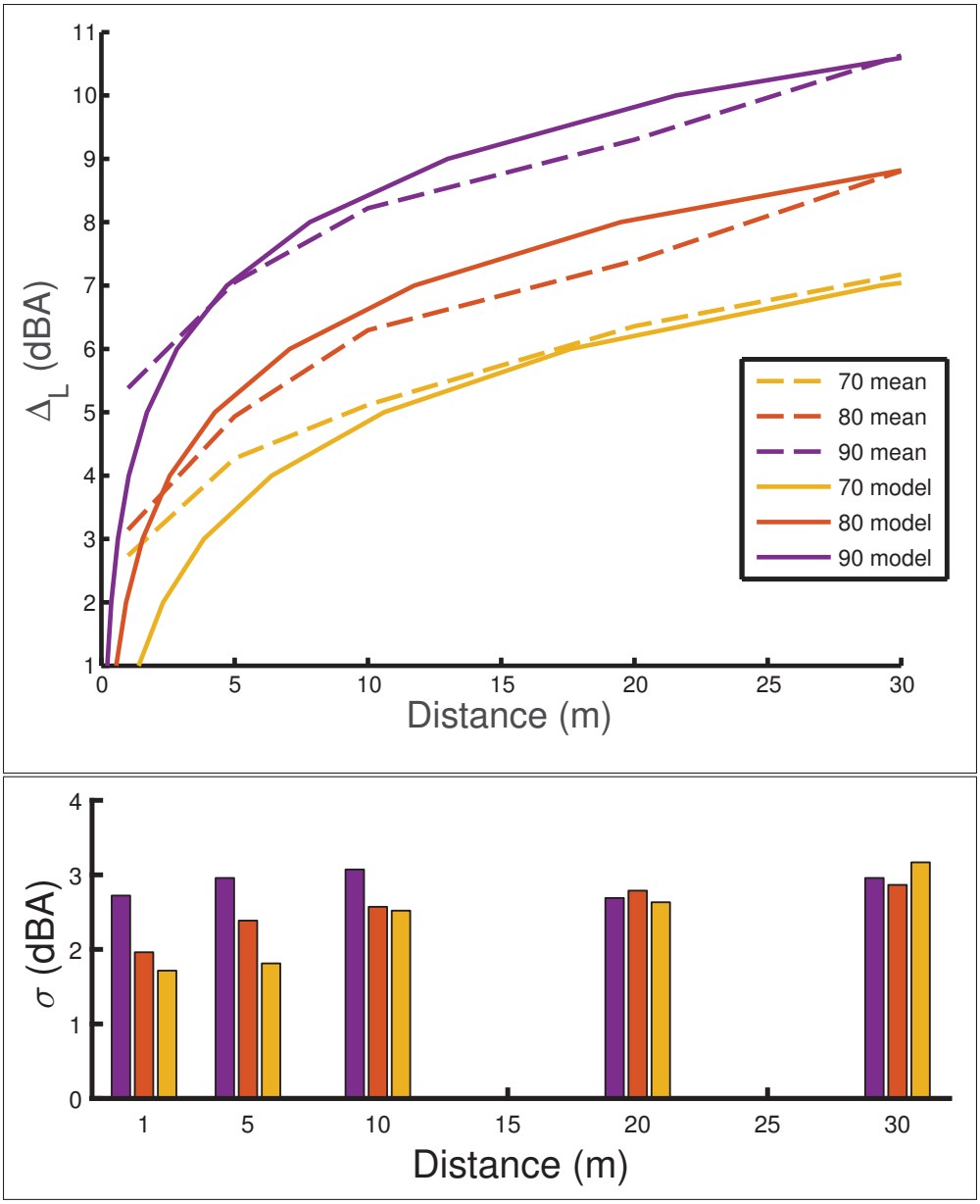

Figure 4.4 The mean curves for the 70, 80 and 90 dB(SPL) conditions compared to their respective model curves with $a = -1.659$ $c = 0.18$ and $\varepsilon = -0.0675$ (a), and the respective standard deviation at each noise level (b).

levels of occlusion which affect the perception of one's own voice: high levels of occlusion caused by shallow HPDs could cause an increase in speech levels compared to the open-ear condition. The parameter $c = 0.18$ is in accordance with the recent study showing that talkers wearing HPDs increase their speech level by 1.8 dB for every 10 dB increase in noise. The low

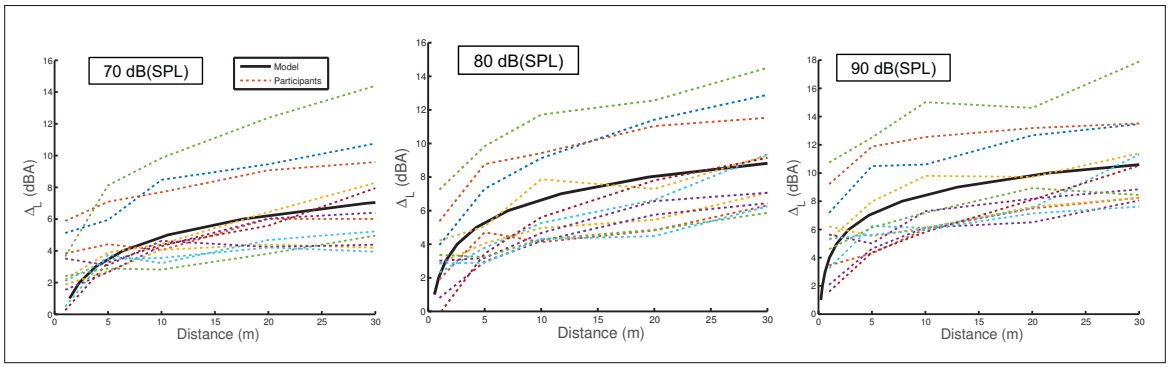

Figure 4.5 Model curves vs. individual participant curves.

Table 4.3 Mean (μ) and standard deviation (σ) in Δ_L for each noise level and distance.

Noise (dB(SPL))	Distance (m)	μ (dBA)	σ (dBA)
70	1	2.7	1.7
	5	4.3	1.8
	10	5.1	2.3
	20	6.4	2.6
	30	7.2	3.2
80	1	3.2	2.0
	5	4.9	2.4
	10	6.3	2.6
	20	7.4	2.8
	30	8.8	2.9
90	1	5.4	2.7
	5	7.1	3.0
	10	8.2	3.1
	20	8.3	2.7
	30	10.6	3.0

value observed for the error factor $\varepsilon = -0.07$, implies that a reverberant room environment best resembles an occluded condition as previously hypothesized.

The model can well describe the average trend between speech sound level, talker-to-listener distance and background noise level when the mean values are considered. However, it appears from Figure 4.4 and Table 4.3 that there is a large variability between talkers. From Figure 4.5,

it can be seen that three participants were consistently to the right of the models and could have been considered outliers. However, considering the small number of participants they constitute 30% of all the participants and cannot be discarded from the analysis. In addition, the combination of the large variability between speakers as well as the exponential nature of the model limit the predictive capabilities of this model. However, with a better understanding of the individual's trend correction factors may be added to better fit the model to the individual user and could be then used as a predictive tool.

4.5 Conclusions

There is a clear relationship between speech level, background noise level and talker-to-listener distance for persons wearing HPDs. Even with a large inter-talker variability the relationship is captured with the model. This model is an improvement on the model presented by Pelegrín-García *et al.* (2011) as it includes the effects of noise and occlusion on the variations in speech levels with changing talker-to-listener distance. Talker-dependent correction factors may be added to the model to be used to predict the intended communication distance of a talker given the background noise level and speech level. A predictive model as such can be integrated into HPDs equipped with radio communication to create a smarter transmission system of relevant verbal communication for talkers in noisy environments .

Acknowledgment

This work was made possible via funding from the Centre for Interdisciplinary Research in Music Media and Technology, the Fonds de recherche du Québec – Nature et technologies (2015-NC-181280), the Natural Sciences and Engineering Research Council of Canada (402126-2012, 402237-2011), the Erasmus Mundus student exchange program in Auditory Cognitive Neuroscience, and the EERS Industrial Research Chair in In-Ear Technologies.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

The problem of communication in noise while wearing HPDs is an ongoing issue. Concerns of proper levels of attenuation of the noise and innovative ways of providing verbal communication of good quality make this research project very relevant. In this doctoral work, an intra-aural HPD equipped with an IEM, a miniature loudspeaker, an OEM, wireless capabilities and signal processing abilities is used. Bandlimited IEM speech is picked up from inside the occluded ear, it is denoised using a novel adaptive filtering technique whose the adaptation is triggered on and off as a function of the ratio between filter coefficients. Once the IEM speech signal is denoised, it is enhanced using speaker independent, low complexity BWE techniques that utilize the nonlinear characteristics of cubing the excitation signal. Reaching an enhanced IEM signal fulfills two of the three objectives for this work identified in the Introduction. Lastly, the third objective is realized by coding the variations in speech levels alongside the background noise level to determine an intended talker to listener distance. Once this distance is determined the talker's enhanced speech signal is transmitted only to listeners within that spatial distance.

Thus far, this work provides the fundamentals needed to achieve a 'Radio Acoustical Virtual Environment' (RAVE) mimicking a natural acoustical environment. However, further work could optimize and enhance the performance of RAVE.

5.2 Future Work

5.2.1 Intelligibility of IEM speech

Throughout this work only the quality of the IEM speech was assessed and enhanced. Anecdotally the intelligibility of the IEM speech was considered to be relatively high and thus priority

was given to the enhancement of the quality of the IEM speech. However, it is important to formally study the intelligibility of the IEM speech signal and assess whether or not the proposed denoising and BWE technique improves upon the intelligibility. This can be assessed in a similar manner to the quality assessment discussed in Chapter 2. The intelligibility of the clean IEM speech can be compared to that of the OEM speech. The noisy OEM speech intelligibility can then be compared to the respective IEM speech containing the residual noise. Finally, the intelligibility of the denoised bandwidth extended speech can be compared to the two previous intelligibility scores. There are many ways to evaluate intelligibility both subjectively and objectively. However, it would be most useful to use subjective tests such as the one proposed by Ellaham *et al.* (2014) to analyze the IEM speech intelligibility since objective tests are not designed to evaluate BC speech.

5.3 Distance model as a predictive tool

As presented in Chapters 3 and 4, there is a large variability in speech levels and fundamental frequency between talkers speaking in noise while wearing HPDs. The exponential nature of the distance model results in small variations in the model causing large errors in predicted distance. There are several steps that could be taken to refine this model and use it as a predictive tool.

a. A larger sample size: the model was based on only 12 participants, thus a larger sample size can help refine the model and can be used to make further conclusions.

b. A formal monitoring of hearing health: no audiogram was performed on the participants. Participants were merely asked to report if they have any known hearing loss. Hearing impairment affects how noise is perceived and can thus affect speech production in noise altering the data.

c. Monitoring noise level under the HPD: currently, the model is based on the speech levels as a function of the ambient noise level. The residual noise inside the ear is realized by filtering the ambient noise using the transfer function of each participant's individual

earplug transfer function. This means that even though the ambient noise level is the same the residual noise inside the ear is different for each participant based on the fit of the participant's earplug. It would be more relevant, however, to base the model on the residual noise level under the earplug, as it is actually what contributes to the changes in vocal effort. This requires controlling the level of the residual noise inside the ear for each participant. Controlling for the level of the residual noise inside the ear is feasible with the IEM. It would also be of interest to investigate the importance of not only controlling the level of the residual noise but its spectrum as well.

d. A piecewise model: creating a piecewise model based on the data collected would be more appropriate as a predictive tool. Since the relationship between changes in talker-to-listener distance and speech level are exponential, such a model would greatly reduce errors in prediction of the distance. It would allow for users to be trained to its thresholds which would further reduce prediction errors. This is currently being realized, however, the exact values for the piecewise function are difficult to determine because of the large variability between talkers. Some preliminary validation tests have to be conducted to find the optimal values that would cause the least prediction errors.

5.4 Consideration of hearing-impaired listeners

As discussed in Section 5.3, the participants in the study were all assumed to have normal hearing. It is important, however, to consider the effects of hearing loss on speech production especially while wearing HPDs in noise. Since hearing loss varies greatly between individuals it is expected that a unified model could only be used among normal hearing talkers, while hearing-impaired listeners will have to train a speaker dependent model. Hearing-impaired listeners could be asked to speak while occluded to someone 1 m away, then 5 m away, then 10 m away at different levels of noise. The data could then be quickly fitted to a model that better describes the hearing-impaired talker's variation trends. This would allow RAVE to be used and validated with both hearing-impaired and normal hearing talkers.

5.5 Implementation and validation

So far only the enhancement of the IEM speech has been implemented and validated. No formal validation of the model nor of the overall RAVE system has been conducted. The last step to achieve a high performance RAVE would be to validate the work with normal hearing and hearing-impaired participants. Any optimization and tuning can be done at this stage. Once optimized and implemented, the RAVE algorithm integrated with the ARP will be benchmarked against other commercially available communication headsets.

5.6 Contributions

The work done in this doctoral project has scientific contributions, technical contributions and contributions relating to occupational safety and health as follows:

- **Scientific Contributions:** three journal articles and four conference proceedings in which a low complexity speaker independent way of denoising and enhancing IEM speech was introduced, new insight on how occluded talkers adjust their speech level to changes in talker-to-listener distance as well as background noise level was presented, and knowledge of the relationship between speech captured using an OEM, an IEM and a microphone placed in front of the mouth is detailed.
- **Technical Contributions:** a patent (Bouserhal *et al.*, 2016) on the denoising of signals captured with an IEM, algorithms of denoising and enhancement of the IEM speech to be used and implemented with the ARP, and the fundamentals for the realization of RAVE.
- **Occupational Safety and Health Contributions:** implementation of the aforementioned algorithms and utilization of the speech model for occluded talkers to enhance the verbal communication experience of talkers wearing HPDs in noisy environments which can promote the use of HPDs in highly noisy environments which may in turn reduce occupational NIHL.

