ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MANUSCRIPT-BASED THESIS PRESENTED TO
ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

BY
Narimene LEZZOUM

DEVELOPMENT OF ALGORITHMS FOR SMART HEARING PROTECTION DEVICES

MONTREAL, "16 MARCH 2016"

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS:

Mr. Jérémie Voix, Thesis Supervisor
Département de génie mécanique, École de technologie supérieure

Mr. Ghyslain Gagnon, Co-supervisor
Département de génie électrique, École de technologie supérieure

Mr. Chakib Tadj, President of the Board of Examiners
Département de génie électrique, École de technologie supérieure

Mr. Frederic Laville, Member of the jury
Département de génie mécanique, École de technologie supérieure

Mrs. Rebecca Reich, External Examiner
MITACS, Inc

Mr. Daniel Massicotte, Independant External Examiner
Département de génie électrique et génie informatique, Université du Québec à Trois Rivières

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON "9 FEBRUARY 2016"

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Jérémie Voix for the continuous support during my Ph.D, for his enthusiasm, motivation, and patience. Thank you Jérémie for helping me make this thesis happen. Of course, I would like to thank my co-supervisor, Professor Ghyslain Gagnon for his meticulous work, dedication and for always giving me challenges. Jeremie, Ghyslain, it has been an honour to be your first PhD student.

I would like to gratefully acknowledge the industrial research chair in in-ear technologies CRITIAS for its financial support during the four years.

I would like to thank the jury members for examining my thesis.

Thanks to all my colleagues and friends in the CRITIAS lab for making in a wonderful experience

Last but not least, I would like to thank my friends and family for their support and endless love.

"Parler est un besoin, écouter est un art."

—Johann Wolfgang von Goethe.

DÉVELOPPEMENT D'ALGORITHMES POUR PROTECTEURS AUDITIFS INTELLIGENTS

Narimene LEZZOUM

RÉSUMÉ

Dans les milieux industriels, le port de protecteurs auditifs est nécessaire pour protéger l'audition contre les bruits à niveaux élevés et prévenir la perte auditive. Évidemment, les protecteurs auditifs bloquent également d'autres types de signaux, même si ces derniers ne sont pas désagréables ou gênants pour la personne, mais plutôt utiles et commodes. De ce fait, si des personnes veulent communiquer entre elles et échanger des informations, elles doivent retirer les protecteurs, chose qui n'est pas très pratique, voire dangereux.

Afin de pallier aux problèmes rencontrés avec les protecteurs auditifs traditionnels passifs, le travail de cette thèse présente les étapes et le processus suivis pour le développement d'un nouveau type de protecteur auditif qui permet la protection contre les bruits extérieurs ainsi que la communication orale entre les usagers. Ce nouveau protecteur auditif est appelé le "protecteur auditif intelligent".

Le protecteur auditif intelligent est un protecteur auditif traditionnel dans lequel un processeur numérique du signal miniature est embarqué afin de traiter les signaux, en plus d'un microphone externe miniature pour capter les signaux et un haut-parleur interne miniature pour transmettre les signaux traités à l'oreille protégée.

Afin de permettre aux porteurs de protecteurs auditifs intelligents de communiquer sans enlever leurs protecteurs, des algorithmes de traitement du signal doivent être développés. Par conséquent, l'objectif de cette thèse consiste à développer un algorithme de détection d'activité vocale robuste dans les environnements à faible rapport signal/bruit ainsi qu'un algorithme de réduction du bruit afin d'améliorer la qualité et l'intelligibilité de la parole.

La méthodologie suivie pour le développement du protecteur auditif intelligent est divisée en trois étapes: en premier lieu, les algorithmes de détection de la parole et de réduction du bruit doivent être développés, en second lieu, ces algorithmes doivent être évalués et validés dans le logiciel, et en troisième lieu , ils doivent être implémentés dans le processeur numérique du signal pour valider leur faisabilité pour l'application visée.

Lors du processus de développement des deux algorithmes, des contraintes devaient être prises en compte et respectées. Ces contraintes sont dues au fait que le processeur numérique du signal embarqué dans le protecteur auditif soit limité en termes de ressources matérielles (mémoires, nombre d'opérations par seconde), et que le temps de traitement des algorithmes ne doit pas dépasser un certain seuil pour ne pas générer un délai entre la voie active et la voie passive du protecteur ou bien un délai entre le mouvement des lèvres et de la perception de la parole.

D'un point de vue scientifique, la thèse permet premièrement de déterminer les seuils que le processeur numérique du signal ne doit pas dépasser afin de ne pas générer un délai perceptible entre la voie active et la voie passive du protecteur. Ces seuils ont été obtenus par une étude subjective, où il a été trouvé que ce délai dépend de différents paramètres: (a) du degré d'atténuation du protecteur auditif, (b) de la durée du signal, (c) du niveau de bruit, et (d) du type de bruit dans lequel le signal est noyé. Cette étude a montré que lorsque le protecteur auditif offre une forte atténuation, 20% des participants commencent à percevoir un délai après 8 ms pour un signal de cloche (transitoire), 16 ms pour un signal de parole sans bruit, 22 ms pour un signal de parole noyé dans un bruit de type "babillage". Cependant, pour un protecteur auditif offrant une forte atténuation, il a été trouvé que le délai entre les deux voies est de 18 ms pour le signal de cloche, 26 ms pour le signal de parole sans bruit, et aucun délai lorsque la parole est noyée dans un bruit de type "babillage", montrant qu'une meilleure atténuation permettrait un temps plus grand pour le traitement numérique des signaux.

Deuxièmement, ce travail présente un nouvel algorithme de détection d'activité vocale dans lequel une caractéristique à faible complexité déterminant la présence de la parole a été extraite. Cette caractéristique a été calculée comme étant le rapport entre l'énergie du signal dans la bande fréquentielle qui contient le premier formant afin de caractériser le signal de parole, et les basses ou hautes fréquences pour caractériser les signaux de bruit. L'évaluation de cet algorithme et sa comparaison à un autre algorithme de référence a montré sa capacité de sélectivité avec un taux de faux positifs moyenné sur trois rapports signal/bruit (10, 5, et 0 dB) de 4.2% et un taux de vrais positifs de 91.4% comparé à 29.9% de faux positifs et 79.0% de vrais positifs pour l'algorithme de référence.

Troisièmement, ce travail montre que l'extraction de l'enveloppe du signal afin de générer un gain non-linéaire et adaptatif permet de réduire le bruit, améliorer la qualité du signal de parole et génère le moins de son musical comparé à trois autres algorithmes de référence.

Le développement des algorithmes de détection de parole et réduction du bruit, leurs évaluations objectives et subjectives dans différents types de bruits, ainsi que leurs implémentations dans des processeurs numériques du signal ont permis de valider leur efficacité ainsi que leur faible complexité pour l'application de protection auditive intelligente.

Mots clés: Protecteur auditif intelligent, détection d'activité vocale, réduction du bruit, amélioration de la qualité de la parole, processeur numérique du signal.

DEVELOPMENT OF ALGORITHMS FOR SMART HEARING PROTECTION DEVICES

Narimene LEZZOUM

ABSTRACT

In industrial environments, wearing hearing protection devices is required to protect the wearers from high noise levels and prevent hearing loss. In addition to their protection against excessive noise, hearing protectors block other types of signals, even if they are useful and convenient. Therefore, if people want to communicate and exchange information, they must remove their hearing protectors, which is not convenient, or even dangerous.

To overcome the problems encountered with the traditional passive hearing protection devices, this thesis outlines the steps and the process followed for the development of signal processing algorithms for a hearing protector that allows protection against external noise and oral communication between wearers. This hearing protector is called the "smart hearing protection device".

The smart hearing protection device is a traditional hearing protector in which a miniature digital signal processor is embedded in order to process the incoming signals, in addition to a miniature microphone to pickup external signals and a miniature internal loudspeaker to transmit the processed signals to the protected ear.

To enable oral communication without removing the smart hearing protectors, signal processing algorithms must be developed. Therefore, the objective of this thesis consists of developing a noise-robust voice activity detection algorithm and a noise reduction algorithm to improve the quality and intelligibility of the speech signal.

The methodology followed for the development of the algorithms is divided into three steps: first, the speech detection and noise reduction algorithms must be developed, second, these algorithms need to be evaluated and validated in software, and third, they must be implemented in the digital signal processor to validate their feasibility for the intended application.

During the development of the two algorithms, the following constraints must be taken into account: the hardware resources of the digital signal processor embedded in the hearing protector (memory, number of operations per second), and the real-time constraint since the algorithm processing time should not exceed a certain threshold not to generate a perceptible delay between the active and passive paths of the hearing protector or a delay between the lips movement and the speech perception.

From a scientific perspective, the thesis determines the thresholds that the digital signal processor should not exceed to not generate a perceptible delay between the active and passive paths of the hearing protector. These thresholds were obtained from a subjective study, where

it was found that this delay depends on different parameters: (a) the degree of attenuation of the hearing protector, (b) the duration of the signal, (c) the level of the background noise, and (d) the type of the background noise. This study showed that when the fit of the hearing protector is shallow, 20 % of participants begin to perceive a delay after 8 ms for a bell sound (transient), 16 ms for a clean speech signal and 22 ms for a speech signal corrupted by babble noise. On the other hand, when having a deep hearing protection fit, it was found that the delay between the two paths is 18 ms for the bell signal, 26 ms for the speech signal without noise and no delay when speech is corrupted by babble noise, showing that a better attenuation allows more time for digital signal processing.

Second, this work presents a new voice activity detection algorithm in which a low complexity speech characteristic has been extracted. This characteristic was calculated as the ratio between the signal's energy in the frequency region that contains the first formant to characterize the speech signal, and the low or high frequencies to characterize the noise signals. The evaluation of this algorithm and its comparison to another benchmark algorithm has demonstrated its selectivity with a false positive rate averaged over three signal to noise ratios (SNR) (10, 5 and 0 dB) of 4.2 % and a true positive rate of 91.4 % compared to 29.9 % false positives and 79.0 % of true positives for the benchmark algorithm.

Third, this work shows that the extraction of the temporal envelope of a signal to generate a nonlinear and adaptive gain function enables the reduction of the background noise, the improvement of the quality of the speech signal and the generation of the least musical noise compared to three other benchmark algorithms.

The development of speech detection and noise reduction algorithms, their objective and subjective evaluations in different noise environments, and their implementations in digital signal processors enabled the validation of their efficiency and low complexity for the the smart hearing protection application.

Keywords: smart hearing protection device, voice activity detection, noise reduction, speech quality improvement, digital signal processor.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

ADC	Analog to Digital Converter
AGC	Adaptive Gain Control
AMR	Adaptive Multi-Rate
ANOVA	Analysis of Variance
ANR	Active Noise Control
ARHL	Age-Related Hearing Loss
ARP	Auditory Research Platform
CRITIAS	Chaire de recherche industrielle en technologies intra-auriculaires Sonomax-ETS
DAC	Digital to Analog Converter
DFT	Discrete Fourier Transform
DNR	Digital Noise Reduction
DSP	Digital Signal Processor
ERB	Equivalent Rectangular Bandwidth
ETS	École de technologie supérieure
ETSI	European Telecommunication Standard Institute
FNR	False Negative Rate
FPR	False Positive Rate
GSM	Global System for Mobile
HATS	Head and Torso Simulator

HINT	Hearing in Noise Test
HPD	Hearing Protection Device
ICASSP	International Conference on Acoustics, Speech and Signal Processing
ICSV	International Conference on Sound and Vibration
IEEE	Institute of Electrical and Electronics Engineers
ISCA	International Speech Communication Association
ITU-T	International Telecommunications Union
JND	Just Noticeable Difference
LLR	Log Likelihood Ratio
MAC	Multiply-Accumulate
MB-DNR	Modulation Based Digital Noise Reduction
MFCC	Mel Frequency Cepstral Coefficients
MIPS	Millions of Instructions Per Second
NIDCD	National Institute on Deafness and Other Communication Disorders
NIHL	Noise Induced Hearing Loss
NIOSH	National Institute of Occupational Safety and Health
OSHA	Occupational Safety and Health Administration
PESQ	Perceptual Evaluation of Speech Quality
POLQA	Perceptual Objective Listening Quality Analysis
S-HPD	Smart Hearing Protection Device

SNHL	Sensorineural Hearing Loss
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
SRT	Speech Reception Threshold
TCAPS	Tactical Communications and Protection Systems
TNR	True Negative Rate
TPR	True Positive Rate
VAD	Voice Activity Detection
VOIP	Voice Over IP
ZCR	Zero Crossing Rate

INTRODUCTION

The studies contained in this doctoral thesis were conducted between June 2011 and December 2015 under the "Chaire de Recherche Industrielle en Technologies Intra-Auriculaires Sonomax-ETS (CRITIAS)" (Sonomax-ETS Industrial Research Chair in In-Ear Technologies). The main objective of this thesis was to enable face-to-face communication for wearers of electronic hearing protection devices. To achieve this objective, the development of low complexity speech-based algorithms that could be implemented into a digital hearing protection device (HPD) is required.

This chapter is organized as follows: Section one defines the problem. Section two and three describe the sub-objectives and scope of the thesis. Section five introduces the challenges and opportunities and Section six presents the thesis contributions.

0.1 Problem definition and context

Since the 17th century, the industrial revolution has led to a tremendous surge of invention of machines, motors and devices that can be used in different areas such as textile manufacturing, metallurgy, agriculture, transportations and leisure. While these machines, motors and devices mainly improved the standard quality of life, they contributed to the rise of some important drawbacks. One of the most important drawbacks, which is affecting human health and is considered as a dangerous environmental pollutant is **noise**.

According to the National Institute on Deafness and Other Communication Disorders (NIDCD, 2015), approximately 15 % of North Americans between the ages of 20 and 69 suffer from hearing loss due to noise exposure either at work or during leisure activities. Furthermore, the National Institute of Occupational Safety and Health (NIOSH, 1998) reported that occupational hearing loss is the most common work injury in North America with more than 22 million workers exposed daily to noise levels exceeding a 85 dBA time-weighted average (TWA) for an 8-hour work day, which corresponds to the limit of noise exposure in workplaces recommended by the Occupational Safety and Health Administration (OSHA, 1983).

Hearing loss originating from noise exposure is known as Noise Induced Hearing Loss (NIHL), and represents the second form of sensorineural hearing loss (SNHL) after presbyacusis (age-related hearing loss (ARHL)) (Bao *et al.*, 2013). Hearing loss is permanent, and the wearing of hearing aids currently represents the only solution for hearing-impaired people to improve the intelligibility of speech in noise, with somehow limited success since the speech intelligibility is not completely improved.

Although NIHL is currently an incurable hearing deficiency, it is 100% preventable. The best ways to prevent NIHL is to limit noise exposure either by controlling the noise at the source, controlling the transmission paths of the noise by adding an enclosure or isolating screens between the machine and worker, or limiting the duration of noise exposure administratively. While these solutions can be challenging and expensive, the wear of HPD represents the best and less expensive solution to protect the ear from hazardous noise levels.

Unfortunately, while offering protection from noise, traditional HPDs isolate the wearers from their environment and limit their communication abilities. Indeed, traditional HPDs block not only unwanted sounds such as background noise, they also block wanted sounds such as speech and warning signals. For instance, if an alarm is triggered, the HPD wearers may not perceive it, or if HPD wearers want to communicate orally, they have to remove their HPDs, which may be inconvenient in a work situation and downright dangerous since the repetitive removal of HPDs in high noise level environments may limit the cumulative overall efficiency of the HPD.

To facilitate oral communication for HPD wearers and keep them protected from the background noise, the present thesis relies on an electronic HPD that detects speech signals and transmits them to the protected ear, while continuing to block the external noise when no speech signal is present.

The envisioned digital HPD is an active (electronic) HPD with an embedded miniature digital signal processor (DSP) to process the incoming signals in addition to an external miniature microphone and an internal miniature loudspeaker to pick-up and transmit the signals to the protected ear. Figure 0.1 illustrates the hardware components embedded in this digital HPD.

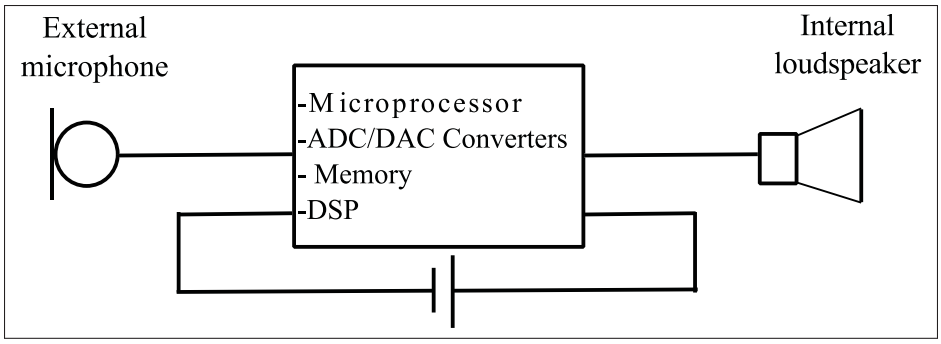

Figure 0.1 The hardware components embedded in the envisioned digital hearing protection device.

This hearing device was referred as the Smart HPD (S-HPD) in (Voix and Laville, 2005), and the steps followed for the development of its software components will be described in this thesis. Figure 0.2 illustrates its operating principle.

Contrary to other technologies that use wireless communication between the devices to transmit the signals via radio frequencies (Kvaloy *et al.*, 2007), the S-HPD does not use such a connection between devices, and is aimed to work as a face-to-face communication device.

0.2 Thesis objectives

The main goal of the S-HPD is to enable face-to-face communication among HPD wearers while keeping them protected from background noise. To do so, two speech-based algorithms are required, one for speech detection in noise and the other to reduce background noise without deteriorating speech. Thus, the two sub-objectives of this thesis involve:

- **Speech Detection**: To enable oral communication between S-HPD wearers, a speech detection algorithm also known as voice activity detection (VAD) has to be developed. Since the S-HPD is intended to work in industrial environments, the VAD has to be robust for low signal to noise ratio (SNR) environments.

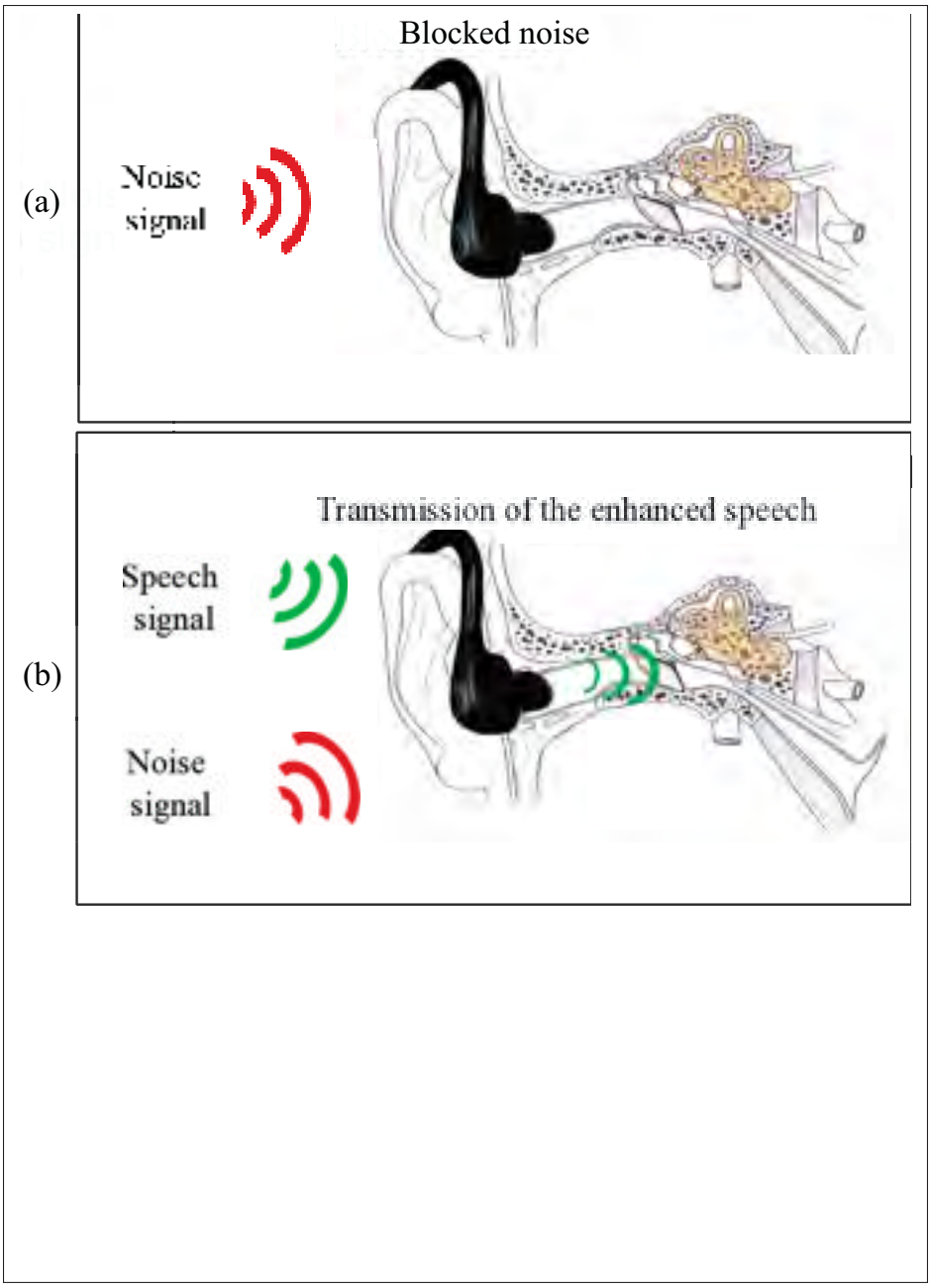

Figure 0.2 The operating principle of the smart hearing protection device: (a) block noise, (b) let enhanced speech and warning signals through.

- **Noise Reduction**: Once the speech signal is detected, the background noise needs to be reduced to enhance the speech's quality and intelligibility. For this purpose, a real-time and low-complexity noise reduction algorithm has to be developed.

Figure 0.3 illustrates the operating principle of the algorithms.

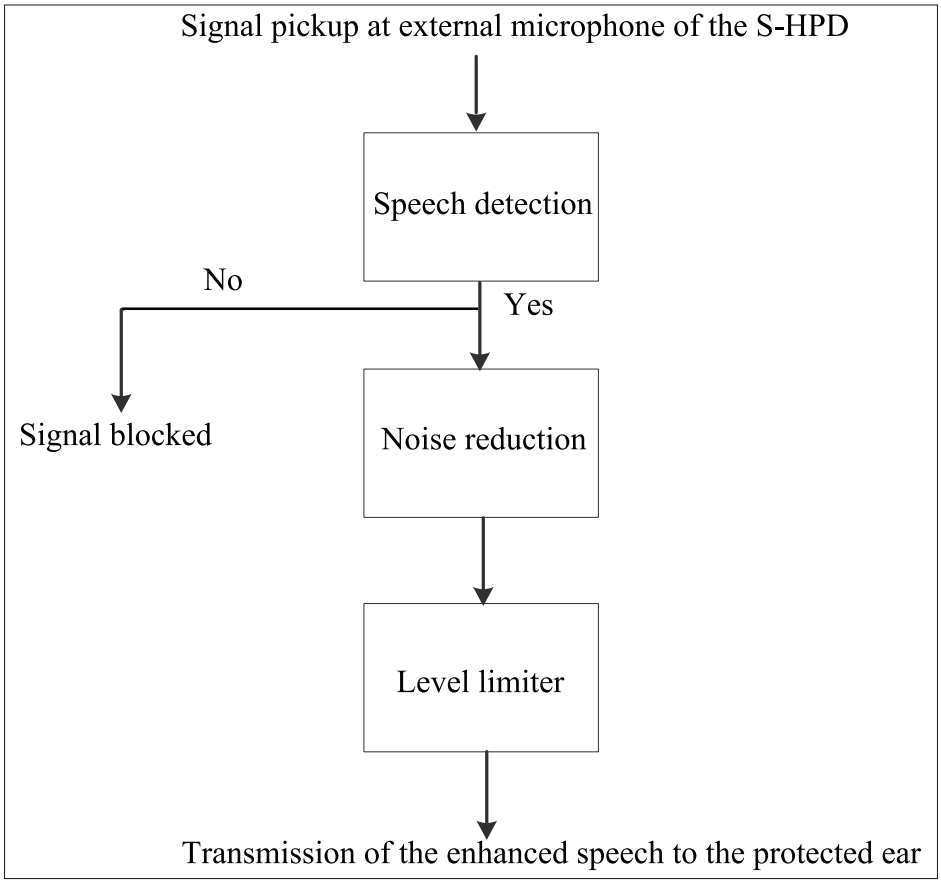

Figure 0.3 The operating principle of the algorithms.

Although warning signal detection is also needed to keep the HPD wearer aware if an alarm is triggered, the current thesis mainly focuses on speech communication in noise. In (Carbonneau *et al.*, 2013), a warning signal detection algorithm has been developed.

0.3 Scope of the thesis

The development the S-HPD is part of a larger project nicknamed “bionic ear”, which represents the next generation of in-ear protection and communication devices. The bionic ear will integrate smart hearing protection, hearing aid, and wireless communication. This multi-functional device is currently under development within CRITIAS.

The S-HPD is a combination of three different technological areas. Figure 0.4 illustrates a scheme that situates the S-HPD technology among the existing technologies. First, algorithms development for the S-HPD application requires principles from *hearing protection technologies* (attenuation, HPD transfer function, sound transmission paths, etc.). Second, since the first objective of the S-HPD will be the discrimination between speech and noise signals, principles used in *telecommunication technologies* such as voice activity detection are required. Third, when a speech signal is detected, it must be enhanced before it is transmitted to the protected ear to reduce the background noise and improve its quality and intelligibility. For this, concepts from *hearing aid technologies* are also required to develop a low complexity speech enhancement algorithm that can be implemented in the low power DSP.

0.4 Challenges and opportunities

0.4.1 Challenges

Developing a S-HPD that permits the detection and enhancement of speech signals in noisy environments is a challenging task. The three main challenges facing the development of the S-HPD algorithms are:

- **Real-time processing**: the S-HPD must process the incoming signals in real-time and transmit useful signals to the ear if present while protecting the wearers when only background noise is present. Above the real time processing, the developed algorithms must be able to process the incoming signals within time constraints so as not to create a perceptible difference between the passive and digital paths of the HPD, nor produce lip sync errors.
- **Low SNR robustness**: the S-HPD could be used in transportation (air-plane, train, etc.), in industrial environments, thus, in noisy environments where the SNR can be very low. Thus, the development of noise robust algorithms is needed. The VAD algorithm has to detect speech in low SNRs, while the speech enhancement algorithm needs to reduce the background noise with as little impact on the speech signal as possible. Another challenge

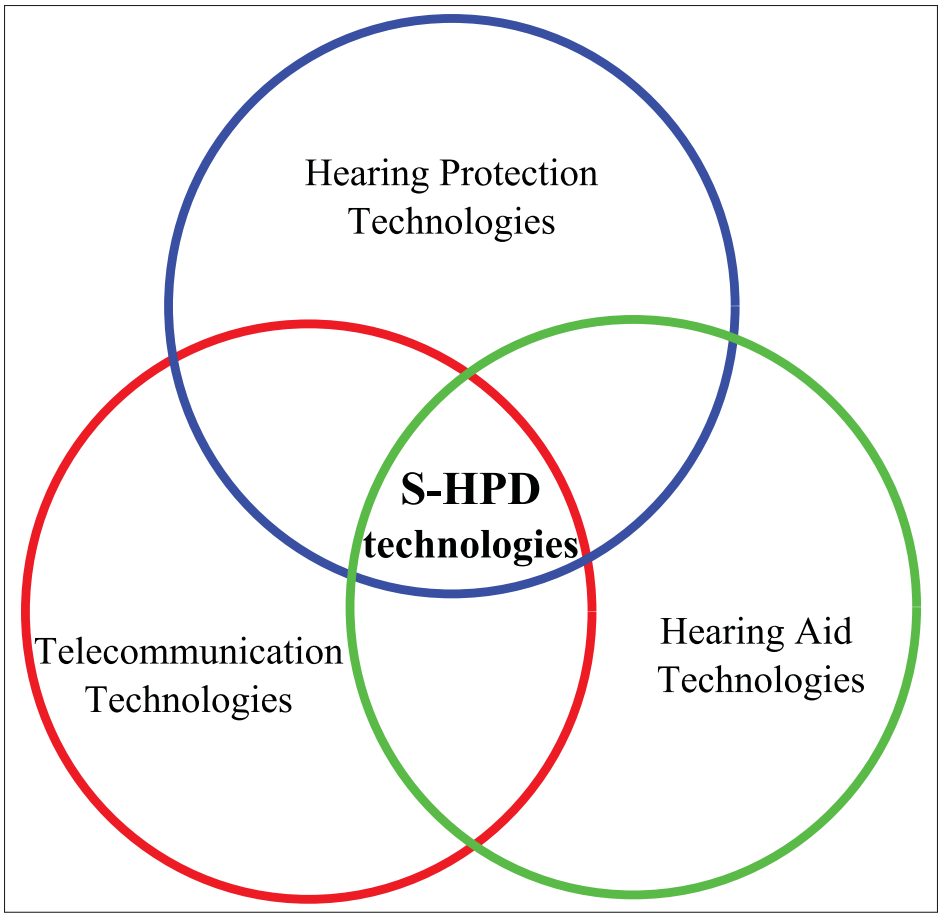

Figure 0.4 The combination of three technologies involved in the development of signal processing algorithms for the smart hearing protection application.

related to the robustness of the algorithms in low SNRs is the presence of only one channel (one external microphone) in the S-HPD, which does not provide spatial information such as the Direction Of Arrival (DOA) of the sound, and renders speech detection and noise reduction arduous.

- **Hardware constraints**: since a miniature DSP will be embedded in the traditional HPD to process the incoming signals, the algorithms have to take into consideration the DSP capabilities such as its memory capacity and the number of instructions per second.

These challenges are highly interdependent: for instance, while the literature presents different VAD algorithms (see chapter 1), these are either dedicated for high SNR environments and are of low complexity, or are dedicated for low SNR environments and are not suitable for low power DSPs or real-time applications. Likewise, various noise reduction algorithms for speech enhancement have been proposed in the literature, however, they either need knowledge of the background noise, or use long time frames to reduce the noise depending on some long-term features, which is not as suitable for the S-HPD real-time processing.

0.4.2 Opportunities

Although the development of algorithms for the S-HPD faces several challenges, today's emerging technologies are making it possible to overcome these challenges. With the rapid advancements in micro and nano-technology, miniature DSPs have become as powerful as small computers, performing millions of instructions per second (MIPS) and offering different functionalities that are optimized for real-time signal processing, such as the filterbank analysis/synthesis techniques that enable an incoming signal to be divided into different frequency bands without requiring a Fourier transform (Semiconductor, 2012).

0.5 Methodology

Three main steps are planned for the development of the S-HPD: first, the speech detection and noise reduction algorithms need to be developed, second, these algorithms need to be evaluated and validated in the software, and third they need to be implemented in the hardware and evaluated.

0.5.1 Algorithm development

The first step in the S-HPD algorithms development plan consists in developing the VAD and noise reduction algorithms. This first step will be done in Matlab^{TM} (Mathworks, MA) using

frame-based processing, that is, the incoming signals will be cut into frames upon which the processing will be performed.

Thus, the extraction of speech characteristics, the determination of decision rules that enable the discrimination between speech and noise segments and the reduction of noise with as little impact on the speech signal as possible are required.

0.5.2 Algorithm evaluation and validation

The evaluation of the VAD will be done using objective metrics such as the true positive rate (TPR) which corresponds to the number of speech frames detected as speech, the false positive rate (FPR), which corresponds to the number of noise frames detected as speech, the false negative rate (FNR) which corresponds to the number of speech frames detected as noise and the true negative rate (TNR) which corresponds to the number of noise frames detected as noise. In this step, the developed VAD algorithm will also be compared to another benchmark VAD.

The noise reduction algorithm will be evaluated in terms of speech intelligibility and speech quality using objective and subjective metrics. The subjective evaluation is very important for noise reduction algorithms since it reflects the human perception of the enhanced speech compared both to the noisy speech and to the performance of other noise reduction algorithms.

In this step, the subjective evaluation will be conducted using speech databases to which noise is artificially added with different SNRs.

0.5.3 Algorithm hardware implementation and system evaluation

Once the algorithms have been objectively and subjectively evaluated, they will be optimized and implemented within the selected hardware. In the optimization step, the different parameters used by the algorithms have to be optimized such as the frame length to avoid producing a perceptible delay between the passive and active paths of the S-HPD and the number of

instructions per second that the algorithms should not exceed to be implemented in the targeted hardware.

Thereafter, another real-time evaluation of the hardware-embedded algorithms needs to be performed to validate the S-HPD system. In this step, speech and noise signals played from loudspeakers in an audiometric booth and captured through the S-HPD microphone will be used to mimic a more realistic environment.

0.6 Contributions

The contributions of the current thesis can be divided in two parts: the scientific and technological contributions.

0.6.1 Scientific contributions

The scientific contributions consist of three journal articles in which the student is first author: one published and two submitted, all in international peer-reviewed journals. Two conference proceedings in which the student is first author were also presented in international peer-reviewed conferences. Finally, a peer-reviewed article, in which the student is second author, has also been published.

The three journal articles and two conference proceedings in which the student is first author are listed below:

- Journal article 1 entitled “Echo Threshold between Passive and Electro-Acoustic Transmission Paths in Digital Hearing Protection Devices” submitted to the International Journal of Industrial Ergonomics in March 2015. This journal article corresponds to chapter 2 and determines the time processing limits for the VAD and noise reduction algorithms to not generate an echo.

- Journal article 2 entitled “Voice Activity Detection System for Smart Earphones” published in the IEEE Transactions on Consumer Electronics in November 2014, Volume 60, Issue 4. This journal article corresponds to chapter 3 and presents the proposed VAD algorithm.

- Journal article 3 entitled “Noise reduction of speech signal using time-varying and multi-band adaptive gain control” accepted for publication in « Applied Acoustics» Elsevier Journal in March 2016. This journal article corresponds to chapter 4 and presents the proposed noise reduction algorithm.