APPENDIX I

INTEGRATION OF A DISTANCE SENSITIVE WIRELESS COMMUNICATION PROTOCOL TO HEARING PROTECTORS EQUIPPED WITH IN-EAR MICROPHONES

Rachel E. Bou Serhal[1,3], Tiago H. Falk[2,3], Jérémie Voix [1,3]

[1]École de technologie supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Article presented at the International Congress in Acoustics (ICA) in Montréal, Canada on June 2-7, 2013

90 **Proceedings of Meetings on Acoustics**

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Noise
Session 1pNSa: Advanced Hearing Protection and Methods of Measurement II

1pNSa5. Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones.

Rachel E. Bou Serhal, Tiago H. Falk and Jérémie Voix*

***Corresponding author's address: Ecole de Technologie Superieure, Universite du Quebec, 1100 rue Notre-Dame Ouest, Montréal, H3C 1K3, QC, Canada, jeremie.voix@etsmtl.ca**

Using radio communication in noisy environments is a practical and affordable solution allowing communication between workers wearing Hearing Protection Devices (HPD). However, typical radio communication systems have two main limitations when used in noisy environments: first, the background noise is disturbing the voice signal picked-up and transmitted, and second, that voice signal goes to all listeners on the same radio channel regardless of their physical proximity. A new concept of a so-called " Radio Acoustical Virtual Environment" (RAVE) addressing these two issues is presented. Using an intra-aural instantly custom molded HPD equipped with both an in-ear microphone and miniature loudspeaker, undisturbed speech is captured from inside the ear canal and transmitted over the wireless radio to the remote listener. The transmitted signal will only be received by listeners within a given spatial range, such range depending on the user's vocal effort and background noise level. This paper demonstrates the technological challenges to overcome and the methodology involved in the implementation of RAVE.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4800452]
Received 22 Jan 2013; published 2 Jun 2013
Proceedings of Meetings on Acoustics, Vol. 19, 040013 (2013)

Introduction

Hearing protection has been widely discussed and researched. Several Hearing Protection Devices (HPD) have been developed to protect workers' hearing from noisy environments. HPDs come in several different shapes and sizes and can be made from a variety of materials. The two main types of HPDs are intra-aural i.e. earplugs, and supra-aural i.e. earmuffs (Berger, 2003). Depending on the type of HPD worn, as well as, the spectrum of the noise and the wearer's hearing ability, wearing HPDs could limit communication (Berger, 2003). Good communication in a work environment is vital. Unfortunately, workers must make compromises between protecting their hearing and maintaining good communication. There are several different ways that are used to communicate in noise, one could:

a) Remove the HPD: get closer to a listener and adjust vocal effort to communicate

b) Use passively filtered HPD: flat attenuation HPDs could be beneficial for speech communication as they do not attenuate high frequencies as much as other HPDs.

c) Use a hand-held radio device: use of a walkie-talkie allows for distance communication with multiple people while remaining stationary (with HPDs or without).

d) Use of a communication headset: usually an earmuff with a miniature loudspeaker and an external boom microphone. The voice picked up by the boom microphone is transmitted through either a wired or wireless network to a remote listener.

Although these techniques are feasible and commonly utilized, their performance is unsatisfactory. Removing an HPD to communicate is counter-productive, potentially harmful to the worker's hearing and requires the workers to be in close proximity. Passively filtered HPDs do not require the user to remove the HPD for communication, but the speaker must still be in close range for the listeners to understand. As a result of the excessive levels of background noise, the persons communicating will naturally increase their vocal effort to compensate for such conditions in comparison to a quite environment. Using a hand-held radio overcomes the problem of proximity but still requires the removal of the HPD. The best current alternative is the use of HPDs that are equipped with an external microphone called a boom microphone and connected to a personal radio system. Although a step in the right direction, these headsets still present the following inconvenience: the external microphone will not only pick up the user's voice but background noise as well, which dramatically affects intelligibility.

Another issue associated with using any kind of radio transmitter, is that it does not distinguish a receiver and all communication is sent to everyone on the same radio channel. Thus, the users' radio is often flooded with irrelevant conversation that could be annoying and somewhat loud and thus contributing to the noise dose. Clearly there is a need for a device that provides good noise attenuation as well as good communication without compromising the performance of one or the other.

Proposed Approach

We propose a new concept called "Radio Acoustical Virtual Environment" (RAVE) in which workers in noisy environments can achieve intelligible communication without hindering their hearing protection. RAVE uses an advanced intra-aural instantly custom molded HPD, shown in Figure 1, equipped with an In-Ear Microphone (IEM), a miniature loudspeaker, a Digital Signal Processor (DSP), an Outer-Ear Microphone (OEM) and Wireless Radio (WR) capabilities. Such a device can capture a somewhat undisturbed speech signal from inside the ear (referred to as IEM speech). Because the signal captured originates from bone conducted vibrations, it lacks higher frequencies. Thus, the IEM signal must first be enhanced in its high frequency content. Once

enhanced, the IEM signal is coded and sent to an appropriate radius of listeners based on the acoustical features of the produced speech and the level of background noise.

This paper introduces the design of RAVE and the methodology involved in realizing such a protocol. The next section discusses different techniques available for the enhancement of the IEM speech signal followed by the concept of vocal effort coding. Then we discuss the envisioned experimental work required to obtain RAVE and the final section presents our conclusions.

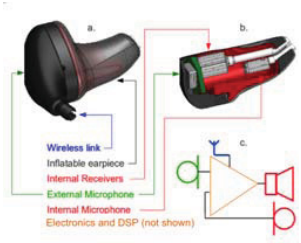

FIGURE 1: Overview of digital custom earpiece (a), its electroacoustical components (b), and equivalent schematic (c).

ENHANCEMENT OF THE IEM SPEECH

When speech is captured conventionally (with a boom microphone), to be sent over a radio network in a noisy environment, it is disturbed and contains the noise picked up by the exposed microphone, even when using a directional microphone. On the other hand, capturing speech from inside the protected ear allows for the transmission of a less-disturbed speech signal that will not require extra de-noising usually achieved by the electronics within the radio. When the ear canal is blocked by an in-ear device, there is a regeneration of the speech inside the ear canal and one experiences what is called the occlusion effect (Berger, 2003). The occlusion effect allows for the capturing of speech inside the ear, which is useful in noisy environments. Because of cranial bone conduction, this signal is "boomy", containing most of its energy in the lower frequencies while missing important high frequency content (Bernier and Voix, 2010). The difference between the frequency content of the IEM speech and the OEM speech (referred to as REF) of the utterance /u/, for a male speaker, is demonstrated in Figure 2. In Figure 2, we notice that above 1.8 kHz, the IEM signal is missing important high frequency content. As a consequence of the IEM signal's limited bandwidth, fricative consonants such as /s/ and /f/, and nasals such as /n/ and /m/ are unintelligible. The IEM signal is thus perceived as having lower quality and intelligibility than "free air speech", or speech that is recorded near the mouth. To solve this, the IEM signal could be expanded using Bandwidth Extension (BWE) of the speech signal as will be reviewed in the next section.

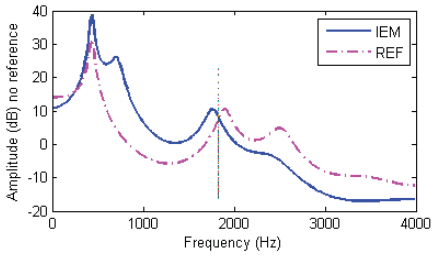

FIGURE 2: IEM vs. REF spectral envelopes of the utterance /u/ from the word 'canoe', showing the increased low frequency content and the missing high frequency content.

Bandwidth Extension (BWE)

In this section, we introduce some BWE techniques commonly utilized in the field of speech signal processing (a good reference of common speech terms can be found in (O'shaughnessy, 2000)). Many different BWE techniques exist, and the proper choice depends on the desired results and available resources. BWE can range from spectral estimation and expansion through excitation signal extension, to Vector Quantization (VQ) and codebook mapping. Iser et al. give a good review of the basics of such techniques (Iser, Bernd *et al.*, 2008). In the past, the need for BWE arose because of the limited bandwidth of the telephone network. The narrow bandwidth of a telephone is about 3.5 kHz leaving some significant parts of human speech unrepresented. In this context, wideband signals refer to signals that can represent the entire vocal range while narrowband signals can only represent a limited part of the vocal range. With access to an IEM and an OEM, BWE can be used for our purposes by treating the IEM signal as the narrowband signal and the free-air speech captured by the OEM as the wideband signal.

One BWE technique is *excitation signal extension*. This technique involves three main procedures: *envelope extraction*, *excitation signal extraction*, and *excitation signal extension* (Iser, Bernd *et al.*, 2008). The envelope extraction technique depends on the Linear Predictive Coding (LPC) analysis of the narrowband signal. The excitation signal extraction and extension could be done using several methods: non-linear characteristics approach, spectral shifting approach and the function generator approach. Another way BWE can be achieved, is by *wideband spectral envelope expansion*. To estimate the wideband spectral envelope, several methods are available, such as neural networks, linear mapping, and codebooks (Iser, Bernd *et al.*, 2008). Statistically based methods also exist, such as the statistical recovery function used by Cheng et al. (Cheng *et al.*, 1994), and Gaussian Mixture Models (GMM) (Park and Kim, 2000). Wideband spectral envelope estimation differs from excitation signal extension in that it requires a training data set. While only the narrowband input is required for excitation signal extension, the estimation of the wideband spectral envelope requires a sufficiently large training data set that contains the desired sampling rate and bandwidth (Iser, Bernd *et al.*, 2008).

With all these available techniques, listed in Figure 3, it is important to assess the resources available to choose a practical and efficient technique with good performance. Some things to consider are the computational complexity and cost of the algorithm, power consumption and whether the algorithm will be speaker dependent or speaker independent. *Excitation signal extension* and *spectral envelope expansion* could be used for speaker independent BWE. Quality may be increased with speaker dependent techniques using spectral envelope expansion at the cost of some practicality. When speaker dependent algorithms are used the user must train the algorithm. Although speaker dependent algorithms may lead to better quality reconstructed speech, they are less robust when compared to speaker independent algorithms. Small variations in speech for instance, caused by a common cold, may lead to undesirable results. This could be palliated by making the algorithm re-trainable. However, this is impractical and may lead users to abandoning the use of the device. It is thus important to evaluate such adverse effects and assure that the BWE algorithm used is practical, efficient, and reliable.

Vocal Effort Coding

In this section we discuss the various vocal modes and their relationship with physical distance between a speaker and a listener. Naturally, human beings adjust their vocal effort to compensate for changes in their environment. One can whisper a confidential message, call out for a meeting or shout out for help. It is important to distinguish "vocal effort" from "vocal level". The latter suggests a change in Sound-Pressure Level (SPL) while vocal effort involves a lot more than just changes in SPL (Traunmüller and Eriksson, 2000). Zhang et al. (2007) classified

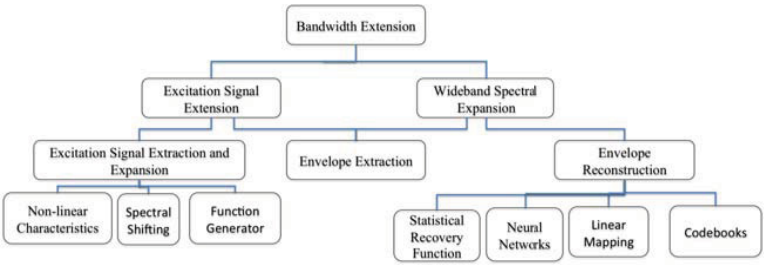

FIGURE 3: Classification of different bandwidth extension techiniques applicable to in-ear microphone signal pickup inside workers' ears.

5 speech modes: (1) whispered, (2) soft, (3) neutral, (4) loud, and (5) shouted. Each of these speech modes is characterized by its deviations from the neutral speaking condition. Many studies have been done to characterize each speech mode as to enhance speaker recognition systems and other applications. In particular, whispered and shouted speech require the most dramatic change in excitation (Zhang and Hansen, 2007) and have thus received a lot of attention. Our interest lies mostly with the shouted speech mode and the changes in acoustical features that occur.
As documented by many, as the vocal effort increases so does the fundamental frequency, *F0*. Another widely accepted change in the formants is the increase of the first formant, *F1* (Liénard and Di Benedetto, 1999) (Elliot, 2000) (Garnier *et al.*, 2008). Liénard and Di Benedetto (1999), however, also claim an increase in the second formant, *F2*, for females but this has not yet been widely accepted. Shouted sentences have increased initial *F0* slope but a decreased final *F0* (Fux *et al.*, 2011) and a decreased spectral slope (Zhang and Hansen, 2007). They are longer in duration which is caused by longer word duration, but have a decreased silence duration (Zhang and Hansen, 2007). Typically, shouted speech is detected based on *F0*, *F1* and the spectral tilt (Nanjo *et al.*, 2009). A summary of these changes can be seen in Table 1.

Traunmüller et al. describe vocal effort as *"the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance"* (Traunmüller and Eriksson, 2000). As distance increases so does the vocal effort. In fact, Brungart et al. report that as distance doubles the intensity increases by 8 dB, while Liénard et al. report that *F0* increases at 3.5 Hz/dB (Fux *et al.*, 2011) (Liénard and Di Benedetto, 1999). Distance, however, is not the only time we adjust our vocal effort. When our ability to hear our own voice changes, as a result of background noise for example, our vocal effort changes (Junqua, 1993). This is known as the *Lombard* effect. Although Lombard speech may share some characteristics with shouted speech, it is unique and cannot be treated the same way as shouted speech. Speakers vary their vocal effort based on the spectrotemporal properties of the background noise. In fact, significant differences of adjustments in the presence of white noise and babble noise have been reported (Traunmüller and Eriksson, 2000). A summary of the acoustical changes cause by Lombard speech as found by Junqua (1993) is shown in Table 1 .