- Conference proceeding article 1: “A Low-Complexity Voice Activity Detector for Smart Hearing Protection of Hyperacusic Persons” presented at the « Interspeech» conference which is the annual conference held by the International Speech Communication Association (ISCA), in 2013, Lyon, France. This article presents an early version of the proposed VAD algorithm using the inter-quartile range statistic feature compared to chapter 3, where a simpler energy-based feature is used for an efficient implementation in a low-power DSP.

- Conference proceeding article 2: “A Demonstration of a Single Channel Blind Noise Reduction Algorithm with Live Recordings” presented at the International Conference in Acoustics, Speech and Signal Processing (ICASSP) in 2014 in the show and tell session, Florence, Italy. This article presents an early version of the noise reduction algorithm, compared to chapter 4 where objective and subjective evaluations of the proposed algorithm are performed.

In addition to these contributions, during the four years of the thesis, the project and student received different awards and distinctions:

- Third place at the joint ACFAS (Association canadienne-française pour l’avancement des sciences) and RESMIQ (Regroupement Stratégique en Micro-systèmes du Québec) Conference in the poster and oral presentation competition 2014, Montréal, Qc, Canada.

- Best student reviewer at the ÉTS article reviewing competition (March 2014), Montréal, Qc, Canada.

- Second place for the best student poster competition, ÉREST (Équipe de recherche en santé et sécurité du travail) ÉTS, Montréal, Qc, Canada (2013 and 2014).
- First place in the ÉTS competition "Your Thesis in 180 Seconds", and finalist in the Canadian competition (2012).

In addition, an important hands-on experience has been acquired in an internship that has been completed by the student from January to August 2015 in the audio hardware department at Apple Inc in Cupertino, California.

0.6.2 Technological contributions

From a technological point of view, the scientific contributions lead to the development of the first S-HPD that transmits noise-reduced speech to the ear while keeping the wearer protected from background noise. This S-HPD can be used in industries, in the military, and other environments and for other cases where a hearing device is used such as hearing protections, earphones and headphones.

0.7 Outline of the thesis

This thesis is divided into five chapters and three appendices. The first chapter corresponds to a literature review, the second, third and fourth chapters correspond to three journal articles that describe the work that has been done, and chapter five concludes and synthesizes the work.

a. **Chapter 1** consists of a literature review for the development of the S-HPD. This literature review gathers principles from three technologies that will be used for the S-HPD development: hearing protection technologies (from passive HPDs to current active HPDs), since the S-HPD is a traditional HPD in which a DSP and transducers are embedded. Then a literature review on speech-based technologies used in telecommunications especially for speech detection in noise will be presented since the S-HPD will be mainly used

to detect speech in a noisy environment and transmit it to the protected ear. Thereafter, a literature review on hearing aid technologies will be approached, knowing that hearing aids embed a miniature DSP and help to enhance the quality and intelligibility of speech for the hearing impaired.

b. **Chapter 2** consists of the first journal article which presents a study conducted to determine the maximum delay that the S-HPD should not exceed so as not to create a perceptible difference between the passive and digital paths of the HPD. This study establishes the time processing limits for the VAD and noise reduction algorithms.

c. **Chapter 3** consists of the second journal article which presents the developed real-time VAD algorithm, its evaluation and hardware implementation within a miniature DSP. Although the current thesis is mainly focused on the hearing protection of workers, this article shows that the idea of embedding a miniature DSP and an external microphone in addition to the loudspeaker can also be done in traditional earphones/headphones to allow wearers to hear external speech signals such as public announcements or oral communication while listening to music, without removing their listening devices.

d. **Chapter 4** consists of the third journal article which presents the developed noise reduction algorithm and includes the description of the algorithm, its objective and subjective evaluation and its hardware implementation.

e. **Chapter 5** synthesizes the work, shows the achieved objectives, and presents the recommendations and future work for the S-HPD.

Three appendices are included in this thesis:

- Appendix I contains an article that was presented at INTERSPEECH 2013 (the annual conference held by the International Speech Communication Association (ISCA)). This article is entitled “A Low-complexity voice activity detector for smart hearing protection of hyperacusic persons” and presents an early version of the proposed VAD algorithm, in which the inter-quartile statistic feature is used.

- Appendix II provides an article that was presented in the International Conference in Acoustics Speech and Signal Processing (ICASSP) 2014 (Show and Tell session) entitled “A Demonstration of a Single Channel Blind Noise Reduction Algorithm with Live Recordings”. This article shows an early version of the noise reduction algorithm using objective evaluation and real-time demonstration.

- Appendix III introduces an article that was presented in the International Congress on Sound and Vibration 2015 (ICSV) entitled “Evaluation of a digital earplug featuring a multi-band adaptive gain control noise reduction algorithm for enhanced audibility in noisy environments”. This article presents the intelligibility evaluation of the noise reduction algorithm (it complements the article presented in chapter 4, which includes only the subjective and objective quality evaluation).

CHAPTER 1

SMART HEARING PROTECTION DEVICES: A LITERATURE REVIEW

1.1 Introduction

The first alternatives to prevent NIHL in workplaces lie in the limitation of noise exposure either by controlling the noise at the source, (when designing the machines, which requires acoustic design engineers in the product design stage), by controlling the noise transmission paths by adding isolating screens between the machine and worker, or by limiting the duration of noise exposure administratively. However, these first alternatives are costly since the design will vary from one machine to another, besides the fact that noise control at the source needs some yet undeveloped or un-used technology, or as in some cases, such as for professional musicians, the solution is still not optimal (Voix and Laville, 2005).

The other most common alternative to prevent NIHL, is to have exposed workers wear Hearing Protection Devices (HPDs) (Voix and Laville, 2005), (Berger and Voix, 2015).

The wearing of HPDs was first regulated by the United States Air Force in 1948, due to the fact that numerous soldiers from World War II returned home with hearing loss (Berger, 2000). In the late 1940s and early 1950s, hearing conservation programs were introduced in the industry such as in aviation and metal industries. In 1970, the Occupational Safety and Health Act was enacted, and in 1971, the noise standard was promulgated by OSHA. About ten years later, OSHA produced the hearing conservation amendment (OSHA, 1981), (OSHA, 1983) which details the rules of noise exposure and hearing conservation, such as the limit of noise exposure, which is of 85 dBA for an 8 hour work day.

While hearing protection devices protect the wearers from external noise, they hinder oral communication (Abel, 2008), (Burrell and Abel, 2009). For example, in a factory, if a worker wearing a pair of HPDs wants to communicate with his co-workers, he needs to remove the HPD, which is not convenient. Furthermore, this may expose the worker to high noise levels

and induce hearing loss. Moreover, workers may dismiss the use of hearing protectors temporarily or permanently if these compromise their safety and ability to communicate, all of which may induce hearing loss (Berger, 2010), (Hong *et al.*, 2008) .

In addition to the fact that they hinder oral communication, HPDs may also prevent wearers from hearing useful sounds such as sirens and warning signals, which may lead to accidents (Hong *et al.*, 2008), (Abel, 2008), (Carbonneau *et al.*, 2013), (Brammer *et al.*, 2015).

Studies conducted with workers exposed to high levels of noise such as industry workers, firefighters, and soldiers (Voix and Laville, 2005) (Hong *et al.*, 2008) (Abel, 2008) showed the need for more technologically advanced hearing protection that would allow wearers to exchange verbal face-to-face communication, hear if someone is calling in addition to keeping the wearer aware if a warning signal is triggered, while otherwise being protected from the hazardous noise.

For this purpose, the authors propose the development of a smart HPD (S-HPD) that enables protection against harmful noise and communication in high levels of noise. The S-HPD includes a miniature external microphone, a miniature internal loudspeaker and a miniature digital signal processor (DSP). In the DSP, real-time signal processing algorithms will be implemented. These algorithms consist of:

- Speech detection: speech detection allows S-HPD wearers to exchange oral communication without removing their devices.
- Speech enhancement: noise reduction for speech quality and intelligibility enhancement is very important to the S-HPD to improves speech quality and reduce the listening effort.

The current chapter presents a literature review of the three technological areas illustrated in Figure 0.4 upon which the development of the S-HPD will be based on. Section 1.2 “Hearing protection technologies” presents a literature review of traditional HPDs (passive and active) in addition to the existing advanced HPDs. Section 1.3 “Speech in telecommunication tech-

nologies" reviews speech-based technologies used in the field of telecommunications, while Section 1.4 "Hearing aid technologies" presents an overview of hearing aid technologies.

In addition, this chapter presents in section 1.5 a short literature review on the different assessment protocols that can be used to evaluate and validate the developed algorithms and subsequently the S-HPD.

1.2 Hearing protection technologies

This Section presents, in chronological order, the development of hearing protection technology. It starts by presenting passive HPDs that reduce the background noise by means of a material barrier, then active HPDs that embed basic electronic circuits (frequency limiter, etc.), and finally the recent HPDs that perform more advanced signal processing techniques.

In (Berger and Voix, 2015), a complete overview and comparison between the different types of HPD can be found.

HPDs come in various forms. There are earplugs, which must be placed within or against the entrance of the ear canal (insert, semi-insert) and are currently available in different sizes since ear canals can be different from one user to another, and are made with different materials from one manufacturer to another. In the last decade, custom-molded earplugs, which are made to adapt to the ear-canal by means of the material used, have increased in popularity in North America and Europe (Voix and Laville, 2009). Different techniques and materials can be used to generate this kind of earplug, such as in (Voix and Laville, 2009), where a soft medical-grade silicon rubber is injected between a rigid core and an expendable envelope.

HPDs can also come in the form of earmuffs, which fit around the ear, or in the form of helmets that encase the entire head.

Irrespective of the form, HPDs can be grouped into two categories depending on their operating mode: passive HPDs and active HPDs (Casali, 2010).

1.2.1 Passive HPDs

Passive HPDs reduce the background noise mechanically based on their shape and material composition. The amount of attenuation procured by passive HPDs (in the form of earplugs or earmuffs) depends on the material used and the shape (or length for earplugs) of the HPD. Some HPDs are designed for uniform attenuation over the frequency range, while others have a level-dependent attenuation.

Passive HPDs with uniform attenuation are designed to reduce the external signals by about 10 dB from 125 to 8000 Hz uniformly (Berger and Voix, 2015). The advantage of uniform HPDs is their identical attenuation across all the frequency components contrary to other HPDs which attenuate the high frequencies more than the low frequencies and which may cause a coloration of the passively transmitted sounds (Berger and Voix, 2015).

Passive HPDs with level dependent attenuation reduce highly transient noises when their level exceeds a certain threshold. This kind of HPD is more likely used in the military. It uses a non-linear component, such as a valve, diaphragm, or sharp-edged orifice in an earplug. It may also change the amount of attenuation by opening into a duct within an earmuff cup, which takes advantage of the fact that low-level sound waves predominantly exhibit laminar airflow and pass relatively unimpeded through the aperture, whereas high-level waves involve turbulent flow and are attenuated to a greater extent with increased acoustic resistance ((Allen and Berger, 1990); (Berger and Hamery, 2008), (Berger and Voix, 2015)).

1.2.2 Active HPDs

Active HPDs are also known as electronic HPDs and are equipped with a powered digital or analog circuit to process the external sounds before their transmission/obstruction to the protected ear. (Casali, 2010) divided active HPDs into four types, three of which will be described here after: Active Noise Reduction (ANR) devices, Electronically-Modulated Sound Transmission Devices and Electronic Tactical Communications and Protection Systems (TCAPS). The fourth type includes HPDs with dosimetry measurements.

The main and most popular task that current Active HPDs offer is Active Noise Reduction (ANR) (also known as Active Noise Control or Noise Cancellation), which can be done using analog or digital circuits and consists of cancelling the external sound wave by adding to it the same sound wave but phase inversed. ANR technology is currently becoming an important functionality not only in hearing protection, but also in communication systems such as headphones and wireless communication devices (Brimhall *et al.*, 2002).

The Electronically-Modulated Sound Transmission Devices incorporate a limiting/amplifier, in addition to an external microphone to pick-up the sound, and an internal loudspeaker to transmit the limited-amplified signals to the protected ear. These active HPDs can be used to boost certain frequency ranges in the incoming signal before it is transmitted to the wearer, such as the critical frequency range of speech or other useful signals.

The TCAPS which are also part of active HPDs since they incorporate transducers and an electronic circuit, are mainly used in the military domain. They feature either passive or active hearing protection in addition to radio communication. In this type of HPD, voice pick-up is performed using a microphone located either in front of the wearer's mouth, on a throat-mounted fixture, in the ear canal, or within a bone-conduction pick-up held against the skull.

A complete study on passive and active HPDs can be found in (Casali, 2010).

1.2.3 Advanced electronic HPDs

Recently, advanced hearing protection devices have been developed. They are currently being manufactured. Among these advanced devices are smart earplugs from Sensear Pty Ltd (Belmont, Australia) that perform noise reduction to enhance the quality of speech and improve its intelligibility. However, these smart earplugs do not block the external signal if no speech is present, but only reduce it to keep the wearer aware of the external environment. The continuous transmission of processed signal to the ear is annoying the wearer and may yield to an excess in noise exposure since the wearer can increase and reduce the volume of the processed sound.

Other advanced HPDs have been developed such as the Nacre QUIETPRO Intelligent Hearing System (Trondheim, Norway) which plays the role of: an amplifier when the external sound is low to transmit to the ear, the role of a level limiter when the level increases, and noise canceller (ANR) when the level of the external noise exceeds a certain threshold. However, since these HPDs act like an active noise canceller in high levels of noise, they cannot detect useful signals such as speech or warning signals, which may hinder safety and face-to-face oral communication.

1.2.4 Summary

Passive HPDs attenuate the external sounds by means of a material barrier, active HPDs cancel the external sounds electronically or amplify a certain frequency range of the sound and advanced HPDs enhance the speech and transmit the reduced noise when no speech is present. The S-HPD will enable protection when no speech signal is detected by acting as a passive HPD, and face-to-face oral communication when speech is present, and this, by transmitting the enhanced speech signal to the ear within a certain limited level so as not to cause hearing loss due to the transmission of high levels of sound.

1.3 Speech in telecommunication technologies

For the purposes of the current thesis, telecommunication technologies involving speech, and more particularly speech detection in noise, are investigated since the techniques used in telecommunication are embedded in DSP and work in real-time such as in mobile telephony.

Speech represents the first tool that humans use to communicate with each other. In the last decade, speech has also been used by humans to communicate with machines that are provided with speech technology such as automatic speech recognition (ASR) (O'Shaughnessy, 2000).

The following subsections present an overview of speech characteristics, speech analysis, and a literature review of speech detection techniques.

1.3.1 Basics of speech production

A speech signal is produced by a speaker in the form of a wave that propagates through the air to reach the ear of the listener. The characteristics of the speech signal vary depending on the movement of the articulators which are located in the vocal tract (O'Shaughnessy, 2000). Among these articulators we have: the lips, the tongue, the teeth, the palate, the jaw, etc.

The speech production process can be seen as a filtering operation in which a sound source excites the filter. In the human speech production system, this filter corresponds to the vocal tracts and helps to attenuate and amplify certain frequencies (modulate the sound). The sound source or (exciter) can be periodic and generate a voiced speech signal, or aperiodic and generate an unvoiced speech signal. Voiced speech is generated in the larynx where the air is interrupted by the vibrations of the vocal cords, which open and close periodically, whereas for unvoiced speech, the air is not interrupted by the vibration of the vocal cords, but passes through the vocal tract directly. Unvoiced speech is a random signal similar to a noise signal modulated by the vocal tracts.

Speech sounds are divided into two types: (a) the vowels, which are voiced sounds and have high energy, and (b) the consonants which can be voiced or unvoiced, and have lower energy compared to vowels. The consonants are divided into three types: fricatives, stops and nasals. The voiced sounds have a spectrum that consists of a fundamental frequency that is perceived as the pitch, in addition to the harmonics that are multiples of the fundamental frequency.

As illustrated in Figure 1.1, in the spectrum of a voiced speech there are peaks (maxima), which are called formants and are different from one speech sound to another (O'Shaughnessy, 2008). Formants therefore characterize the spectrum of the speech signal by their high energy.

1.3.2 Speech analysis

Speech analysis is the determination and extraction of relevant speech information. Thus speech analysis enables the characterization of speech signals and their comparison to other

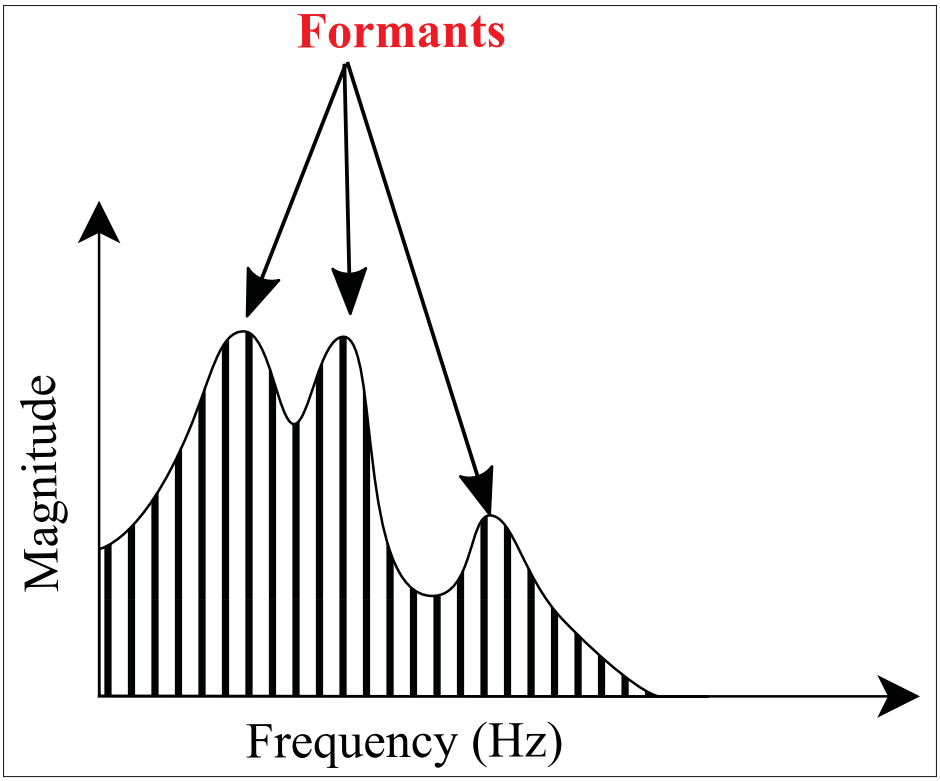

Figure 1.1 Spectral representation of the vowel /a/ showing the formants.

signals. The analysis of speech signals may take place in the time domain, in the frequency domain, or in the time and frequency domains simultaneously, depending on the needs, applications and resources. In most speech-based applications, the signal must first be cut into successive analysis segments or frames, in order to extract the short time information that it carries. Depending on the application, speech processing can also be done in a sample-based approach, which means, instead of dividing the signal into different frames, the digital signal can be processed sample by sample so that each processed sample can be sent to the output. For example, frame-based processing is used when the Fourier transform is needed to characterize the signal, and the sample-based approach is used in applications that require a simple time domain filter.

Speech analysis can be done using different features, the following sections introduce some features used in speech detection. Other features such as the Teager Energy Operator (TEO)

(Jabloun and Çetin, 1999), the entropy (Shen *et al.*, 1998), MFCC (Mel Frequency Cepstral Coefficients) (Davis and Mermelstein, 1980), and LPC (Linear Prediction Coefficients) (Atal, 1974) are not addressed.

1.3.2.1 Time domain speech analysis

The speech signal is a dynamic signal that varies over time. Its analysis in the time domain is simple and low cost. There are several temporal speech characteristics currently used in various applications. Among these characteristics we can mention:

- **Energy**: The energy of the signal is calculated by the following equation:

$$E = \sum_{i=1}^{k} x(i)^2 \tag{1.1}$$

  The calculation of the energy has been widely used in speech-based applications, such as speech recognition, where the energy is calculated in several speech frames to recognize syllables, or discrimination between a voiced and unvoiced sound, or simply to determine the end of the speech signal when produced in silence.

- **Zero Crossing Rate**: The zero crossing rate (ZCR) is calculated when the signal changes its algebraic sign. ZCR is one of the basic methods for calculating the fundamental frequency $F0$ of the signal (as shown by equation 1.2), which means that this characteristic can provide frequency information without passing by the frequency domain (Fourier transform).

$$F0 = \frac{ZCR \times Fs}{2} \tag{1.2}$$

  This characteristic can also be used to discriminate between voiced and unvoiced speech: a high ZCR corresponds to an unvoiced sound, and a low ZCR corresponds to a voiced

sound. Likewise, the ZCR has been widely used in voice activity detection algorithms by assuming that the ZCR of a noise signal is above a certain threshold, while the ZCR of speech is below that threshold.

- **Autocorrelation**: The short-term autocorrelation is computed as follows:

$$\mathbf{R}_{xx}(k) = \sum_{i=1}^{k} x(i)x(i-k) \tag{1.3}$$

The maximum value of the Autocorrelation represents the energy of the signal and is located at $k = 0$. Thus $R_{xx}(0)$ represents the energy of the signal. If the signal is periodic with a period P, the maximum of R_{xx} will be located at $k = 0$, P, $2P$, etc. This analysis method allows us to determine if a signal is periodic (voiced) and to calculate its period (fundamental frequency).

1.3.2.2 Frequency domain speech analysis

Even though the analysis of speech in the time domain is very simple and has a low complexity, the analysis in the frequency domain is used more often since it gives access to the different frequency components of the speech signal. Some of the frequency domain analysis methods are listed below:

- **The Discrete Fourier Transform**: The discrete Fourier transform (DFT) is used to represent the amplitude and the phase of a signal in the frequency domain. Therefore, the DFT informs on the frequency characteristics of the signal: the fundamental frequency, the harmonics, the formants, and other more speech-specific characteristics such as the MFCC (Mel Frequency Cepstral Coefficients) which are extracted from the spectrum.

- **Analysis using filterbanks**: An analysis with a filterbank is used in digital speech processing to divide the signal into different frequency bands using several filters to analyse each frequency band separately. This analysis method is simple and can be used to extract the

frequency components of the signal without performing the DFT. Currently, filterbanks are optimized and implemented within miniature DSPs to perform real time frequency analysis without going through the DFT (ON Semiconductors, 2009).

The filterbank characteristics (bandwidth, cut-off frequencies, order, etc.) can be different from one application to another. In speech coding and transmission applications, for example, the octave and third octave bands have been widely used in addition to other well-defined critical bands (E. Zwicker, 1961). In hearing aid applications, the Bark scale, which contains 24 frequency bands is then used.

Nowadays, different types of frequency bands have been developed to mimic the auditory system such as the the Gammatone filterbank (Aertsen and Johannesma, 1981) and the equivalent rectangular bandwidth (ERB) (Glasberg and Moore, 1990).

1.3.2.3 Time-Frequency speech analysis

Speech analysis can also be performed jointly in the temporal and spectral domains using Time-Frequency Representation (TFR). Among the TFR techniques, the Wavelet Transform (WT), which correspond to a windowing technique with variable-sized frames: It uses long time frames when low-frequency information is needed, and short time frames when high-frequency information is needed.

The WT has been used in different applications such as speech enhancement, speech recognition and speech detection. For example, in speech enhancement, wavelet shrinkage is used to denoise the signal based on the thresholding of the wavelet coefficients: the wavelet coefficients lower than a certain threshold are set to zero with the assumption that the speech coefficients dominate the noise coefficients. Although the use of wavelet for speech detection and enhancement has been widely used (Bahoura and Rouat, 2001) (Stegmann and Schroder, 1997) and (Chen *et al.*, 2010), this technique is not suitable for the targeted application due to the limited resources available in the DSP.

1.3.3 Speech in noise

Although speech technology has seen important advancements, it still represents a challenging task when occurring in a noisy environment. The higher the noise level is, the lower the SNR and the harder it is for communication and speech intelligibility, for humans and machines alike.

To increase the intelligibility of their speech in high levels of noise, speakers tend to increase the SNR by augmenting their vocal effort (Pickett, 1958). However, there is a limit at which the intelligibility of speech starts to reduce when the level of speech increases to shouting (Stedmon, 1997).

Since the human hearing mechanism is more effective with low and moderate levels of sound, it has been found that passive HPDs help the understanding of speech in noise when the level of noise is at 85 dBA or greater. This phenomenon is due to the fact that in such environments, distortions are introduced in the cochlea which reduce the clarity of speech (Berger, 2010). Thus the wearer of the HPD will reduce the level of the passively transmitted signal and thus better understand the speech utterance. This phenomenon has been assimilated to the phenomenon experienced when wearing sunglasses in intense sunlight (Berger, 2010). However, when the level of the environmental sound becomes lower than 85 dBA, speech understanding begins to decrease.

Consequently, speech understanding depends on the level of the background noise in addition to the HPD's attenuation. For this purpose, the development of an algorithm that detects speech in noisy environments to transmit it to the HPD wearer is more suitable for workers and people exposed to high noise levels.

1.3.4 Voice activity detection

Nowadays, almost all the existing speech-based applications rely on speech detection algorithms to discriminate between speech and non-speech segments. In the literature, speech de-

tection is known as Voice Activity Detection (VAD). The first VAD algorithm with widespread usage was developed for mobile telephony in 1989 (Freeman *et al.*, 1989) for the pan-European digital cellular mobile telephone service (Global System for Mobile: GSM). The principle behind the need for VAD within a mobile telephony application was to detect non-speech segments which could then be encoded with lower bit rates, relative to speech segments, thus lowering the overall transmission rate. Since then, more advanced VAD algorithms have been proposed, such as the one used in the International Telecommunications Union (ITU-T) Recommendation G.729b in 1996 (ITU T, 1996) and the one used in the adaptive multi-rate (AMR) codec from the European Telecommunication Standard Institute (ETSI) (ETSI, 1999). The latter VAD system is based on a combination of different speech features and exemplifies one of the major shifts from the earlier generation energy- or periodicity-based solutions.

After becoming mainstream within mobile telephony, the use of VADs expanded to other speech-based applications, such as speech recognition for human/machine interaction applications (Chuangsuwanich and Glass, 2011) and personal digital assistants in mobile environments (Lee and Yook, 2009). Within these more recent applications, VAD is performed to reduce false alarm rates due to the use of noise segments in the recognition process. In addition, in hearing aids and cochlear implants (Chung, 2004) (Bentler and Chiou, 2006) (Cornelis *et al.*, 2011), speech detection is also performed before noise reduction, an important step needed to assure intelligible speech is transmitted to the user.

Over the years, different methods have been used for voice activity detection. Earlier generations of VAD algorithms were based on the estimation of an inverse filter using the speech signal's first frames (Freeman *et al.*, 1989) or on the extraction of a periodicity measure (Tucker, 1992). Later, more advanced VAD algorithms were based on the combination of different speech features such as zero-crossing rate, energy, and pitch (ITU T, 1996) (ETSI, 1999). This notwithstanding, these VADs are known for their inefficiency in low signal to noise ratios (SNRs) (Beritelli *et al.*, 2002) where they tend to detect speech as well as noise. To overcome this problem, statistical VADs have been developed, such as Sohn's VAD (Sohn *et al.*, 1999), which relies on the estimation of the *a posteriori* and *a priori* SNR using the signal's first few

frames, which are assumed to contain only noise. Afterwards, other statistical methods have been proposed such as Davis et al.'s VAD (Davis *et al.*, 2006). To increase their robustness against low SNRs and non-stationary noise environments, some VAD algorithms use a noise reduction algorithm at their front end. For example in (Lei *et al.*, 2009), the Wiener filter is used before the VAD.

Recently, pattern recognition tools, such as hidden Markov models, support vector machines and neural networks, have also been applied for voice activity detection (Liu *et al.*, 2010), (Wu and Zhang, 2011), (Saon *et al.*, 2013), and (Segbroeck *et al.*, 2013), in order to increase the efficiency of VADs in pre-known environments. Furthermore, modulation characteristics such as modulation frequency and modulation depth are commonly used nowadays to distinguish between speech and noise knowing that the speech signal is characterized by modulation frequency content between 4 and 16 Hz (Drullman *et al.*, 1994b), (Evangelopoulos and Maragos, 2006).

Despite the progress that this field has seen, detection of speech in low SNRs and non-stationary noise environments is still a challenging and rather unsolved task. This task becomes even more complex once VADs become embedded in applications with limited hardware resources, such as in the S-HPD.

1.3.5 Summary

This review showed the basics of speech production and analysis. It also highlighted the pros and cons of the existent speech detection techniques, and showed that speech detection in noise is a very challenging task especially when the information on the background noise is not known, when the SNR is low, and when the VAD algorithm needs to be implemented in a low power DSP.

1.4 Hearing aid technologies

In this Section, hearing aid technologies are presented, from the analog hearing aids that enable the amplification of the external signals to the current digital hearing aids that perform advanced noise reduction to enhance the quality and intelligibility of speech and increase the listening comfort for the hearing impaired.

1.4.1 Hearing aid principles

As a consequence of excessive noise exposure, NIHL is an irreversible disease. While the intelligibility of speech signals in a noisy environment is already a difficult task for normal hearing people, it is much more difficult for the hearing impaired. In noisy environments, the speech signal must be 30 dB higher than the speech signal for persons with hearing loss than for a person with normal hearing for them to perceive it in the same way (Baer and Moore, 1994).

Currently, the only solution for hearing impaired people is to wear hearing aids. Primarily, hearing aids have been developed to amplify the incoming signals, and more recently, they have been developed to increase speech understanding and listening comfort in noisy environments where speech tends to be masked by noise.

In the last 20 years, developments in hearing aid technology have been growing and evolving with the progress in circuits technology. Likewise to Moore's law which stipulates that the number of transistors in an integrated circuit will double each 18 months, meaning that the speed of computers doubles approximately each one year and a half, advancements in hearing aids is also following the same path since the functionalities of a hearing aid depend first and foremost on the capabilities of the integrated circuit that it embeds.

1.4.1.1 Analog hearing aids

In the early 1900, electrical hearing aids embedding an analog circuit were developed. In these first generation hearing aids, an external microphone was used to pick-up the audio signal, then

send it to the analog filter to modify the frequency components of the signal to reduce the effect of background noise, by for example, using a low-pass filter to reduce the low frequency components assimilated to background noise, then send the filtered signal to the ear. Later, analog hearing aids performed adaptive filtering, adaptive compression and other frequency dependent compression techniques (Bentler and Chiou, 2006). Thereafter, analog programmable hearing aids were developed. These hearing aids contained a digital control circuit in which some parameters are programmed in order to set the amplification or compression variables.

1.4.1.2 Digital hearing aids

With the growth of technology, digital hearing aids came to be. These circuits enable the sampling, quantization, and digital conversion of the incoming signals before processing them. Thus, processing occurs in the digital domain using digital filters and algorithms, then the processed signals are converted to the analog domain using a digital to analog converter and transmitted to the ear.

Currently, digital hearing aids perform advanced signal processing techniques for the following purpose: improve speech quality and increase speech intelligibility and listening comfort. These signal processing techniques include noise reduction and adaptive dynamic range compression (Blamey, 2005).

Since the main objective of the current thesis is the development of signal processing algorithms for the S-HPD to be used by normal hearing people, the following Section will only review noise reduction methods used in hearing aids.

1.4.2 Noise reduction in hearing aids

Noise reduction is performed to reduce the background noise, increase listening comfort, and thus, increase speech understanding for the hearing impaired (Kuk *et al.*, 2002), (Mueller *et al.*, 2006) (Bentler *et al.*, 2008). When speech is corrupted by noise, the frequency components of

noise may overlap the frequency components of speech which leads to frequency masking of speech by noise.

In order to perform noise reduction, all hearing aids manufacturers follow the same steps:

- Signal detection,
- Signal analysis,
- Decision and noise reduction.

The detection stage is generally composed of several sub-blocks. The first block consists of dividing the signal into different frequency bands. Frequency-band processing is currently used by all hearing aid manufacturers to analyse each frequency band independently and extract the relevant characteristics for speech. These characteristics are subsequently compared to reference characteristics to make a decision.

In addition to the use of the same steps to perform noise reduction, almost all hearing aid manufacturers use modulation characteristics to detect speech and reduce the background noise. Figure 1.2 shows an example of the envelope of a speech signal.

This noise reduction technique is known as modulation-based digital noise reduction (MB-DNR) (Lamm *et al.*, 2011). The differences between various brands of hearing aids lie in the parameters used for speech detection such as the number of frequency bands in addition to the decision rules used to determine the presence/absence of speech.

When someone speaks, the vocal tracts are in movement. This movement determines the amplitude of modulation of the temporal envelope of the speech signal. The modulation rate for sentences is approximately between 4 and 16 Hz (Drullman *et al.*, 1994b), (Drullman *et al.*, 1994a), which corresponds to the syllable rate. Thus, speech can be discriminated from noise by its modulation frequency since environmental noise is often unmodulated, or has a modulation rate not within the range of 4 to 16 Hz.

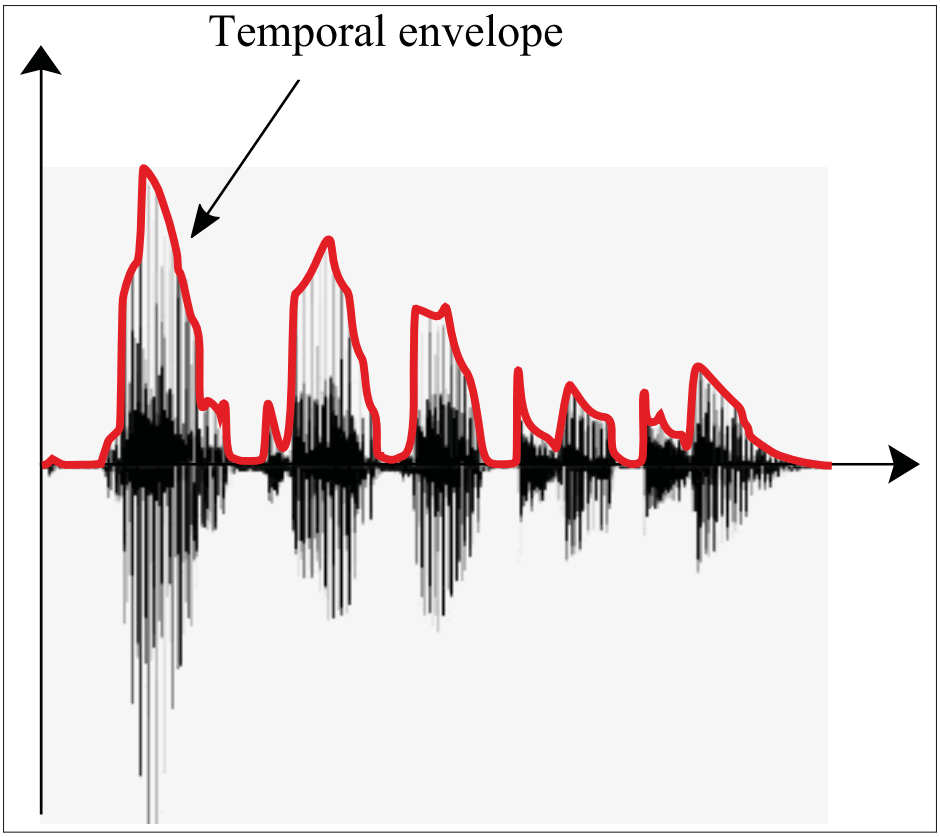

Figure 1.2 Speech signal and its temporal envelope.

In addition to the modulation frequency, another modulation characteristic is used to determine the presence/absence of speech and it can be used to estimate the SNR in each frequency band. This feature is called the modulation depth and is calculated in dB. The modulation depth is the difference between the maximum and the minimum of the temporal envelope of the signal (Sandlin, 2002). High modulation depth indicates that speech is produced in a quiet environment while low modulation depth indicates that speech is produced in a noisy environment. The modulation depth is used to estimate the SNR in the signal and thereafter set a gain for noise reduction. The amount of reduction gain is often inversely proportional to the estimated SNR in each frequency band.

Some manufacturers use only the modulation depth to determine the presence of speech and estimate the SNR, while others rely on several characteristics including: the sound pres-

sure level of the input signal, the sound pressure level of the noise, the modulation rate, etc. (Mueller and Ricketts, 2005).

The decision rules and reduction gain are different from one manufacturer to another, yet, they all tend to not reduce the noise when the speech signal is detected, in order to preserve the quality of speech without degrading it.

Once the reduction gain is determined, noise reduction goes into effect depending on the time constants which permit to reduce the noise without affecting the useful signals. These time constants are:

- The attack time, which represents the time between the detection of the presence of noise and the start of its reduction.
- The speed of gain reduction, which represents the time that the technique takes to reduce the maximum gain.
- The release time, which represents the time the noise reduction algorithm detects the absence of noise and the time that the gain starts to recover.
- The speed of gain recovery, which represents the time between the start of the gain recovery and the end of gain reduction.

After reducing the noise in each frequency band, a reconstruction of the signal is performed by summing all the bands. This reconstructed signal is the one transmitted to the ear.

Levitt (2001), Chung (2004) and Bentler and Chiou (2006) presented wide literature reviews of hearing aid technologies from those using a single microphone to the most advanced using multiple microphones to take into consideration the spatial information to determine the location of the source by calculating some parameters such as the Direction Of Arrival (DOA).

As the use of MB-DNR methods in hearing aids was successful and very promising, in 2009 Chung et al. (Chung *et al.*, 2009) proposed its evaluation for hearing protection in order to assess if this technology can be implemented in active HPDs.

The results obtained by Chung et al. (Chung *et al.*, 2009) showed that the noise reduction using the MB-DNR method seems promising in the field of hearing protection. However, the noise reduction was very low in four types of industrial noises, which is due to the fact that the MB-DNR method was developed and dedicated to a hearing aid application where the gain reduction was small, contrary to a hearing protection application, which requires a higher reduction in gain.

1.4.3 Summary

This Section presented an overview of noise reduction techniques currently used in hearing aids as their hardware resources are very similar in size to the hardware resources available for the S-HPD. It showed that the MB-DNR method is promising for hearing protection and the use of modulation characteristics for noise reduction can be considered for the present work.

1.5 Methodology for the evaluation and validation of speech based systems

The evaluation of speech-based systems needs to be performed in two steps. In the first step the developed algorithms must be evaluated objectively in a simulation (software) and subjectively using clean speech databases mixed artificially with noise in different SNRs. The second step consists in evaluating in real-time the algorithms implemented in the hardware.

Since the first objective of the S-HPD is the transmission of enhanced speech signals to the ear, the following subsections describe some techniques for the evaluation of VAD and noise reduction algorithms.

1.5.1 Algorithm evaluation

1.5.1.1 Evaluation of voice activity detection algorithm

A VAD algorithm is said to be efficient if it is able to detect only speech segments in a given environment. For the assessment of the developed VAD algorithm, three objective evaluation metrics can be used: the True Positive Rate (TPR) which is the rate of speech segments detected as speech, the False positive Rate (FPR) which is the rate of noise segments detected as speech, and the F1 score.

In the literature, TPR and FPR are usually calculated using noisy speech signals to discriminate between speech and noise signals. However, for the S-HPD application, TPR must be calculated using noisy speech signals, whereas FPR must be calculated using only noise signals to see how accurate the algorithm is when no speech signal is present.

The F1 score measures the accuracy of the algorithm in terms of precision and recall:

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{1.4}$$

with

$$\mathrm{precision} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FPR}} \tag{1.5}$$

$$\mathrm{recall} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FNR}} \tag{1.6}$$

The choice of this metric is based on its combination of TPR, FPR and False Negative Rate (FNR). (C. J. van Rijsbergen, 1979).

The Receiver Operating Characteristic (ROC) curve is widely used for VAD algorithms assessment. It shows how the sensitivity and specificity of an algorithm changes when one parameter varies such as a decision threchold). Thus, it enables the determination of the optimal param-

eter. In the current work, the ROC curve will not be used since more than one parameter may be included in the decision (Metz, 1978).

The robustness of the proposed VAD algorithm will be compared to Sohn et al's VAD. The choice of this algorithm as a benchmark is based on: one, its wide use by the community for comparison, two, its hangover scheme which is calculated using hidden markov model (HMM) to maintain the detection of low energy speech frames, and three its wide use as front-end to in most of the speech enhancement techniques in which the noise parameters are estimated during speech-free frames (Hu and Loizou, 2007a).

1.5.1.2 Evaluation of noise reduction algorithm

Assessing the robustness of a noise reduction/speech enhancement algorithm has to be done with different tests that consist of evaluating the enhanced signal in terms of speech quality and speech intelligibility. These tests can be done first using objective metrics, then a subjective protocol has to be determined and tests will be conducted in an audiometric booth with human participants.

The first and low cost evaluation procedure can be done using objective metrics. For example, Perceptual Evaluation of Speech Quality (PESQ) (ITU-T P.862, 2001), Perceptual Objective Listening Quality Analysis (POLQA) (ITU-T P.862, 2011), Log Likelihood ratio (LLR), Cepstral Distance, segmental SNR, Speech Intelligibility Index, etc. (Hu and Loizou, 2008), (Ma *et al.*, 2009), can be used to evaluate speech quality or intelligibility.

Speech quality is defined as the overall impression of the listener on how "good" the quality of the speech is while speech intelligibility is defined as the accuracy with which the listener can understand what is being said (Kondo, 2012). Speech quality and intelligibility may diminish as the speech gets distorted or corrupted (masked) by noise. In some cases, speech perceived as being of "good" quality also has high intelligibility, while in other cases the quality of a speech signal can be rated as "bad" because of the presence of noise (low SNR) or other distortions, but can still have a high intelligibility.

From the first speech intelligibility evaluation performed in 1978 using the spectral subtraction by (Lim, 1978), to the recent speech intelligibility evaluation of eight existing noise reduction algorithms performed by (Hu and Loizou, 2007b), no improvement in speech intelligibility has been found compared to the unprocessed signals, even if the speech quality was improved. Thus, speech quality and intelligibility are two different and uncorrelated aspects of speech and need to be assessed separately.

The robustness of the proposed noise reduction algorithm will be compared to three algorithms, namely: the Wiener filter based on an *a priori* SNR estimation (Scalart and Fiho, 1996), the spectral subtraction (Boll, 1979), and band-pass modulation filtering (Falk *et al.*, 2007). The Wiener filter and spectral subtraction codes were taken from (Loizou, 2007) (wiener-as and SpecSub), while the code of the modulation filtering was obtained directly from the authors of (Falk *et al.*, 2007). In the spectral subtraction, the noise spectrum was estimated and updated from non-speech frames detected using a simple VAD based on segmental SNR, while in the Wiener filter, non-speech frames were detected using *a priori* SNR estimation. The Wiener filter and spectral subtraction were chosen because of their wide use as benchmark algorithms (e.g. (Ming *et al.*, 2011), (Chen and Loizou, 2010), (Paliwal *et al.*, 2010))

Other recent algorithms (e.g. (Westerlund *et al.*, 2004), (Parikh *et al.*, 2009), (Shahid *et al.*, 2011)) were not selected as benchmarks because, unlike the selected benchmarks, the code was not available to the authors. Implementation intricacies such as non-optimal parameter settings could have potentially led to biased comparisons.

- **Subjective speech quality evaluation:** Different subjective speech evaluation tests can be used to assess the quality of the enhanced speech (Grancharov and Kleijn, 2008) and (ITU-T, 2003a). For example, the evaluation of the enhanced signal can be compared to the noisy signal and other benchmark algorithms using a rating scale (From 1 to 5) for different categories: speech quality (SIG), background noise intrusiveness (BAK) and overall quality (OVRL). Another subjective test can also be done to evaluate the quality of a speech signal in terms of musical noise. This test is very important when evaluating speech enhancement

algorithms since most noise reduction algorithms perform in the spectral domain such as the spectral subtraction algorithm and tend to generate spurious peaks in the processed spectrum which sounds like musical noise (Cappé, 1994), (Esch and Vary, 2009), (Inoue *et al.*, 2011), (Leitner and Pernkopf, 2012).

- **Subjective speech intelligibility evaluation:** Speech intelligibility is defined as the amount of speech understood by the listener. Thus, speech intelligibility evaluation is done by calculating the number of words recognized and successfully repeated by the listener compared to the total number of key words produced by the speaker. For example, in the sentence "The **birch canoe slid** on the **smooth planks**", five words (in bold) are considered as key words. Thus, the intelligibility is calculated as the rate of recognized key words divided by the total number of key words.

  Recently, a new evaluation technique and environment have been developed to evaluate the speech intelligibility of the enhanced speech while taking into consideration the Speech Reception Threshold (SRT) (Ellaham *et al.*, 2014). In this work, a toolbox method using the Matlab Speech Testing Environment (MSTE) was designed to assess the SRT using various testing methods, either using a fixed speech level presented at typical levels of 45-55 dB HL (Hearing Loss) and adapting the level of the masking noise ("adaptive masking level"), or using a fixed masking noise level and adapting the speech level ("adaptive speech level") or even using fixed speech and masking noise levels while adapting other metrics, such as the distortion threshold ("adaptive distortion threshold"). This evaluation toolbox also calculates the number of words correctly recognized by the listeners and finds the SNR at which each listener identifies correctly 50 % of the words from the Hearing In Noise Test (HINT) database (Nilson *et al.*, 1994), which contains 25 phonetically balanced lists of sentences, with 10 sentences per list. The HINT sentences were first designed in American English then adapted to Canadian French (Lamothe *et al.*, 2002).

1.5.2 System evaluation and validation

After evaluating, validating and implementing the algorithms in a the DSP, the evaluation of the complete system can be performed.

In the system's evaluation the same speech intelligibility and quality tests can be performed using real world signals or signals played through loudspeakers. This evaluation step also provides an opportunity to tune some parameters to maximize the system's efficiency.

1.6 Literature review synthesis

This chapter presented literature reviews of three different technological areas, from which the development of the S-HPD algorithms will draw its inspiration and build upon. It first presented the previous and current types of HPDs and showed the need for a new active HPD that enables face-to-face communication and hearing protection. To enable face-to-face communication, a speech detection technique must be developed. Thus, this chapter presented a literature review of different speech detection algorithms and highlighted the ones used in the telecommunication area such as in mobile telephony, since the algorithm that will be developed will also be targeted for a real-time and embedded application. Furthermore, this chapter presented an overview of the noise reduction techniques used in manufactured hearing aids, and showed that the use of the modulation characteristics for noise reduction is promising for hearing protection.

Moreover, this chapter showed that the advances reached in digital signal processing and integrated circuits will permit the development of a S-HPD able to detect and transmit enhanced speech while protecting the wearer's hearing.

CHAPTER 2

ECHO THRESHOLD BETWEEN PASSIVE AND ELECTRO-ACOUSTIC TRANSMISSION PATHS IN DIGITAL HEARING PROTECTION DEVICES

Narimene Lezzoum[1], Ghyslain Gagnon [1], Jérémie Voix [1]

[1] École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted in the « International Journal of Industrial Ergonomics», an Elsevier Journal, in March 2015.

2.1 Abstract

Electronic hearing protection devices are increasingly used in noisy environments. These devices feature a miniaturized external microphone and internal loudspeaker in addition to an analog or digital electronic circuit. They can transmit useful audio signals such as speech and warning signals to the protected ear and can reduce the sound pressure level using dynamic range compression. In the case of a digital electronic circuit, the transmission of audio signals may be noticeably delayed because of the latency introduced by the digital signal processor and by the analog-to-digital and digital-to-analog converters. These delayed audio signals will hence interfere with the audio signals perceived naturally through the passive acoustical path of the device. The proposed study presents an original procedure to evaluate, for two representative passive earplugs, the shortest delay at which human listeners start to perceive two sounds composed of the signal transmitted through the electro-acoustic circuit and the passively transmitted signal. This shortest delay is called the *echo threshold* and represents the delay between the time of perception of *one fused sound* from *two separate sounds*. In this study, a transient signal, a clean speech signal, a speech signal corrupted by factory noise, and a speech signal corrupted by babble noise are used to determine the echo thresholds of the two earplugs. Twenty untrained listeners participated in this study, and were asked to determine the echo thresholds using a test software in which attenuated signals are delayed from the original

signals in real-time. The findings show that when using hearing devices, the echo threshold depends on four parameters: (a) the attenuation function of the device, (b) the duration of the signal, (c) the level of the background noise and (d) the type of background noise. Defined here as the shortest time delay at which at least 20% of the participants noticed an echo, the echo threshold was found to be 8 ms for a bell signal, 16 ms for clean speech and 22 ms for speech corrupted by babble noise when using a shallow earplug fit. When using a deep fit, the echo threshold was found to be 18 ms for a bell signal and 26 ms for clean speech and 68 ms for speech in factory. No echo threshold could be clearly determined for the speech signal in babble noise with a deep earplug fit.

2.2 Introduction

Occupational hearing loss is the most common work injury in North America with approximately 22 million workers exposed daily to hazardous noise (NIOSH, 1998). To prevent hearing loss, wearing Hearing Protection Devices (HPD) becomes a necessity in industrial workplaces. In fact, wearing HPDs is also required nowadays for professional musicians since they too are exposed to loud sounds and thus vulnerable to hearing loss (Macdonald *et al.*, 2008).

HPDs come in various forms. There are earplugs, which must be placed within or against the entrance of the ear canal, and earmuffs, which either fit around the ear, or in the form of helmets, encasing the entire head (Berger *et al.*, 2003b). HPDs can be grouped in two types of operating mode: passive HPDs and active (or electronic) HPDs (Casali, 2010).

Passive HPDs are the traditional HPDs. They reduce the background noise mechanically based on their shape and material composition, while electronic HPDs are equipped with an external microphone to capture the signals, an internal loudspeaker to playback the signals under the protected ear and an analog circuit or a Digital Signal Processor (DSP) in order to process the incoming signals in real-time (Casali, 2010). Electronic HPDs are increasingly used by workers, musicians, and the military for their high flexibility and multiple functionalities such as active noise control (Brimhall *et al.*, 2002) or adaptive gain control (Hotvet, 1996).

Recently, some advanced functionalities have been developed for electronic HPDs as listed in (Voix, 2014), such as background noise reduction (Chung, 2007) and (Chung *et al.*, 2009), warning signals detection (Carbonneau *et al.*, 2013), and voice activity detection (Lezzoum *et al.*, 2014a) for the development of a smart HPD (S-HPD) or smart earphones to guarantee protection and to discriminate between speech and noise, allowing the transmission of enhanced speech signals to the ear.

Electronic HPDs process the incoming signals in real-time for retransmission to the ear. Real-time processing is defined as the continuous generation of an output signal within time constraints (Kuo *et al.*, 2014). These time constraints depend on the targeted application for which the processing is dedicated. For example, in Voice over IP (VOIP) communications, the time that elapses between the moment the talker utters the words and the moment the listener hears them is referred as the *mouth to ear delay* (Janssen *et al.*, 2002), and represents the maximum delay between the input and the output signals. As mentioned in the ITU-T recommendation (ITU-T, 2003b), mouth to ear delays of less than 150 ms for the transmission of speech or non-speech signals will experience essentially transparent interactivity. However, in other applications where visual information is also available in addition to the audio, such as teleconferencing, the audio signal should never be delayed by more than 45 ms from the video signal, while the video signal should never be delayed by more than 15 ms from the audio signal as demonstrated in (Cooper, 2003) and (Younkin and Corriveau, 2008) to avoid introducing lip-sync errors.

Digital HPDs may introduce a delay between the signals transmitted through the passive path of the HPD and the signals processed and transmitted through the internal loudspeaker. The passively transmitted signals reach the protected ear through the bone conduction and HPD material. When the processing delay increases, the processed signal will be heard as an echo of the passively transmitted signal, thus two signals will be heard. The delay at which the perception of *one fused sound* becomes *two separate sounds* is called the *echo threshold* (Litovsky *et al.*, 1999), or the delay of the *Just Noticeable Difference* (JND) (Quené, 2007), which are widely used psycho acoustic metrics. In (Haas, 1972), the influence of a single echo

on the audibility of clean speech has been studied depending on different parameters such as the intensity, the timbre, the angle of incidence and the room reverberation, concluding that when the echo sound is at the same intensity as the original sound, the critical delay (the delay where 10-20% of participants felt disturbed) is about 68 ms, while when the echo sound is attenuated by 3 dB, the critical delay rises to 108 ms, and when the echo sound is attenuated by 10 dB, no echo is felt. Furthermore, (Haas, 1972) showed that the attenuation of the high frequencies of the echo increases the tolerable delay.

The echo threshold can also be determined when a sound from one direction is followed by the same sound coming from another direction (Yang and Grantham, 1997). This phenomenon is known as the *precedence effect*. The precedence effect has been widely studied in the last decades and the influence of an echo on the audibility of clicks (transient signals) coming from different spatial locations has also been studied such as in (Freyman *et al.*, 1991), (Yang and Grantham, 1997), and (Saberi and Antonio, 2003). These studies showed that when the click sound echo has equal intensity as the original click sound, the echo threshold is around 5 to 10 ms.

Studies and experiments reported to date on the determination of the echo thresholds have been conducted with clean speech (Haas, 1972), or with transient signals (Yang and Grantham, 1997), (Litovsky *et al.*, 1999), (Saberi and Antonio, 2003). However, non-ideal real-world conditions such as noisy speech signals have not been investigated yet. For transient signals, the echo threshold was only determined with equal intensities. In addition, the motivation of almost all the previous studies was to understand how the auditory system processes and perceives the same signal coming from different directions such as reverberant spaces. However, the determination of the echo threshold for applications such as electronic HPDs, including the effect of their specific frequency response and resonances, has not been addressed yet, despite the fact that these electronic devices inevitably generate a processing delay.

The current study investigates the influence of frequency-dependent attenuation functions obtained from two representative fits of a custom earplug to evaluate the echo threshold depen-

dence on the attenuation function. Furthermore, this study tends to mimic real-world environments using clean speech signals and speech corrupted by two types of noise environments: factory and babble noise. In addition, a bell ringing sound is used as a transient signal.

This study was conducted on 20 human participants. Each participant was asked to determine the echo threshold between the passively and digitally transmitted signals using a real-time test software where the delay between the two signals could be user-controlled.

The present article is organized as follows: Section 2.3 models the sound transmission paths in digital HPDs. Section 2.4 describes the materials and methods used for the attenuation functions calculation, stimuli generation, and subjective test protocol. Section 2.5 presents the analysis of the stimuli signals using the spectrograms and the results from the subjective test and Section 2.6 discusses the findings and concludes this work.

2.3 Digital hearing protection device

2.3.1 Sound transmission paths

The digital HPD is a traditional passive HPD in which electro-acoustic hardware is embedded (Fig. 2.1). To capture signals, a miniature external microphone is connected to the audio input of an ultra-low power DSP. The DSP output is connected to a miniature loudspeaker to transmit the desired signals to the ear.

In addition to the digital path, the external sound is also transmitted through the HPD's material and, to a lesser extent, through bone conduction. Figure 2.2 illustrates the three sound transmission paths for a digital HPD.

The transmission through the HPD material highly depends on the fit of the earplug. As an example, Figure 2.3 shows the attenuation function of a shallow and deeply fitted HPDs, where differences of up to 20 dB can be observed.

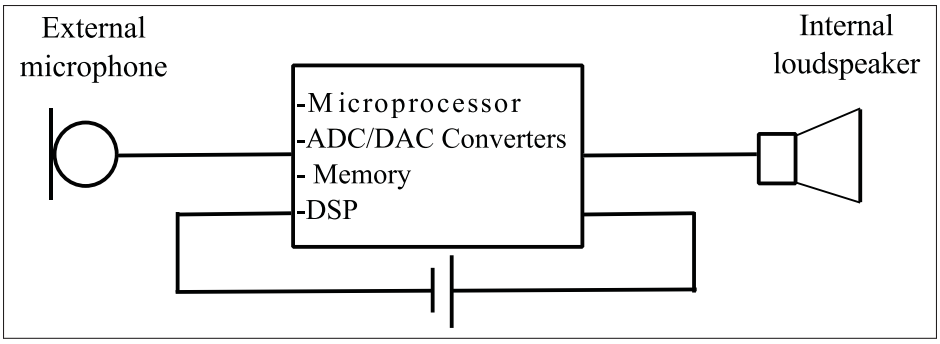

Figure 2.1 The hardware resources embedded in the digital hearing protection device.

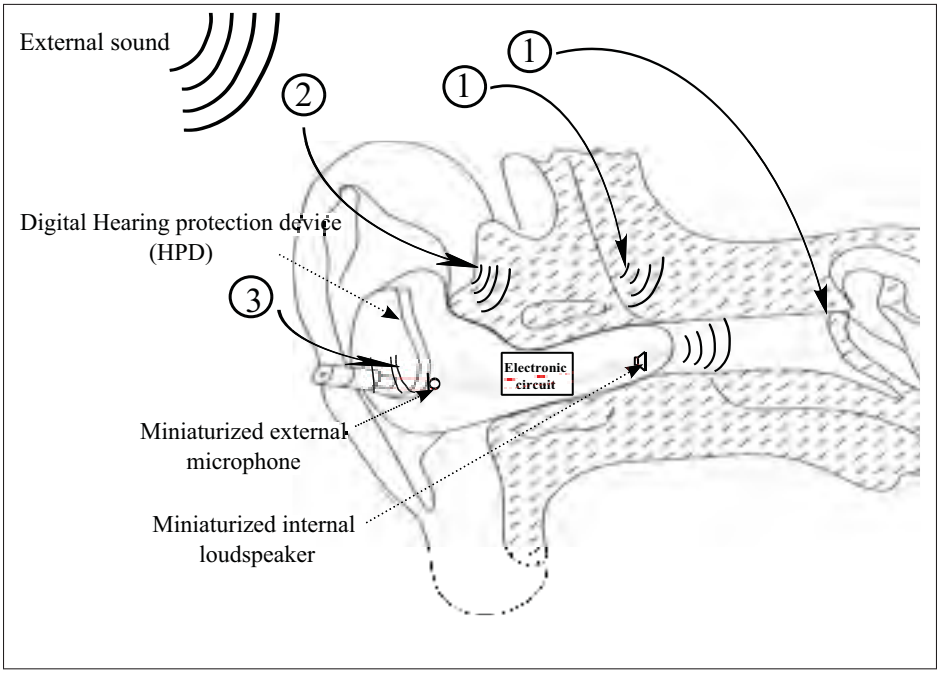

Figure 2.2 Sound transmission pathways through a digital HPD: (1) bone conduction path, (2) passive transmission through the earplug material, and (3) digital transmission through the active path of the earplug. This figure has been adapted from (Voix and Laville, 2009).

The signal path through the human skull (bone conduction) is highly attenuated (from 45 to 55 dB) making it a negligible secondary path (Berger *et al.*, 2003a). Therefore, in the rest of this paper, bone conduction is ignored and passively transmitted signals denote only the signals transmitted by means of HPD material.

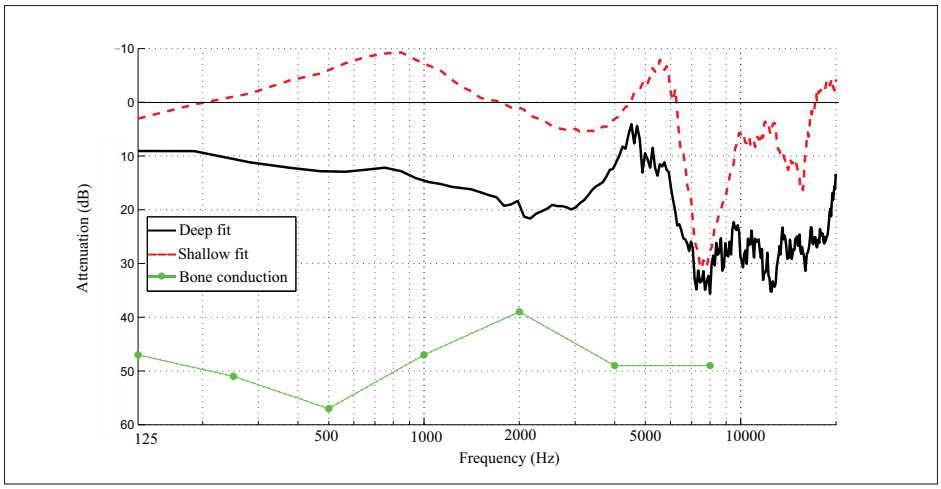

Figure 2.3 Attenuation functions of two custom-molded earplugs: shallow fit represents the low attenuation function, and deep fit represents the high attenuation function.

2.3.2 HPD characteristics

Custom-molded earplugs are a type of HPD that fit instantly to the user's ear canal, by injecting a soft expendable medical silicon rubber agent between a rigid core and an expendable envelope (Voix and Laville, 2009).

The prediction of the attenuation function of these earplugs is conducted as described in (Voix and Laville, 2009): an internal microphone is used to capture the passively transmitted sound to the ear, and an external microphone is used simultaneously to capture the external sound. The attenuation functions are computed from the internal and external sound pressure level. In a previous study (Nadon *et al.*, 2015), eight human participants were fitted with custom-molded earplugs. The corresponding transfer functions were assessed using white noise. From this dataset, two transfer functions have been selected for the purposes of the current study. These transfer functions represent two extreme cases: the first transfer function has a low attenuation and was obtained from a participant with a shallow fitted earplug, while the second has a high attenuation and was obtained from a participant with a deeply fitted earplug. The magnitudes of these transfer functions are illustrated in Figure 2.3. This figure shows the frequency-dependent attenuations that both (shallow and deep) fits exhibit.

It also shows that the shallow fit has two resonance frequencies, the first one corresponds to a Helmholtz resonator resulting from the leaking earplug, while the second one corresponds to the longitudinal resonance of the occluded ear canal. Figure 2.3 shows that the attenuation function corresponding to the deep fit attenuates the signal by 10 dB below 3500 Hz, and by 5 dB around 5000 Hz, while it attenuates the high frequencies by about 30 dB.

2.4 Methodology

In the first part of this Section, we present the stimuli signals used for this study, while in the second part, subjective tests conducted with 20 untrained human participants using the generated stimuli signals and a test software are presented.

2.4.1 Stimuli generation

2.4.1.1 Types of signals

Two types of signals are used in this study. These signals are considered as desired signals for an S-HPD application use case, thus their unaltered transmission through the digital earplug to the protected ear is important. These signals are:

- Speech signals: one speech sentence uttered by a male speaker in Canadian French from the HINT (Hearing in Noise Test) database (Lamothe *et al.*, 2002) was used. The length of this clean speech signal is around 2 seconds, and the sampling frequency is 22 kHz. Two different scenarios are considered: the first, consists of presenting clean speech to the participants. In the second scenario, noisy speech signals were presented to the participants by artificially adding, to the same clean speech signal, babble and factory noise obtained from the Aurora database (Hirsch and Pearce, 2000) with a 5 dB signal to noise ratio. This situation mimics noisy environments, such as workplaces or restaurants, in which wearing HPDs or other smart in-ear devices is beneficial.

- Transient signals: transient signals are characterized by their abrupt high energy peaks with a period varying between 5 and 10 ms followed by decaying oscillations with a longer period. Hearing an echo of the transient signal can be annoying to the HPD wearer. For this purpose, a bell ring obtained from a free online database (FreeSound, 2014) is used. The sampling frequency is 44 kHz.

2.4.1.2 Signal processing

The two attenuation functions, corresponding to two fits of the earplugs, are applied to the four signals presented in Section 2.4.1.1 for the generation of the passively transmitted signals $y(n)$:

$$y(n) = x(n) * h(n) \tag{2.1}$$

with * for the convolution, $x(n)$ for the original signal, $h(n)$ the impulse response of the attenuation function, and n the sample number. The impulse response of the attenuation function $h(n)$ is adjusted to 44 μs, corresponding to the delay of the passively transmitted signal through the earplug (15 mm traveled at 340 m/s).

Thus, eight signals are generated: four signals for the shallow fit, and four signals for the deep fit. Figure 2.4 illustrates a block diagram for stimuli generation. The stimulus $s(n)$ is generated by adding the passively transmitted signal $y(n)$ to the digitally transmitted signal $x(n)$:

$$s(n) = y(n) + x(n-d) \tag{2.2}$$

with $n > d$ and d is the number of taps that represents the delay difference between the original signal transmitted via the digital path of the earplug and the passively transmitted signal.

Varying the delay d between the two signals will lead to the determination of the echo threshold using a subjective tracking procedure described in the next subsection.

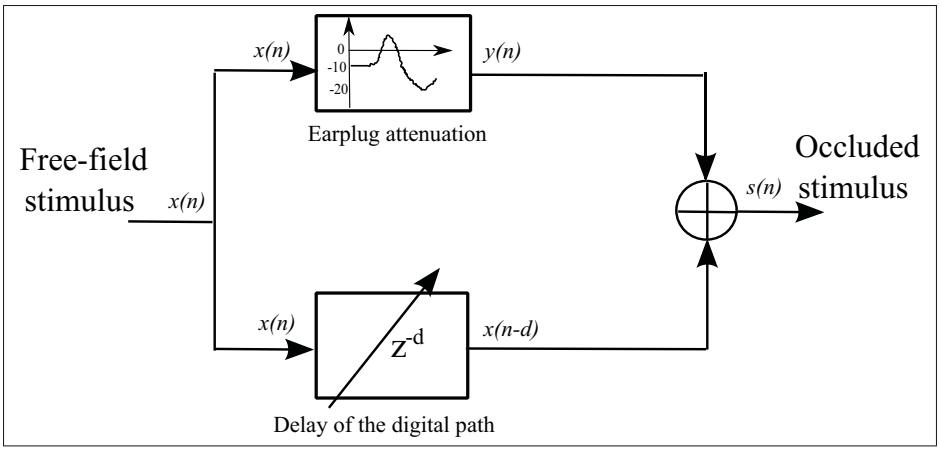

Figure 2.4 Block diagram for the stimuli generation.

2.4.2 Subjective test protocol

The test was conducted in an ANSI S3.1 compliant audiometric booth with 20 French speaking and normal hearing participants: 17 males and 3 females aged between 22 and 35 years of age with an average age of 25 years. All signed a consent form prior to participation. The subjective tests presented in this paper were approved by the internal review board of ETS (Comité d'éthique de la Recherche of École de technologie supérieure) (CER, 2014).

This subjective test is conducted with untrained participants as the electronic HPDs are aimed to be used by a large population, with no knowledge or experience in the field of acoustics.

Participants were outfitted with professional headphones and placed in front of a computer screen equipped with a test interface which allows the user to change the delay between the passively attenuated and the non-attenuated signals in real-time to determine the echo threshold. The test interface was created using the open source software PureData (PureData, 2014). Figure 2.5 illustrates a screen shot of this interface which has been developed in our labs.

Before the test, participants were instructed to find the echo threshold that corresponds to the delay at which they start to hear an echo of the first signal. To do so, participants were instructed to vary the delay by moving the wheel of the computer mouse: when the wheel is moved up, the delay increased by 2 ms, when moving it down, the delay decreased by 2 ms steps. The delay

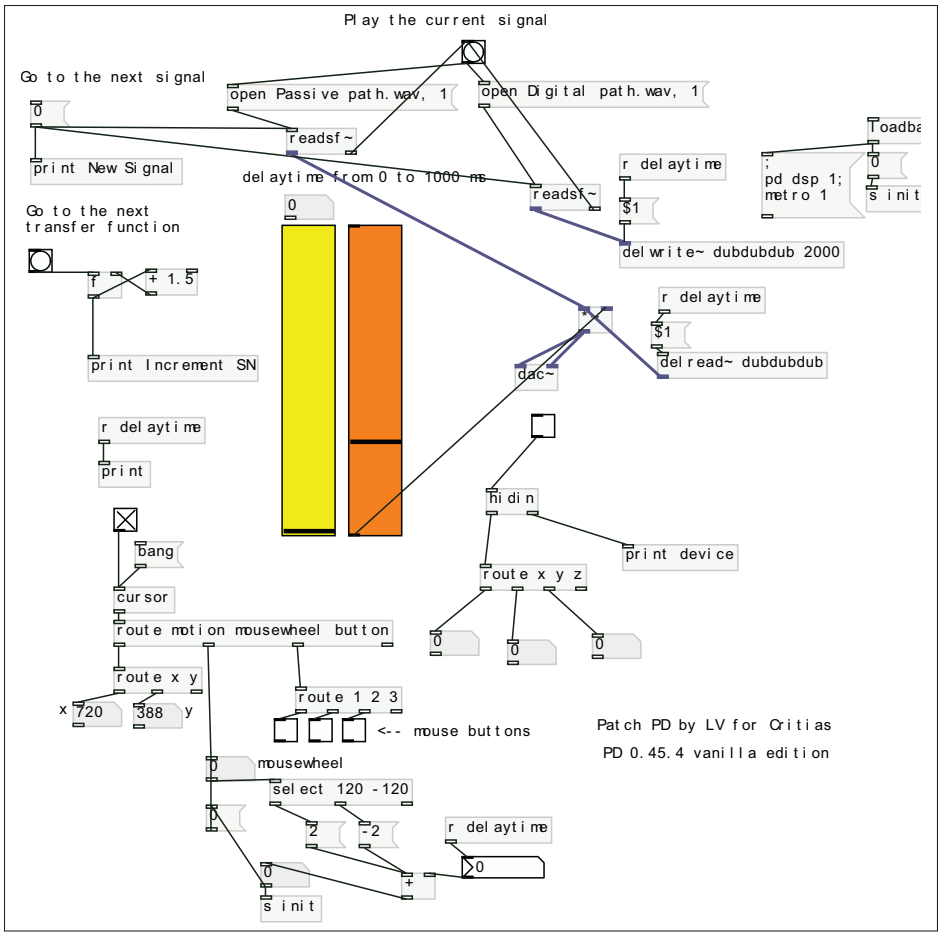

Figure 2.5 Screen shot of the test interface designed with PureData for real-time signal delaying.

could vary between 0 and 1000 ms. Before validating their response, participants were asked to decrease the delay to be more accurate and detect the threshold of the just noticeable difference. Once this threshold was found, they were asked to validate their response by pressing a button and pass to the following stimuli.