When wearing HPDs, the Lombard effect is also a contributing factor in decreased speech intelligibility, from the perspective of both the speaker and the listener. Wearing hearing protection in noise not only affects the way speech is heard, it changes the way speech is produced. At the speaking end, Tufts and Frank (2003) studied the differences in speech acoustics when produced in noise while wearing hearing protection. For people with normal hearing, the level of adjustment in vocal effort as the level of noise increased was less when hearing protection was worn than without. As the level of noise increased from 60 dB to 100 dB SPL, speakers not wearing hearing protection increased their speaking leved by about 40dB, while those wearing hearing protection increased their vocal level by only 3 dB to 15 dB (Tufts and Frank, 2003). At the hearing end, studies by Candido Fernandes (2003) report that in environments with +5 dB Signal-to-Noise Ratio (SNR) and +10 dB SNR , wearing hearing protection decreases the intelligi-

TABLE 1: Summary of acoustical differences between shouted speech and Lombard when compared to neutral speech

Acoustical Feature	Shouted speech	Lombard speech
F0	Increased frequency	Increased frequency (more dominant in male speakers)
F1	Increased frequency	Increased frequency (more dominant in female speakers)
F2	Increased frequency (females only)	Increased frequency (females only)
Sentence Duration	Increased duration	Increased duration
SPL	Increased level	Slightly increased level

bility of speech. However, at -5 dB and -10 dB SNR, wearing hearing protection increases speech intelligibility by up to 10% (Candido Fernandes, 2003). It is also useful to note that studies by Giguère and Dajani (2009) report that persons wearing HPDs prefer an SNR of 13.5 dB when listening to speech in noise. This can be utilized in the experimental procedures as discussed in the next section.

The studied changes that characterize the different vocal efforts could be utilized in our application. However, as can be concluded from the preceding discussion, to correctly make a link between vocal effort and intended communication distance while wearing HPDs in noise, the effects of the Lombard effect along side the occlusion effect must be considered.

ENVISIONED EXPERIMENTAL APPROACH

In order to reach our goal to provide workers with intelligible communication without compromising their hearing protection through our proposed "Radio Acoustical Virtual Environment", we must answer the following research questions:

(1) *What is the relationship between vocal effort and communication distance in noise with and without HPDs? What are the changes in relevant acoustical features between the case where HPDs are worn and when they are not?*

(2) *To what extent does post processing of the IEM speech enhance intelligibility? Are there ways to train a speaker in noise to further enhance intelligibility?*

To answer the questions above, several tests must be performed on a large control group of human subjects for data collection. Below, is a list of tests that we anticipate could be helpful in advancing our research. For the following tests the speaker will be in a quiet environment but exposed to background noise through the in-ear device and asked to communicate at different levels of this background noise. This will leave the OEM free of noise enabling it to pick up a clean speech signal. Since the transfer function between the IEM and REF will always be the same, the respective IEM level can be figured out from the found REF signal. The control group for theses tests will consist of normal hearing people and have an equal number of females and males. There are two main types of tests that we envision carrying out:

(1) The first test will involve two normal hearing human subjects. One subject will be assigned as the speaker the other subject will be assigned as the listener. The speaker will be asked to relay a set of actions to the listener, for example, *"Pick up the hammer"*. The listener will have to correctly perform the requested action. Once the listener is successful the speech from the speaker is saved for analysis and annotated with the distance between speakers

and level of background noise. The same test will be repeated for multiple speakers and listeners, with and without hearing protection, in silence and different levels of background noise.

(2) The second test will utilize a moving cardboard target (equipped with a measurement microphone) that, again gradually moves farther away. At a fixed background noise level the speaker will be asked to speak so that her/his speech is intelligible to the moving target. Using objective speech intelligibility measures such as the one presented by Giguère *et al.* (2009), the speaker will receive a cue of whether or not the information was understood by analysis of the measurement microphone signal. The speaker's own speech will be played back to them and the speaker will be asked to adjust their speech to make it more intelligible.

Each test will contribute to a certain aspect of our research. The first test will help us collect the necessary data to map vocal effort and variations in relevant acoustical features of speech, with and without the use of HPDs, to intended communication distance. This test will also allow us to assess what acoustic features of speech are robust enough to be used in coding the vocal effort. An interesting consideration is the SPL of the speech. Conventionally SPL is not used to characterize different levels of vocal effort because of the unfixed position of the microphone. However, in our application the microphone location is stationary and could be used along with other features to code the vocal effort. The second test could indicate whether training a speaker how to speak in noise could further increase the intelligibility. It could also provide significant data on how much facial cues and gestures from a human listener are useful to the speaker and the listener. For example, Erber (1969) reports that lip reading in -10 dB SNR can increase speech intelligibility in noise by about 60%.

At the conclusion of these tests we envision producing a relationship as portrayed in Figure 4. The green blocks represent distances where speech is intelligible for the given vocal effort and background noise level. The yellow blocks represent areas of reduced intelligibility or areas where intelligibility is achieved only with reinforcement from facial cues or gestures. Red blocks represent areas where speech is unintelligible. Note, the numbers in this table are strictly for illustrative purposes and do not yet come from research data. Once this table is compiled, the vocal effort of the speaker may be coded and sent to an appropriate radius of intended listeners through an *ad-hoc* radio system such as cognitive radios (Li *et al.*, 2011). Figure 5 demonstrates the anticipated performance of RAVE. If a worker is speaking at 70 dBA SPL in a quiet environment the radio signal will be transmitted to anyone within a 20 m radius. As the level of noise increases and the vocal effort of the speaker remains constant the transmitting distance will decrease. Therefore, in an extremely noisy environment the transmitting distance of the radio will only be 5 m to compensate for such phenomena as the Lombard effect.

	Residual Background Noise (dBA SPL)				
	<60	60-70	70-80	80-90	>90
Whispered	2 m	unintelligible	unintelligible	unintelligible	unintelligible
Soft	4 m	1 m	reduced intelligibility	reduced intelligibility	unintelligible
Neutral	15 m	8 m	1 m	reduced intelligibility	unintelligible
Loud	20 m	10 m	1 m	reduced intelligibility	unintelligible
Shouted	40 m	20 m	10 m	5 m	unintelligible

FIGURE 4: Illustrative table of relationship between vocal effort and communication distance in the presence of background noise while wearing HPDs.

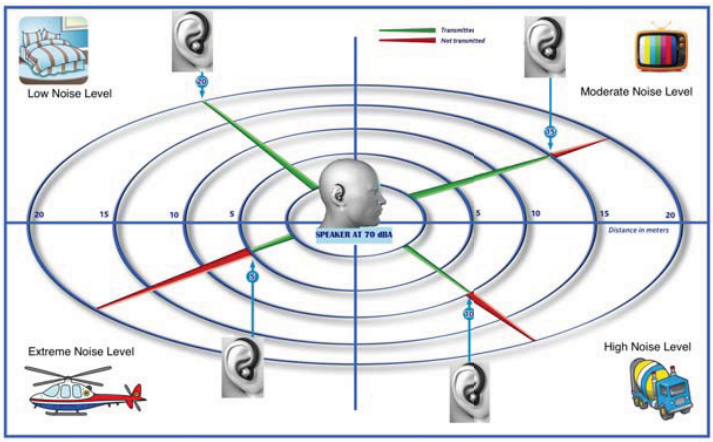

FIGURE 5: Illustration of functionality of RAVE. The green and red lines represent the areas where the signal is transmitted and not transmitted, respectively.

CONCLUSIONS

Good hearing protection is currently achieved at the cost of decreased communication while good communication is achieved at the cost of jeopardizing good hearing protection. Providing workers with satisfactory hearing protection and communication is still difficult and requires the compromise of one or the other. In this paper, we propose a new distance sensitive protocol that provides intelligible speech to workers wearing hearing protection. Using changes in acoustical features of speech the vocal effort will be coded and the speech signal will be sent in a way that mimics a natural acoustical environment. The "Radio Acoustical Virtual Environment" discussed will allow workers to communicate without the need to remove their HPDs and without having to move closer to their listener. Undisturbed speech from inside the ear canal will be captured and transmitted over wireless radio to the remote listener. The transmitted signal will only be received by listeners within a given spatial range, this range depending on the user's vocal effort and background noise level. Providing workers with such a device will enhance their work experience and potentially promote the use of HPDs in noisy work environments.

ACKNOWLEDGMENTS

The authors would like to thank Sonomax Technologies Inc. and its *Sonomax-ÉTS Industrial Research Chair in In-Ear Technologies* (CRITIAS) for their financial support. The authors also acknowledge the financial support from MITACS- Accelerate research internship program.

REFERENCES

Berger, E. (**2003**). *The noise manual* (Aiha).

Bernier, A. and Voix, J. (**2010**). "Signal characterization of occluded ear versus free-air voice pickup on human subjects", Canadian Acoustics **Vol. 38**, pp. 78–79.

Candido Fernandes, J. a. (**2003**). "Effects of hearing protector devices on speech intelligibility", Applied Acoustics **64**, 581–590.

Cheng, Y., O'Shaughnessy, D., and Mermelstein, P. (**1994**). "Statistical recovery of wideband

speech from narrowband speech", Speech and Audio Processing, IEEE Transactions on **2**, 544–548.

Elliot, J. (**2000**). "Comparing the Acoustic Properties of Normal and Shouted Speech: A Study in Forensic Phonetics", in *8th Aus. Int. Conf. Speech Sci. & Tech*, 154–159.

Erber, N. (**1969**). "Interaction of audition and vision in the recognition of oral speech stimuli", Journal of Speech, Language and Hearing Research **12**, 423.

Fux, T., Feng, G., and Zimpfer, V. (**2011**). "Talker-to-listener distance effects on the variations of the intensity and the fundamental frequency of speech", Cognition 4964–4967.

Garnier, M., Wolfe, J., Henrich, N., and Smith, J. (**2008**). "Interrelationship between vocal effort and vocal tract acoustics : a pilot study Music Acoustics Group , School of Physics , University of New South Wales , Sydney , Australia Département Language and Cognition , GIPSA-Lab , Grenoble , France", **2**, 3–6.

Giguère, C. and Dajani, H. R. (**2009**). "Noise exposure from communication headsets: the effects of external noise, device attenuation and effective listening signal-to-noise ratio", INTER-NOISE . . . .

Giguère, C., Laroche, C., Vaillancourt, V., and Soli, S. D. (**2009**). "A predictive model of speech intelligibility in noise for normal and hearing-impaired listeners wearing hearing protectors", INTER-NOISE . . . .

Iser, Bernd, Schmidt, G., and Minker, W. (**2008**). *Bandwidth Extension of Speech Signals* (ISBN 978-0-387-68898-5).

Junqua, J. C. (**1993**). "The Lombard reflex and its role on human listeners and automatic speech recognizers.", The Journal of the Acoustical Society of America **93**, 510–24.

Li, J., Zhou, Y., Lamont, L., and Gagnon, F. (**2011**). "A Novel Routing Algorithm in Cognitive Radio Ad Hoc Networks", in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, 1–5 (IEEE).

Liénard, J. S. and Di Benedetto, M. G. (**1999**). "Effect of vocal effort on spectral properties of vowels.", The Journal of the Acoustical Society of America **106**, 411–22.

Nanjo, H., Nishiura, T., and Kawano, H. (**2009**). "Acoustic-Based Security System: Towards Robust Understanding of Emergency Shout", 2009 Fifth International Conference on Information Assurance and Security **1**, 725–728.

O'shaughnessy, D. (**2000**). *Speech communications: human and machine* (Universities press).

Park, K.-y. and Kim, H. S. (**2000**). "Narrowband to wideband conversion of speech using GMM based transformation", Spectrum 1843–1846.

Traunmüller, H. and Eriksson, A. (**2000**). "Acoustic effects of variation in vocal effort by men, women, and children.", The Journal of the Acoustical Society of America **107**, 3438–51.

Tufts, J. B. and Frank, T. (**2003**). "Speech production in noise with and without hearing protection", The Journal of the Acoustical Society of America **114**, 1069.

Valin, J.-M. (**2002**). "Extension spectrale d'un signal de parole de la bande téléphonique à la bande AM", Ph.D. thesis, Sherbrooke University.

Zhang, C. and Hansen, J. H. L. (**2007**). "Analysis and Classification of Speech Mode : Whispered through Shouted", in *Proceedings of the Interspeech*, 2289– 2292.

APPENDIX II

PROTECTING MINERS' HEARING WHILE FACILITATING COMMUNICATION

Rachel E. Bou Serhal[1,3], Tiago H. Falk[2,3], Jérémie Voix [1,3]

[1]École de technologie supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Article presented at the World Mining Congress in Montréal, Canada on August 11-13, 2013

PROTECTING MINERS' HEARING WHILE FACILITATING COMMUNICATION

R. E. Bou Serhal, *J. Voix
École de technologie supérieure
1100 Rue Notre-Dame Ouest
Montréal, QC H3C 1K3
*(*Corresponding author: jeremie.voix@etsmtl.ca)*

T. Falk
Institut national de la recherche scientifique
800, de La Gauchetière Ouest
Portail Nord-Ouest, bureau 6900
Montréal, H5A 1K6

PROTECTING MINERS' HEARING WHILE FACILITATING COMMUNICATION

ABSTRACT

Many miners are exposed to dangerously high levels of noise on a daily basis. Over the past 15 years, a continually increasing number of occupational hearing loss has been reported from the mining community in the United States, of which, more than 95% is attributed to prolonged noise exposure. Although the noise exposure levels may differ between coal miners and metal and non-metal miners, in the absence of noise control at the source, the solution is the same: use of personal hearing protection devices (HPD). While protecting the miners' hearing it is also essential to no longer hinder their ability to communicate. With access to an advanced HPD that is customized to the miners' ears we are able to combine these two requirements. Using an intra-aural instantly custom molded HPD miners are protected from high levels of noise. The HPD is equipped with wireless capabilities, and contains both a speaker and an In-Ear Microphone (IEM). Therefore, the miners' speech may be captured from inside the ear and transmitted to the remote listener. This IEM signal is relatively noise-free since it is isolated from the background noise. The IEM speech signal, however, is "boomy" and is missing some high frequency content, making fricative consonants hard to understand. Nonetheless, the IEM speech signal is correlated with the natural speech signal and may be manipulated through statistical techniques to more closely resemble natural speech. By improving the intelligibility and quality of the IEM signal, numerous applications may be enabled. One use of the enhanced IEM signal will be for radio communication. Using wireless radio communication in a noisy mining environment is sometimes the only practical and affordable solution to allow communication between miners equipped with personal hearing protection devices. Traditionally, one of the weaknesses of such wireless radio communication lies in the lack of designating receivers: whether they are the intended receiver or not, all those carrying a radio receiver are subjected to the broadcasted signal. The current work will detail a new concept of a "radio-acoustical virtual environment" where the radio signal will only be received by miners within a given spatial range, such range depending on the user's vocal effort as well as the ambient and perceived background noise levels.

KEYWORDS

Hearing Protection Devices, Communication, Wireless, Radio

INTRODUCTION

Miners are among over 30 million workers in North America who are exposed to excessive levels of noise that put them at risk of losing their hearing (National Institute of Occupational Safety and Health, 1998). A study of metal and non metal miners across the united states reported over 95% of hearing loss reported to the Mine Safety and Health Administration (MSHA) was caused from prolonged noise exposure (Valoski, 1997). This is unfortunate as noise-induced hearing loss is a serious yet preventable health hazard. The Occupational Safety and Health Administration (OSHA) proposes the following three methods of protecting workers from noise exposure (Katz et al., 2009):

1. engineered reduction of the noise
2. limiting exposure time
3. use of personal hearing protection

The mining environment and current practices have made it difficult to prevent hazardous noise exposure to miners. Noise control i.e. the engineered reduction of noise is expensive and requires the attention of the higher management. New materials and enclosures have been developed to decrease the noise levels of some equipment. Even making sure that equipment is well maintained can aid in controlling the noise at

the source (McBride, 2004). However, noise control can only go so far in limiting noise exposure as some exposure to loud noise, such as the impact from a drill bit, are inevitable. Limiting exposure time has also been unsuccessful. The allowable limit as set by the National Institute of Occupational Safety and Health (NIOSH) is 85 dBA for eight hours exposure (Berger, 2003). Yet studies have shown that 80% of U.S miners are exposed to a Time-Weighted Average (TWA) over 85 dBA, and of those, 25% are exposed to a TWA exceeding 90 dBA (McBride, 2004). The final solution is the use of personal Hearing Protection Devices (HPD). HPDs come in many different shapes and sizes and can be made from a variety of materials. The two main types of HPDs are intra-aural i.e. earplugs, and supra-aural i.e. earmuffs (Berger, 2003). There are a couple of points to consider when discussing HPDs: the comfort and the effectiveness of the personal HPD. Using HPDs that are comfortable to wear for an extended period of time is vital because an uncomfortable fit is more likely to drive the user to remove the HPD. It is also important to properly wear HPDs because an improper fit leads to misrepresented attenuation, causing the user to be unknowingly unprotected. Both of these issues may be resolved by a custom molded HPD that allows for a way to monitor the real attenuation inside the ear (Voix and Laville, 2009). The problems that arise with the use of HPDs in mining environments is twofold: the acoustical environment of mining and, as a consequence, difficulties in communication. Depending on the the Signal-to-Noise Ratio (SNR), HPD's can be detrimental to communication. Fernandes (2003) reports that in environments with +5 dB SNR and +10 dB SNR , wearing hearing protection decreases the intelligibility of speech. However, at -5 dB and -10 dB SNR, wearing hearing protection can increase speech intelligibility by up to 10%. Therefore, for environments where noise is intermittent, such as mining, wearing HPDs deteriorates communication and users are more likely to seek out forms to better communicate. Currently there are several different ways that are used to communicate in noise while using HPDs, one could:

1. Remove the HPD: get closer to a listener and adjust vocal effort to communicate. Removing an HPD to communicate is problematic as the effectiveness of HPDs is greatly reduced with non-continuous use (Berger, 2003). It also requires the miners to be in close proximity of one another to communicate.
2. Use passively filtered HPD: flat attenuation HPDs could be beneficial for speech communication as they do not attenuate high frequencies as much as other HPDs. However, in noise, these HPDs are not as effective as they usually do not provide sufficient attenuation. In quiet, they also decrease speech intelligibility, which would compel the wearer to remove the HPD for communication.
3. Use a hand-held radio device over HPDs: use of a walkie-talkie allows for distance communication with multiple people while remaining stationary. Using a hand-held radio overcomes the problem of proximity but still requires the removal of the HPD.
4. Use of a communication headset: usually an earmuff with a miniature loudspeaker and an external boom microphone. The voice picked up by the boom microphone is transmitted through either a wired or wireless network to a remote listener. Although these are the best current alternative, these headsets still present the following inconvenience: the external microphone will not only pick up the user's voice but it will as well pick up the background noise, which dramatically affects intelligibility.

Another issue associated with using any kind of radio transmitter, is that it does not distinguish a receiver and all communication is sent to everyone on the same radio channel. Therefore, the users' radio is often flooded with irrelevant conversation that could be annoying and somewhat loud and thus contributing to the noise dose. For underground miners, using radio communication in general is problematic. Electromagnetic wave propagation in underground mines is complex, rendering wireless communication a difficult task (Moutairou et al., 2009). Clearly there is a need for a device that provides good noise attenuation as well as good communication without compromising the performance of one for the other.

Proposed Approach

We propose a new concept called "Radio Acoustical Virtual Environment" (RAVE) in which miners can achieve intelligible communication without hindering their hearing protection. RAVE uses an advanced intra-aural instantly custom molded HPD, shown in Figure 1, equipped with an In-Ear Microphone (IEM), a miniature loudspeaker, a Digital Signal Processor (DSP), an Outer-Ear Microphone (OEM) and Wireless Radio (WR) capabilities. Such a device can capture a somewhat undisturbed speech signal from inside the ear (referred to as IEM speech). Because the signal captured originates from bone conducted vibrations, it lacks higher frequencies. Thus, the IEM signal must first be enhanced in its high frequency content. Once enhanced, the IEM signal is coded and sent to an appropriate radius of listeners based on the acoustical features of the produced speech and the level of background noise. Figure 2 demonstrates the anticipated performance of RAVE. If a miner is speaking at 70 dBA SPL in a quiet environment the radio signal will be transmitted to anyone within a 20 m radius. As the level of noise increases and the vocal effort of the speaker remains constant the transmitting distance will decrease. Therefore, in an extremely noisy environment the transmitting distance of the radio will only be 5 m to compensate for such phenomena as the Lombard effect.

This paper introduces the concept of RAVE and the methodology involved in realizing such a protocol. The next section discusses different techniques available for the enhancement of the IEM speech signal followed by the concept of vocal effort coding. Finally, limitations and promising avenues are discussed.

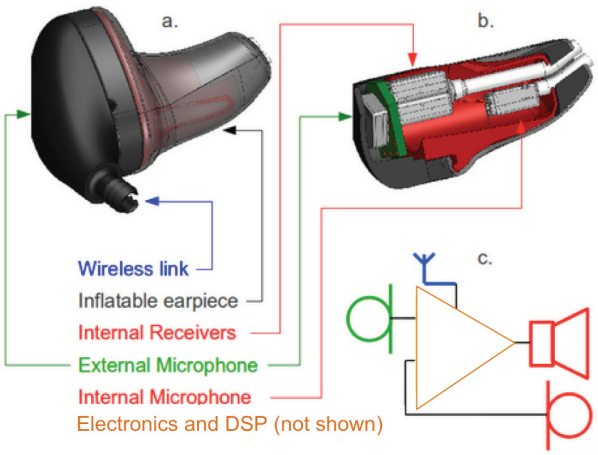

Figure 1 – Overview of digital custom earpiece (a), its electroacoustical components (b), and equivalent schematic (c).

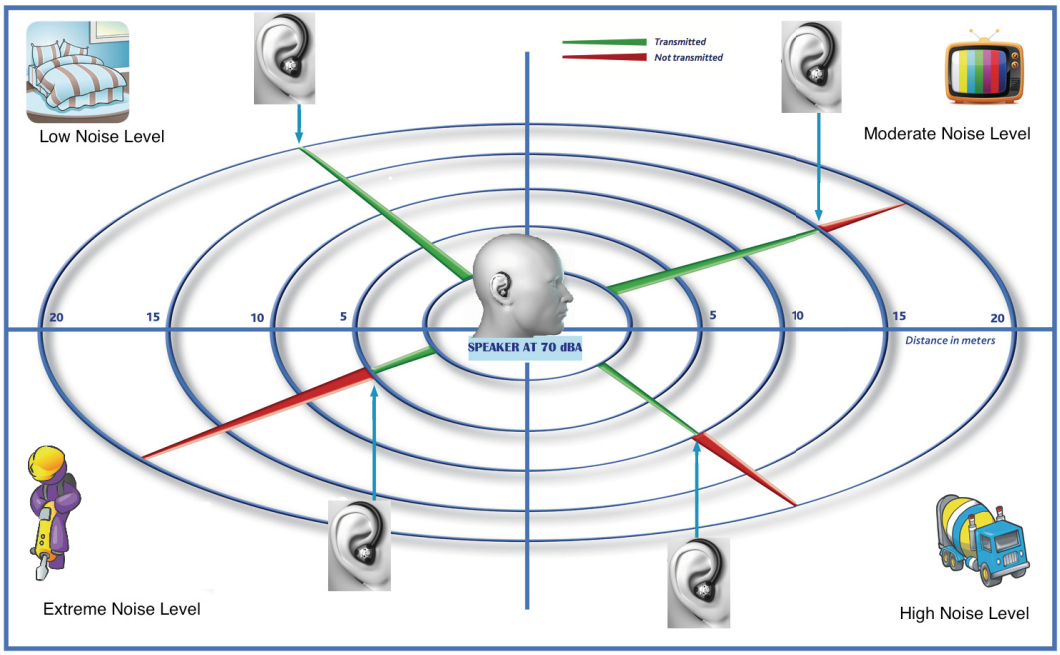

Figure 2 – Illustration of functionality of RAVE. The green and red lines represent the areas where the signal is transmitted and not transmitted, respectively.

ENHANCEMENT OF THE IEM SPEECH

When speech is captured conventionally (with a boom microphone), to be sent over a radio network in a noisy environment, it is disturbed and contains the noise picked up by the exposed microphone, even when using a directional microphone. On the other hand, capturing speech from inside the protected ear allows for the transmission of a less-disturbed speech signal that will not require extra de-noising, usually achieved by the electronics within the radio. When the ear canal is blocked by an in-ear device, there is a regeneration of the speech inside the ear canal and one experiences what is called the *occlusion effect* (Berger, 2003). The *occlusion effect* allows for the capturing of speech inside the ear, which is useful in noisy environments. Because of cranial bone conduction, this signal is "boomy", containing most of its energy in the lower frequencies while missing important high frequency content (Bernier and Voix, 2010). The difference between the frequency content of the IEM speech and the OEM speech (referred to as REF) of the utterance /u/, for a male speaker, is demonstrated in Figure 3. From Figure 3, it we notice that above 1.8 kHz the IEM signal is missing important high frequency content. As a consequence of the IEM signal's limited bandwidth, fricative consonants such as /s/ and /f/, and nasals such as /n/ and /m/ are unintelligible. The IEM signal is thus perceived as having lower quality and intelligibility than "free air speech", or speech that is recorded near the mouth. To solve this, the IEM signal could be expanded using Bandwidth Extension (BWE) of the speech signal. Many different BWE techniques exist, and the proper choice depends on the desired results and available resources. BWE can range from spectral estimation and expansion through excitation signal extension, to Vector Quantization (VQ) and codebook mapping. Iser et al. (2008) give a good review of the basics of such techniques (Iser et al., 2008). In the past, the need for BWE arose because of the limited bandwidth of the telephone network. The narrow bandwidth of a telephone is about 3.5 kHz leaving some significant parts of human speech unrepresented. In this context, wideband signals refer to signals that can represent the entire vocal range while narrowband signals can only represent a limited part of the vocal range. With access to an IEM and an OEM, BWE can be used for our purposes by treating the IEM signal as the narrowband signal and the free-air speech captured by the OEM as the wideband signal. All available techniques for BWE are listed in Figure 4. It is important to assess the resources available to choose a practical and efficient technique with good performance. Some things to consider are the computational complexity and cost of the

algorithm, power consumption and whether the algorithm will be speaker dependent or speaker independent. *Excitation signal extension* and *spectral envelope expansion* could be used for speaker independent BWE. Quality may be increased with speaker dependent techniques using spectral envelope expansion at a cost of some practicality. When speaker dependent algorithms are used the miner must train the algorithm. Although speaker dependent algorithms may lead to better quality reconstructed speech, they are less robust when compared to speaker independent algorithms. Small variations in speech that may be caused by a common cold may lead to undesirable results. This could be palliated by making the algorithm re-trainable. However, this is impractical and may lead miners to abandoning the use of the device. It is thus important to evaluate such adverse effects and assure that the BWE algorithm used is practical, efficient, and reliable.

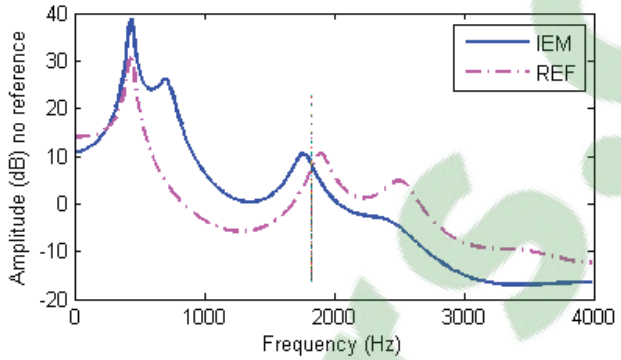

Figure 3 – IEM vs. REF spectral envelopes of the utterance /u/ from the word 'canoe', showing the increased low frequency content and the missing high frequency content.