Participants were instructed to fix the threshold to the maximum value (1000 ms) if the passively transmitted signal could not be distinguished from the digitally transmitted signal, i.e. if no echo was perceptible.

2.5 Data analysis and results

The test signals were presented to all the participants in the same order: starting with the stimuli generated from the shallow earplug fit then the stimuli generated from the deep earplug fit following this order: the bell ring signal, the clean speech, the speech corrupted by factory noise, then the speech corrupted by babble noise.

Afterwards, the data collected from this test is subjected to statistical analysis.

2.5.1 Spectrogram analysis

Figure 2.6 illustrates the spectrograms of the bell signal, clean speech signal, and speech corrupted by factory noise (each with the two fits) with no delay (d=0 ms) and with a delay of 80 ms. This figure shows that for the bell signal, a difference is observed between the two spectrograms (d=0 and d=80 ms) for the shallow and deep fits. The same observations are also noticed for the clean speech with the two fits. For speech corrupted by factory noise with deep fit, we notice that there is no difference between the two spectrograms. This is due to the low energy of the passively transmitted speech signal (as shown in Figure 2.6), which is masked by the factory noise. However, with the shallow fit, we observe a difference in the spectrograms between t=0.6 and t=0.8 seconds where there is an obvious redundancy of the speech segment.

2.5.2 Descriptive statistics

The minimum, the first quartile, the median, the third quartile, and the maximum were calculated upon the echo threshold determined by the participants depending on the stimuli and are illustrated in the box-and-whisker plot in Figure 2.7.

Figure 2.7 shows that for the bell ring the echo threshold median for the shallow fit (16 ms) is close to the median of the deep fit (26 ms). It also shows that for the clean speech, the median is almost the same for the shallow (36 ms) and deep fits (39 ms). However, for the speech corrupted by factory noise, the echo threshold medians are distant for the two fits (39 ms for the

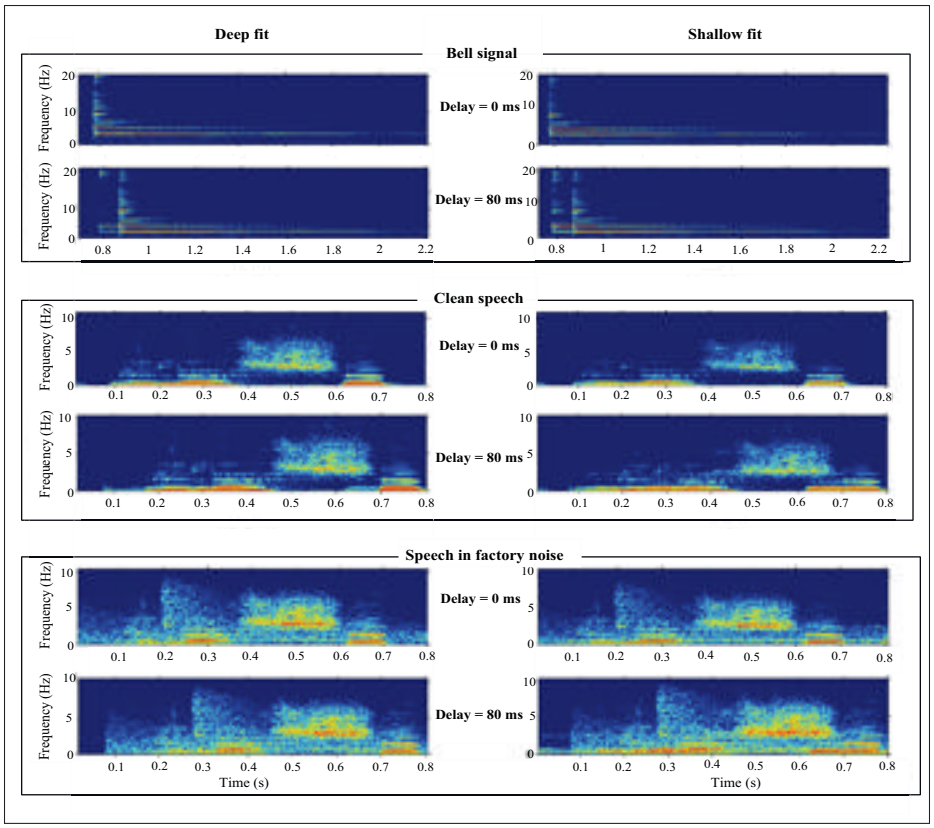

Figure 2.6 Spectrograms of the bell, clean speech, and speech corrupted by factory noise with a delay of 0 and 80 ms between the passively and the digitally transmitted signals for the shallow and deep fits.

shallow fit and 76 ms for the deep fit), knowing that during the test, four participants notified that they did not perceive an echo even if the maximum value was reached (1000 ms) for the deep fit, which is due to the background noise which masks the passively transmitted speech signal. With the last stimuli (speech corrupted by babble noise), a big difference is noticed between the shallow and deep fit stimuli. The median echo threshold for the shallow fit was found at 43 ms, while for the deep fit, 15 subjects among the 20 did not notice any difference between the passively and numerically transmitted signals. With the speech corrupted by babble noise (deep fit), one participant fixed the echo threshold at 46 ms. This participant is a musician and is very sensitive to changes in the frequency components of a signal. He commented that his choice was very influenced by frequency components change in the signal.

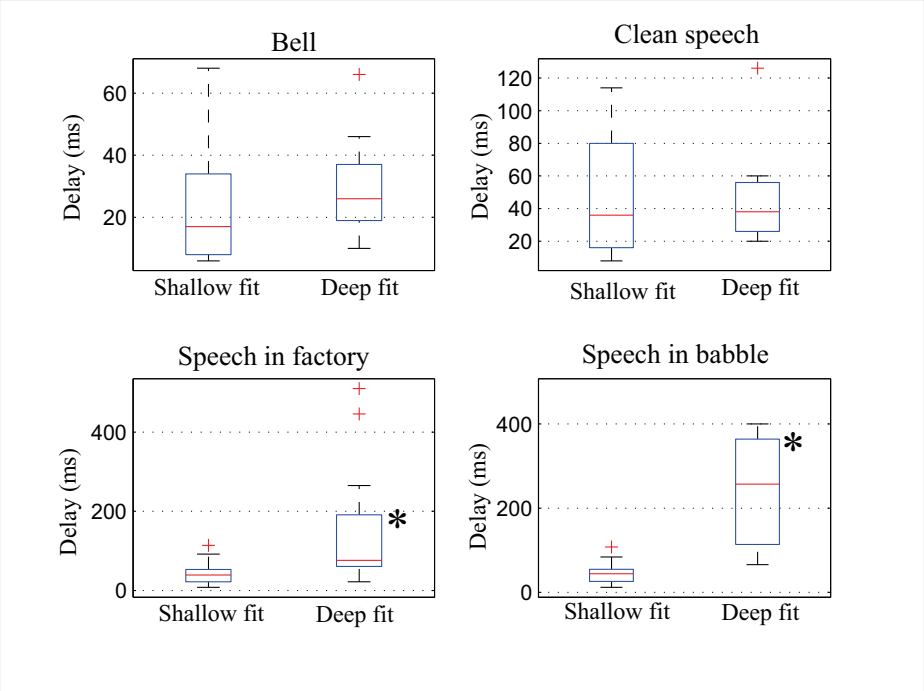

Figure 2.7 Echo threshold box and whisker plot. The asterix (*) symbol upon the deep fit results of the speech corrupted by factory noise reflects that 20 % of the subjects did not perceive any difference between the passively and digitally transmitted signals respectively, and the asterix (*) symbol upon the deep fit results of the speech corrupted by babble noise reflects that 75 % of the subjects did not perceive any difference between the passively and digitally transmitted signals respectively.

2.5.3 Analysis of the variance

In order to assess significant differences between the echo thresholds obtained with each attenuation function, signal and participant, we subjected the echo threshold determined by the 20 participants to statistical analysis. For this purpose, a three way analysis of variance (ANOVA) was conducted in Matlab^{TM} (Mathworks, MA). The model used is not with repeated measurements and the alpha value is 0.05 which corresponds to 95% confidence. In this analysis, two factor interactions with three levels have been used and consist of: first, the interaction between the fit type and the signal type; second, the interaction between the fit type and the participants,

and third, the interaction between the signal type and the participant. Table2.1 illustrates the details of the ANOVA analysis with the p-values

Results from the ANOVA confirm the previous obtained results and show that there was a significant interaction between the type of the fit and the type of signal ($F(1.5)$=10.55, $p < 0.05$), while no significant interaction was found between the type of the fit and the participant ($F(1.5)$=0.97) as well as between the type of the signal and participant ($F(1.5)$=0.79).

Source	F	p values
Fit type	35.03	0
Signal type	13.93	0
Participants	2.12	0.0237
Fit type * Signal type	10.55	0
Fit type * Participants	0.97	0.5087
Signal type * Participants	0.79	0.7945

Table 2.1 Results of the Analysis of variance ANOVA.

2.5.4 Determination of the echo threshold

The echo threshold is defined here as the minimum delay at which at least 20% of the participants perceive two distinct signals, as was done in (Haas, 1972). The echo threshold for each stimuli was determined by plotting the Cumulative Density Functions (CDFs) which are shown in Figure 2.8 The results are summarized in Table 2.2, which highlights the dependence on the fit of the earplug (from 8 to 18 ms for the bell signal, and 16 to 26 ms for the clean speech). With a deep fit, most participants could not distinguish the echo from the original sound for the speech signals in babble noise.

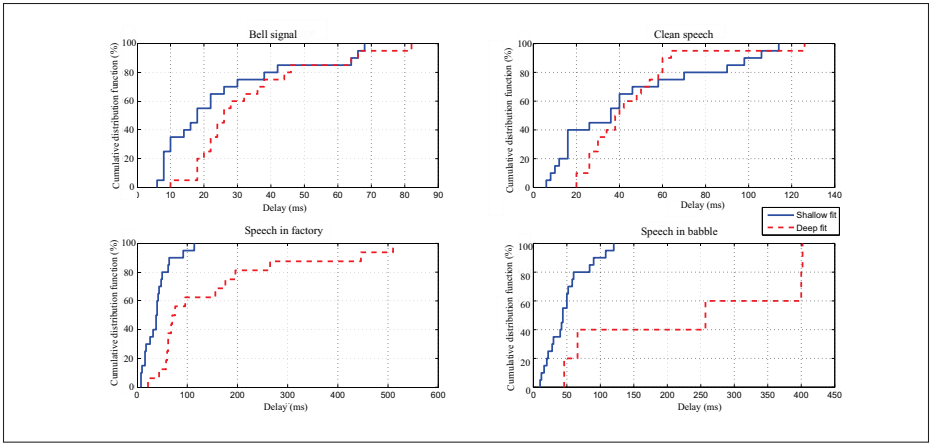

Figure 2.8 Cumulative Density Functions for the eight stimuli.

Table 2.2 The echo thresholds for the eight stimuli. For the deep fit with factory noise, 20% of the subjects did not perceive any difference between the passively and digitally transmitted signals. In babble noise 75% of the subjects did not perceive any difference between the passively and digitally transmitted signals.

Signal	**Shallow (ms)**	**Deep (ms)**
Bell	8	18
Clean speech	16	26
Speech in factory	16	68
Speech in babble	22	*

2.6 Discussions and conclusions

As described in this paper, with the miniaturization of microelectronic devices, it is now possible to include a DSP in a HPD to perform real-time signal processing on incoming audio signals. However, this signal processing introduces some delay which can be annoying to the user. Determining the echo threshold in real-world conditions allows to set the allowable processing delay of the DSP in such devices.

The allowable processing delay for the electronics represents the time difference between the echo threshold and the delay of the acoustic path through the earplug. The delay of the acoustic path through the earplug is around 44 μs (15 mm travelled at 340 m/s), which is non-significant

when compared to the echo threshold reported in the study. Therefore, we conclude that the delay introduced by the entire electronic path (from the microphone to the loudspeaker) should be made lower than the echo threshold.

The subjective results presented in this paper showed that when the earplug has a shallow fit and presents a resonance frequency in the critical frequency range of speech (between 200 and 1000 Hz), the echo threshold of clean speech stimuli is almost the same as the echo threshold of speech corrupted by a stationary (factory) or non-stationary speech-shaped noise (babble) (for 20% of the participants the echo threshold for the three stimuli is 16, 16, and 22 ms respectively). However, when the earplug has a deep fit without resonance frequency in the critical frequency range of speech, the echo threshold of clean speech stimuli is lower than the echo threshold of speech corrupted by factory noise, while when the speech is corrupted by speech-shaped noise (babble), there is no perceptible difference between the passively and digitally transmitted signals.

From the current study, we conclude that the echo threshold between the passively and digitally transmitted signals depends on four parameters:

- The attenuation function: the amount of attenuation of the in-ear device is a very important parameter for the determination of the echo threshold between the passively and the digitally transmitted signals. The higher the attenuation is, the higher the delay.
- The duration of the signal: the delay depends on the duration of the signal, if the signal has a short duration such as transients, the delay is low and it increases when the incoming signal duration increases.
- The presence of background noise: the current study showed that when background noise is present, the echo threshold increases compared to clean speech conditions. For instance, with a deep fit, the clean speech stimuli gave a median echo threshold of 38 ms, while when speech is corrupted by factory noise, the median echo threshold was found at 96 ms.

- The type of background noise: when the incoming signal is corrupted by background noise, the delay increases since the background noise masks the passively transmitted signal. The delay not only depends on the presence of background noise, but also on the type of noise: if the background noise is non-stationary such as babble noise, the delay is higher than when the background noise is stationary such as factory noise.

The delay between the passively and digitally transmitted signals depends not only on one criterion but on the combination of the four criteria.

Our findings suggest that manufacturers of electronic HPDs and the next generation of digital in-ear devices should set the processing time depending first on the attenuation function of the device. In addition, the processing time should be chosen as a function of the type of the desired signal to be sent through the digital path of the HPD: if the electronic HPD is designed to transmit signals with short periods such as transient signals, then the processing time should be lower (between 8 and 18 ms) than if the device was designed to transmit other signals such as speech signals (between 16 and 26 ms). Furthermore, if the device is developed to be used in noisy environments, the processing time can be higher and depend on the nature of the background noise. In situations where the processing time can be sufficiently long, other sophisticated modules such as speech recognition, speaker recognition, signal identification or background noise classification can be embedded in the in-ear devices. Nevertheless, in situations where the visual information is also provided to the in-ear device wearer, the processing time should be determined as a function of this information and should not exceed a certain amount of time (45 ms) to avoid generating a lip-sync error.

CHAPTER 3

VOICE ACTIVITY DETECTION SYSTEM FOR SMART EARPHONES

Narimene Lezzoum[1], Ghyslain Gagnon [1], Jérémie Voix [1]

[1] École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in the « IEEE Transactions on consumer electronics» in November 2014,

Volume 60, Issue 4.

3.1 Abstract

This paper presents a real-time voice activity detection (VAD) algorithm implemented in a miniature Digital Signal Processor (DSP) for in-ear listening devices such as earphones or headphones. This system allows consumers to hear external speech signals such as public announcements or oral communication while listening to music without removing their listening devices. The proposed algorithm uses two normalized energy features that compare the energy in the frequency region containing speech information with the frequency regions typically containing noise. The extraction of the normalized features represents the key of the proposed VAD since it eliminates the need for a signal-to-noise ratio (SNR) estimator. The VAD's decision is made using two threshold comparison rules computed from the normalized features and a hangover scheme triggered after a given number of observations. The algorithm parameters, namely the frequency regions' boundaries, number of observations, two decision thresholds and hangover's duration, have been optimized off-line using a genetic algorithm. The performance of the proposed VAD is compared to a benchmark algorithm in four noise environments and three SNRs. Results show that the average false positive rate (FPR) of the proposed algorithm is 4.2% and the average true positive rate (TPR) is 91.4 % compared to the benchmark algorithm which has a FPR average of 29.9% and a TPR average of 79.0%. The proposed VAD is implemented in hardware to validate its reliability and complexity.

3.2 Introduction

Nowadays, smart-phones, mp3 players, and other portable audio player devices are ubiquitous. Wearing earphones or headphones for listening to music in public places such as airports, airplanes, or railway stations causes sensory and cognitive distractions and isolates the wearer from the external environment. For example, in a railway station when train departures are announced, earphone wearers may miss this announcement and consequently miss their train. Similarly, in an airplane, passengers must remove their earphones when a steward is addressing them.

To palliate problems caused by the wearing of earphones in public places, several tools have been developed to enable consumers to hear external signals, ranging from push-to-hear electronic devices to dedicated wireless systems.

Earphone manufacturers have developed systems which include a microphone and a push button that allows the users to mute the music and transmit external sounds to the ear, thus allowing communication without the need to remove the earphones. These devices are either available as external dangles or included directly into the headphones. Since the users must manually push a button, they must know that a spoken message is addressed, which is unsuitable in situations where no visual cue is available (public announcement, for example). Software tools for external signals transmission are also available in smart-phones. They enable consumers to hear the external environment while listening to the music when the loudness of the external environment exceeds a certain threshold. Although these tools let the earphone wearers remain aware of their external environment, they can be annoying since all signals (useful and not-useful) are transmitted to the ear whenever they reach the predetermined loudness threshold.

Sophisticated wireless systems have also been developed to address this problem. These systems transmit the announcements to the wearer's audio device via a network, and then play relevant announcements in the earphones (Desai, 2014). This method requires a specific infrastructure in a given location, and the user cannot benefit from this technology where the infrastructure has not been developed. The present paper describes a real-time Voice Activity

Detection (VAD) system for smart earphones that can be integrated to current advanced communication earpieces (Voix *et al.*, 2014). The proposed system discriminates between a speech (useful) signal and noise (not-useful) signal to transmit speech signals through the earphones while blocking noise signals. A miniature Digital Signal Processor (DSP) is integrated in the earphones for real-time speech and noise discrimination.

Voice activity detection is commonly used in various speech-based applications. In voice over IP transmission and GSM communication, a VAD is used to encode non-speech segments with a lower bit rate than speech segments and thus reduce the transmission rate (ITU T, 1996). It is also widely used in human/machine interaction applications (Cho and Kim, 2011), (Lee *et al.*, 2009) for speech recognition or speaker identification and verification to reduce false alarm rates due to the use of noise segments in the recognition process. Likewise, VADs are used for noise reduction in hearing aids (Chung, 2004) and recently for smart hearing protection (Lezzoum *et al.*, 2013).

The performance of VADs relying solely on the extraction of one or several features (ITU T, 1996), (Tucker, 1992), (ETSI, 1999) degrades when the signal-to-noise ratios (SNR) decreases (Beritelli *et al.*, 2002). To palliate this problem, other VADs have been developed and rely on the estimation of the a posteriori and a priori SNR using the signal first frames, assumed to contain only noise signals (Sohn *et al.*, 1999). Unfortunately, these VADs become sensitive to changes in the SNR (Moattar and Homayounpour, 2009). Learning techniques or modeling algorithms have also been applied to VADs (Wu and Zhang, 2011), (Liu *et al.*, 2010) making the VAD efficient but more complex and difficult to implement in a DSP with limited hardware resources for real-time applications.

Recently, Hsu et al (Hsu *et al.*, 2013) proposed an energy-based VAD where the decision is made using a threshold upon the energy of the frequency modulation of harmonics. This VAD has shown its effectiveness in low SNRs and requires low computational resources. However its response delay makes it unsuitable for real-time low-latency applications.

While a relatively low-complexity VAD has been proposed based on the inter-quartile range statistic feature (Lezzoum *et al.*, 2013), the current approach proposes improvements, using simpler energy-based features, for an efficient implementation in a low-power DSP. The proposed VAD is implemented in a miniature DSP for smart earphones or headphones applications. The proposed solution can be integrated into active noise control headphones, which are already equipped with external microphone and other electronics. It can even be retrofitted to traditional headphones or earphones by integrating a miniature external microphone and DSP.

This paper is organized as follows: Section 3.3 presents the proposed smart earphones and their operating principle. Section 3.4 describes the proposed VAD algorithm. In Section 3.5, the parameters used in the VAD's decision are defined and their off-line optimization using a genetic algorithm is performed. Section 3.6 presents the validation of the proposed VAD, and Section 3.7 describes its implementation in a low-power DSP and its real-time validation in the embedded system. Finally, Section 3.8 concludes the paper.

3.3 The smart earphone

Smart earphones are traditional earphones, in which a field-programmable electronic hardware is embedded (Figure 3.1). To capture signals, a miniature external microphone is connected to the audio input of an ultra-low power DSP. The DSP output is connected to a miniature loudspeaker to transmit the desired signals to the ear (Carbonneau *et al.*, 2013).

The main task of the proposed system is the discrimination between speech and noise signals to allow speech signals to get through the earphones while blocking noise signals when speech is absent to enable the wearer to listen to music. Figure 3.2 illustrates the operating principle of the whole system.

3.4 The proposed VAD algorithm

A study conducted by Parikh et al. (Parikh and Loizou, 2005) on the influence of noise on vowels and consonants concluded that when the speech signal is corrupted by noise, the first

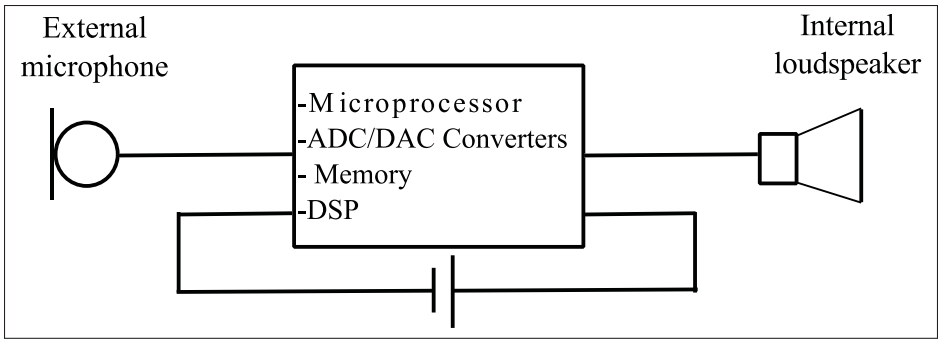

Figure 3.1 The hardware resources embedded in the smart earphones

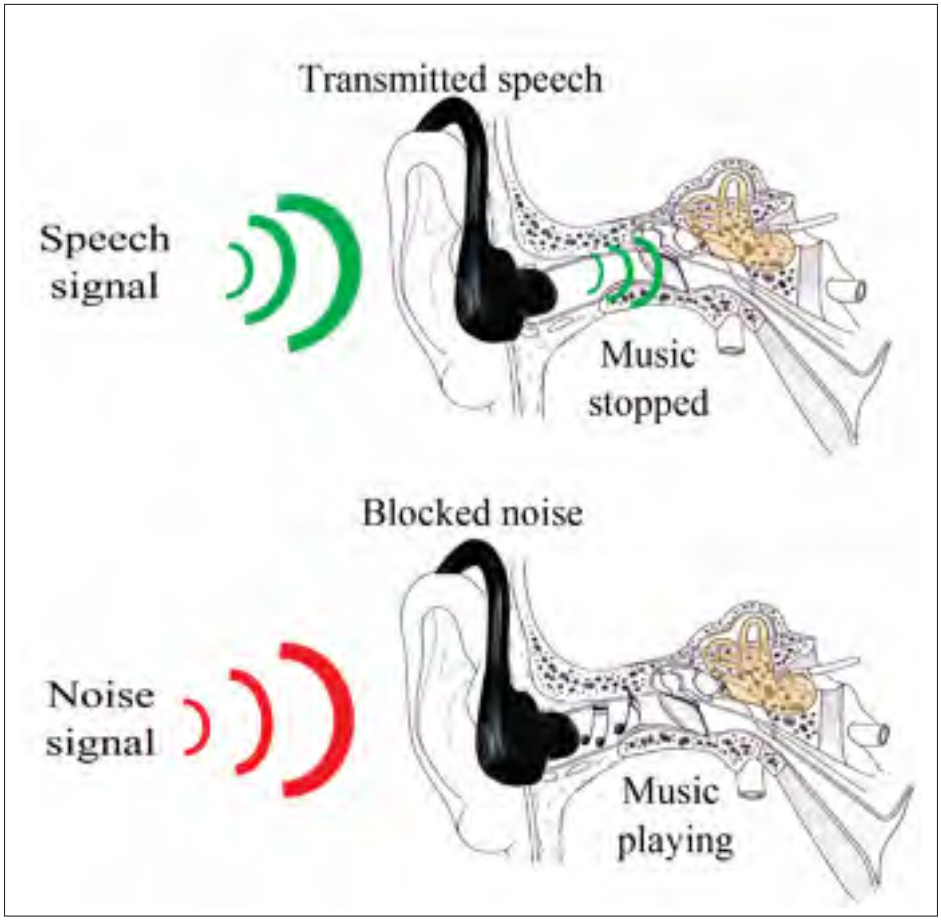

Figure 3.2 The selective operating principle of the system.

formant can be reliably detected compared to the second formant, which is heavily masked by noise in low SNR (0 dB). Based on these findings, we propose the use of an energy feature

which is extracted from the frequency region containing the first formant for speech characterization. Thereafter, this feature is normalized using two noise features extracted from the frequency regions containing typical noise information. The normalization of the energy feature eliminates the need for an SNR estimator. The VAD's decision is made after multiple observations using two decision thresholds, determined from the normalized energy features in addition to a hangover scheme to consider the "long time" information, knowing that the speech signal is highly time-correlated (Davis *et al.*, 2006). The value of the two thresholds, the frequency bounds, the number of observations and the hangover parameters are optimized off-line using a genetic algorithm. The optimization increases the performance of the proposed VAD by maximizing the $F1$ score (C. J. van Rijsbergen, 1979). Figure 3.3 illustrates the detailed architecture of the proposed VAD algorithm. The signal is first time-windowed into i frames. Features are extracted and the decision $D(i)$ is made after N observations based on two thresholds and a hangover scheme.

3.4.1 Windowing

The entire signal is cut into frames using a Hamming window. The length of each frame is 25 ms with 80 % overlap.

3.4.2 Feature extraction

3.4.2.1 Filterbank

The incoming signal is filtered into M=3 frequency bands using 4th order Butterworth filters. Cut-off frequencies of the 3 bands (15-153 Hz, 153-1323 Hz, 1323-1944 Hz) have been optimized off-line using a Genetic algorithm (see Section 3.5.3).

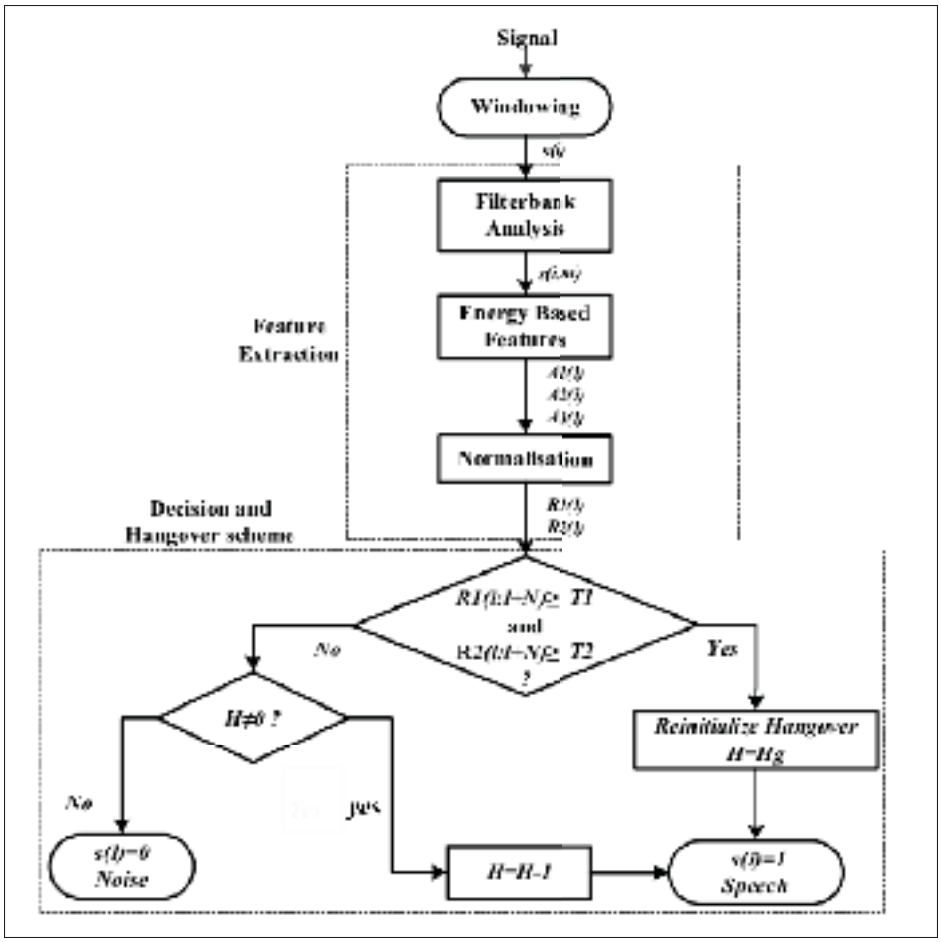

Figure 3.3 Block diagram of the proposed VAD algorithm.

3.4.2.2 Energy based feature

Parikh et al (Parikh and Loizou, 2005) concluded that when the speech signal is corrupted by noise; the first formant can be reliably detected. Based on the conclusions of this study, the energy of each frequency band is calculated. Figure 3.4 illustrates an example of the energy in the three frequency bands for one speech frame produced by a male speaker corrupted by car noise with 10, 5 and 0 dB SNRs. $A1$, $A2$, $A3$ denote the energy in the first, second, and third frequency bands respectively. One can see that in the second frequency band, which contains the first formant of the speech frame (a voiced phoneme), the energy of the speech is significantly higher than the energy of the noise in this band, whereas the energy of the noise in the first band (especially in 0 dB SNR) is higher for noise than speech.

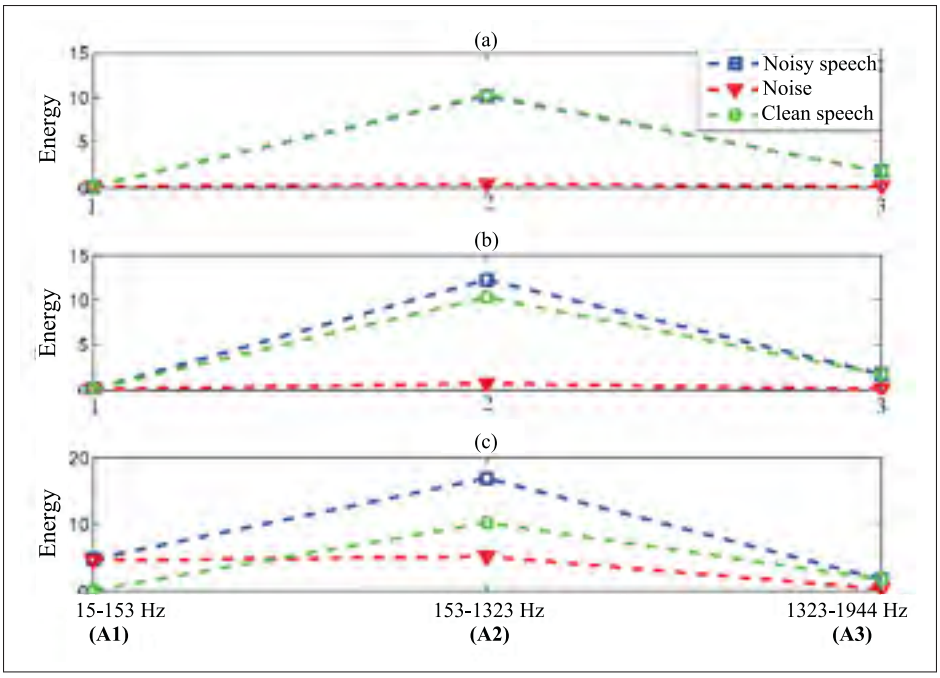

Figure 3.4 Energy in three frequency bands for one signal frame with (a) 10 dB, (b) 5 dB and (c) 0 dB SNR.

3.4.2.3 Normalization

While Figure 3.5 shows that $A2$ is a reliable indicator of the presence of speech, it cannot be used directly with a decision threshold in the VAD because it is dependent on the input signal level. Thus, the following normalized ratios, which increase the VAD's performance by taking advantage of the different frequency content of speech and noise ($A1$ and $A3$), are proposed:

$$\mathrm{R1} = \frac{A2}{A1} \tag{3.1}$$

$$\mathrm{R2} = \frac{A2}{A3} \tag{3.2}$$

$R1$ is normalized by the low-frequency components, knowing that noise signals generally have more energy in the lower frequencies than speech signals (Levitt, 2001). $R2$ is normalized

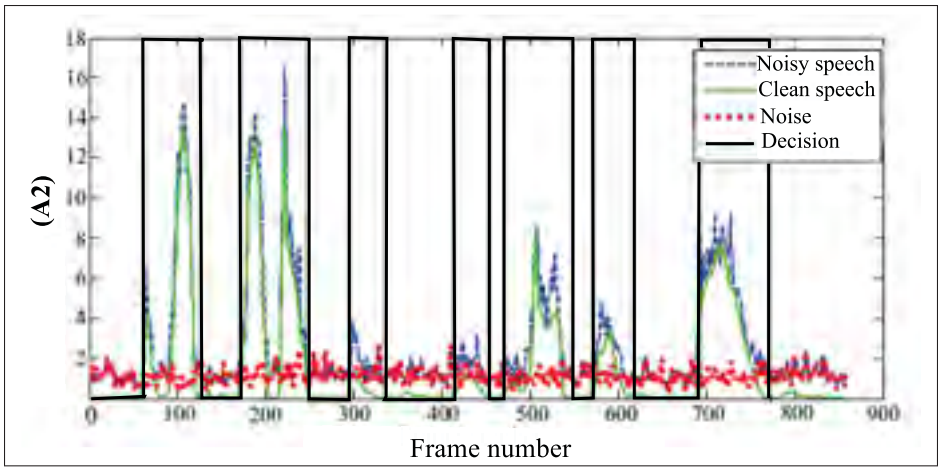

Figure 3.5 $A2$ in speech, noise and noisy speech signal with 0 dB SNR, in addition to the hand-labeled decision on clean speech.

by the high-frequency components that characterize high frequency noise signals. The VAD's decision is based on ratios $R1$ and $R2$, thus eliminating the need for an SNR estimator.