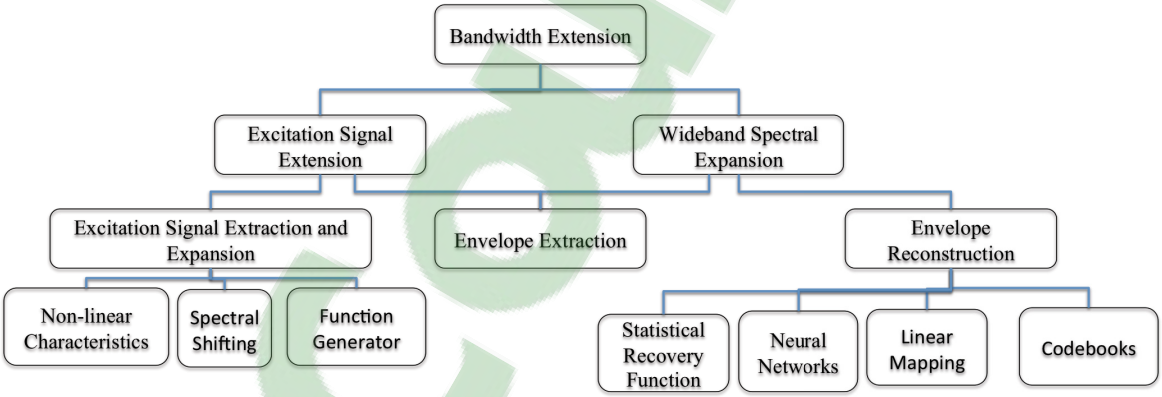

Figure 4 – Classification of different bandwidth extension techniques applicable to in-ear microphone signal pickup inside miners' ears

VOCAL EFFORT CODING

In this section we discuss the various vocal modes and their relationship with physical distance between a speaking miner and a listening miner. Naturally, human beings adjust their vocal effort to compensate for changes in their environment. One can whisper a confidential message, call out for a meeting or shout out for help. It is important to distinguish "vocal effort" from "vocal level". The latter suggests a change in Sound-Pressure Level (SPL) while vocal effort involves a lot more than just changes in SPL (Traunmüller and Eriksson, 2000). Zhang et al. classified 5 speech modes: (1) whispered, (2) soft, (3) neutral, (4) loud, and (5) shouted (Zhang and Hansen, 2007). Each of these speech modes is characterized by its deviations from the neutral speaking condition. Many studies have been done to characterize each speech mode as to enhance speaker recognition systems and other applications. In particular, whispered and shouted speech require the most dramatic change in excitation (Zhang and

Hansen, 2007) and have thus received a lot of attention. Our interest lies mostly with the shouted speech mode and the changes in acoustical features that occur. As documented by many, as the vocal effort increases so does the fundamental frequency, *F0*. Another widely accepted change in the formants is the increase of the first formant, *F1* (Liénard and Di Benedetto, 1999; Elliot, 2000; Garnier et al., 2008). Liénard et al. (1999), however, also claim an increase in the second formant, *F2*, for females but this has not yet been widely accepted. Shouted sentences have increased initial *F0* slope but a decreased final *F0* (Fux et al., 2011) and a decreased spectral slope (Zhang and Hansen, 2007). They are longer in duration which is caused by longer word duration, but have a decreased silence duration (Zhang and Hansen, 2007). Typically, shouted speech is detected based on *F0*, *F1* and the spectral tilt (Nanjo, 2009). A summary of these changes can be seen in Table 1.

Traunmüller et al. describe vocal effort as *"the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance"* (Traunmüller and Eriksson, 2000). As distance increases so does the vocal effort. In fact, Brungart et al. report that as distance doubles the intensity increases by 8 dB, while Liénard et al. report that *F0* increases at 3.5 Hz/dB (Fux et al., 2011; Liénard and Di Benedetto, 1999). Distance, however, is not the only time we adjust our vocal effort. When our ability to hear our own voice changes, as a result of background noise for example, our vocal effort changes (Junqua, 1993). This is known as the *Lombard* effect. Although Lombard speech may share some characteristics with shouted speech, it is unique and cannot be treated the same way as shouted speech. Speakers vary their vocal effort based on the spectrotemporal properties of the background noise. In fact, significant differences of adjustments in the presence of white noise and babble noise have been reported (Traunmüller and Eriksson, 2000). A summary of the acoustical changes cause by Lombard speech as found by Junqua (1993) is shown in Table 1.

Bringing together the acoustical changes caused by vocal effort and those caused by the Lombard effect, will bring about a relationship between vocal effort while wearing HPD in noise and intended communication distance. Scheduled tests on a group of normal-hearing human subjects will be the starting point in the data collection involved in achieving the aforementioned relationship. Once adequate data collection is reached, we envision producing a relationship as portrayed in Table 2. The green blocks represent distances where speech is intelligible for the given vocal effort and residual background noise level under the HPD. The yellow blocks represent areas of reduced intelligibility or areas where intelligibility is achieved only with reinforcement from facial cues or gestures. Red blocks represent areas where speech is unintelligible. Note, the numbers in this table are strictly for illustrative purposes and do not yet come from research data. Once this table is compiled, the vocal effort of the speaker may be coded and sent to an appropriate radius of intended listeners through an *ad-hoc* radio system such as cognitive radios (Li et al., 2011).

Table 1 – Summary of acoustical differences between shouted speech and Lombard when compared to neutral speech .

Acoustical Feature	**Shouted Speech**	**Lombard Speech**
F0	Increased frequency	Increased frequency (more dominant in males)
F1	Increased frequency	Increased frequency (more dominant in females)
F2	Increased frequency (females only)	Increased frequency (females only)
Sentence Duration	Increased duration	Increased duration
SPL	Increased level	Slightly increased level

Table 2 – Illustrative table of relationship between vocal effort and communication distance in the presence of background noise while wearing HPD.

		Residual Background Noise (dBA SPL)				
		<60	60-70	70-80	80-90	>90
Vocal effort of speaker	Whispered	2 m	unintelligible	unintelligible	unintelligible	unintelligible
	Soft	4 m	1 m	reduced intelligibility	reduced intelligibility	unintelligible
	Neutral	15 m	8 m	1 m	reduced intelligibility	unintelligible
	Loud	20 m	10 m	1 m	reduced intelligibility	unintelligible
	Shouted	40 m	20 m	10 m	5 m	unintelligible

DISCUSSION

The "Radio-Acoustical Virtual Environment" discussed will allow miners to communicate without the need to remove their HPDs and without having to move closer to their listener. Undisturbed speech from inside the ear canal will be captured and transmitted over wireless radio to the remote listener. The transmitted signal will only be received by listening miners within a given spatial range, this range depending on the speaking miner's vocal effort and background noise level. This solves most of the issues that are currently faced by miners trying to communicate and protect their hearing, however, a few problems persist and require further discussion. Access to the auditory platform shown in Figure 1, can open up the door to a more adaptive hearing protection device.

Previously, we mentioned that wearing HPDs in quiet environments decreases intelligibility and with the current design of RAVE this problem persists. In this case, we could take advantage of the OEM and the DSP by utilizing them to monitor the environmental SPL (Mazur and Voix, 2012). If the level is safe, the internal speaker could be used to reproduce what is picked up by the OEM and bypass the HPD. If the OEM registers that the levels are unsafe then no bypass occurs and the HPD functions as previously discussed. It would be useful to have a way to manually enable the bypass of the HPD and allow the signal picked up by the OEM to pass through for communication between those wearing the HPD described and those that are not. Another issue to consider when the environment is quiet is the annoyance caused by the *occlusion effect*. In noise, we depend on the *occlusion effect* for communication, which is not problematic because the high levels of noise counteract the predominance of the *occlusion effect*. However, when trying to communicate in quiet, even when the HPD is bypassed, one's own speech is predominantly what is heard which makes it annoying for the speaker. To solve this, an active *occlusion effect* reduction system can be implemented (Bernier and Voix, 2012). The last foreseen difficulty is the use of a wireless radio for distant communication in underground mining. Research in this area is growing and many new protocols are developing. Advancements in this area (Ndoh and Delisle, 2005; Srinivasan, Ndoh, and Kaluri, 2005; Moutairou et al., 2009) could be further investigated to be implemented with our radio system, to offer the most efficient radio system available.

CONCLUSIONS

RAVE already addresses many of the issues that are faced by miners communicating in noise and is thus a better alternative to what is presently available. Good hearing protection is currently achieved at the cost of decreased communication while good communication is achieved at the cost of jeopardizing good hearing protection. Providing miners with satisfactory hearing protection and communication is still difficult and requires the compromise of one or the other. In this paper, we propose a new distance sensitive protocol that provides intelligible speech to miners wearing hearing protection. Using changes in

acoustical features of speech the vocal effort will be coded and the speech signal will be sent in a way that mimics a natural acoustical environment. Providing miners with such a device will enhance their work experience and potentially promote the use of HPDs in noisy environments.

ACKNOWLEDGMENTS

The authors would like to thank Sonomax Technologies Inc. and its Sonomax-ÉTS Industrial Research Chair in In-Ear Technologies (CRITIAS) for their financial support. The authors also acknowledge the financial support from l'Équipe de recherche en sécurité du travail et en analyse des risques industriels (ÉREST) as well as from MITACS- Accelerate research internship program.

REFERENCES

Berger, E.H. 2003. *The Noise Manual*. AIHA.

Bernier, Antoine, and Jérémie Voix. 2010. "Signal Characterization of Occluded Ear Versus Free-air Voice Pickup on Human Subjects." *Canadian Acoustics* Vol. 38 (Num. 3): pp. 78–79.

Bernier, Antoine, and Jérémie Voix. 2012. "Une Nouvelle Protection Auditive Adaptée Aux Musiciens." *Fameq* 27 (1): 39–41.

Elliot, Jennifer. 2000. "Comparing the Acoustic Properties of Normal and Shouted Speech: A Study in Forensic Phonetics." In *8th Aus. Int. Conf. Speech Sci. & Tech*, 154–159.

Fernandes, J. C. 2003. "Effects of Hearing Protector Devices on Speech Intelligibility." *Applied Acoustics* 64: 581–590.

Fux, Thibaut, Gang Feng, and Véronique Zimpfer. 2011. "Talker-to-listener Distance Effects on the Variations of the Intensity and the Fundamental Frequency of Speech." *Cognition*: 4964–4967.

Garnier, Maëva, Lucie Bailly, Marion Dohen, Pauline Welby, and Hélène Lœvenbruck. "An Acoustic and Articulatory Study of Lombard Speech : Global Effects on the Utterance": 2–5.

Garnier, Maëva, Joe Wolfe, Nathalie Henrich, and John Smith. 2008. "Interrelationship Between Vocal Effort and Vocal Tract Acoustics : a Pilot Study" 2: 3–6.

Iser, Bernd, Gerhard Schmidt, and Wolfgang Minker. *Bandwidth Extension of Speech Signals*. ISBN 978-0-387-68898-5. doi:10.1007/978-0-387-68899-2.

Junqua, J C. 1993. "The Lombard Reflex and Its Role on Human Listeners and Automatic Speech Recognizers." *The Journal of the Acoustical Society of America* 93 (1) (January): 510–24.

Katz, Jack, Larry Medwetsky, Robert Bukard, and Linda Hood. 2009. *Handbook of Clinical Audiology*. 6th ed. Lippincott Williams & Wilkins.

Li, Jun, Yifeng Zhou, Louise Lamont, and Francois Gagnon. 2011. "A Novel Routing Algorithm in Cognitive Radio Ad Hoc Networks." In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, 1–5.

Liénard, J S, and M G Di Benedetto. 1999. "Effect of Vocal Effort on Spectral Properties of Vowels." *The Journal of the Acoustical Society of America* 106 (1) (July): 411–22.

Mazur, Kuba, and Jérémie Voix. 2012. "Development of an Individual Dosimetric Hearing Protection Device." In *INTER-NOISE*.

McBride, David I. 2004. “Noise-induced Hearing Loss and Hearing Conservation in Mining.” *Occupational Medicine (Oxford, England)* 54 (5) (August): 290–6. doi:10.1093/occmed/kqh075.

Moutairou, Manani, Student Member, Gilles Y Delisle, and Michel Misson. 2009. “Coverage Efficiency of Narrow-Band Wave Propagation in Mining Environments” 51 (2): 391–400.

Nanjo, Hiroaki. 2009. “Acoustic-based Security System : Towards Robust Understanding of Emergency Shout” 1: 725–728. doi:10.1109/IAS.2009.121.

National Institute of Occupational Safety and Health. 1998. “Occupational Noise Exposure”. U.S Department of Health and Human Services.

Ndoh, M., and G.Y. Delisle. 2005. “Propagation Characteristics for Modern Wireless System Networks in Underground Mine Galleries.” *First International Workshop on Wireless Communicaion in Underground and Confined Area*.

Srinivasan, K., M. Ndoh, and K. Kaluri. 2005. “Advanced Wireless Networks for Underground Mine Communications.” *First International Workshop on Wireless Communicaion in Underground and Confined Area*.

Traunmüller, H, and A Eriksson. 2000. “Acoustic Effects of Variation in Vocal Effort by Men, Women, and Children.” *The Journal of the Acoustical Society of America* 107 (6) (June): 3438–51.

Valoski, Michael P. 1997. “REPORTED NOISE-INDUCED HEARING LOSS AMONG MINERS.” *Applied Occupational and Environmental Hygiene* 12 (12): 1055–1058.

Voix, Jérémie, and Frédéric Laville. 2009. “The Objective Measurement of Individual Earplug Field Performance.” *The Journal of the Acoustical Society of America* 125 (6) (June): 3722–32. doi:10.1121/1.3125769.

Zhang, Chi, and John H L Hansen. 2007. “Analysis and Classification of Speech Mode : Whispered Through Shouted.” In *Proceedings of the Interspeech*, 2289– 2292.

APPENDIX III

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

Rachel E. Bou Serhal[1,3], Tiago H. Falk[2,3], Jérémie Voix [1,3]

[1]École de technologie supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Article presented at the International Conference on Acoustics, Speech and Signal Processing, in Brisbane, Australia on April 19-25, 2015

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
[◇] Centre for Interdisciplinary Research in Music Media and Technology, 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].

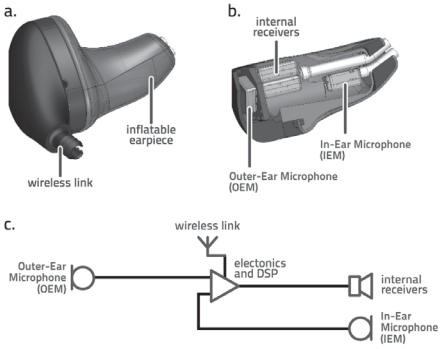

Fig. 1. Overview of communication headset (a), its electro-acoustical components (b), and equivalent schematic (c).

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 of 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

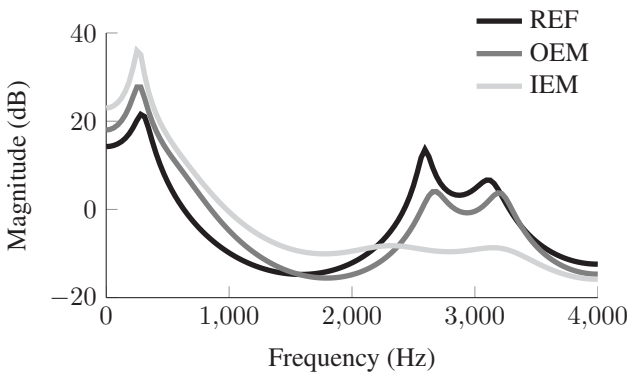

Fig. 2. The LPC spectral envelope of the phoneme /i/ recorded with the REF, the OEM and the IEM simultaneously.

(MFCC) as they provide a good representation of human 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 number 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 frequency 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:

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

where N is a very large number. 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 sub-bands.

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. Measuring 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 [13]. 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 [14] 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.

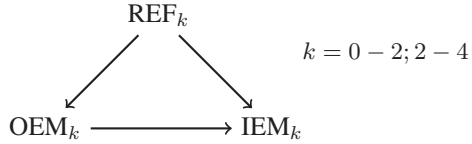

Fig. 3. Schematic showing the signal pairs used in the mutual information calculation, for each tested SNR value.

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.

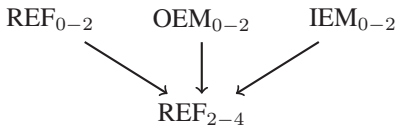

Fig. 4. Schematic showing the cross-band signal pairs used in the mutual information calculation for each tested SNR value.

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 [15], 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

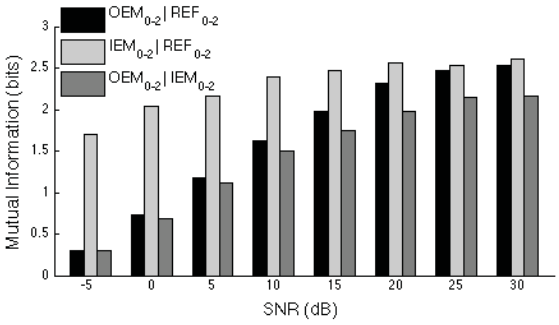

Fig. 5. Mutual information of the low-band between the REF, OEM and IEM signals.

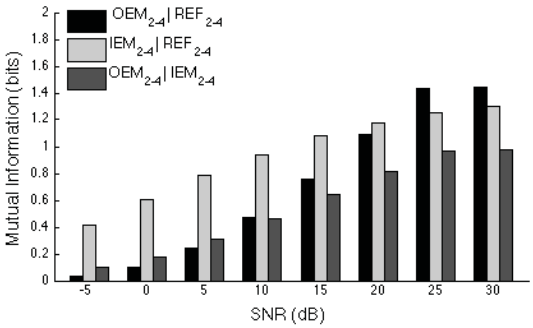

Fig. 6. Mutual information of the high-band between the REF, OEM and IEM signals.

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 [16] could be used to extend the bandwidth of the IEM signal. The GMM can be trained offline in

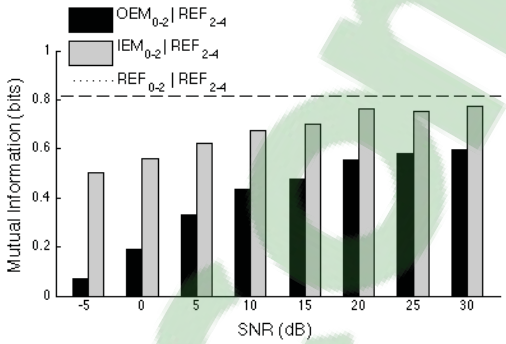

Fig. 7. Cross-band mutual information between the OEM, IEM and REF signals compared with the average cross-band mutual information within the REF signal.

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 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 Centre for Interdisciplinary Research in Music Media 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] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," *Sixth IEEE International Symposium on Signal Processing and Information Technology, ISSPIT*, pp. 426–431, 2007.

[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] V. Nadon, A. Bockstael, D. Botteldooren, J.M. Lina, and J. Voix, "Individual monitoring of hearing status: Development and validation of advanced techniques to measure otoacoustic emissions in suboptimal test conditions," *Applied Acoustics*, vol. 89, pp. 78–87, 2015.

[14] 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.

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

[16] 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.

APPENDIX IV

MODELING SPEECH PRODUCTION IN NOISE FOR THE ASSESSMENT OF VOCAL EFFORT FOR USE WITH COMMUNICATION HEADSETS

Rachel E. Bou Serhal[1,3], Tiago H. Falk[2,3], Jérémie Voix [1,3]

[1]École de technologie supérieure, Montréal, Canada

[2]Institut national de la recherche scientifique, Montréal, Canada

[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Article presented at Euronoise, in Maastricht, Netherlands on May 31 to June 3, 2015

Modeling Speech Production in Noise for the Assessment of Vocal Effort for Use with Communication Headsets

Rachel E. Bouserhal
École de technologie supérieure, Université de Québec, Montréal, Canada.
Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada.

Jérémie Voix*
École de technologie supérieure, Université de Québec, Montréal, Canada.
Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada.

Tiago H. Falk
Institut national de la recherche scientifique, Centre EMT, Montréal, Canada.
Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada.

Summary
A Radio Acoustical Virtual Environment (RAVE) is being developed to address issues occurring when communicating in noise while wearing Hearing Protection Devices (HPD). RAVE mimics a natural acoustical environment by transmitting the speaker's voice signal only to receivers within a given radius, the distance of which is calculated by considering the speaker's vocal effort and the level of background noise. To create a genuine RAVE, it is necessary to understand and model the speech production process in noise while wearing HPDs. Qualitative open-ear and occluded-ear models of the vocal effort as function of background noise level, exist. However, few take into account the effect of communication distance on the speech production process and none do so for the occluded-ear. To complement these models, quantitative data is used to generate quantitative open-ear and occluded-ear models, representing the relationship between vocal effort, communication distance, background noise level and type of HPD. These models can later be implemented within radio-communication headsets used in the proposed RAVE. Speech production models for occluded-ear accounting for the intended communication distance are presented in qualitative terms.

PACS no. 43.70.+i, 43.72.+q

1. Introduction

Using radio communication in noisy environments is a practical and affordable solution allowing communication between people with Hearing Protection Devices (HPD). Traditionally, one of its weaknesses lies in the lack of designating receivers: all those carrying a personal radio (walkie-talkie, etc.) are subjected to the broadcasted signal regardless of whether or not they are the intended listeners. Receiving irrelevant communication is annoying and contributes to the daily accumulated noise dose [1]. A new concept of a "Radio-Acoustical Virtual Environment" (RAVE) is being developed [2]. RAVE intends to mimic a natural acoustical environment by transmitting a communication signal only to people within a specific spatial range. This range is defined as the intended communication distance of the speaker.

Speakers with normal hearing adjust their vocal effort in the presence of noise [3], when trying to communicate at a distance [4] and to express emotion [5]. These adjustments still occur when wearing HPDs, however, they are altered as a function of the type of HPD worn [6, 7]. These changes in vocal effort as a function of the background noise and the type of HPD have been studied and modelled [8]. Interestingly, none of the studies include the effect of the intended communication distance to the model.

This paper presents a review of the current known work of the speech production process in noise both for the open-ear and the occluded-ear condition in Sections 2 and 3 respectively. We also propose a new model that includes the effect of communication distance to the speech production process in noise with

* Corresponding author: jeremie.voix@etsmtl.ca

HPDs in Section 4. Finally, the conclusions are presented in Section 5.

2. Open-Ear Condition

Naturally, speakers raise their voice when speaking in noise. This is called the Lombard Effect [3]. The Lombard effect has been well studied for the open-ear condition. Studies have shown that speakers raise the level of the their voice by 1-6 dB for every 10 dB of noise increase [9]. Multiple studies have found that Lombard speech, i.e. speech produced in noise, increases the speaker's fundamental frequency, f_0, [10, 11] by 0.6-2.5 semitones [12]. A gender-dependant increase in the spectral center of gravity also occurs during Lomabrd speech [6].

In quiet conditions, in turn, speakers raise their vocal effort to reach farther distances. As the communication distance doubles speakers raise their vocal level between 1.3-6 dB [13, 14, 15]. A study done by Zahorik et. al showed that speakers adjust their vocal effort according to their environment as well as the communication distance [15]. The speakers' f_0 as well as first formant, $F1$, also increase as a function of distance. As the vocal intensity increases, f_0 increases by 5 Hz/dB while $F1$ increases by 3.5 Hz/dB [16]. The changes in f_0, Δf_0, caused by increase in communication distance, and thus vocal intensity, was studied to be unique and telling of an increase in effort as a consequence of the increase in distance [4].

It is evident from previous studies that adjustments in the vocal effort as a consequence of either increase in communication distance or the presence of noise varies from speaker to speaker but follows the same trend across speakers. Vocal intensity, and changes in the speaker's f_0 are good indicators of raising the vocal effort. Zahorik et al. suggested that speakers adequately try to match the degradation in their vocal intensity due to propagation loss over distances [15]. Let us consider the model presented in [14] for the vocal power level as a function of the distance. In this case, we choose the model created for the speech produced in an anechoic room. This is because it eliminates any corrections from reverberation and it is the model that best fits data collected from other studies. The model is as follows:

$$L_w = 59.54 + 2.96 \times \log_2\left(\frac{d}{1.5}\right) \tag{1}$$

where L_w is the speech power level in decibels (dB) and d is the communication distance in meters. As a function of distance, the vocal power level can be graphed as shown in Fig. 1. Combining the model in Eq. 1 and what we know about communication in noise we can create a model that incorporates the presence of noise as a correction factor to the model. This will lead to a model that relates the vocal effort to the level of background noise and the intended

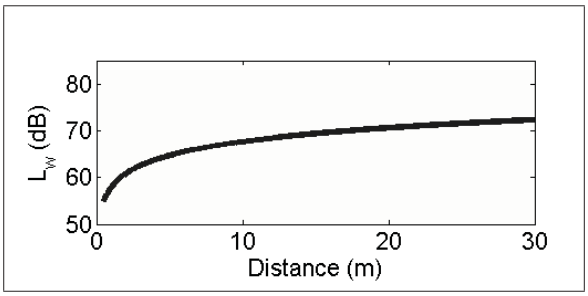

Figure 1. Vocal power level as a function of communication distance as presented in [14].

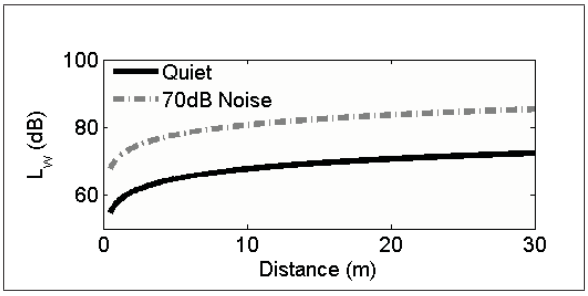

Figure 2. Comparing vocal power level of a speaker in quiet and a speaker in noise as a function of communication distance.

communication distance. If we consider the presence of noise to be anything above 60 dB(SPL) for the and average that a speaker's level will increase by 3 dB for every 10 dB of noise, then the modified model becomes:

$$L_w = 59.54 + 2.96 \times \log_2\left(\frac{d}{1.5}\right) + n \times [10 + 0.3 \times (N - 60)] \tag{2}$$

where n is 0 in quiet and 1 if the noise is greater than 60 dB and N represents the level of background noise in dB(SPL). The addition of the 10 dB accounts for an initial increase at the onset of noise that can be estimated from [6]. For example, the vocal power of a speaker exposed to 70 dB of noise can be compared to that of a speaker in quiet as shown in Fig. 2. From Eq. 1, a speaker in quiet trying to reach a distance of 50 m will speak at an estimated power level, L_w, of 74.5 dB while, from Eq. 2, a speaker in 70 dB noise trying to reach the same communication distance will speak at 87.5 dB. It is important to keep in mind that the model presented in equation Eq. 2 has not been validated but merely a prediction based on the already available data from previous studies. In the next section we review the effects on wearing hearing protection devices on the speech production process.

120

3. Occluded-Ear Condition

Blocking the ear canal path causes a resonance of the bone conducted vibrations caused by speech, causing speakers to hear an amplified 'boomy' version of their voice as they speak. This phenomenon is called the "occlusion effect" [17]. The contributions of the occlusion effect on changes in speech production while wearing HPDs is arguable. In fact, it's one's perception of his/her own voice that greatly affects the speech production process in noise [6]. A speaker's perception of their own voice level compared to the level of noise is the driving factor in the speech production process. Studies have shown that speakers wearing HPDs do not react to increase in noise levels as much as speakers not wearing HPDs. Tufts et al. report a 4-11 dB decrease in the level of speech produced in noise while wearing earplugs compared to speech produced in noise without HPDs. In the presence of 60 dB(SPL) of noise, while wearing earplugs, speakers did not increase their vocal effort from the quiet condition. Also, overall speech level increased by only 5 dB even though the noise increased 40 dB [6]. In other words, while wearing HPDs speakers adjust their vocal effort by only 1.25 dB for every 10 dB increase in noise. In quiet, however, speakers wearing earplugs did not significantly alter their overall speech level [6, 18] from their open-ear level. None of the studies performed on the occluded ear looked at the effects of the communication distance.

If we assume that the model from [14] presented in Eq. 1 still holds for speech production as a function of communication distance and we treat the use of HPDs as a correction factor just as we did in Eq. 2 then the model becomes:

$$L_w = 59.54 + 2.96 \times \log_2\left(\frac{d}{1.5}\right) + n \times 0.125 \times (N - 60) \qquad (3)$$

where again n is 0 in quiet and 1 if the noise is greater than 60 dB and N represents the level of background noise. The three conditions, a speaker in quiet with open ears, a speaker in noise with open ears and a speaker in noise wearing HPDs, are compared in Fig. 3. This model would imply two assumptions:

1. In noise wearing HPDs does not greatly affect the speech production process as a function of the communication distance from the open-ear condition.
2. In quiet wearing HPDs would not affect the speech production process as a function of distance.

Based on the studies of speech production in noise while wearing HPDs the first assumption seems reasonable. However, intuitively, wearing HPDs might still alter the speech production process as a function of the communication distance, making assumption 2 invalid. The effects of communication distance and wearing HPDs in noise on the speech production process must be better studied. In the next section we present an experimental protocol to model the speech production process while wearing HPDs as a function of the background noise level, the speaker's vocal effort and the intended communication distance.