3.4.3 VAD's decision

3.4.3.1 The decision thresholds

Two decision thresholds $T1$ and $T2$ are fixed upon the ratios $R1$ and $R2$. The VAD's decision is made after N observations:

$$D(i) = \begin{cases} 1 & \text{if } R1(i:i+N) \geq T1 \text{ and } R2(i:i+N) \geq T2 \\ 0 & \text{else} \end{cases} \tag{3.3}$$

with N being the number of consecutive observations, i the frame number and $D(i)$ the decision in the current frame.

3.4.3.2 Start and end of speech confirmation parameters

The VAD's decision is made after multiple observations (start of speech confirmation parameter). These observations are defined by the number of consecutive frames having ratios $R1$ and $R2$ higher than thresholds $T1$ and $T2$ respectively and after which the decision is to be set to 1 (speech). Ramirez et al.(Ramírez *et al.*, 2005) demonstrated that taking several frames into account in the VAD improves the reliability of its decision. In the proposed VAD, the number of consecutive frames should not exceed 8 frames to not exceed a delay of 40 ms. Hangover schemes (end of speech confirmation parameter) have been widely used in VADs to reduce the false rejection rate attributable to the non-detection of low energy speech frames containing consonants such as fricatives and unvoiced stops (Sohn *et al.*, 1999) (Davis *et al.*, 2006). In the adaptive Multi-Rate (AMR) VAD (ETSI, 1999), the hangover was set to 2 seconds if the signal is of a complex nature.

3.5 Off-line parameters optimization

The choice of the two decision thresholds $T1$ and $T2$ depends on the desired specificity and sensitivity of the VAD. High decision thresholds make the VAD more specific than sensitive, which minimizes both the False Positive Rate (FPR) and True Positive Rate (TPR). Low decision thresholds make the VAD more sensitive by maximizing the TPR and FPR. The two decision thresholds $T1$ and $T2$, the number of consecutive frames (start of speech confirmation), and the hangover (end of speech confirmation), in addition to the frequency bands' boundaries are optimized off-line using a genetic algorithm approach by maximizing an objective function.

3.5.1 Objective function

In the literature, VAD performance evaluation can be performed using various metrics (Beritelli *et al.*, 2002). Nevertheless, solving an optimization problem requires the use of one metric reflecting the entire performance of the VAD algorithm. For this purpose, the F1 score is used as the objective function (C. J. van Rijsbergen, 1979):

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{3.4}$$

with

$$\mathrm{precision} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FPR}} \tag{3.5}$$

$$\mathrm{recall} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FNR}} \tag{3.6}$$

The F1 score combines the TPR, FPR and False Negative Rate (FNR). It reflects the VAD's accuracy by considering its precision and recall. In VAD algorithms, TPR, FPR and FNR are respectively: the ratio of speech frames classified as speech, the ratio of noise frames classified as speech, and the ratio of speech frames classified as noise. In the existing VAD algorithms, these rates are calculated in noisy speech signals to distinguish between speech and noise frames. However, for a smart earphone application, the TPR and FNR are calculated for noisy speech signals and the FPR for noise signals. This evaluation method focuses on the fact that once the speech frames have been detected, the detection of the next non-speech frames does not have any detrimental effect on the performance of the proposed VAD. Whereas when no speech signal is present, the detection of noise frames and their transmission to the protected ear is significantly detrimental on the performance of the proposed VAD.

3.5.2 Audio signals used for off-line optimization

Off-line parameters optimization is conducted to maximize the F1 score, using a small number of noisy speech signals. In the envisioned application, noise signals typical of everyday environments are to be used. Thus 20 speech signals (14 speech signals produced by male speakers and 6 speech signals produced by female speakers) from the TIMIT database (Zue *et al.*, 1990) corrupted by "Airport" noise recorded in real world environment with 5 dB SNR are used. Speech and noise were artificially mixed together with 5 dB SNR. The TIMIT database was chosen for the envisioned application because the speech signals in this database are not altered by filters such as the ITU MIRS or ITU G.712, that tend to consider the realistic frequency

characteristics of terminals and equipment in the telecommunication area (Hirsch and Pearce, 2000).

3.5.3 Genetic algorithm for off-line parameters optimization

Genetic algorithms (Goldberg, 1989) are randomized search and optimization techniques based on the mechanism of natural selection and natural genetics. They are robust and efficient, they adapt to a wide variety of environments and they produce a near optimal solution when solving an optimization problem. The genetic algorithms are used to optimize the frequency bands' boundaries, the hangover, the number of consecutive observations, and the decision thresholds. In the optimization process, the lower and upper frequency bounds variations for the three band-pass filters are illustrated in Table 3.1. The lower bound of the second and third frequency bands correspond to the upper bound of the first and second frequency bands respectively.

Table 3.1 Frequency bands' lower and upper bounds for the optimization process

Bounds	**Lower bound (Hz)**	**Upper bound (Hz)**
1	10	20
2	50	250
3	250	1500
4	1500	6000

The hangover varies from 50 to 300 frames with a step of one frame (0.25 to 1.5 second). The number of observations varies from 4 to 8 consecutive frames, which is equivalent to a decision delay varying from 20 to 40 ms. After 10 generations, the genetic algorithm reached an optimal solution with an F1 score of 98.5%. Figure 3.6 shows a plot of the function's best and mean penalty values in each generation with each generation being composed of 40 individuals. The optimization process gave a hangover value of Hg=1.26 seconds and a number of consecutive frames N=7. The optimized cut-off frequencies for the three band-pass filters are: [15, 153, 1323, 1944] Hz. These parameters are then used for the decision-making and the validation of the proposed VAD algorithm using a validation database.

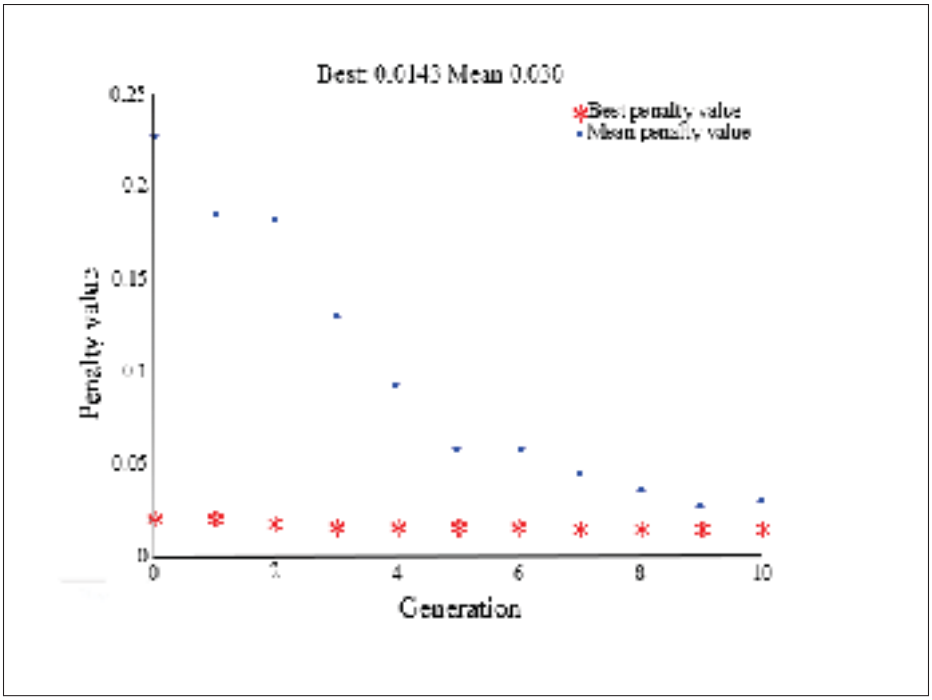

Figure 3.6 Penalty values (1-F1) of the optimization process using Genetic Algorithm.

3.6 Experiments and validation

3.6.1 Validation database

The validation database is composed of 10 sentences produced by 630 speakers (439 male speakers and 191 woman speakers) from the TIMIT database (Zue *et al.*, 1990). Signals are sampled at 8 kHz. All 10 sentences are concatenated into one signal. Noisy speech signals were created by adding the same noise at three SNRs (10, 5, and 0 dB) to each concatenated speech signal. Four noise signals obtained from real world recordings were used. These noises are representative of everyday environments to which consumers may be exposed to:

- Car: this environment tends to mimic the noise of the wind perceived by car passengers with opened windows.

- Airport: this noise was recorded in the hall of an airport, with talking crowds and baggage trolleys passing by.
- Hammer: this noise contains transient noises. It is used to mimic some scenarios such as renovations in the neighbourhood, or constructions in the street.
- Train: this noise was recorded near a railway with sounds of trains passing by.

3.6.2 Performance evaluation

The performance evaluation is conducted using the F1 score, in addition to the TPR and FPR. The proposed algorithm is compared to Sohn's VAD (Sohn *et al.*, 1999) which uses the first signal's frames to estimate the a posteriori and the a priori SNR to make the decision. Figure 3.7 illustrates the F1 score results of both algorithms in all noise conditions. As it can be seen in this figure, the F1 score of the proposed algorithm outperforms the F1 score of Sohn's algorithm in all noise environments and SNRs. In applications such as the smart earphones (to simultaneously enable the wearer to listen to music and transmit speech signals when present), the less desirable situation is the detection of short-time noise. This situation occurs when the false positive rate is high. Table 3.2 presents the true positive rate and false positive rate for the two VADs.

The FPR average of the proposed VAD is 4.2% compared to Sohn's VAD which has a FPR average of 29.9%. The same FPR is found in the three SNRs of each noise since both VADs are insensitive to the level of the incoming signal. Furthermore, the TPR of the proposed algorithm is higher than the TPR of Sohn's algorithm in all noise environments in the range of 5 and 10 dB SNR. This is due to the hangover scheme presented previously, which permits the detection of almost all the speech frames without interruptions or mid-speech clipping.

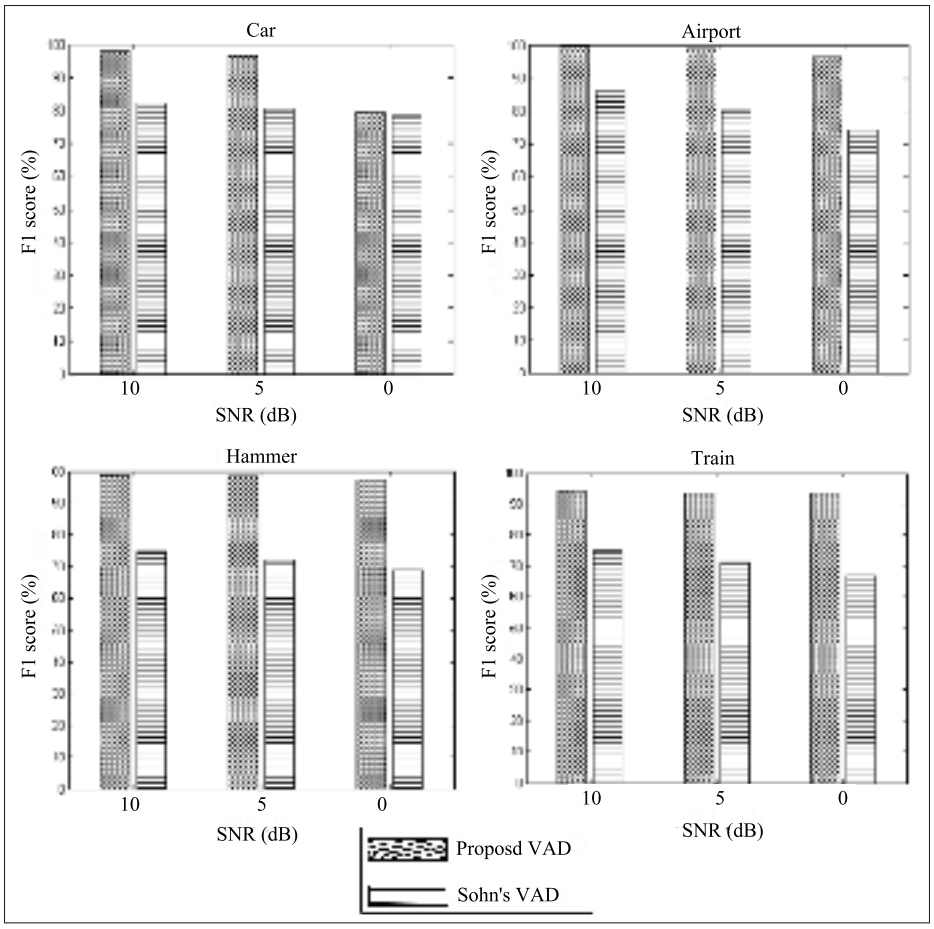

Figure 3.7 F1 scores of Sohn's and the proposed VAD in four noise environments with 10, 5, and 0 dB SNR.

3.7 Hardware implementation

3.7.1 DSP overview

The DSP used for the implementation of the VAD is a stream-oriented DSP core provided in a small 32-lead, 5 mm x 5 mm package. The Analog to Digital Converter (ADC) and the Digital to Analog Converter (DAC) are high quality 24 bit stereo audio converters, and can operate at sampling frequencies ranging from 8 kHz to 96 kHz. The DSP core consist of a simple multiply-accumulate (MAC) unit with a data source and a coefficient source.

Table 3.2 Performance evaluation of the proposed and Sohn's VADs in four noise environments and three SNRs

Noise Environment		Proposed VAD (%)		Sohn's VAD (%)	
Noise	SNR	TPR	FPR	TPR	FPR
Car	10	97.6	0	87.5	20.9
	5	91.3	0	76.1	20.9
	0	73.4	0	60.2	20.9
Airport	10	98.4	0	85.5	14.9
	5	97.0	0	72.9	14.9
	0	88.4	0	55.6	14.9
Hammer	10	98.7	0	91.7	50.2
	5	98.4	0	85.3	50.2
	0	96.7	0	77.6	50.2
Train	10	97.7	16.8	92.7	33.9
	5	91.1	16.8	86.4	33.9
	0	68.4	16.8	76.5	33.9
Average		91.4	4.2	79.0	29.9

Three RAMs are encompassed in the address space of the DSP: the program RAM, the coefficient RAM, and the data RAM. The program RAM governs the execution of the instructions in the core, and cannot exceed 1024 instructions per audio frame. The parameter RAM stores the initial coefficients of the program and cannot exceed 1024 coefficients, while the data RAM stores audio data-words for processing in addition to some run-time parameters. The data RAM is divided into two memory addressing types: modulo and non-modulo memories. Each of the modulo and non- modulo data RAM offer 4096 memory words.

3.7.2 Hardware implementation

The Auditory Research Platform (ARP) (Mazur and Voix, 2013) integrates the DSP in addition to other associated electronics such as audio inputs, audio outputs, and battery. It is used to implement the proposed VAD in real-time. Figure 3.8 illustrates the ARP with two earpieces, in each ear-piece an external miniature microphone and an internal miniature loudspeaker are integrated for external sound acquisition and VAD's decision transmission respectively.

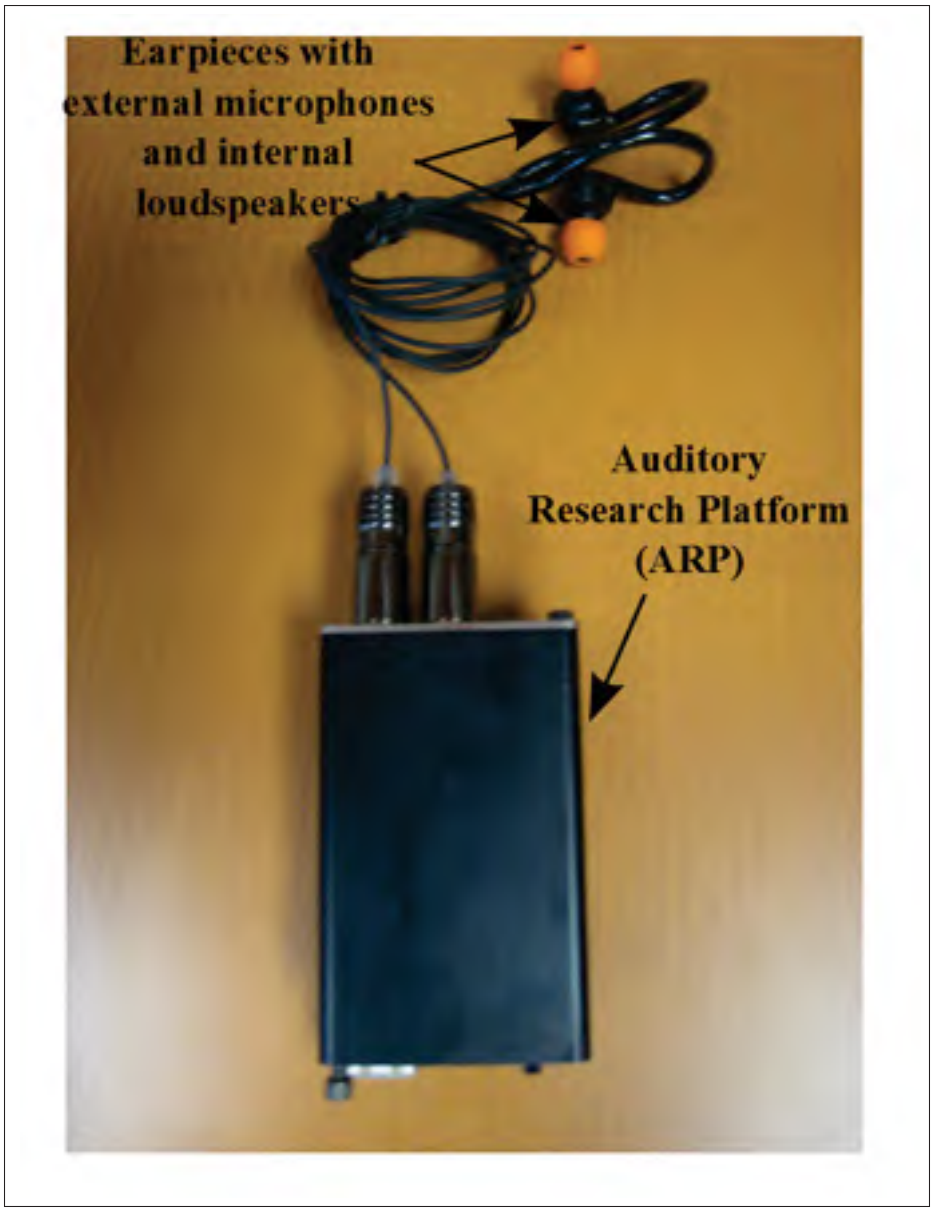

Figure 3.8 The auditory research platform in which the VAD is implemented for real-time processing connected to two earpieces for audio signal acquisition and VAD's decision transmission.

The hardware VAD implementation is made following the steps described in Section 3.4. The resulting number of instructions per audio frame is 890, which is equivalent to a rate of 87% from the entire program RAM. The data RAM used by the VAD is 346 (8% from the entire modulo data RAM, and 0% from the non-modulo data RAM), while the coefficient RAM used is 240 (23% of the coefficient RAM).

3.7.3 VAD real-time tests and validation

The real-time validation of the proposed algorithm is performed using some of the noisy audio samples used in the first validation process presented in Section 3.6. For this purpose, the audio input of the ARP was connected to the audio output of a computer in which the noisy signals were playing, while the output of the VAD's decision was saved in the computer to compare it with the result of the first validation presented in Section V. Figure 3.9 illustrates an example of the comparison between the results of the VAD before its hardware implementation and the VAD implemented in the DSP. This comparison is made using a signal composed of 3 s of airport noise, 3 s of speech corrupted by airport signal with 5 dB SNR, and 3 s of airport noise signal. The decision of the VAD in simulation and its hardware implementation are equivalent.

3.8 Discussions and conclusions

In this paper, a robust and yet simple real time VAD for smart earphones is presented. This VAD uses an energy-based feature for the characterization of speech and noise signals. The speech and noise characteristics are thereafter normalized and two decision thresholds are determined. The decision is made after multiple observations and triggers a hangover scheme. The algorithm parameters are optimized off-line using a genetic algorithm by maximizing the F1 score which represents the global performance of the VAD. The parameters optimization is performed using 20 speech signals corrupted by airport noise with 5 dB SNR. The first experiment for the validation of the proposed VAD was conducted using 10 sentences produced by 439 male speakers and 191 female speakers corrupted by four noise environments. These experiments showed that the proposed VAD is more efficient than a benchmark VAD. Coupling multiple observations and the hangover scheme in the decision process shows that the proposed VAD detects almost all speech signals without interruption since the true positive rate average is 91.4%. The entire VAD system was validated for the smart earphones application. The proposed VAD was implemented in a miniature low-power DSP integrated in a research platform in which the audio inputs, battery, and other electronics were selected for real-time implementation. The hardware resources show that other tasks can be combined to the VAD

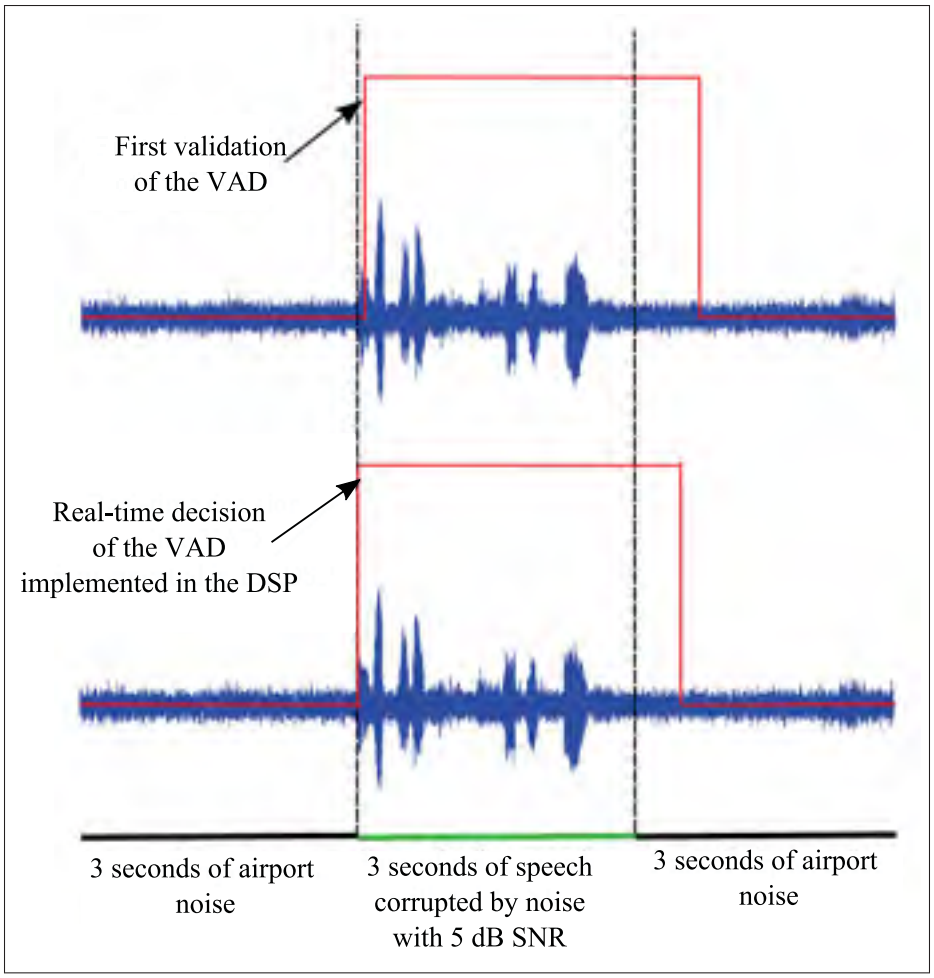

Figure 3.9 Comparison between the VAD decision on the computer and the VAD decision obtained from the output of the DSP.

such as a low complexity on-line parameters optimization algorithm to allow to the VAD to adapt for each noise environment in which the smart earphones are used.

CHAPTER 4

NOISE REDUCTION OF SPEECH SIGNAL USING TIME-VARYING AND MULTI-BAND ADAPTIVE GAIN CONTROL

Narimene Lezzoum[1], Ghyslain Gagnon [1], Jérémie Voix [1]

[1] École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted to « Applied Acoustics», an Elsevier Journal, in September 2015 and published in March 2016

4.1 Abstract

In this paper, a single-channel speech enhancement algorithm based on non-linear and multi-band adaptive gain control (AGC) is proposed. The algorithm requires neither Signal-to-Noise Ratio (SNR) nor noise parameters estimation. It reduces the background noise in the temporal domain rather than the spectral domain using a non-linear and automatically adjustable gain function for multi-band AGC. The gain function varies in time and is deduced from the temporal envelope of each frequency band to highly compress the frequency regions where noise is present and lightly compress the frequency regions where speech is present. Objective evaluation using the PESQ (Perceptual Evaluation of Speech Quality) metric shows that the proposed algorithm performs better than three benchmarks, namely: the spectral subtraction, the Wiener filter based on *a priori* SNR estimation and a band-pass modulation filtering algorithm. In addition, blind subjective tests show that the proposed algorithm introduces less musical noise compared to the benchmark algorithms and was preferred 78.8% of the time in terms of signal quality. The proposed algorithm is implemented in a miniature low power digital signal processor to validate its feasibility and complexity for smart hearing protection in noisy environments.

4.2 Introduction

Nowadays, wearing Hearing Protection Devices (HPD) in workplaces and noisy environments becomes a necessity for people exposed to high noise levels on a daily basis, to protect them from what otherwise would damage the inner ear and induce hearing loss. Yet, in reality, most wearers will not use their HPD when oral communication is needed (Hong *et al.*, 2008). To palliate this problem, we intend to develop a smart HPD (S-HPD) that guarantees protection and discriminates between speech and noise to allow the transmission of enhanced speech signals to the protected ear. For this purpose, the integration of a Digital Signal Processor (DSP), an external microphone and an internal loudspeaker in a passive HPD are required (Carbonneau *et al.*, 2013). With one external microphone, single-channel speech enhancement can be performed.

While multi-channel speech enhancement takes advantage of the spatial audio information, single-channel speech enhancement does not benefit from these information and therefore remains a challenging task, especially in low signal-to-noise ratio (SNR). Single channel speech enhancement algorithms can be grouped into four main types (Loizou, 2007): spectral subtractive, linear estimators, non-linear estimators, and subspace algorithms. Despite significant differences in the type of estimated parameters, most of the aforesaid speech enhancement algorithms use the signal's first frames to estimate the noise parameters and update these parameters in non-speech segments using a voice activity detector (VAD). However, most existing VAD algorithms are unreliable in low SNRs (Beritelli *et al.*, 2002).

Single-channel speech enhancement algorithms usually perform in the spectral domain (Boll, 1979), and (Scalart and Fiho, 1996). However, reducing the noise in the spectral domain may generate musical noise due to a random amplification of frequency bins that varies over time (Leitner and Pernkopf, 2012) and (Cappé, 1994). In some cases, musical noise is more annoying than the background noise itself. Much research has been conducted for the reduction of the musical noise using post-filtering techniques or image processing approaches such as in (Leitner and Pernkopf, 2012), (Esch and Vary, 2009), and (Hasan and Hasan, 2009).

Other single-channel speech enhancement algorithms based on modulation filtering have been proposed such as (Hermansky and Morgan, 1994), (Falk *et al.*, 2007), and (Paliwal *et al.*, 2010). These methods require the constant computation of Discrete Fourier Transforms (DFTs) and Inverse DFTs (IDFTs), making their use incompatible for real-time, low-latency applications in embedded systems.

In (Westerlund *et al.*, 2004) a time domain speech enhancement algorithm based on an Adaptive Gain Equalizer (AGE) has been proposed. In this algorithm, a gain is applied for each frequency band based on an SNR estimation to boost the speech signal when the SNR in the frequency band is high. While this algorithm proved to be effective at enhancing speech, it does not significantly reduce the background noise when speech is absent. In the targeted application, where the user wears hearing protection in noisy environments, continuous reduction of background noise is an important feature. It would also be desirable in other applications such as noise-canceling ear-buds. In (Parikh *et al.*, 2009), an Adaptive Gain Control (AGC) based on an SNR estimator was proposed. Unfortunately, it was also shown in (Parikh *et al.*, 2009) that the proposed SNR estimation method is not accurate in low SNR environments (0 dB) and adds artefacts to the enhanced signal. In (Shahid *et al.*, 2011), an AGE applied to the multi-band temporal envelopes was proposed to boost the signal when speech is present. In this method, the gain function is applied to the temporal envelope which is afterwards multiplied by the carrier of the signal.

The authors introduced in (Lezzoum *et al.*, 2014b) a single-channel speech enhancement algorithm with a live demonstration using recordings. This paper extends this work with objective and subjective evaluations of this algorithm, its comparison with three other state of the art methods, in addition to its implementation in a miniature low-power DSP for smart hearing protection applications.

The proposed algorithm calculates a time-varying and frequency-band dependent gain function from the temporal envelope of each frequency-band. This function enables high compression of frequency bands containing noise and light compression of frequency-bands containing speech.

The proposed algorithm operates without any knowledge or estimation of the noise parameters, only assuming that the background noise is additive. It will be shown that this gain function reduces the background noise and improves the quality of the speech signal.

The paper is organized as follows. Section 4.3 details the proposed noise reduction algorithm. In Section 4.4, the experimental methodology is presented. Section 4.5 discusses the objective and subjective results. Section 4.6 presents the hardware implementation of the method, and Section 4.7 concludes the paper.

4.3 Proposed algorithm

Figure 4.1 illustrates the architecture of the proposed algorithm. The incoming noisy speech signal $y(n)$ is composed of clean speech $x(n)$ and additive noise $w(n)$:

$$y(n) = x(n) + w(n) \tag{4.1}$$

The incoming signal is divided into M=16 frequency bands using fourth order band-pass Butterworth filters. Filter bandwidth are characterized by the equivalent rectangular bandwidth (ERB) (Glasberg and Moore, 1990). The centre frequency of the first and last frequency bands are 125 and 3700 Hz respectively.

The output of each filter is given by:

$$y_m(n) = (y * h_m)(n) \tag{4.2}$$

with $h_m(n)$ the impulse response of the m^{th} band-pass filter, and the symbol "*" denoting convolution.

The Hilbert envelope of the signal $y_m(n)$ is extracted as per the following equation:

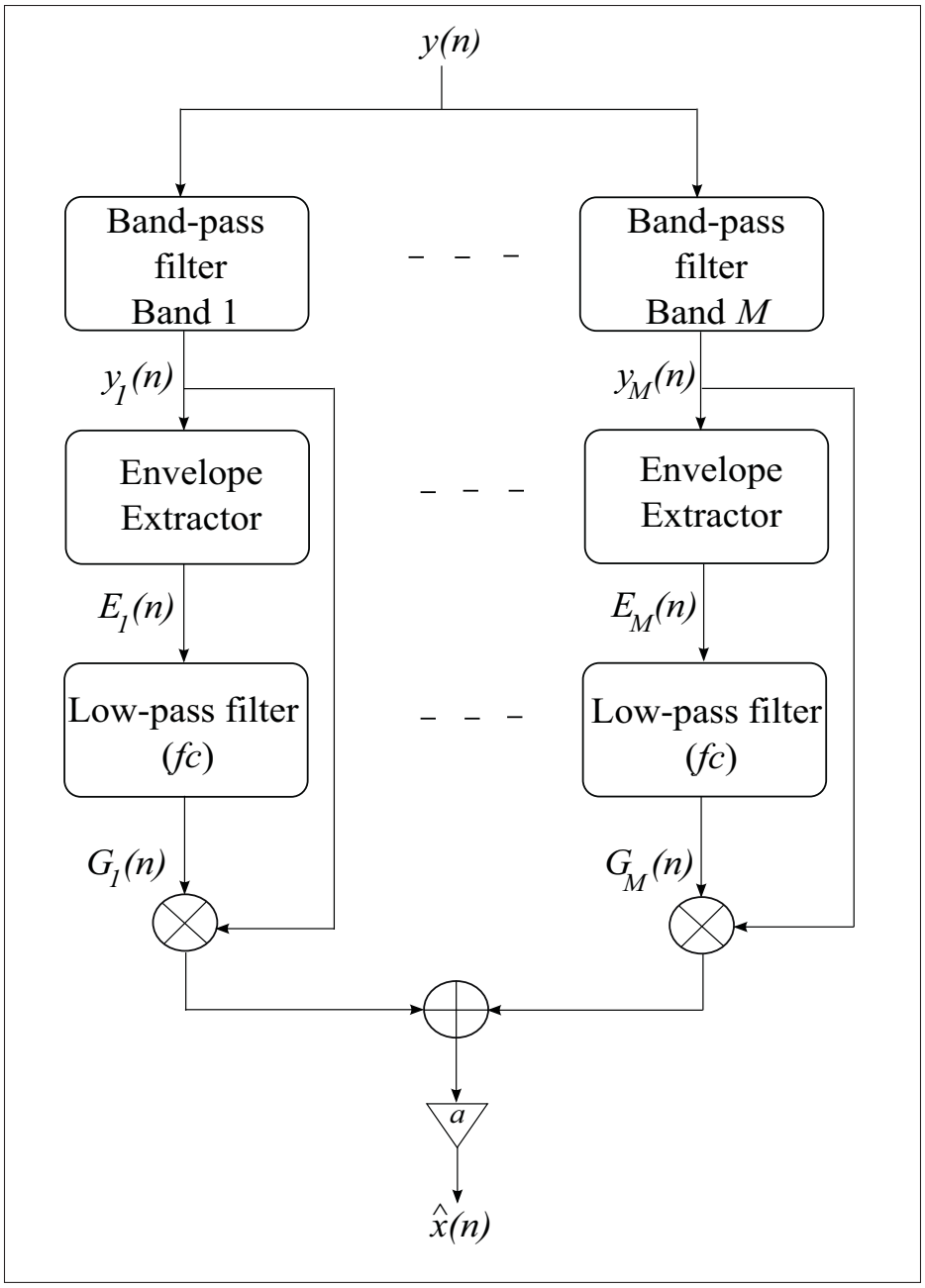

Figure 4.1 Block diagram of the proposed speech enhancement algorithm.

$$E_m(n) = \sqrt{y_m(n)^2 + \tilde{y}_m(n)^2} \tag{4.3}$$

with $\tilde{y}_m(n)$ the Hilbert transform of $y_m(n)$, defined as (Choi and Jiang, 2008):

$$\tilde{y}_m(n) = y_m(n) * \frac{1}{\pi n} \tag{4.4}$$

The proposed technique achieves noise reduction using multi-band time-varying gain functions. Our investigation shows that these gain functions must meet three criteria:

- The gain function of each frequency band must be smooth and continuous to avoid abrupt changes in the enhanced signal.
- The gain function must be chosen as a function of the temporal envelope $E_m(n)$ in order to preserve the quality of speech without adding artefacts.
- The gain function should be near 1 in the frequency bands containing speech and near 0 in the frequency bands containing noise, in order to preserve speech components and attenuate noise components.