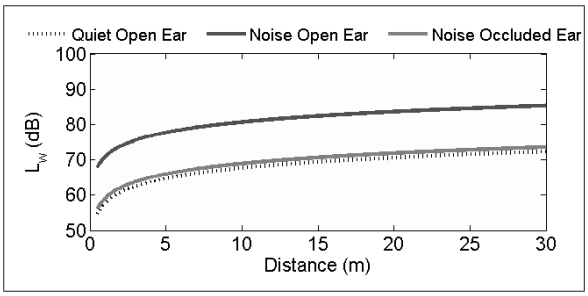

Figure 3. Comparing vocal power level of a speaker in quiet, a speaker in noise, and a speaker in noise wearing HPDs as a function of communication distance.

4. Proposed Experimental Protocol

To model the speech production process while wearing HPDs as a function of the background noise as well as the intended communication distance, an experimental protocol must be designed. Normal hearing individuals will be recruited to perform an instruction task. Each participant will be equipped with the intra-aural communication earpiece shown in Fig. 4. This communication earpiece is chosen for several reasons:

1. It is intra-aural, so it can be fitted into a participant's ear using different tips (roll-down foam plug, rounded flanged tips, malleable silicon wax, custom molded earpiece) and, thus, causing different levels of the occlusion effect.
2. It contains a microphone and miniature loudspeaker inside the ear as well as a microphone outside the ear.
3. It is the earpiece used for the radio-acoustical environment described in Section 1.

Having a miniature loudspeaker inside the ear allows us to play noise inside the ear directly. This leaves the speech signal captured with the outer-ear microphone free of noise and easier to process. The in-ear microphone can capture a noisy speech signal from inside the ear. This allows us to look at the difference in speech level between the outside and the inside of the ear and, in quiet, to measure the occlusion effect.

4.1. Experiment

Participants will be asked to give instructions to a listener in a corridor at 5 different communication distances: 0.3 m, 5 m, 10 m, 15 m, and 20 m. These distances were chosen to cover a wide range of distances and vocal efforts. Since the participants will be wearing HPDs the effects of reverberation can be ignored. At each distance, the speaker will be asked

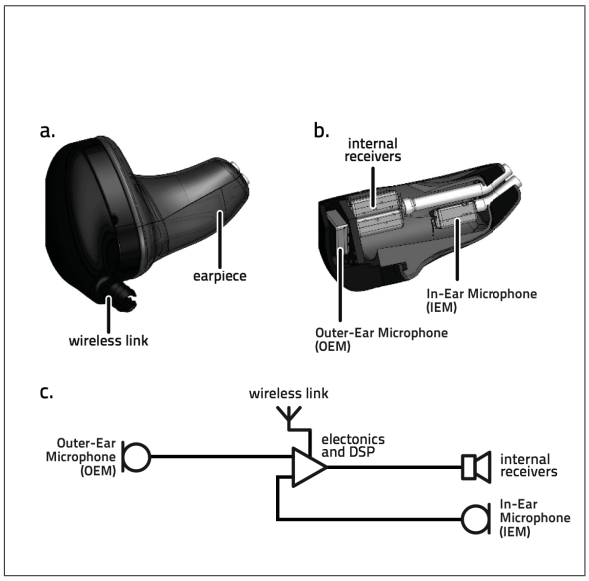

Figure 4. Intra-aural communication earpiece (a), its electroacoustical components (b), and equivalent schematic (c).

to instruct the listener to show him/her a color and a digit, 20 different times. The speaker will have 4 different colors (Red, Green, Blue, Yellow) to choose from and 10 different digits (0-9). The speaker can choose any combination he/she desires and can even repeat combinations. This is done so that the speech is natural and not read, mimicking a realistic situation. This procedure will be repeated for 5 different conditions: in quiet and in pink noise ranging from 60 dB to 90 dB at increments of 10 dB. Since the noise will be played directly inside the ear, only residual noise (after the passive attenuation of the plug) will be played. Once the participant is fitted with the earpiece the transfer function of the earpiece will be measured by playing white noise at a high level (∼85dB(SPL)) using a loudspeaker outside the ear which will be recorded using both the OEM and IEM. After the transfer function is found, the noise is filtered and played inside the ear. During the recordings the level of the speech, as well as the speaker's fundamental frequency, f_0, will be recorded at all conditions. The data is then collected from all the participants and a model will be found to fit it. This will give a relationship between vocal effort, background noise level and intended communication distance while wearing HPDs.

5. Conclusions

Communication is a key part of any workplace. Unfortunately, the use of currently available HPDs tends to affect communication. Modelling speech production while wearing HPDs as a function of the noise level and the intended communication distance can aid in alleviating the communication problem for personal radio systems. In this paper, we review the existing models of speech production in quiet, as a function of distance and in noise with and without the use of hearing protection devices. We also propose an experimental procedure to model speech production in noise while wearing HPDs to include the effects of the communication distance, which is currently not found in the literature.

Acknowledgement

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

References

[1] K. Mazur and J. Voix, "A case-study on the continuous use of an in-ear dosimetric device," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3274–3274, 2013.

[2] 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.

[3] J.C. Junqua, S. Fincke, and K. Field, "The Lombard effect: a reflex to better communicate with others in noise," *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pp. 2083–2086 vol.4, 1999.

[4] T. Fux, G. Feng, and V. Zimpfer, "Talker-to-listener distance effects on the variations of the intensity and the fundamental frequency of speech," *Cognition*, pp. 4964–4967, 2011.

[5] M. Schröder, "Emotional speech synthesis: A review," *Source*, vol. 1, pp. 2–5, 2001.

[6] J.B. Tufts and T. Frank, "Speech production in noise with and without hearing protection," *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 1069, 2003.

[7] J.G. Casali, M.J. Horylev, and J.F. Grenell, "A pilot study on the effects of hearing protection and ambient noise characteristics on intensity of uttered speech," *Trends in Ergonomics/Human Factors IV, edited by SS Asfour, Elsevier Science Pub.,(North-Holland)*, pp. 303–310, 1987.

[8] D. Byrne, *Influence of ear canal occlusion and air-conduction feedback on speech production in noise*, Ph.D. thesis, University of Pittsburgh, 2014.

[9] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *Journal of Speech, Language, and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[10] J.C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers.," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–24, Jan. 1993.

[11] M. Garnier and N. Henrich, "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?," *Computer Speech and Language*, vol. 28, no. 2, pp. 580–597, 2014.

[12] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise.," *The Journal of the Acoustical Society of America*, vol. 124, no. November 2008, pp. 3261–3275, 2008.

[13] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children.," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[14] D. Pelegrín-García, B. Smits, J. Brunskog, and C. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments.," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 1981–90, Apr. 2011.

[15] P. Zahorik and J. W. Kelly, "Accurate vocal compensation for sound intensity loss with increasing distance in natural environments.," *The Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. EL143–50, Nov. 2007.

[16] J. S. Liénard and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels.," *The Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–22, July 1999.

[17] 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.

[18] R. Navarro, "Effects of ear canal occlusion and masking on the perception of voice," *Perceptual and motor skills*, vol. 82, no. 1, pp. 199–208, 1996.

BIBLIOGRAPHY

3M^{TM}. 2012. *3M^{TM} Peltor^{TM} ORA TAC In-Ear Communications Headset*.

Abel, S. M., P. W. Alberti, C. Haythornthwaite, and K. Riko. March 1982. "Speech intelligibility in noise: effects of fluency and hearing protector type.". *The Journal of the Acoustical Society of America*, vol. 71, n° 3, p. 708–15.

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, et al. 1991. "The HCRC map task corpus". *Language and speech*, vol. 34, n° 4, p. 351–366.

Békésy, G. V. 1949. "The structure of the middle ear and the hearing of one's own voice by bone conduction". *The Journal of the Acoustical Society of America*, vol. 21, n° 3, p. 217–232.

Berger, E., 2003. *The Noise Manual*. 796 p.

Berger, E. H. and J. Voix. 2016. Hearing Protection Devices. *The Noise Manual*, p. 106 p. American Industrial Hygiene Association, ed. 6th Edition. 00177.

Bernier, A. and J. Voix. 2013. "An active hearing protection device for musicians". In *Proceedings of Meetings on Acoustics*. p. 040015. Acoustical Society of America.

Blauert, J., H. Els, J. Schr, et al. 1980. "A Review of the Progress in External Ear Physics Regarding the Objective Performance Evaluation of Personal Ear Protectors". In *Proceedings of INTER-NOISE 80*. p. 653–658. Noise-Control Found.

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

Bouserhal, R. E., T. H. Falk, and J. Voix. 2015a. "On the potential for artificial bandwidth extension of bone and tissue conducted speech: a mutual information study". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. p. 5108–5112. IEEE.

Bouserhal, R. E., J. Voix, and T. H. Falk. 2015b. "Modeling Speech Production in Noise for the Assessment of Vocal Effort for Use with Communication Headsets". p. 1633-1637.

Bouserhal, R. E., J. Voix, and T. H. Falk. 05 2016. "Device and Method for Improving the Quality of In-Ear Microphone Signals in Noisy Environments, USPTO 62/332861".

Bouserhal, R. E., E. N. Macdonald, T. H. Falk, and J. Voix. 2016a. "Variations in voice level and fundamental frequency with changing background noise level and talker-to-listener distance while wearing hearing protectors: A pilot study". *International journal of audiology*, p. 1–8.

Bouserhal, R. E., T. H. Falk, and J. Voix. 2016b. "Improving the Quality of In-Ear Microphone Speech Via Adaptive Filtering and Artificial Bandwidth Extension". *The Journal of the Acoustical Society of America*.

Bouserhal, R. E., T. H. Falk, and J. Voix. 2016c. "Modeling the talker-to-listener distance as a function of background noise level and speech level for talkers wearing hearing protection devices". *The Journal of Speech, Language, and Hearing*.

Brammer, A. J., G. Yu, D. R. Peterson, E. R. Bernstein, and M. G. Cherniack. 2008. "Hearing protection and communication in an age of digital signal processing: Progress and prospects". In *ICBEN International Congress on Noise as a Public Health Problem*. p. 1–9.

Brookes, M. et al. 1997. "Voicebox: Speech processing toolbox for matlab". *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*.

Brumm, H. and S. A. Zollinger. 2011. "The evolution of the Lombard effect: 100 years of psychoacoustic research". *Behaviour*, vol. 148, n° 11-13, p. 1173–1198.

Brummund, M. K., F. Sgard, Y. Petit, and F. Laville. 2014. "Three-dimensional finite element modeling of the human external ear: Simulation study of the bone conduction occlusion effect". *The Journal of the Acoustical Society of America*, vol. 135, n° 3, p. 1433–1444.

Byrne, D. 2014. "Influence of ear canal occlusion and air-conduction feedback on speech production in noise". PhD thesis, University of Pittsburgh.

Casali, J. G. 2010. "Passive Augmentations in Hearing Protection Technology Circa 2010 Including Flat-Attenuation, Passive Level-Dependent, Passive Wave Resonance, Passive Adjustable Attenuation, and Adjustable-Fit Devices: Review of Design, Testing, and Research". *Int J Acoust Vib*, vol. 15, n° 4, p. 187–195.

Casali, J. and E. Berger. 1996. "Technology advancements in hearing protection circa 1995: Active noise reduction, frequency/amplitude-sensitivity, and uniform attenuation". *American Industrial Hygiene Association*, vol. 57, n° 2, p. 175–185.

Casali, J., M. Horylev, and J. Grenell. 1987. "A pilot study on the effects of hearing protection and ambient noise characteristics on intensity of uttered speech". *Trends in Ergonomics/Human Factors IV, edited by SS Asfour, Elsevier Science Pub.,(North-Holland)*, p. 303–310.

Chennoukh, S., A. Gerrits, G. Miet, and R. Sluijter. 2001. "Speech enhancement via frequency bandwidth extension using line spectral frequencies". In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. p. 665–668. IEEE.

Concha-Barrientos, M., D. Campbell-Lendrum, and K. Steenland. 2004. "Occupational noise: assessing the burden of disease from work-related hearing impairment at national and

local levels". *Geneva, World Health Organization (WHO Environmental Burden of Disease Series, No. 9)*.

Davis, G. M., 2002. *Noise reduction in speech applications*, volume 7.

Dekens, T. and W. Verhelst. 2013. "Body Conducted Speech Enhancement by Equalization and Signal Fusion".

Dreher, J. J. and J. O'Neill. 1957. "Effects of ambient noise on speaker intelligibility for words and phrases". *The Journal of the Acoustical Society of America*, vol. 29, n° 12, p. 1320–1323.

Ellaham, N. N., C. Giguère, and W. Gueaieb. 2014. "A new research environment for speech testing using hearing-device processing algorithms". *Canadian Acoustics*, vol. 42, n° 3.

Falk, T. H., E. Sejdic, T. Chau, and W.-Y. Chan, 2010. *Spectro-temporal analysis of auscultatory sounds*.

Falk, T. H., V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie. 2015. "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools". *Signal Processing Magazine, IEEE*, vol. 32, n° 2, p. 114–124.

Fastl, H. and E. Zwicker, 2001. *Psychoacoustics, facts and models*.

Fux, T., G. Feng, and V. Zimpfer. 2011. "Talker-to-listener distance effects on the variations of the intensity and the fundamental frequency of speech". In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. p. 4964–4967. IEEE.

Gan, W. S., S. Mitra, and S. M. Kuo. 2005. "Adaptive feedback active noise control headset: implementation, evaluation and its extensions". *IEEE Transactions on Consumer Electronics*, vol. 51, n° 3, p. 975–982.

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

Garnier, M. and N. Henrich. 2014. "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?". *Computer Speech and Language*, vol. 28, n° 2, p. 580–597.

Garnier, M., M. Dohen, H. Loevenbruck, P. Welby, and L. Bailly. 2006. "The Lombard Effect: a physiological reflex or a controlled intelligibility enhancement?". In *7th International Seminar on Speech Production*. p. 255–262.

Garnier, M., N. Henrich, and D. Dubois. 2010. "Influence of sound immersion and communicative interaction on the Lombard effect". *Journal of Speech, Language, and Hearing Research*, vol. 53, n° 3, p. 588–608.

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

Giguère, C., C. Laroche, A. J. Brammer, V. Vaillancourt, and G. Yu. 2011. "Advanced hearing protection and communication : progress and challenges", . p. 225–233.

Giguère, C., C. Laroche, V. Vaillancourt, S. D. Soli, and S. M. Abel. 2008. "Modelling the Effect of Personal Hearing Protection and Communications Devices on Speech Perception in Noise". *Contract Report CR 2008-178 DRDC Toronto*.

Giguere, C., C. Laroche, V. Vaillancourt, and S. S. Soli. 2009. "A predictive model of speech intelligibility in noise for normal and hearing-impaired listeners wearing hearing protectors", vol. 2009, n° 6. p. 1013–1021.

Giguere, C., C. Laroche, V. Vaillancourt, and S. D. Soli. 2010. "Modelling speech intelligibility in the noisy workplace for normal-hearing and hearing-impaired listeners using hearing protectors". *International Journal of Acoustics and Vibration*, vol. 15, n° 4, p. 156.

Giguère, C., A. Behar, H. R. Dajani, T. Kelsall, and S. E. Keith. 2012a. "Direct and indirect methods for the measurement of occupational sound exposure from communication headsets". *Noise Control Engineering Journal*, vol. 60, n° 6, p. 630–644.

Giguère, C., C. Laroche, V. Vaillancourt, E. Shmigol, T. Vaillancourt, J. Chiasson, and V. Rozon-Gauthier. 2012b. "A multidimensional evaluation of auditory performance in one powered electronic level-dependent hearing protector". In *Proceedings of the 19th International Congress on Sound and Vibration*. p. 8–12.

Girard, S., T. Leroux, M. Courteau, M. Picard, F. Turcotte, and O. Richer. 2015. "Occupational noise exposure and noise-induced hearing loss are associated with work-related injuries leading to admission to hospital". *Injury Prevention*, vol. 21, n° e1, p. e88–e92.

Hager, L. D. et al. 2011. "Fit-testing hearing protectors: An idea whose time has come". *Noise and Health*, vol. 13, n° 51, p. 147.

Hansen, J. H. L. and V. Varadarajan. 2009. "Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition". *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, n° 2, p. 366–378.

Henry, P. and T. R. Letowski. 2007. *Bone conduction: Anatomy, physiology, and communication*. Technical report.

Honeywell International, I. 2016. *Honeywell, Safety Products QUIETPRO QP400*.

Huang, B., Y. Xiao, J. Sun, G. Wei, and H. Wei. 2014. "Speech enhancement based on FLANN using both bone-and air-conducted measurements". In *Asia-Pacific Signal and*

Information Processing Association, Annual Summit and Conference (APSIPA). p. 1–5. IEEE.

Hughson, G., R. Mulholland, and H. Cowie, 2002. *Behavioural studies of people's attitudes to wearing hearing protection and how these might be changed*.

IEEE . 1969. "Harvard Sentences". *IEEE Transactions on Audio and Electroacoustics*, vol. 17, n° IEEE Recommended Practices for Speech Quality Measurements, p. 227–46.

Iser, B. and G. Schmidt. 2008. Bandwidth extension of telephony speech. *Speech and Audio Processing in Adverse Environments*, chapter 5, p. 135–184.

ITU-R, R. 2001. "BS. 1534-1. Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)". *International Telecommunications Union, Geneva*.

ITU-T, R. 2011. "P. 863," Perceptual Objective Listening Quality Assessment (POLQA)". *International Telecommunication Union, CH-Geneva*.

Junqua, J.-C. jan 1993. "The Lombard reflex and its role on human listeners and automatic speech recognizers.". *The Journal of the Acoustical Society of America*, vol. 93, n° 1, p. 510–24.

Junqua, J.-C., S. Fincke, and K. Field. 1999. "The Lombard effect: A reflex to better communicate with others in noise". In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. p. 2083–2086. IEEE.

Katz, J., W. Gabbay, and D. Ungerleider, 1994. *Handbook of clinical audiology*.

Kondo, K., T. Fujita, and K. Nakagawa. 2006. "On equalization of bone conducted speech for improved speech quality". *Sixth IEEE International Symposium on Signal Processing and Information Technology, ISSPIT*, p. 426–431.

Kryter, K. 1946. "Effects of ear protective devices on the intelligibility of speech in noise". *The Journal of the Acoustical Society of America*, vol. 18, n° 2, p. 413–417.

Kuo, S. and D. Morgan. 1999. "Active noise control: a tutorial review". *Proceedings of the IEEE*, vol. 87, n° 6, p. 943–975.

Ladefoged, P., 1972. *Three areas of experimental phonetics: Stress and respiratory activity, the nature of vowel quality, units in the perception and production of speech*, volume 15.

Lane, H. and B. Tranel. 1971. "The Lombard sign and the role of hearing in speech". *Journal of Speech, Language and Hearing Research*, vol. 14, n° 4, p. 677.

Lehnert, H. and F. Giron. 1995. "Vocal communication in virtual environments". *Virtual Reality World*, p. 279–293.

Li, M., I. Cohen, and S. Mousazadeh. 2014. "Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones". In *IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*. p. 1–5. IEEE.

Liénard, J. S. and M. G. Di Benedetto. jul 1999. "Effect of vocal effort on spectral properties of vowels.". *The Journal of the Acoustical Society of America*, vol. 106, n° 1, p. 411–22.

Liu, Z., Z. Zhang, A. Acero, J. Droppo, and X. Huang. 2004. "Direct filtering for air-and bone-conductive microphones". In *6th Workshop on Multimedia Signal Processing*. p. 363–366. IEEE.

Liu, Z., A. Subramanya, Z. Zhang, J. Droppo, and A. Acero. 2005. "Leakage Model and Teeth Clack Removal for Air-and Bone-Conductive Integrated Microphones.". In *ICASSP (1)*. p. 1093–1096.

Lu, Y. and M. Cooke. 2008. "Speech production modifications produced by competing talkers, babble, and stationary noise.". *The Journal of the Acoustical Society of America*, vol. 124, n° November 2008, p. 3261–3275.

Manolakis, D., V. Ingle, and S. Kogon, 2005. *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, volume 46.

Martinek, R. and J. Zidek. 2010. "Use of adaptive filtering for noise reduction in communications systems". In *Applied Electronics (AE), 2010 International Conference on*. p. 1–6. IEEE.

Maurer, D. and T. Landis. 1990. "Role of bone conduction in the self-perception of speech". *Folia Phoniatrica et Logopaedica*, vol. 42, n° 5, p. 226–229.

Mazur, K. and J. Voix. 2013. "A case-study on the continuous use of an in-ear dosimetric device". *The Journal of the Acoustical Society of America*, vol. 133, n° 5, p. 3274–3274.

McBride, M., P. Tran, T. Letowski, and R. Patrick. 2011. "The effect of bone conduction microphone locations on speech intelligibility and sound quality". *Applied ergonomics*, vol. 42, n° 3, p. 495–502.

Nadon, V., A. Bockstael, D. Botteldooren, J. Lina, and J. Voix. 2015. "Individual monitoring of hearing status: Development and validation of advanced techniques to measure otoacoustic emissions in suboptimal test conditions". *Applied Acoustics*, vol. 89, p. 78–87.

Navarro, R. 1996. "Effects of ear canal occlusion and masking on the perception of voice". *Perceptual and motor skills*, vol. 82, n° 1, p. 199–208.

Neitzel, R. and N. Seixas. 2005. "The effectiveness of hearing protection among construction workers.". *Journal of occupational and environmental hygiene*, vol. 2, n° 4, p. 227–38.

Nilsson, M., H. Gustaftson, S. Andersen, and W. Kleijn. 2002. "Gaussian mixture model based mutual information estimation between frequency bands in speech". In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. p. I–525. IEEE.

NIOSH. 1998. "Occupational Noise Exposure".

NIOSH. 2005. *Advanced hearing protector study*. Technical report.

NIOSH. May 2015. "A story of impact: measuring how well earplugs work". <http://www.cdc.gov/niosh/docs/2015-181/>.

Ogata, S. and T. Shimamura. 2001. "Reinforced spectral subtraction method to enhance speech signal". In *TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*. p. 242–245. IEEE.

OSHA, U. S., 1983. *Occupational Noise Exposure: Hearing Conservation Amendment, Final Rule*.

Paliwal, K. and W. Kleijn. 1995. "Quantization of LPC parameters". *Speech Coding and Synthesis*, p. 433–466.

Park, K. and H. Kim. 2000. "Narrowband to wideband conversion of speech using GMM based transformation". In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. p. 1843–1846. IEEE.

Pelegrín-García, D., B. Smits, J. Brunskog, and C. Jeong. 2011. "Vocal effort with changing talker-to-listener distance in different acoustic environments.". *The Journal of the Acoustical Society of America*, vol. 129, n° 4, p. 1981–90.

Pörschmann, C. 2000. "Influences of bone conduction and air conduction on the sound of one's own voice". *Acta Acustica united with Acustica*, vol. 86, n° 6, p. 1038–1045.

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

Reddy, R. K., D. Welch, P. Thorne, S. Ameratunga, et al. 2012. "Hearing protection use in manufacturing workers: A qualitative study". *Noise and Health*, vol. 14, n° 59, p. 202.

Schröder, M. 2001. "Emotional speech synthesis: a review.". In *Proceedings of INTERSPEECH*. p. 561–564.

Seltzer, M., A. Acero, and J. Droppo. 2005. "Robust Bandwidth Extension of Noise-corrupted Narrowband Speech". *Interspeech 2005*, p. 1509–1512.

Sensear. 2016. *smartPlug USER GUIDE*. Sensear, ed. DOC00071 - Rev.00.

Shimamura, T. and T. Tamiya. 2005. "A reconstruction filter for bone-conducted speech". In *Circuits and Systems, 2005. 48th Midwest Symposium on*. p. 1847–1850. IEEE.

Shimamura, T., J. Mamiya, and T. Tamiya. 2006. "Improving bone-conducted speech quality via neural network". In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*. p. 628–632. IEEE.

Shin, H. S., H.-G. Kang, and T. Fingscheidt. 2012. "Survey of speech enhancement supported by a bone conduction microphone". In *Proceedings of Speech Communication; 10. ITG Symposium*. p. 1–4. VDE.

Studebaker, G. 1985. "Directivity of the human vocal source in the horizontal plane". *Ear and hearing*, vol. 6, n° 6, p. 315–319.

Subramanya, A., Z. Zhang, Z. Liu, and A. Acero. 2008. "Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling". *Speech Communication*, vol. 50, n° 3, p. 228–243.

Summers, W. V., D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. a. Stokes. 1988. "Effects of noise on speech production: acoustic and perceptual analyses.". *The Journal of the Acoustical Society of America*, vol. 84, n° 3, p. 917–928.

Sundberg, J. and M. Nordenberg. 2006. "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech.". *The Journal of the Acoustical Society of America*, vol. 120, n° 1, p. 453–457.

Tamiya, T. and T. Shimamura. 2004. "Reconstruction filter design for bone-conducted speech.". In *INTERSPEECH*.

Tat Vu, T., K. Kimura, M. Unoki, and M. Akagi. 2006. "A study on restoration of bone-conducted speech with MTF-based and LP-based models". *Journal of signal processing*, vol. 10, n° 6, p. 407-417.

Tat Vu, T., M. Unoki, and M. Akagi. 2008. "An LP-based blind model for restoring bone-conducted speech". In *Communications and Electronics, 2008. ICCE 2008. Second International Conference on*. p. 212–217. IEEE.

Titze, I. R. and J. Sundberg. 1992. "Vocal intensity in speakers and singers". *The Journal of the Acoustical Society of America*, vol. 91, n° 5, p. 2936–2946.

Tran, P., T. Letowski, and M. McBride. 2008. "Bone conduction microphone: Head sensitivity mapping for speech intelligibility and sound quality". *ICALIP 2008 - 2008 International Conference on Audio, Language and Image Processing, Proceedings*, p. 107–111.

Traunmüller, H. and A. Eriksson. 2000. "Acoustic effects of variation in vocal effort by men, women, and children". *The Journal of the Acoustical Society of America*, vol. 107, n° 6, p. 3438–3451.

Tufts, J. and T. Frank. 2003. "Speech production in noise with and without hearing protection". *The Journal of the Acoustical Society of America*, vol. 114, n° 2, p. 1069.

Tufts, J. B., M. A. Hamilton, A. J. Ucci, J. Rubas, et al. 2011. "Evaluation by industrial workers of passive and level-dependent hearing protection devices". *Noise and Health*, vol. 13, n° 50, p. 26.

Tufts, J. B., K. N. Jahn, and J. P. Byram. 2012. "Consistency of attenuation across multiple fittings of custom and non-custom earplugs". *Annals of occupational hygiene*, p. mes096.

Turan, M. and E. Erzin. 2013a. "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. p. 7049–7053. IEEE.

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

Varga, A. and H. Steeneken. 1993. "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, n° 3, p. 247–251.

Voix, J. and F. Laville. 2004. "Method and apparatus for determining in situ the acoustic seal provided by an in-ear device". *Journal of the Acoustical Society of America*, vol. 116, n° 1, p. 28.

Voix, J. and F. Laville. 2009. "The objective measurement of individual earplug field performance.". *The Journal of the Acoustical Society of America*, vol. 125, n° 6, p. 3722–32.

Voix, J., C. Le Cocq, and E. H. Berger. 2014. "Intra-subject fit variability using field microphone-in-real-ear attenuation measurement for foam, pre-molded and custom molded earplugs". *The Journal of the Acoustical Society of America*, vol. 136, n° 4, p. 2135–2135.

Yu, J., L. Zhang, and Z. Zhou. 2005. "A novel voice collection scheme based on bone-conduction". In *IEEE International Symposium on Communications and Information Technology (ISCIT)*. p. 1164–1168. IEEE.

Zahorik, P. and J. W. Kelly. nov 2007. "Accurate vocal compensation for sound intensity loss with increasing distance in natural environments.". *The Journal of the Acoustical Society of America*, vol. 122, n° 5, p. EL143–50.

Zannin, P. H. T. and S. N. Gerges. 2006. "Effects of cup, cushion, headband force, and foam lining on the attenuation of an earmuff". *International journal of industrial ergonomics*, vol. 36, n° 2, p. 165–170.

Zheng, Y., Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang. 2003. "Air- and bone-conductive integrated microphones for robust speech detection and enhancement". *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, p. 3–8.

Zollinger, S. A. and H. Brumm. 2011. "The Lombard effect". *Current Biology*, vol. 21, n° 16, p. R614–R615.