A time-varying gain function that fulfils these criteria is the low-pass filtered temporal envelope $E_m(n)$ of the signal. The gain function is thus:

$$G_m(n) = (E_m * L)(n) \tag{4.5}$$

with $L(n)$ the impulse response of a fourth order low-pass filter. The optimal cut-off frequency f_c of the low-pass filter is later determined in Section 4.4.1.

Enhancing the signal in each frequency band consists of multiplying the signal by its smoothed envelope:

$$\hat{x}_m(n) = G_m(n).y_m(n) \tag{4.6}$$

This can be seen as non-linear compression: frequency bands with high energy will barely be compressed and frequency bands with low energy will be highly compressed.

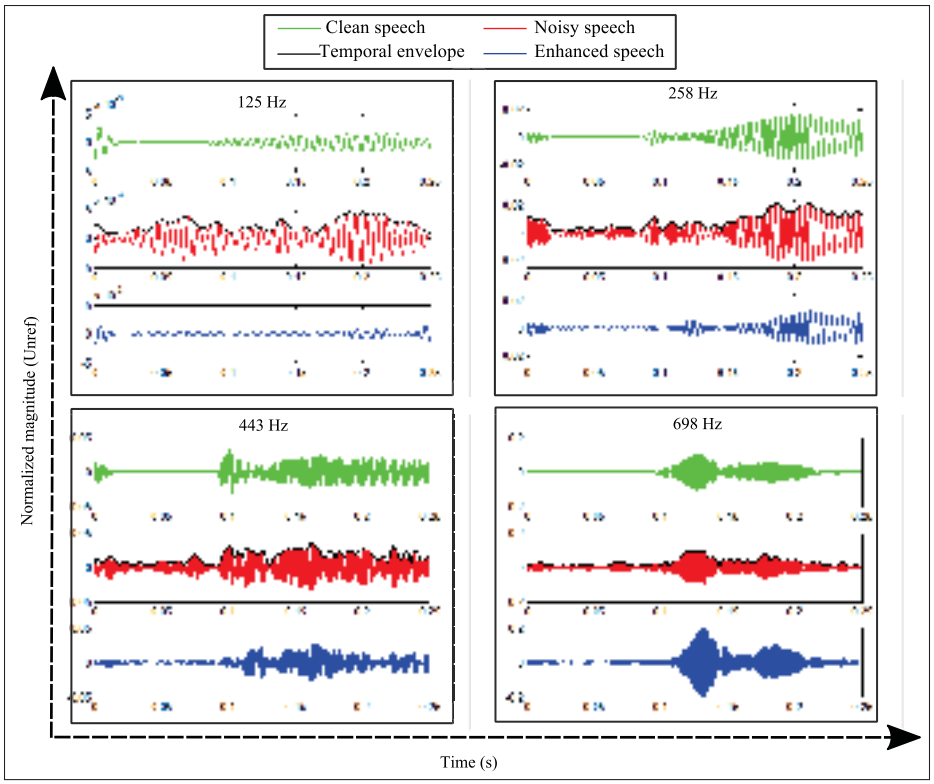

Figure 4.2 An example of the effect of the gain function for noise reduction of a voice phoneme in four frequency bands, centered at: 125 Hz, 258 Hz, 443 Hz, and 698 Hz. The noisy speech is corrupted by car noise in 0 dB SNR.

The enhanced signal $\hat{x}(n)$ of each frame is reconstructed by summing the M frequency bands, and rescaled using a gain constant a. In this paper, "a" is the ratio between the RMS (Root Mean Square) values of the noisy and enhanced signals. This gain constant could also be set by the user to adjust the desired listening level.

As an illustrative purpose, Figure 4.2 shows the noise reduction effect of the gain function on a 250 ms speech signal corrupted by car noise. The clean speech signal is considered as the reference to see how the gain function reduces the background noise continuously in the temporal domain and in the different frequency bands. When speech is absent (from 0 to 0.1 second in the four frequency bands), the background noise is highly compressed. Figure 4.2 also shows that speech signal is amplified in the frequency band centered at 698 Hz due to the presence of the first formant.

Figure 4.3 illustrates three waveform and spectrogram charts: clean speech produced by a male speaker, speech corrupted by car noise in 5 dB SNR, and the enhanced speech signal. This figure shows, in the temporal and spectral domains, the background noise-reducing effect of the proposed algorithm.

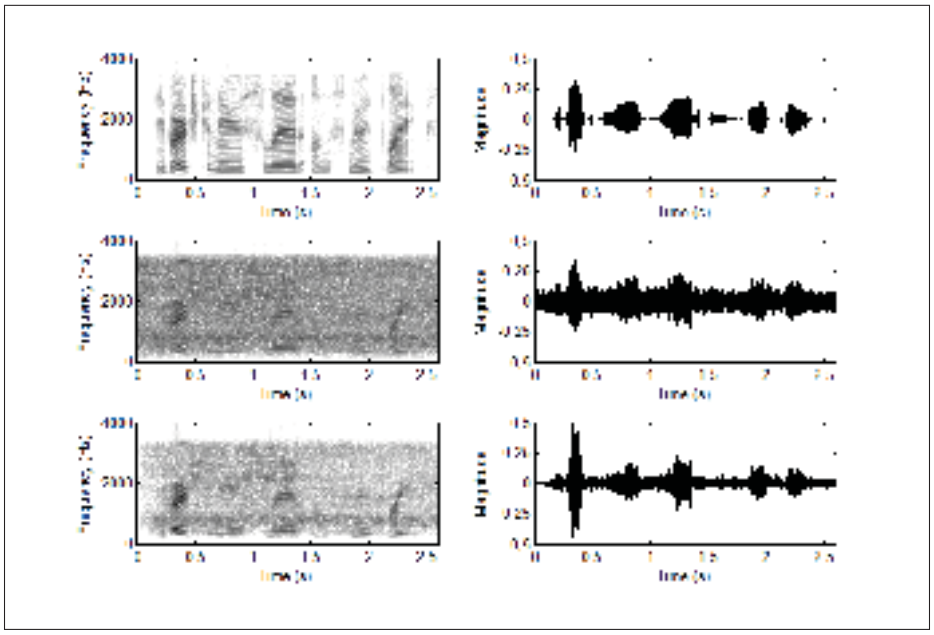

Figure 4.3 On the left are the spectrograms and on the right their corresponding waveforms: top panel, the clean speech signal (a male speaking: “the birch canoe slid on the smooth planks”), middle panel, the same speech signal corrupted by car noise in 0 dB SNR, bottom panel, the enhanced signal with the proposed method.

4.4 Experimental methodology

Although the proposed algorithm can perform in a sample-based approach, it was implemented in Matlab using 250 ms frames with 80% overlap for ease of simulations. Objective and subjective quality tests were conducted to evaluate the performance of the proposed noise reduction algorithm. The results are shown in Section 4, and some audio samples are available online (Lezzoum *et al.*, 2015).

4.4.1 Optimal cut-off frequency of the gain function

The fluctuation rate of the temporal envelope is called the modulation frequency and represents one of the characteristics of speech signal. In (Drullman *et al.*, 1994a), a study on the impact of the modulation frequency on speech intelligibility was performed: the speech signal was divided into different frequency bands, and the temporal envelopes and fine structures of each frequency band were extracted. The temporal envelope has been low-pass filtered with different cut-off frequencies (0, 0.5, 1, 2, 4, 8, 16, 32 and 64 Hz) to determine the most important modulation frequency range for speech intelligibility, knowing that the cut-off frequency are frequency-band independent. In these studies, it has been found that with a modulation frequency of 16 Hz, the speech intelligibility remains the same, and when reducing it, the speech intelligibility starts decreasing.

In this work, the same evaluation has been performed to find the optimal fluctuation rate of the gain function that represents the cut-off frequency f_c of the low pass filtered envelope. This was achieved by using the perceptual evaluation of speech quality (PESQ) (ITU-T P.862, 2001) as the objective function to maximize.

Signals from the Noizeus corpus (Hu and Loizou, 2008) were used: 30 speech utterances corrupted by two noise environments, babble and car, in three SNRs (5, 0, and -5 dB). All signals were sampled at 8 kHz. Figure 4.4 illustrates the PESQ metric obtained using 6 low-pass filters with cut-off frequencies at 4, 8, 16, 32, 64, and 128 Hz. This figure shows that higher PESQ scores are obtained at 16 Hz in both noise environments and SNRs. In 4 and 8 Hz the temporal envelope is almost at a constant value. Thus, the gain functions calculated using these low cut-off frequencies are equivalent to a constant gain in each frequency band. Low-pass cut-off frequencies of 32, 64, and 128 Hz also gave a PESQ score lower than the 16 Hz, due to their high fluctuation rate, which brings up artefacts in the enhanced signal. From Figure 4.4, we conclude that 16 Hz is the optimal cut-off frequency for the gain function low-pass filter.

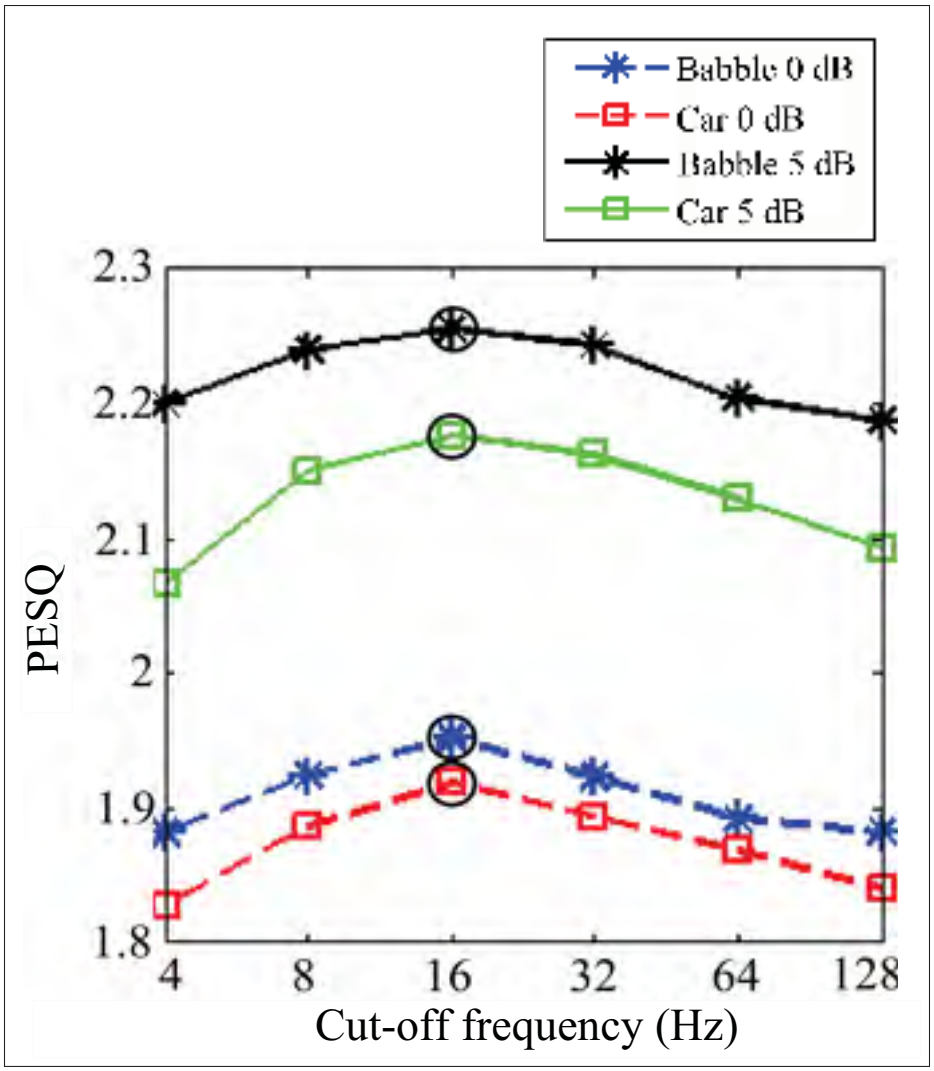

Figure 4.4 The PESQ metric calculated with different cut-off frequencies for speech signal corrupted by car and babble noise in 5 and 0 dB SNR.

4.4.2 Objective evaluation

Although the ITU.T P862 standard (ITU-T P.862, 2001) mentions that the developed PESQ metric has not been validated for noise reduction algorithms, it was shown in (Hu and Loizou, 2008) that among seven objective speech quality metrics for the evaluation of speech enhancement algorithms, the PESQ metric is the most correlated with overall quality and signal distortion. Thus, in this paper, the PESQ is used as the objective metric, using the same signals from

the NOIZEUS database used in Section 4.4.1. The performance of the proposed algorithm was compared to three noise reduction algorithms implemented in Matlab, namely: the Wiener filter based on an *a priori* SNR estimation (Scalart and Fiho, 1996), the spectral subtraction (Boll, 1979), and band-pass modulation filtering (Falk *et al.*, 2007). The Wiener filter and spectral subtraction codes were taken from (Loizou, 2007) (wiener-as and SpecSub), while the code of the modulation filtering was obtained directly from the authors of (Falk *et al.*, 2007). In the spectral subtraction, the noise spectrum was estimated and updated from non-speech frames detected using a simple VAD based on segmental SNR, while in the Wiener filter, non-speech frames were detected using *a priori* SNR estimation. The Wiener filter and spectral subtraction were chosen because of their wide use as benchmark algorithms (e.g. (Ming *et al.*, 2011), (Chen and Loizou, 2010), (Paliwal *et al.*, 2010)).

Other recent algorithms (e.g. (Westerlund *et al.*, 2004), (Parikh *et al.*, 2009), (Shahid *et al.*, 2011)) were not selected as benchmarks because, unlike the selected benchmarks, the code was not available to the authors. Implementation intricacies such as non-optimal parameter settings could have potentially led to biased comparisons.

4.4.3 Subjective evaluation

In addition to the objective evaluation procedure described above, two series of tests were performed to subjectively compare the proposed algorithm to benchmarks, in terms of level of musical noise and overall quality. To conduct these tests, 20 participants were recruited.

4.4.3.1 Musical noise assessment

In this test, the proposed algorithm and the three aforementioned benchmark algorithms are evaluated in terms of musical noise generation. Similarly, 30 speech signals corrupted by “car” and “babble” noise with 5, 0 and -5 dB SNR were used. The signals were identified by numbers and presented in a random order to 10 trained participants. These participants were asked to choose from the 4 signals processed with the four algorithms, the signal in which musical noise

was the least perceptible. Before this test, participants were trained by being exposed to several speech signals heavily corrupted by musical noise.

4.4.3.2 Overall quality evaluation

The overall quality was subjectively evaluated to determine user preference among the proposed algorithm, three benchmarks and unprocessed signals. This evaluation was performed using 30 speech signals corrupted by "car" and "babble" noise with 5, 0 and -5 dB SNR, with 10 participants. These participants were asked to pick the signal that has the best overall quality. During the tests, participants were allowed to repeat the same signal as often as needed.

4.5 Results and discussions

4.5.1 Objective test results

Figure 4.5 presents a comparison of PESQ results obtained in the two noise environments and three SNRs using the noisy signals, the Wiener filter, spectral subtraction, band-pass modulation algorithm and the proposed algorithm. In car noise, the band-pass modulation filtering, the Wiener filter, and the proposed algorithm improve the quality of speech almost equally. However, in babble noise, the proposed algorithm shows better performances than the other three benchmark algorithms. For instance, in -5 dB SNR with babble noise, the proposed algorithm PESQ score was 1.62 while the Wiener algorithm scored 1.47, the modulation filtering scored 1.36 and the spectral subtraction scored 1.19.

4.5.2 Subjective test results

4.5.2.1 Musical noise results

Subjective results conducted for the evaluation of the proposed and benchmark algorithms in terms of musical noise perception were averaged over all participants and signals. The

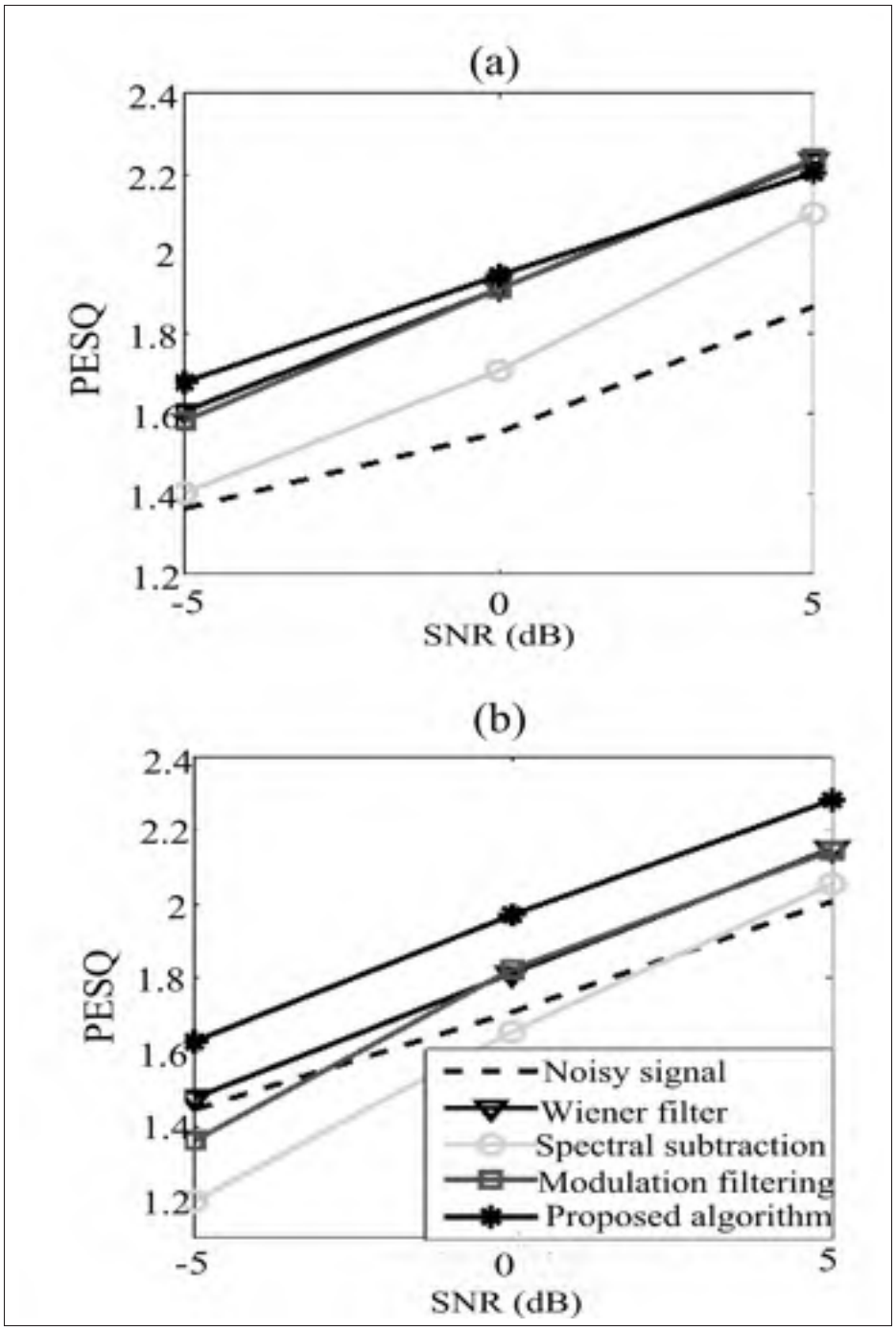

Figure 4.5 PESQ results for (a) car noise and (b) babble noise in -5, 0, and 5 dB SNRs using the unprocessed signals, the Wiener algorithm, spectral subtraction, band-pass modulation filtering and the proposed algorithm.

proposed algorithm was chosen 96.3% of the time to be the algorithm with the least musical noise, while spectral subtraction, the Wiener filter and modulation filtering were chosen 1.1%, 1.5% and 1.1% of the time respectively.

4.5.2.2 Overall quality results

The results of the subjective test evaluating the overall quality of processed and unprocessed signals are illustrated in Table 4.1. Overall, in both noise environments (car and babble) and the three SNRs (5, 0 and -5 dB), the proposed algorithm was chosen to be the algorithm with the best overall quality 78.8% of the time.

Table 4.1 Overall quality results for the proposed algorithm, the three benchmarks (SS corresponds to the spectral subtraction, MF to the modulation filtering, and W to the Wiener filter) and the noisy signals, in car and babble noise with 5, 0 and -5 dB SNRs. Results indicate the proportion of users who preferred each algorithm for a given combination of noise and SNR conditions.

Noise Environments		Proposed (%)	SS (%)	MF (%)	W (%)	Noisy (%)
Car	5	82.2	0.0	15.5	2.3	0.0
	0	80.0	4.4	8.9	0.0	6.7
	-5	66.7	8.9	17.8	4.4	2.2
Babble	5	97.7	0.0	2.3	0.0	0.0
	0	86.6	4.4	2.3	6.7	0.0
	-5	60.0	6.7	20	4.4	8.9
Average		78.8	4.1	11.1	3.0	3.0

4.6 Hardware implementation

The proposed algorithm was designed to be implemented in real-time on an embedded system. To validate this goal, we show in this Section implementation details of the algorithm on a low power DSP. The main purpose of this hardware implementation is the development of a smart hearing protection device that enables enhanced speech signals to be transmitted to the ear while protecting the S-HPD wearer from background noise.

4.6.1 DSP overview

The DSP used for the implementation of the proposed algorithm was provided in a small 32-lead, 5 mm x 5 mm package. The Analog to Digital Converter (ADC) and the Digital to Analog Converter (DAC) are 24 bit stereo audio converters. They can operate at sampling frequencies ranging from 8 kHz to 96 kHz. In the adress space of the DSP three RAMs are encompassed: a program RAM, a coefficient RAM, and a data RAM. The program RAM cannot exceed 1024 instructions per audio frame and governs the execution of the instructions in the core. The parameter RAM stores the initial coefficients of the program and cannot exceed 1024 coefficients. The data RAM is divided into two memory addressing types: modulo and non-modulo memories. Each of the modulo and non- modulo data RAM offer 4096 memory words. This RAM stores audio data-words for processing in addition to some run-time parameters.

4.6.2 Hardware implementation

The DSP and other associated electronics such as audio inputs, audio outputs, and battery are integrated in an Auditory Research Platform (ARP) (Mazur and Voix, 2013). This platform is illustrated in Figure 4.6. Two earpieces are connected to this platform, and in each earpiece, an external miniature microphone and an internal miniature loudspeaker are integrated for external sound acquisition and sound transmission.

The hardware implementation of the noise reduction algorithm is made following the steps described in Section 4.3. The resulting number of instructions per audio frame is 333, which is equivalent to a rate of 32.5 % from the entire program RAM. The data RAM used by the algorithm is 140 (3.4 % from the entire modulo data RAM, and 0 % from the non-modulo data RAM), while the coefficient RAM used is 124 (12.1 % of the coefficient RAM).

4.6.3 Real-time test

Real-time tests of the proposed algorithm were performed using some noisy speech signals. For this purpose, the audio input of the ARP was connected to the audio output of a computer in

Figure 4.6 The auditory research platform in which the speech enhancement algorithm is implemented for real-time processing connected to two earpieces for enhanced signals transmission.

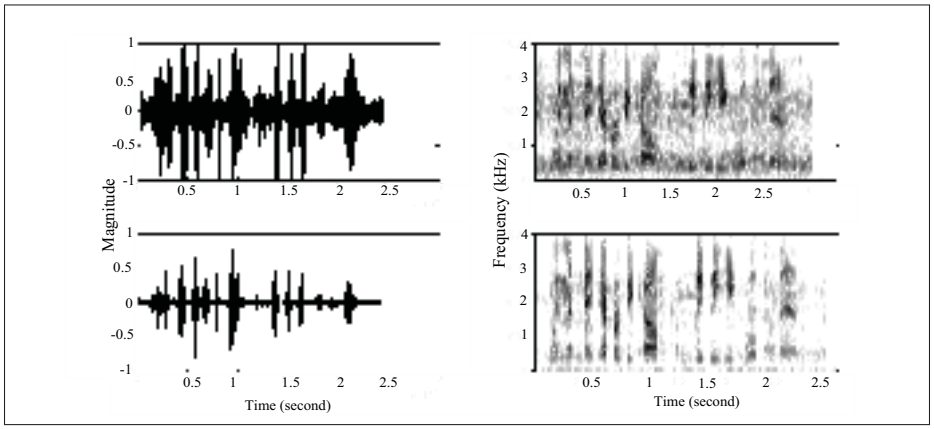

Figure 4.7 Waveform and spectrogram of speech corrupted by babble noise with 5 dB SNR (top panel) and the enhanced speech signal using the algorithm implemented in the hardware platform (bottom panel).

which the noisy signals were playing, while the enhanced signals were saved in the computer. Figure 4.7 shows the enhancement in the temporal and spectral domains.

4.7 Conclusions

This paper presented a noise reduction algorithm for the development of a smart hearing protection device that enables the transmission of enhanced speech while protecting the S-HPD wearer from noise. It demonstrated that noise reduction and speech quality improvement can be performed using a time-varying and frequency-band dependent gain function estimated from the low-pass filtering of the temporal envelope. The proposed method overcomes two types of problems in speech enhancement: one, the musical noise generally associated with processing in the frequency domain, and two, the amplification and attenuation distortions caused by an imperfect SNR and noise parameter estimation. Objective and subjective results show that the proposed algorithm improves the perceptual quality of speech signals without prior estimation of the noise or speech parameters. The hardware implementation of the proposed algorithm validates its reliability and low complexity for the intended real-time application. The hardware resources show that other tasks can be combined to the noise reduction method such as a voice activity detection algorithm to discriminate between speech and noise and transmit enhanced speech signals to the protected ear, while keeping the ear protected from noise when speech is not present. The proposed solution can also be integrated into active noise control headphones, which are already equipped with external microphone and other electronics. Future work includes subjective speech quality and intelligibility tests of the developed prototype for its real world validation.

Acknowledgments

The authors would like to thank Prof Tiago H. Falk for sharing the code of the modulation filtering algorithm. This work was supported by Sonomax Technologies Inc. and its “Industrial Research Chair in In-ear Technologies".

CHAPTER 5

SYNTHESIS, CONTRIBUTIONS, RECOMMENDATIONS AND FUTURE WORK

This thesis presented in chronological order the steps followed for the fulfillment of the PhD program and the development of speech-based algorithms for a smart hearing protector that enables face-to-face communication by transmitting noise reduced speech signals to the wearer while keeping him protected from high levels of noise. The following Sections present a synthesis of the research work in addition to recommendations and future work.

5.1 Synthesis of the research work

The main objective of the thesis was to enable face-to-face communication for wearers of hearing protection devices, by developing low complexity speech-based algorithms that could be implemented into a new electronic hearing protection device: the smart HPD (S-HPD).

To achieve this objective, two sub-objective needed to be reached: first, the development of a noise-robust voice activity detection (VAD) algorithm, and second the development of a real-time noise reduction algorithm for speech quality and intelligibility enhancement.

Before to the development of the algorithms, a subjective study has been conducted to determine the maximum processing delay that the algorithms should not exceed knowing that, in the case of a digital electronic circuit, the transmission of audio signals may be noticeably delayed because of the latency introduced by the digital signal processor and by the analog-to-digital and digital-to-analog converters (ADC) and (DAC). These delayed audio signals will hence interfere with the audio signals perceived naturally through the passive acoustical path of the device. For this purpose, two representative passive earplugs have been used to evaluate the shortest delay at which human listeners start to perceive two sounds composed of the signal transmitted through the electro-acoustic circuit and the passively transmitted signal: a shallow earplug fit and a deep earplug fit. The shortest delay is called the *echo threshold* and represents the delay between the time of perception of *one fused sound* from *two separate sounds*.

A transient signal, a clean speech signal, a speech signal corrupted by factory noise, and a speech signal corrupted by babble noise were used to determine the echo thresholds of the two earplugs. Twenty untrained listeners participated to the study, and were asked to determine the echo thresholds using a test software in which attenuated signals were delayed from the original signals in real-time. The findings showed that when using hearing devices, the echo threshold depends on four parameters: (a) the attenuation function of the device, (b) the duration of the signal, (c) the level of the background noise and (d) the type of background noise. Defined here as the shortest time delay at which at least 20 % of the participants noticed an echo, the echo threshold was found to be 8 ms for a bell signal, 16 ms for clean speech and 22 ms for speech corrupted by babble noise when using a shallow earplug fit. When using a deep fit, the echo threshold was found to be 18 ms for a bell signal and 26 ms for clean speech and 68 ms for speech in factory. No echo threshold could be clearly determined for the speech signal in babble noise with a deep earplug fit. Thus, this study showed that for speech signal, the processing delay that the algorithms should not exceed is 16 ms, when assuming that the earplug has a shallow fit.

After the determination of the processing delay, the first sub-objective consisted of developing a VAD algorithm that distinguishes between speech and noise. Two challenges were facing the development of the VAD: first, the low complexity challenge since the VAD had to be implemented in a low power DSP, and two the robustness against noise since the S-HPD is dedicated to be used in noisy environments which implies low signal to noise ratio (SNR).

The developed VAD uses two normalized energy features that compare the energy in the frequency region containing speech information with the frequency regions typically containing noise. The extraction of the normalized features represents the key of the proposed VAD since it eliminates the need for an SNR estimator. The VAD's decision is made using two threshold comparison rules computed from the normalized features and a hangover scheme triggered after a given number of observations. The algorithm parameters, namely the frequency regions' boundaries, number of observations, two decision thresholds and hangover's duration, have been optimized off-line using a genetic algorithm. This VAD was evaluated in different

noise environments and with different SNRs (10, 5, and 0 dB) and showed that over the three SNRs, the proposed VAD detects 91.4 % of the speech signal and 4.2 % of noise signal, which shows its robustness against noise compared to a benchmark algorithm which detects 79.0 % of speech and 29.9 % of noise in the same environments. In addition to the results obtained in Chapter 3, Table 5.1 presents a comparison between the proposed VAD and two other VAD algorithms namely: the G.729 from the ITU.T standard (ITU T, 1996) which combines different features to decide if speech is present, and Segbroeck et al's VAD in which a neural network is used to classify speech segments and noise segments (Segbroeck *et al.*, 2013). The comparison is done in terms of TPR, FPR, use of speech for training and the use of a priori information about the background noise. This comparison was performed using speech corrupted by babble noise with 0 dB SNR from the NOIZEUS database (Hu and Loizou, 2007a).

Table 5.1 Comparison between two voice activity detection algorithms in terms of true positives, false positives, use of speech for training and the use of a priori information about the background noise.

Algorithm	**TPR**	**FPR**	**Training**	**Priori knowledge of noise**
G.729 (ITU T, 1996)	94.1	82.5	No	No
Segbroeck (Segbroeck *et al.*, 2013)	73.3	70.5	Yes	Yes
proposed VAD	73.4	0	No	No

In addition to its robustness against the background noise, the low complexity of this VAD enabled its implementation in a low power digital signal processor (DSP) offering 50 million instructions per second, which led to the validation of the algorithm and the achievement of the first sub-objective in the thesis.

The second sub-objective consisted of developing a noise reduction algorithm to enhance the detected speech signal before its transmission to the protected ear. The developed algorithm is based on non-linear and multi-band Adaptive Gain Control (AGC) and requires neither SNR

nor noise parameters estimation, which eliminates the need for a speech/noise discrimination to perform the noise reduction. Objective evaluation showed that the proposed algorithm performs better than three benchmarks, and blind subjective tests showed that the proposed algorithm introduces less musical noise compared to the benchmark algorithms and was preferred 78.8 % of the time in terms of signal quality. In addition, other subjective tests showed that the proposed algorithm increases the intelligibility of speech compared to the noisy signals.

In the hardware implementation, the noise reduction algorithm uses sample based processing. Thus, the algorithms will have a processing delay much lower than 16 ms (around 6 ms due to the filter delay) which corresponds to the maximum delay that the algorithms should not exceed to not produce a perceptual delay between speech signals transmitted via the passive and active paths of the HPD.

5.2 Recommendations and future work

5.2.1 Recommendations

The current thesis enabled the development of low complexity speech detection and noise reduction algorithms for the S-HPD application, which shows that all the objectives of the thesis were successfully fulfilled.

Knowing that this project is a part of the bigger project nicknamed “the bionic ear”, and was funded by Sonomax which represents the future technology manufacturer, other steps have to be followed for the validation of the technology. These steps consist of:

5.2.1.1 Algorithms benchmarking

The current thesis presented a VAD and noise reduction algorithm. The performance of the VAD algorithm was compared to Sohn’s VAD and the performance of the noise reduction algorithm was compared to three benchmarks namely: the Wiener filter based on an *a priori*

SNR estimation (Scalart and Fiho, 1996), the spectral subtraction (Boll, 1979), and band-pass modulation filtering (Falk *et al.*, 2007).

As future work, the proposed algorithm need to be compared to other benchmarks such as (Wei *et al.*, 2010), (Kamath and Loizou, 2002), (Chen *et al.*, 2014) in terms of efficiency and computational cost.

5.2.1.2 VAD and noise reduction combination

While the first step in the S-HPD development consisted of developing a VAD algorithm, and the second step consisted of developing a noise reduction algorithm, another step has to be performed in order to combine the two algorithms for the S-HPD application.

5.2.1.3 VAD hardware parameters optimization

Chapter 3 presented the hardware implementation of the VAD algorithm where the parameters have been optimized manually. However, in the next step an automatic method for VAD parameters' optimization in the hardware needs to be developed, knowing that in the algorithm development step the parameters were optimized using the genetic algorithms in Matlab.

5.2.1.4 Adaptive Dynamic Range Compression: Parameters Optimization

Instead of using a level limiter, an adaptive dynamic range compressor (DRC) needs to be developed to amplify low level sounds and compress high level sounds. For the best efficiency of the S-HPD, the DRC has to be tested and its parameters optimized in the hardware.

5.2.1.5 Objective and Subjective evaluation of the S-HPD

After combining and implementing the VAD and noise reduction algorithms in the DSP, their real-time objective and subjective evaluation need to be conducted using real-world noise environments.

5.2.1.6 Smart hearing protection device for hearing impaired people

In North America, approximately 15 % of the population between the ages of 20 and 69 suffer from hearing loss due to noise exposure (NIDCD, 2015). To not worsen their hearing loss, keep them protected when exposed to high levels of noise, and enable face-to-face oral communication to them, the S-HPD has to be evaluated with hearing impaired persons.

APPENDIX I

A LOW-COMPLEXITY VOICE ACTIVITY DETECTOR FOR SMART HEARING PROTECTION OF HYPERACUSIC PERSONS

Narimene Lezzoum[1], Ghyslain Gagnon [1], Jérémie Voix [1]

[1] École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article presented at the « Interspeech» conference in 2013, Lyon, France.

A Low-Complexity Voice Activity Detector for Smart Hearing Protection of Hyperacusic Persons

Narimene Lezzoum, Ghyslain Gagnon and Jérémie Voix

École de technologie supérieure, 1100 Notre Dame West, Montréal (Qc) H3C 1K3 Canada

Abstract

In this paper, a Voice Activity Detector (VAD) is proposed for smart hearing protection applications where speech is to get through the hearing protector while ambient noise is to be blocked out. The VAD calculates a short-term statistical assessment of the temporal envelopes within different frequency bands. This assessment uses the Inter-Quartile Range (IQR) and reflects the dispersion of the envelopes' magnitudes. The VAD's decision is made using two threshold comparison rules and a hangover scheme triggered after a given number of observations. These four parameters have been optimized off-line using a genetic algorithm approach. The performance of the proposed VAD is compared to Sohn's VAD using a database of 90 speech signals corrupted by five real-world noise environments at Signal-to-Noise ratios (SNR) varying from 0 to +10 dB. Results show that the proposed VAD performs better than Sohn's VAD with an 85.9% (compared to 77.5%) F1 score averaged across all SNRs and also minimizes by a factor of three the mid-speech clipping rate. In addition, the evaluation of the proposed VAD's computational cost shows that its implementation on-board a low-power low-consumption DSP is very feasible and would enable smart hearing protection for hypersensitive persons.

Index Terms: Voice activity detection, inter-quartile range, genetic algorithms, temporal envelope

1. Introduction

Hyperacusis is defined as hypersensitivity and intolerance to ordinary environmental sounds [1]. It has been mentioned in [2] that one in 10 people report such sensitivity to sound. Over time, persons with hyperacusis begin to avoid social interaction, withdraw completely from environments that were once pleasant and become socially isolated [3]. The most common treatment for this hearing disorder is desensitization by careful presentation of sounds -limited in level and progressive in time-, as well as wearing passive hearing protection devices (HPDs) during daily activities to prevent the situation from worsening until the desensitization therapy has succeeded [1].

However, wearing passive HPDs is somewhat inconvenient for these patients because HPDs not only block unwanted noise signals, but also wanted speech signals. To palliate this problem, a $smart$ HPD i.e,. an active HPD that guarantees protection while discriminating between speech and noise to allow speech signals to get through to the protected ear is being worked on. For this purpose, the integration of a Digital Signal Processor (DSP) in the traditional passive HPD is required. The smartness of this HPD lies in its capability of transmitting speech signals while protecting the ear from environmental noise.

The discrimination between speech and noise signals is known in the literature as Voice Activity Detection (VAD). Numerous VAD algorithms have been developed; some require the extraction of features such as: the periodicity [4], zero crossing rate, full and low band energy and line spectrum frequencies [5] or pitch [6]. However, the performance of these VADs degrades when the SNR decreases [7]. To palliate this problem, other VADs have been developed and require the characterisation of noise depending on an estimate during noise periods such as the calculation of the a posteriori and a priori SNR [8]. Nevertheless, these VADs are sensitive to changes in the SNR [9]. Therefore, some researchers resort to learning techniques or modelling algorithms in their VAD [10], [11] and [12]. This however, leads to other problems when the intended application must operate in an embedded system with limited hardware resources.

In this paper, we propose the calculation of a short-term statistical assessment of the temporal envelope within different frequency bands. Extracting features from the temporal envelope has been widely used for hearing aids to detect the presence of speech and decide when gain should be reduced [13], [14], [15].

The VAD's decision is made after multiple observations using two thresholds in addition to a hangover scheme to take into consideration "long time" information, knowing that speech signals are highly time-correlated [16]. Thresholds, number of observations and hangover parameters are optimized off-line using a Genetic Algorithm (GA) [17]. The VAD's decision is set after multiple observations and using a hangover scheme to minimize false positives and mid-speech clipping knowing that for hyperacusis patients wearing smart hearing protection, perception of "short time" noise signals is unpleasant.

The paper is organised as follows. Section 2 introduces the proposed VAD algorithm. Section 3 describes the off-line parameters optimization. Section 4 presents the validation and discussions and section 5 the conclusions.

2. Proposed VAD Algorithm

Figure 1 illustrates the detailed architecture of the proposed VAD where N is the number of observations, i the frame number and m the frequency band number.

2.1. Windowing

The entire signal is first cut into frames with a Hamming window. The length of each frame is 25 ms with an 80% overlap.

2.2. Feature Extraction

2.2.1. Filter Bank

Each frame is passed into a filterbank of 16 frequency bands using -for ease on device implementation- a 4^{th} order Butterworth filter. Cut-off frequencies are described in the Bark scale [18] and lie between 20 and 3150 Hz.

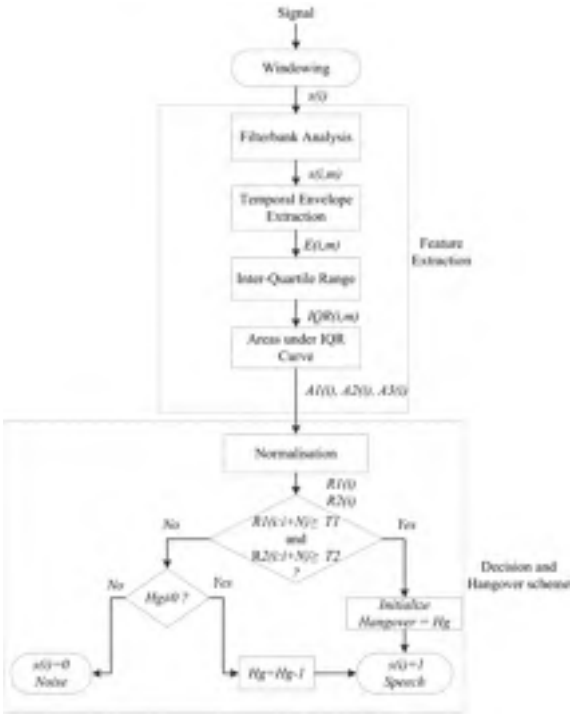

Figure 1: *Diagram block of the proposed VAD algorithm.*

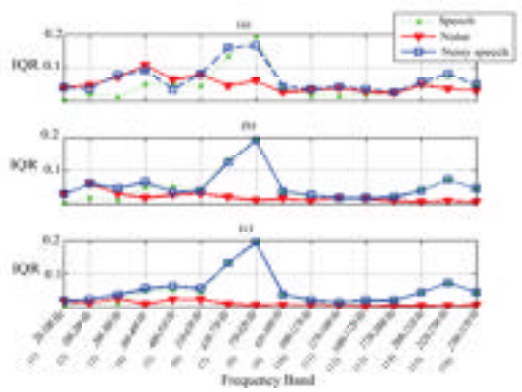

Figure 2: *IQR calculated in the frequency bands of one signal's frame with (a): -5 dB, (b) 0 dB and (c) 5 dB SNR.*

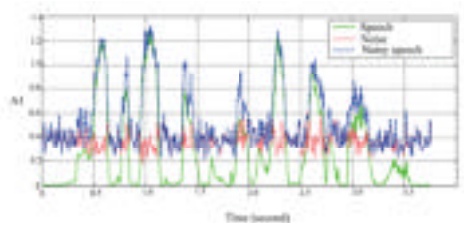

Figure 3: *The area $A1$ calculated for speech in a quiet setting, noisy setting with 0 dB SNR, and separate noise signal.*

2.2.2. *Temporal Envelope Extraction*

For each frame, the temporal envelope of each frequency band is extracted using the Hilbert Transform. Envelope extraction using the Hilbert transform involves the calculation of the analytic signal [19], as illustrated in Eq.1, where $E(t)$ is the Hilbert envelope of $x(t)$.

$$\mathrm{E(t)} = \sqrt{x(t)^2 + \tilde{x}(t)^2} \tag{1}$$

with $\tilde{x}(t)$ the Hilbert Transform of $x(t)$:

$$\tilde{x}(t) = x(t) * \frac{1}{\pi t} \tag{2}$$

2.2.3. *Statistical Assessment of Temporal Envelopes*

The statistical assessment is the Inter-Quartile Range (IQR) and is calculated within the temporal envelopes of the various frequency bands by using the 75^{th} percentile, or third quartile ($Q3$): the value below which 75% of the values in the distribution lie, and the 25^{th} percentile, or first quartile ($Q1$): the value above which 25% of the values lie. The IQR is calculated as shown in equation 3.

$$\mathrm{IQR} = Q3 - Q1 \tag{3}$$

Figure 2 illustrates an example of the IQR in all frequency bands for one signal's frame showing speech produced by a male speaker corrupted by noise with 5, 0 and -5 dB SNRs, speech and then noise. Figure 2 also shows that in the eighth frequency band (770-920 Hz) which represents the first formant of the speech segment (a voiced phoneme), the IQR of speech in a quiet setting is higher than that of the noise signal.

2.2.4. *Area under IQR Curve*

The ascertainment noted from Figure 2 is confirmed by the studies conducted in [20] on the influence of noise on vowels and consonants, which concluded that when the speech signal is corrupted by noise, the first formant can be reliably detected compared to the second formant, which is heavily masked by noise in low SNRs. Based on this conclusion, we choose not to consider only one frequency band to characterize a speech signal, but rather the area under the IQR curve from the third to the ninth frequency band. This area is named $A1$ and its choice is based on the frequency region containing the largest amount of speech information. The area under the IQR curve is calculated to take into consideration both spectral (first formant) and temporal (IQR) characteristics. Figure 3 illustrates $A1$ for a speech signal corrupted by a 'Car' noise in 0 dB SNR and $A1$ for noise and speech separately. Figure 3 illustrates that $A1$ is typically high when a speech signal is present, whereas when only noise is present this area is low. This ascertainment has been validated on different utterances.

2.3. VAD's Decision

2.3.1. *Normalization*

Figure 3 shows that the VADs' decision could be performed using a decision threshold upon $A1$. However, this procedure is not applicable directly on $A1$ since $A1$ depends on the IQR, which itself depends on the scale of the temporal envelope. This yields us to normalize the data by using two other areas under the IQR curve that reflect the noise signal.

Speech and noise signals differ in their frequency compo-

nents: noise signals have generally more energy in the lower frequencies than speech signals, which have a lower energy in these frequencies [21]. This ascertainment yields us to calculate the area $A2$ under the IQR curve from the first frequency band (20-100 Hz) to the second (100-200 Hz). The choice of this area is based on the frequency region containing the most noise information and the least speech information, it has been found empirically to be the most reliable for noise assessment. In addition to the $A2$ area, we added another area under the IQR curve ($A3$) that characterizes high frequency noises. This additional area is calculated in the high frequency bands and represents an alternative choice in the decision.

The three areas show in our testing the same trends: when the signal's level increases, $A1$, $A2$, $A3$ increase and similarly, when the signal's level decreases, $A1$, $A2$ and $A3$ decrease. This trend leads us to calculate the ratios $R1$ and $R2$ (see Eq. 4 and 5), upon $R1$ and $R2$ the first and second decision thresholds $T1$ and $T2$ are determined using the genetic algorithm approach.

$$\mathrm{R1} = \frac{A1}{A2} \tag{4}$$

$$\mathrm{R2} = \frac{A1}{A3} \tag{5}$$

The use of $T1$ and $T2$ as a decision rule eliminates the need for an adaptive decision threshold or an SNR estimator.

$T1$ and $T2$ must be optimized in addition to two other parameters: first, the number of observations that represents the number of consecutive frames having $R1$ and $R2$ higher than $T1$ and $T2$ respectively and after which the decision might be set to 1 (speech) and second, the hangover parameter, which represents the time after which the VAD is reset to 0.

3. Off-Line Parameters Optimization

3.1. Start of Speech Confirmation and Hangover Scheme in Smart Hearing Protection

The start of speech confirmation is defined as the number N of consecutive frames having $R1$ and $R2$ higher than $T1$ and $T2$ and after which the decision is set to 1. They have been used in Ramirez et al's VAD [22], where it was demonstrated that taking several frames into account in the VAD improves the reliability of the decisions.

The value N cannot exceed a certain number of consecutive frames, otherwise lip-sync errors may occur. Lip sync errors are defined by the ITU [23] as the errors between lip movement and the perceived speech signal, and a lip-sync error of 40 ms was considered acceptable. Thus, the maximum number of consecutive frames after which the decision might be set to one in the proposed VAD is eight consecutive frames, which represents a delay of 40 milliseconds.

The hangover scheme or end of speech confirmation has been widely used in VADs to minimize the false rejection rate caused by the non-detection of low energy speech frames containing consonants such as fricatives and unvoiced stops.

3.2. Objective Function

To optimize the thresholds ($T1$ and $T2$), hangover, and number of observations, an objective function should be minimized. This function's role is to evaluate the performance of the VAD algorithm. For this purpose, we used the F1 score measure [24]. This score combines the FPR (False Positive Rate), TPR (True Positive Rate) and FNR (False Negative Rate). Knowing that FPR, TPR, FNR are based on maximum of 100%.

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{6}$$

with

$$\mathrm{precision} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FPR}} \tag{7}$$

$$\mathrm{recall} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FNR}} \tag{8}$$

For the smart hearing protection application, we calculated the TPR and FNR for the noisy speech signals and the FPR for the noise signals. This evaluation method focuses on the fact that once the speech signal has been detected it must be transmitted in its entirety to the Smart HPD's wearer -possibly with a few seconds extra duration- while continuing to protect the wearer from noise when no speech signals are present.

The objective function to be minimized is shown in Equation 9:

$$\mathrm{Penalty} = 1 - F1 \tag{9}$$

3.3. Genetic Algorithm for Off-Line Parameter Optimization

Genetic Algorithms (GAs) [17] are randomized search and optimization techniques based on the mechanisms of natural selection and natural genetics. They are used to optimize the four parameters using an optimization database. For the purposes of this application, we have used a limited samples of five speech signals corrupted by 'Subway' noise, with a 0 dB SNR knowing that many hyperacusis patients are exposed daily to this type of noise. The speech signals are from the TIMIT database [25] and the noise signals from the AURORA database [26], both sampled to 16 kHz. The hangover's duration tends to vary between 50 and 250 frames which represents 0.25 to 1.25 seconds and the number of observations varies from 4 to 8 consecutive frames. The upper boundary of the hangover seems long when compared to the hangover durations used in telecommunications or speech recognition [27] and [16]. However, it was considered that this was not to be included in the objective function since it theoretically does not reduce the performance of the algorithm as it could reduce the performance of VADs for telecommunications or speech recognition. It will not affect the process but permit entire speech signal transmission without interruption.

After 10 generations, the GA reached an optimal solution with a best penalty value of 9%, which is equal to an F1 score of 91%. The optimization process gave a hangover of 250 frames and 6 consecutive frames.

4. Validation and Discussions

In the first part of this section, we present the VAD's performance assessment and in the second part, we quantify the computational cost of the proposed VAD.

4.1. Performance Assessment

The validation database is composed of 90 speech signals corrupted by five everyday noise environments with 3 SNRs (10, 5 and 0 dB). The speech signals are from the TIMIT database and the noise signals from the AURORA database. The average length of each speech signal corrupted by noise is 3.06 seconds and 83.6% of the signal comprises speech.

Noise environment		Sohn VAD		Proposed VAD		
Noise	SNR	F1	MSC	F1	MSC	F1*
Exhibition	0 dB	75.0	15.4	80.0	10.1	85.9
	5 dB	78.9	10.8	91.3	2.3	94.5
	10 dB	79.2	6.7	94.9	0.7	98.1
Babble	0 dB	73.2	9.9	78.6	1.6	76.3
	5 dB	74.9	7.3	82.8	0.9	82.7
	10 dB	76.1	5.6	82.1	0.4	81.0
Subway	0 dB	74.8	13.3	79.1	4.1	84.5
	5 dB	76.2	8.5	91.3	1.8	88.9
	10 dB	78.1	6.4	94.9	0.6	90.9
Airport	0 dB	76.2	9.4	77.5	4.5	77.7
	5 dB	77.7	6.9	86.2	1.5	87.1
	10 dB	79.2	5.2	87.5	0.2	85.5
Car	0 dB	79.7	15.6	77.0	13.2	79.8
	5 dB	81.5	10.8	91.5	2.6	96.0
	10 dB	83.0	7.4	95.1	0.5	98.7
Average	Average	77.5	9.2	85.9	3.0	87.3

Table 1: *Performance evaluation of the proposed VAD compared to Sohn's VAD using the F1 score and the MSC rates.*

As mentioned previously, the F1 score is used to evaluate the VAD's performance. Sohn's VAD [8] has been implemented from the VoiceBox [28]. The proposed VAD is compared to Sohn's VAD, which has proven its effectiveness with standard G729.B [5] AMR1, AMR2 [6] as demonstrated in [22] and [8].

In addition, we calculated the Mid-Speech Clipping rate (MSC) which represents the rate of speech frames classified as noise in the middle of the utterance. This measure is very important for speech intelligibility. The lower it is the more the speech segment is intelligible.

Table 1 illustrates the comparison of the two VADs.

As shown in Table 1, the F1 score of the proposed VAD is higher than the F1 score of Sohn's VAD in all noise environments and SNRs except for the 'Car' noise in 0 dB SNR which gives a F1 score of 77% instead of 79.5% for Sohn's VAD. The performance of the proposed VAD is more noticeable in the range of 5 and 10 dB SNR where the F1 score average in these SNRs has an increase of 11.2% for the proposed VAD.

Furthermore, we note from this table that the proposed VAD minimizes about three times the mid-speech clipping rate in comparison to Sohn's VAD. This leads us to say that the hangover scheme described in this paper is not only simpler but also more efficient than Sohn's hangover.

Moreover, we evaluated the proposed VAD using one speech signal of 150 seconds duration with 77.4% of speech (46 signals concatenated into one signal without additive noise periods between the 46 speech signals) corrupted by five noise environments at three SNR levels. This evaluation was conducted to validate the proposed algorithm with a signal of long duration to ensure that the performance of the proposed VAD is not only due to the hangover's duration. F1 scores are illustrated in the last part of Table 1(F1*). F1* shows almost the same F1 scores found earlier which enables us to validate the proposed VAD for its further implementation.

4.2. Computational Cost

The required hardware resources for the smart hearing protector are quite similar to those presently used in hearing aids and cochlear implants. The first two steps used in the feature extraction stage of the proposed VAD are already optimized to work in DSPs with limited hardware resources. For instance, DSPs for hearing aids are provided with an integrated filterbank coprocessor: the WOLA (Weighted Overlap Add) filterbank coprocessor [29], which allows the splitting of the signal in different frequency bands using an optimized architecture. For this purpose, we evaluated the additional computational cost arising from the IQR and areas calculation, to calculate by how much these two steps increase the number of instructions per second in the entire process.

Data must be sorted to calculate the IQR by using a sorting algorithm. Among the existing sorting algorithms, the Merge sort requires $Nlog_2N$ operations per frame [30]. Furthermore, to calculate $A1$, $A2$ and $A3$, 30 additions and 10 multiplications per frame are required. Table 2 shows the overall resource requirements for these two steps.

Processing step	Op. per frame	Op. per second
IQR	55,337	11,067,400
Areas	40	8,000
Global	55,377	11,075,400

Table 2: *Resource requirements for the 3rd and 4th steps in the feature extraction stage of the proposed VAD (abbreviation Op. defines the number of operations).*

The targeted DSP for smart hearing protection offers typically 60 MIPS (Million Instructions Per Second). Thus, the number of instructions per second required for the IQR and areas is 18.4% of the entire available number of instructions per second. This is reasonable since 81.6% of the entire computational cost could be dedicated to the filterbank, the Hilbert envelope extraction, and other operations such as noise reduction and dynamic range adaptation.

5. Conclusions

In this paper we proposed a new VAD particularly suited for smart hearing protection for hyperacusis patients. The proposed VAD uses a short term statistical assessment of the temporal envelope within different frequency bands. The VAD's decision is made after multiple observations using two decision thresholds and a hangover scheme, all optimized off-line using a genetic algorithm. Experiments conducted using speech signals corrupted by five real-world noise environments show that coupling the multiple observations and the hangover scheme in the decision process permits the maximization of the VAD's performance. Results show that the proposed VAD is more efficient than Sohn's VAD which by itself is more efficient than the Standards G.729b and AMR1, AMR2. This leads us to assume that the proposed VAD outperforms these standards as well. In addition to these satisfactory results, the proposed VAD requires neither assumption nor noise estimation depending on the first signal's frames, and is sufficiently simple to be implemented in a DSP of limited hardware resources. In future work, we intend to validate the proposed VAD with subjective tests, work on noise reduction to render the speech signals intelligible and adapt the dynamic range of the incoming speech signals to send them to the protected ear without damaging it.

6. Acknowledgements

The authors would like to thank Sonomax Technologies Inc. and its "Industrial Research Chair in In-ear Technologies" for its financial support.

7. References

[1] J. Vernon, "Pathophysiology of tinnitus: a special case hyperacusis and proposed treatment," *The American Journal of Otology*, vol. 8, pp. 201–202, 1987.

[2] G. Andersson, N. Lindvall, T. Hursti, and P. Carlbring, "Hypersensitivity to sound (hyperacusis): a prevalence study conducted via the Internet and post," *International Journal of Audiology*, vol. 41, pp. 545–554, 2002.

[3] M. Valente, J. Goebel, D. Duddy, B. Sinks, and J. Peterein, "Evaluation and Treatment of severe Hyperacusis." *Washington University School of Medecine in St. Louis. Paper 15*, vol. 11, no. 6, pp. 295–9, Jun. 2000.

[4] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I Communications, Speech and Vision*, vol. 139, no. 4, pp. 377–380, 1992.

[5] ITU T, "Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," *International Telecommunication Union*, 1996.

[6] ETSI, "Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description (GSM 06.94 version 7.1.0 Release 1998)," Tech. Rep., 1999.

[7] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, Mar. 2002.

[8] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1998–2000, 1999.

[9] M. H. Moattar and M. M. Homayounpour, "A Simple But Efficient Real-Time Voice Activity Detection Algorithm," in *17th European Signal Processing Conference*, 2009, pp. 2549–2553.

[10] E. Chuangsuwanich and J. Glass, "Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation frequency," *INTERSPEECH*, pp. 2645–2648, 2011.

[11] X. Liu, Y. Liang, Y. Lou, H. Li, and B. Shan, "Noise-Robust Voice Activity Detector Based on Hidden Semi-Markov Models," *IEEE, 20th International Conference on Pattern Recognition*, pp. 81–84, Aug. 2010.

[12] J. Wu and X. Zhang, "Efficient Multiple Kernel Support Vector Machine Based Voice Activity Detection," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 466–469, 2011.

[13] L. C. R. Bentler, "Digital noise reduction : An Overview," *Trends in Amplification*, vol. 10, no. 3, pp. 67–82, 2006.

[14] G. Mueller and T. Ricketts, "Digital noise reduction : Much ado about something?" *The Hearing Journal*, vol. 58, no. 1, pp. 10–17, 2005.

[15] K. Chung, J. Tufts, and L. Nelson, "Modulation-Based Digital Noise Reduction for Application to Hearing Protectors to Reduce Noise and Maintain Intelligibility," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 51, no. 1, pp. 78–89, May 2009.

[16] A. Davis, S. Nordholm, and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[17] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, 1989.

[18] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, p. 248, 1961.

[19] S. L. Marple, "Computing the Discrete-Time "Analytic" Signal via FFT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[20] G. Parikh and P. Loizou, "The influence of noise on vowel and consonant cues," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3874–3888, 2005.

[21] H. Levitt, "Noise reduction in hearing aids: a review." *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, 2001.

[22] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical Voice Activity Detection Using Multiple Observation Likelihood Ratio Test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[23] ITU, *International Telecommunication Union Document 11A/47-E*, 1993.

[24] van Rijsbergen, *Information Retrievial*, 2nd ed., Butterworths, Ed., 1979.

[25] S. V.Zue and J.Glass, "Speech Database Development: TIMIT and Beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[26] H.-g. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," *ISCA ITRW ASR2000 "Automatic Speech Recognition : Challenges for the Next Millennium"*, 2000.

[27] D. Vlaj, M. Kos, M. Grašič, and Z. Kačič, "Influence of Hangover and Hangbefore Criteria on Automatic Speech Recognition," in *16th International Conference on Systems, Signals and Image Processing, 2009. IWSSIP*, 2009.

[28] M. Brookes, "VoiceBox," 2004. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[29] ON Semiconductors, "Introduction to Audio Processing Using the WOLA Filterbank Coprocessor," pp. 1–10, 2009.

[30] D. E. Knuth, *The Art of Computer Programming, Sorting and Searching*, 1998.

APPENDIX II

A DEMONSTRATION OF A SINGLE CHANNEL BLIND NOISE REDUCTION ALGORITHM WITH LIVE RECORDINGS

Narimene Lezzoum[1], Ghyslain Gagnon [1], Jérémie Voix [1]

[1] École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article presented at the « ICASSP» conference in 2014 in the show and tell session, Florence, Italy.

A Demonstration of a Single Channel Blind Noise Reduction Algorithm with Live Recordings

Narimene Lezzoum, Ghyslain Gagnon, and Jérémie Voix
École de technologie supérieure
Université du Québec, Montréal (Qc) Canada
narimene.lezzoum@ens.etsmtl.ca, (ghyslain.gagnon)(jeremie.voix)@etsmtl.ca

Abstract

Currently, most noise reduction algorithms are based on an a priori information such as signal-to-noise ratio (SNR) or noise parameters estimation. They are mostly performed in the spectral domain to reduce the background noise at each frequency bin. However noise reduction in the spectral domain may introduce musical noise and artefacts which are in some cases perceptually more annoying than the background noise itself. In this "show and tell", we present a demonstration of a noise reduction algorithm based on dynamic range compression (DRC) using a time-varying and frequency-band dependant gain function deduced from the low-pass filtering of the temporal envelopes. The algorithm is considered as blind since it requires neither SNR nor noise parameters estimation. A graphical user interface (GUI) built under Matlab shows interactively the noise reduction in the temporal (waveform) and spectral (spectrogram) domains using live speech recordings mixed to pre-recorded noise signals.

1. Introduction

Noise reduction algorithms are nowadays used in multiple areas such as hearing aids, cochlear implants, telecommunication systems and human/robot interaction devices. Most of existing noise reduction algorithms perform in the spectral domain in order to reduce the background noise differently in each frequency bin, for instance, the spectral subtraction [1], the Wiener filter [2] [3], and the bandpass modulation filtering [4]. However, enhancing the speech in the spectral domain may introduce musical noise which is well known in the field of speech enhancement, and represents a random amplification of frequency bins [5].

Anderson [6] proposed a frequency-band dependant and time-varying gain function instead of frequency-varying gain function for a fast dynamic range compression (DRC), mentioning that the concept of frequency-band time-varying gain function can be used in some audio processing systems such as noise reduction. However, to our knowledge, methods and results of a such approach has never been demonstrated for noise reduction applications.

In this "show and tell", a noise reduction algorithm using frequency-band dependant and time-varying gain function is proposed. The proposed method employs dynamic range compression (DRC) theory in order to reduce the dynamic range differently in each frequency band using a time-varying gain function. This function is deduced from the temporal envelope and tend to preserve the natural quality of the incoming signal.

In this "show and tell", we demonstrate that the use of a time-varying and frequency-band dependant gain function enables noise reduction and speech quality improvement without introducing musical noise. In addition this demo shows that speech enhancement can be performed without any knowledge, assumption, or estimation of the noise parameters.

2. Scientific and Technical Description

Figure 1 illustrates the architecture of the proposed noise reduction algorithm.

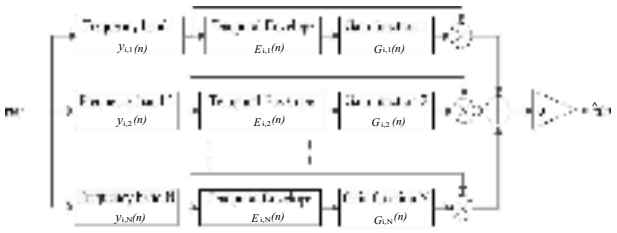

Figure 1. Block diagram of the proposed speech enhancement algorithm.

The proposed algorithm performs in real-time using

250 ms frames with 80% overlap.

The incoming signal is decomposed into N=16 frequency-bands using gammatone filterbank [7]. From each frequency-band the temporal envelope is extracted using the Hilbert Transform [8]:

$$E_{i,m}(n) = \sqrt{y_{i,m}(n)^2 + \tilde{y}_{i,m}(n)^2} \quad (1)$$

with i the frame number and m the frequency band number.

$$\tilde{y}(n) = y(n) * \frac{1}{\pi n} \quad (2)$$

with * denoting the convolution.

A gain function is deduced from the temporal envelope of each frequency-band (see section 2.1), and is thereafter multiplied by the incoming signal $y_{i,m}(n)$ of the same frequency-band. The enhanced signal $\hat{x}(n)$ of each frame is reconstructed by summing the 16 frequency-bands, and amplified by a constant α for the rescaling. Finally, the overlap-add method is used for the reconstruction of the global enhanced signal.

2.1. Time-Varying Gain Function Calculation

When combining the concepts of noise reduction and DRC used in hearing aids [9], [6], a multi-band time-varying noise reduction method can be obtained. According to preliminary results in our research, the multi-band time-varying gain function for noise reduction must meet three criteria:

- The gain function of each frequency-band should be smooth and continuous to avoid abrupt changes in the enhanced signal.
- The gain function must be chosen as a function of the temporal envelope $E_{i,m}(n)$ in order to preserve the quality of speech without adding artefacts.
- The gain function should be near to 1 in the frequency-bands containing speech and near to 0 in the frequency-bands containing noise, in order to preserve speech components and attenuate noise components.

A time-varying gain function that fulfils all the above cited criteria is a low-pass filtered temporal envelope, which represents a smoothed version of the temporal envelope $E_{i,m}(n)$:

$$G_{i,m}(n) = E_{i,m}(n) * L(t) \quad (3)$$

with $L(t)$ the impulse response of a lowpass filter with a 16 Hz cut-off frequency.

3 Objective Validation of the Proposed Method

The proposed method is evaluated using 30 noisy speech signals corrupted by "car" and "babble" noise in 5, 0, and -5 dB SNR from the Noizeus corpus [10]. The performance improvement of the proposed algorithm is compared to noisy signals in addition to a modulation filtering based speech enhancement algorithm [4] (benchmark algorithm) using the Perceptual evaluation speech quality (PESQ) metric [11] (see figure 2).

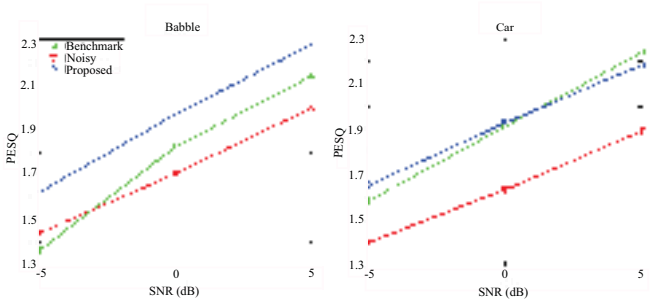

Figure 2. PESQ results for the unprocessed signals, the benchmark algorithm, and the proposed algorithm.

4. System Demonstration

A graphical user interface (GUI) is built in Matlab for an interactive demonstration of the proposed noise reduction algorithm. Figure 3 presents a screenshot of the user interface: part (1) shows the instructions that the user should follow when using pre-recorded speech signals, or recording a speech signal in live. Part (2) shows the experimental settings, while part (3) displays the enhancement in the spectrogram and waveform frame by frame: part (3-a) illustrates the enhanced part while par (3-b) illustrates the noisy part which will be enhanced. The part (4) of the interface presents the results for each noisy/enhanced signal in terms of PESQ.

A video of this demonstration is available in the webpage: http://critias.etsmtl.ca/ts2014. This demonstration runs on a laptop with professional headphones, and a microphone for live speech recordings.

5 Conclusions and Future Developments

In this "show and tell", we demonstrate that the use of a time-varying and frequency-band dependant gain function enables to reduce the background noise and improves the quality of the speech signal. In addition, we tend to show

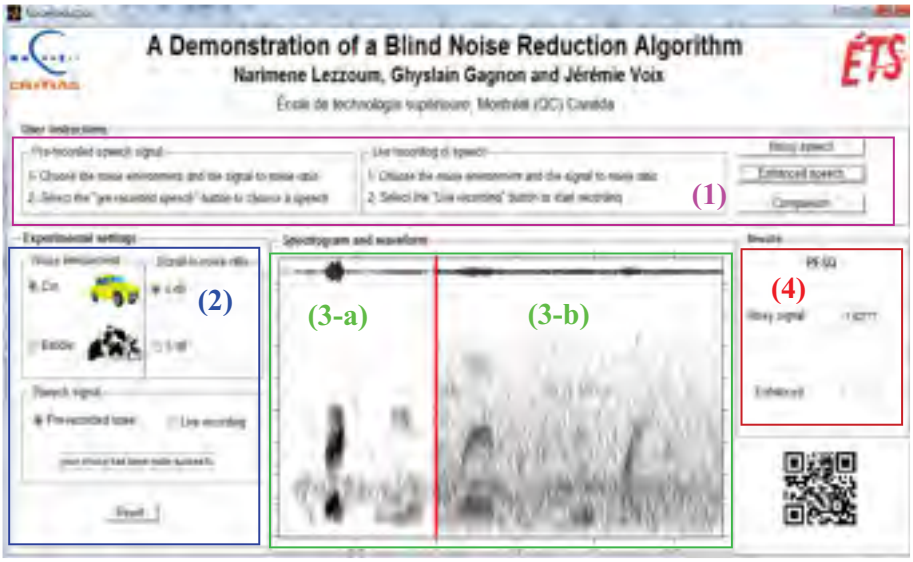

Figure 3. Screen-shot of the graphical user interface

that good noise reduction performance can be achieved without any knowledge, assumption, or estimate of the noise and speech parameters.

As future work, we tend to implement this algorithm in a digital signal processor (DSP) for a real-world embedded application.

Acknowledgement

The authors would like to thank Prof Tiago.H Falk for sharing the code of the benchmark algorithm. This work was supported by Sonomax Technologies Inc. and its "Industrial Research Chair in In-ear Technologies".

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transaction on Acoustics, Speech, and Signal processing*, no. 2, pp. 113–120, 1979.

[2] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications.*, cambrigde, ed., 1949.

[3] J. Scalart, P. and Filho, "Speech enhancement based on a priori signal to noise estimation." in *IEEE Int. Conf. Acoust. , Speech, Signal Processing (ICASSP)*, 1996, pp. 629–632.

[4] T. H. Falk, S. Stadler, W. B. Kleijn, and W.-y. Chan, "Noise suppression based on extending a speech-dominated modulation band," in *INTERSPEECH 2007*, pp. 2–5.

[5] C. Leitner and F. Pernkopf, "Musical noise suppression for speech enhancement using pre-image iteration," in *19th International Conference on Systems, Signals and Image Processing (IWSSIP),*, 2012, pp. 464 – 467.

[6] D. V. Anderson, "A modulation view of audio processing for reducing audible artifacts," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, no. 1, pp. 5474–5477, 2010.

[7] A. Aertsen and P. Johannesma, "The Specto-temporal receptive field: a functional characteristics of auditory neurons," *Biological Cybernetics*, vol. 143, pp. 133–143, 1981.

[8] S. L. Marple, "Computing the Discrete-Time "Analytic" Signal via FFT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[9] G. Kim and P. C. Loizou, "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms." *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1581–96, Sep. 2011.

[10] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[11] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm." Tech. Rep., 2003.

APPENDIX III

EVALUATION OF A DIGITAL EARPLUG FEATURING A MULTI-BAND ADAPTIVE GAIN CONTROL NOISE REDUCTION ALGORITHM FOR ENHANCED AUDIBILITY IN NOISY ENVIRONMENTS

Jérémie Voix[1], Narimene Lezzoum [1], Ghyslain Gagnon[1]

[1] École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article presented at the « ICSV» conference in 2015, Florence, Italy.

EVALUATION OF A DIGITAL EARPLUG FEATURING A MULTI-BAND ADAPTIVE GAIN CONTROL NOISE REDUCTION ALGORITHM FOR ENHANCED AUDIBILITY IN NOISY ENVIRONMENTS

Jérémie Voix, Narimene Lezzoum *and* Ghyslain Gagnon
ÉTS, Université du Québec, 1100 Notre-Dame Ouest, Montréal (QC), H3C 1K3, Canada
email: jeremie.voix@etsmtl.ca

An algorithm for single channel enhancement of noisy speech has been implemented for a digital hearing protection device. The developed algorithm operates without any knowledge or assumption of noise parameters and reduces the noise in the temporal domain using a non-linear and automatically adjustable gain function for multi-band dynamic range compression. The gain function is deduced from the temporal envelope of each frequency-band and compresses the frequency regions where speech is absent, to block ambient noise. Subjective evaluations have already shown that the algorithm improves speech quality. In this work, subjective tests using the Hearing-In-Noise-Test (HINT) approach and measuring the Speech Reception Threshold (SRT) now show that the speech intelligibility was preserved, if not improved, for most listeners.

1. Introduction

For practical and economical reasons, Hearing Protection Devices (HPD) are often used to protect workers from the risk of Noise-Induced Hearing Loss, the number one occupational injury in the workplace. HPDs reduce the sound energy reaching the wearer's eardrum and -in their passive linear form- cannot distinguish between noise and useful signals, such as speech and warning signals. This issue can now be addressed using electronic HPDs that use an external microphone, a digital signal processor (DSP) and an internal loudspeaker [3]. Such electronic could be integrated to an earmuff, or could be integrated to a custom earplug, such as the one illustrated in Figure 1. Figure 2 illustrates the electro-acoustical components and equivalent schematics of a digital version of a custom electronic earplug, featuring a DSP running the speech denoising algorithm. Recent work by the authors [5], presented a noise reduction method that calculates a time-varying and frequency band dependent gain function from the temporal envelopes of each frequency band for Adaptive Gain Compression (AGC) and applies it to the signal in each frequency band. This algorithm, illustrated in a block diagram in Figure 3, enables high compression of frequency bands containing noise and light compression of frequency bands containing speech and operates without any knowledge or estimation of the noise parameters, only assuming that the background noise is additive. The authors have already shown that the proposed use of a gain function that varies over time and frequency bands, does not introduce any of the usual artefacts associated with speech denoising, and effectively reduces the background

noise while improving the perceived quality of the speech signal. While promising, this subjective evaluation is only partial, as the intelligibility of the speech processed by this algorithm has not yet been assessed. Minimally, speech intelligibility should not be altered by the denoising algorithm, and ideally, the proposed processing algorithm should actually enhance the intelligibility of the speech. This paper presents an experimental study of the speech intelligibility achieved by subjects being exposed, in a laboratory environment, to noisy speech signals with and without the proposed processing algorithm, to assess its auditory benefits.

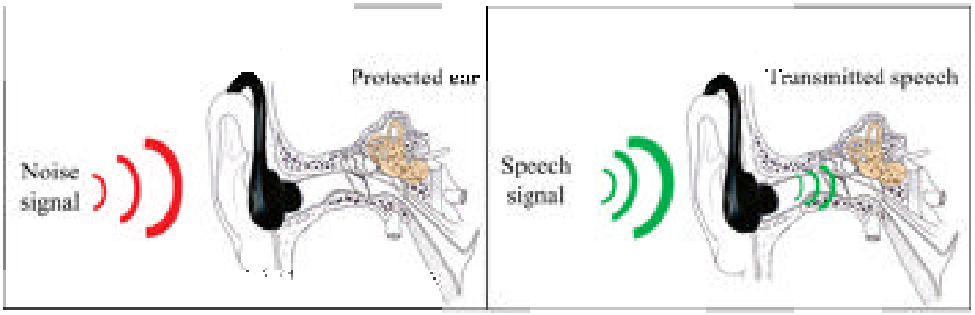

Figure 1: General principle of Digital Hearing Protection Device blocking industrial noises (left) and letting a speech signal through (right)

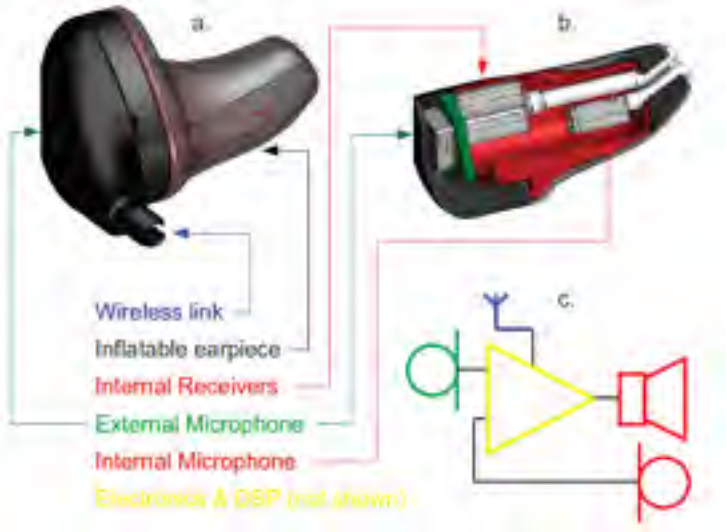

Figure 2: Overview of the digital custom earpiece (a), its electro-acoustical components (b), and equivalent schematics (c).

116

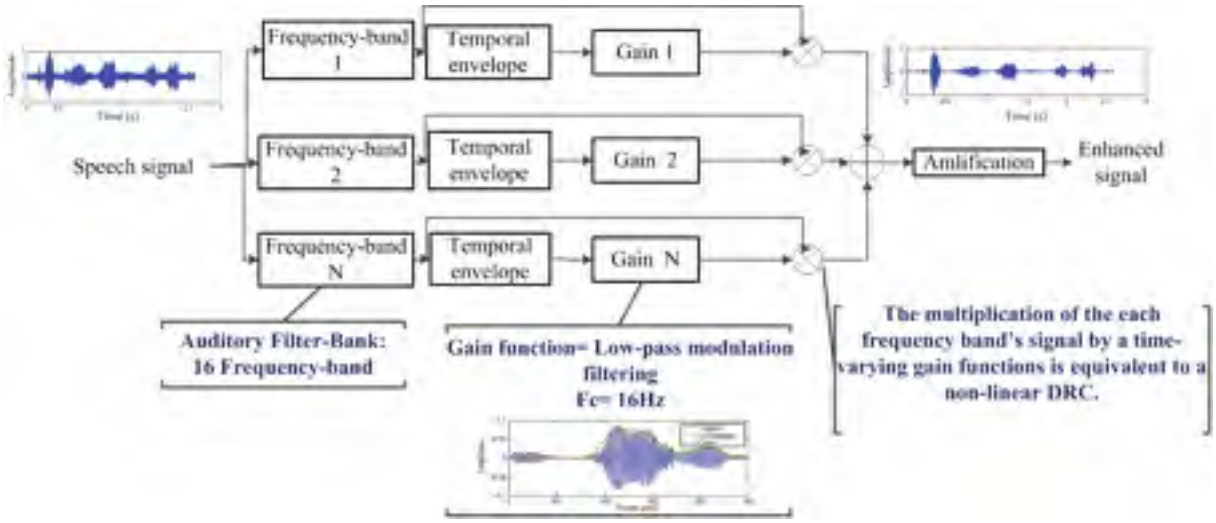

Figure 3: Block diagram of the proposed denoising algorithm being evaluated (from [REF])

2. Method

2.1 Speech Intelligibility Assessment

To assess the effect of the proposed denoising algorithm on speech intelligibility, a psycho-physical testing of a group of normal hearing users has been conducted using high-fidelity headphones under two testing conditions: without the proposed processing algorithm (baseline condition) and with the proposed algorithm (measured condition). For benchmarking and comparison purposes, the proposed processing algorithm (denoted "MBDNR") has been compared to processing algorithms widely used, with source code publicaly available in [2], such as the Multiband spectral subtraction from [4] (denoted "MBSS02") and the Wiener filter using decision-directed SNR estimation from [8] (denoted "Wfscarlat96"). In both test conditions, a clean speech signal is contaminated at a controlled signal-to-noise ratio (SNR) by a choice of two masking noises, taken from from the NOISEX database. The first masking noise is a recording from the inside of a car (denoted "Car"), while the second is babble noise (denoted "Babble"). The compound signal is then presented binaurally under earphones to the test subject, in order to determine the speech reception threshold (SRT) defined as the lowest level at which speech sentences can be correctly identified at least 50 percent of the time. The speech signals are French Canadian sentences available from the Hearing-In-Noise-Test (HINT) [7] and are presented using the Matlab Speech Testing Environment (MSTE) software [6]. As the MSTE software also features a hearing-device simulator, the 3 algorithms tested can be simply simulated as an input-output block, that alters the compounded audio signal being presented binauraly to the test subject.

In the current study, both the speech and masking noises are presented through the earphones and no attempt is made to auralize one signal versus the other, as one would typically do within the HINT paradigm where speech signal would be presented in front of the test-subject while surrounded by the masking noise. The MSTE software is designed to assess the SRT using various testing methods, either using a fixed speech level presented at typical levels of 45-55 dB HL and adapting the level of the masking noise ("adaptive masking level"), or using a fixed masking noise level and adapting the speech level ("adaptive speech level") or even using fixed speech and masking noise levels while adapting other metrics, such as the distortion threshold ("adaptive distortion threshold"). In the current study, the adaptive masking level paradigm has been used and the level of the masking noise is adjusted iteratively until the speech reception threshold is reached, level at which speech sentences

can be correctly identified at least 50 percent. The sentence corpus used was the one adapted for the French-Canadian language [9]. Accordingly, an SNR can be determined in each testing condition, i.e. without and with the processing algorithm. By comparing the SNR achieved by the same individual subject with and without the proposed processing algorithm, for the same SRT, on two different groups of speech sentences, it is possible to calculate the effect of the proposed processing algorithm. If the SNR of the baseline condition (without processing) is higher than the SNR of the measured condition (with processing algorithm), this clearly demonstrates that the proposed processing actually enhances the speech intelligibility, since the same SRT can be achieved while in the presence of more masking noise.

2.2 Experimental Procedures

1. **Test-subject Information** : Each test-subject is welcomed by the experimenter and signs the consent form that has been approved by the CER, the internal review board.
2. **Hearing Threshold Measurement**: Left and right auditory hearing thresholds are measured with a clinical audiometer (Interacoustics, AC40) under calibrated headphones (Telephonics, TDH 39). Pure-tone detection thresholds were assessed using an adaptive method at 250, 500, 1000, 2000, 3000, 4000 and 8000 Hz with supra-auricular earphones. All participants had detection thresholds below 25dB HL at every frequency, which corresponds to normal hearing, and they did not report any speech problems.
3. **Practice Run**: In order for the subject to familiarize him/herself with the testing procedure, the experimenter performed the Adaptive Testing Procedure, described further below, for the reference condition ("without processing");
4. **Test Run** The experimenter performed the Adaptive Testing Procedure, described further below, while ensuring that the order of the 4 test conditions (baseline as well as "MBDNR", "MBSS02"and "Wfscarlat96" processing algorithms) and list presentations were randomized or counter-balanced across the subjects under high-fidelity circumaural headphones.

The MSTE's adaptive tests are based on the one-up one-down procedure used to measure sentence SRTs with the Hearing-In-Noise-Test (HINT) [7]. Each pre-recorded sentence is played to the test-subject using headphones, together with the masking noise presented at the desired SNR, under headphones to the test-subject. At the same time, the words of the sentence are displayed to the experimenter, so that he/she can score the subject's oral response provided via a talk-back system from the sound booth to the experimenter desk. The testing procedure consists of two phases: First, a coarse estimate of the threshold is calculated using a large SNR adjustment step size. Second, a smaller SNR adjustment step size is used to get a more precise threshold. The number of sentences in a test list and the length of each phase can be defined for each test material. This procedure, which has been shown to converge to 50% intelligibility [1], [5], corresponds to a sentence SRT measurement, since all the words of a test sentence must be recognized for a correct response to be registered.

For illustration purposes, following the HINT testing paradigm [7], the adaptive test procedure on a list of 20 sentences is as follows:

- Sentence # 1 is played at the starting SNR and repeated until the subject repeats all words correctly, increasing the (signal or SNR) level by 4 dB for each incorrect response. Once a correct response is received, the SNR is decreased by 4 dB for the next sentence.
- Sentences # 2-4: For these sentences, the SNR is increased by 4 dB after each incorrect response, and decreased by the same step size after each correct response.
- Sentence # 5 is played at a SNR computed as the mean of the signal level of the first sentence, the levels of sentences # 2-4, and the level at which the fifth sentence would have been presented based on the subject's response to sentence # 4.
- Sentences # 6-20: For these sentences, the SNR is increased by 2 dB after each incorrect response, and decreased by the same step size after each correct response. At the end of the

118

list, the SRT for the given test condition is computed as the mean of the levels of sentences # 5-20 and the level at which the twenty first sentences would have been presented based on the subject's response to sentence # 20.

2.3 Results

The data collected during the adaptive test procedure consists in the individual SRT for each test condition (baseline as well as "MBDNR", "MBSS02"and "Wfscarlat96" processing algorithms) for the two masking noises ("car" and "babble") as well as a test condition without any background noise, so that the speech level could be calibrated for the subject. The nine resulting conditions are listed in Table 1, and these conditions are randomized across subjects, to limit the learning effect. The individual results are presented in Table 2, for the ten individuals across the nine conditions, and are expressed as individual SRTs as well as individual standard deviations resulting from the adaptive adjustment. These individual SRTs can be averaged over all subjects for each test condition, along with standard deviation, standard error, and 95% confidence interval. Note, that he masking noise tests signals were not normalized nor calibrated in magnitude, as we were only interested in a differential assessment with and without processing algorithm.

Table 1: The nine test conditions used

Condition	Paradigm	Masking Noise	Algorithm
01	Adapt. Mask. Lvl.	car	none
02	Adapt. Mask. Lvl.	car	MBSS02
03	Adapt. Mask. Lvl.	car	Wfscarlat96
04	Adapt. Mask. Lvl.	car	MBDNR
05	Adapt. Mask. Lvl.	babble	MBSS02
06	Adapt. Mask. Lvl.	babble	Wfscarlat96
07	Adapt. Mask. Lvl.	babble	MBDNR
08	Adapt. Mask. Lvl.	babble	none
09	Fixed	silence	none

Table 2: Individual SRTs and standard deviation (in parenthesis) for the 12 subjects in the 9 conditions, as well as group descriptive statistics

Subj.	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5	Cond. 6	Cond. 7	Cond. 8	Cond. 9
1	-23.3 (9.1)	-20.5 (6.6)	-21.9 (7.6)	-20.9 (6.7)	-2.7 (2.4)	-5.7 (1.7)	-6.2 (2.2)	-7.6 (2.6)	57.6 (4.5)
2	-22.6 (8.3)	-18.9 (6.4)	-21.2 (7.1)	-21.9 (7.6)	-3.3 (2.1)	-3.7 (2.2)	-5.9 (2.5)	-8.0 (3.3)	56.2 (5.1)
3	-21.6 (8.0)	-20.0 (6.7)	-21.6 (7.4)	-21.9 (7.6)	3.3 (1.9)	-3.0 (4.2)	-6.5 (2.9)	-7.7 (2.2)	58.1 (4.0)
4	-23.3 (9.1)	-21.2 (7.0)	-21.9 (7.8)	-22.6 (8.3)	-1.0 (2.0)	-7.4 (2.4)	-10.4 (3.2)	-7.3 (1.7)	58.9 (5.4)
5	-16.2 (4.0)	-12.0 (2.8)	-15.1 (3.6)	-17.3 (6.8)	7.5 (2.6)	-2.5 (1.6)	-2.5 (2.2)	5.2 (4.5)	63.5 (2.2)
6	-21.6 (7.5)	-16.8 (7.2)	-19.1 (8.9)	-22.4 (8.1)	-2.8 (2.1)	-4.8 (1.8)	-6.4 (1.8)	-5.3 (1.9)	57.8 (4.1)
7	-22.1 (7.8)	-17.0 (6.2)	-21.4 (7.1)	-21.4 (7.1)	-2.0 (1.9)	-4.1 (3.0)	-5.2 (2.7)	-5.7 (1.7)	59.2 (3.9)
8	-23.3 (9.1)	-19.5 (5.5)	-21.6 (7.4)	-23.1 (8.8)	-2.3 (1.7)	-3.7 (1.9)	-7.0 (2.6)	-6.7 (2.1)	57.6 (4.6)
9	-19.5 (6.1)	-17.2 (6.2)	-17.9 (7.6)	-20.2 (7.3)	1.7 (2.0)	-4.0 (2.5)	-4.4 (2.9)	-1.5 (3.1)	61.1 (2.5)
10	-22.4 (8.1)	-20.5 (6.3)	-20.9 (7.1)	-22.1 (7.9)	-2.1 (2.1)	-5.3 (2.9)	-8.3 (1.8)	-6.0 (2.7)	60.6 (3.9)
11	-21.9 (8.4)	-20.2 (6.5)	-21.9 (7.6)	-22.4 (8.1)	-3.7 (2.2)	-6.0 (2.4)	-5.5 (3.3)	-11.1 (2.7)	57.4 (4.3)
12	-20.9 (7.4)	-20.7 (6.5)	-21.4 (7.2)	-22.6 (8.4)	-4.1 (2.2)	-6.0 (1.9)	-7.0 (2.4)	-8.6 (2.4)	59.7 (4.4)
AVG	-21.6	-18.7	-20.5	-21.6	-1.0	-4.7	-6.3	-5.9	59.0
STD	2.0	2.6	2.1	1.6	3.5	1.4	2.0	4.2	2.0
STE	1.0	1.3	1.1	0.8	1.7	0.7	1.0	2.1	1.0

2.4 Discussion

The overall sound quality of the proposed and the two benchmarking algorithms were already evaluated in a separate study using 20 speech signals corrupted by the "car" and "babble" noise with 0 and -5 dB SNR, with 10 other participants (3 females and 7 males). Results collected showed that

participants preferred the signals processed with the proposed MBDNR algorithm in terms of overall quality in all noise conditions. While this preliminary result clearly indicates that the proposed algorithm improves the quality of the noisy signals, little could be said of its effect on the speech intelligibility. In the present study, 12 normal hearing subjects (2 females and 10 males) aged from 22 to 47 (mean 30.3 years) were tested for change in SRTs with and without the processing algorithm, so that its effect on speech intelligibility could be assessed. From bottom line of Table 2, it can be seen that in "car" noise the average speech reception threshold, on line "AVG" is similar for the proposed MBDNR algorithm (case # 4) and without algorithm (case # 1), while the two other denoising algorithm appears to have a lightly detrimental effect on the SRT, with respective increases of 2.9 dB and 1.1 dB for the "MBSS02" (case # 2) and "Wfscarlat96" (case # 3) algorithm. In "babble" noise the average speech reception threshold is marginally better, by 0.4 dB, for the proposed MBDNR algorithm (case # 7) than with the baseline test without algorithm (case # 8), while the two other denoising algorithms appears to have a slightly detrimental effect on the SRT, with respective increases of 4.9 dB and 1.2 dB for the "MBSS02" (case # 5) and "Wfscarlat96" (case # 6) algorithm respectively. The bottom lines of Table 2 present the group standard deviation (denoted "STD"), as well as the standard error of the mean (denoted "STE") for further statistical testing. The SRTs obtained in case # 9, in the absence of masking noise indicate that all subjects had a similar ability to understand speech, as assumed from the selection tests outcomes. The standard deviations of the tracking process used in the adaptative test procedure are indicated between parenthesis for each subject in each condition and indicate the ability of one given subject to efficiently converge to the individual SRT reported.

3. Conclusions

An algorithm for single channel enhancement of noisy speech has been implemented in a digital hearing protection device. Subjective evaluation conducted previously have shown that speech quality has been improved, as it uses a non-linear and automatically adjustable gain function for multi-band dynamic range compression that compresses the frequency regions where speech is absent to block ambient noise. The psycho-physical tests conducted in this work, on normal hearing subjects using the Hearing-In-Noise-Test (HINT) approach and measuring the Speech Reception Threshold (SRT), showed that the speech intelligibility is typically unaffected by the proposed algorithm or can even be slightly enhanced in the presence of babble noise. As the proposed MBDNR algorithm has a very limited computational requirements, it is now foreseeable that it could soon be implemented in digital hearing protection devices to protect against the risk of noise induced hearing loss while enabling speech to be perceived.

Acknowledgements

The authors are thankful to Dr. Nicolas Ellaham for sharing his MSTE software code for the current study. They would also wish to acknowledge the financial support of the Sonomax-ÉTS Industrial Research Chair in In-Ear Technologies. Finally, the help from Ms. Christine Turgeon and Ms. Shima Zokaei, trained audiologists, was also greatly appreciated for the experimental collection of the clinical data.

REFERENCES

1. Donald D. Dirks, Donald E. Morgan, and Judy R. Dubno. A Procedure for Quantifying the Effects of Noise on Speech Recognition. *Journal of Speech and Hearing Disorders*, 47(2):114, May 1982. 00126 bibtex: Dirks1982.

2. Esfandiar Zavarehei. Sample Speech Enhancement methods, 2006. 00000 bibtex: EsfandiarZavarehei2006.

120

3. Jérémie Voix. Did you say "bionic" ear? In *Acoustics Week in Canada Proceedings*, volume Vol. 42, No. 3, pages 68–69, Winipeg (MB), 2014. Canadian Acoustics. 00000 bibtex: JeremieVoix2014.

4. Sunil Kamath and Philipos Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *IEEE international conference on acoustics speech and signal processing*, volume 4, pages 4164–4164. Citeseer, 2002. 00312 bibtex: Kamath2002.

5. Narimene Lezzoum, Ghyslain Gagnon, and Jérémie Voix. A Demonstration of a Single Channel Blind Noise Reduction Algorithm with Live Recordings. In *Show Tell Event*, Florence, Italy, May 2014. 00000 bibtex: NarimeneLezzoum2014.

6. Nicolas Ellaham, Christian Giguère, and JR Dubno. A new research environment for speech testing using hearing-device processing algorithms. In *Canadian Acoustics*, volume 42(3), pages 92–93, 2014. 00000 bibtex: NicolasEllaham2014.

7. Michael Nilsson, Sigfrid D. Soli, and Jean A. Sullivan. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2):1085–1099, February 1994. bibtex: Nilsson1994.

8. P. Scalart and J.V. Filho. Speech enhancement based on a priori signal to noise estimation. In *, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings*, volume 2, pages 629–632 vol. 2, May 1996. 00345 bibtex: Scalart1996.

9. Véronique Vaillancourt, Chantal Laroche, Chantal Mayer, Cynthia Basque, Madeleine Nali, Alice Eriks-Brophy, Sigfrid D. Soli, and Christian Giguère. Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *International Journal of Audiology*, 44(6):358–361, January 2005. bibtex: Vaillancourt2005.

BIBLIOGRAPHY

Abel, S.M. 2008. "Barriers to Hearing Conservation Programs in Combat Arms Occupations". *Journal of Aviation, space, and environmental medecine*, vol. 79, n° 6, p. 591–598.

Aertsen, A.M.H.J and P.I.M Johannesma. 1981. "The Specto-temporal receptive field: a functional characteristics of auditory neurons". *Biological Cybernetics*, vol. 143, p. 133–143.

Allen, C. H and E. H Berger. 1990. "Development of a unique passive hearing protector with level- dependent and flat attenuation characteristics". *Journal of Noise Control Engineering*, vol. 34, n° 3, p. 97–105.

Atal, B.S. 1974. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". *The Journal of the Acoustical Society of AmericaJournal of the Acoustical Society of America*, vol. 55, n° 6, p. 1304–1312.

Baer, T and BCJ Moore. 1994. "Effect of spectral smearing on the intelligibility of sentences in the presence of interfering speech". *Journal of Acoustical Society of America*, vol. 94, n° 3, p. 2277–2280.

Bahoura, M. and J.R Rouat. 2001. "Wavelet speech enhancement based on the Teager energy operator". *IEEE Signal Processing Letters*, vol. 8, n° 1, p. 10–12.

Bao, J, M Hungerford, R Luxmore, D Ding, Z Qiu, D Lei, A Yang, R Liang, and KK Ohlemiller. 2013. "Prophylactic and therapeutic functions of drug combinations against noise-induced hearing loss". *Hearing research*, vol. 304, p. 33–40.

Bentler, R. and LK. Chiou. 2006. "Digital Noise Reduction : an overview". *Trends in Amplification*, vol. 10, n° 3, p. 67–82.

Bentler, R, H Wu, J Kettel, and R Hurtig. 2008. "Digital noise reduction: outcomes from laboratory and field studies". *International Journal of Audiology*, vol. 47, n° 8, p. 447–460.

Berger, E. H. 2010. "Hearing through the protectors: how hearing protection changes speech understanding and what to do about it".

Berger, E. H and P Hamery. 2008. "Empirical evaluation using impulse noise of the level-dependency of various passive earplug designs". In *Acoustics*.

Berger, E.H. 2000. Noise Control and Hearing Conservation: Why Do It? *The Noise Manual*.

Berger, E.H and J. Voix. 2015. Chapter 10 : Hearing Protection Devices. *The Noise Manual*, volume 1.

Berger, E.H, R.W Kieper, and D. Gauger. 2003a. "Hearing protection: Surpassing the limits to attenuation imposed by the bone-conduction pathways". *The Journal of the Acoustical Society of America*, vol. 114, n° 4, p. 1955–1967.

Berger, E.H., L.H. Royster, J.D. Royster, D.P Driscoll, and M Layne, 2003b. *The Noise Manual*. ed. Fifth.

Beritelli, F., S. Casale, G. Ruggeri, and S. Serrano. 2002. "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors". *IEEE Signal Processing Letters*, vol. 9, n° 3, p. 85–88.

Blamey, P. J. 2005. "Adaptive Dynamic Range Optimization (ADRO): a digital amplification strategy for hearing aids and cochlear implants". *Trends in Amplification*, vol. 9, n° 2, p. 77–98.

Boll, S.F. 1979. "Suppression of acoustic noise in speech using spectral subtraction". *IEEE Transactions on Acoustics, Speech, and Signal processing*, vol. 27, n° 2, p. 113–120.

Brammer, A.J, G. Yu, and E.R Bernstein. 2015. "Communicating in noise when wearing hearing protection: issues and approaches.". In *International Congress on Sound and Vibration*. p. 12–16.

Brimhall, Owen D, Craig M. Collotzi, and Gregory N. Koshowich. 2002. "Protective hearing devices with multi-band automatic amplitude control and active noise attenuation".

Burrell, C. and S. Abel. 2009. *Enhancing communication in noisy environments*. Technical report.

C. J. van Rijsbergen, 1979. *Information Retrievial*. ed. 2nd.

Cappé, O. 1994. "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor". *IEEE Transactions on Speech and Audio Processing*, vol. 2, n° 2, p. 345–349.

Carbonneau, M., N. Lezzoum, J. Voix, and G. Gagnon. 2013. "Detection of alarms and warning signals on a digital in-ear device". *International Journal of Industrial Ergonomics*, vol. 43, n° 6, p. 503–511.

Casali, J.G. 2010. "Powered electronic augmentations in hearing protection technology Circa 2010 including active noise reduction , electronically-modulated sound transmission, and tactical communications devices: review of design, testing, and research.". *International Journal of Acoustics and Vibration*, vol. 15, n° 4, p. 168–186.

CER. 2014. *Comité d'ethique et de la recherche, Ecole de technologie superieure N H20121004*. Technical report.

Chen, F. and P.C. Loizou. 2010. "Speech enhancement using a frequency-specific composite Wiener function". In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. p. 4726–4729. Ieee.

Chen, S., R.C. Guido, T. Truong, and Y. Chang. 2010. "Improved voice activity detection algorithm using wavelet and support vector machine". *Computer Speech & Language*, vol. 24, n° 3, p. 531–543.

Chen, Y., C. Wei, Y. Fanchiang, Y. Meng, Y. Huang, and S. Jou. 2014. "Neuromorphic Pitch Based Noise Reduction for". *IEEE Transactions on Circuits and Systems*, vol. 61, n° 2, p. 463–475.

Cho, N. and E. Kim. 2011. "Enhanced voice activity detection using acoustic event detection and classification". *IEEE Transactions on Consumer Electronics*, vol. 57, n° 1, p. 196–202.

Choi, S and Z Jiang. 2008. "Comparison of envelope extraction algorithms for cardiac sound signal segmentation". *Expert Systems with Applications*, vol. 34, n° 2, p. 1056–1069.

Chuangsuwanich, E. and J. Glass. 2011. "Robust voice activity detector for real world applications using harmonicity and modulation frequency". In *INTERSPEECH*. p. 2645–2648.

Chung, K. 2004. "Challenges and recent developments in hearing aids: Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms". *Trends in Amplification*, vol. 8, n° 3, p. 83–124.

Chung, K. 2007. "Effective compression and noise reduction configurations for hearing protectors". *The Journal of the Acoustical Society of America*, vol. 121, n° 2, p. 1090–1101.

Chung, K., J. Tufts, and L. Nelson. 2009. "Modulation-based digital noise reduction for application to hearing protectors to reduce noise and maintain intelligibility". *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 51, n° 1, p. 78–89.

Cooper, J.C. 2003. *A Short tutorial on lip sync errors, the sources and solutions*. Technical Report 408. 1–6 p.

Cornelis, B., M. Moonen, and J. Wouters. 2011. "Binaural voice activity detector for MWF based noise reduction in binaural hearing aids.". In *European Signal Processing Conference*. p. 486–490.

Davis, A., S. Nordholm, and R. Togneri. 2006. "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n° 2, p. 412–424.

Davis, S.B. and P. Mermelstein. 1980. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n° 4, p. 357–366.

Desai, V.R. 2014. "Smart tool for headphones".

Drullman, R, J M Festen, and R Plomp. 1994a. "Effect of temporal envelope smearing on speech reception.". *The Journal of the Acoustical Society of America*, vol. 95, n° 2, p. 1053–64.

Drullman, R, J M Festen, and R Plomp. 1994b. "Effect of reducing slow temporal modulations on speech reception.". *The Journal of the Acoustical Society of America*, vol. 95, n° 5, p. 2670–80.

E. Zwicker. 1961. "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)". *The Journal of the Acoustical Society of America*, vol. 33, n° 2, p. 248.

Ellaham, 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.

Esch, T. and P. Vary. 2009. "Efficient musical noise suppresion for speech enhancement systems". In *International conference on Acoustics, Speech and Signal Processing*. p. 4409–4412.

ETSI. 1999. *Digital cellular telecommunications system (Phase 2+); voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels*. Technical report. 1–30 p.

Evangelopoulos, G. and P. Maragos. 2006. "Multiband modulation energy tracking for noisy speech detection". *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14, n° 6, p. 2024–2038.

Falk, T.H, S. Stadler, W B. Kleijn, and W. Chan. 2007. "Noise suppression based on extending a speech-dominated modulation band". In *Interspeech*. p. 2–5.

Freeman, D.K, G Cosier, C.B Southcott, and I Boyd. 1989. "The voice activity detector for the pan-European digital cellular mobile telephone service". In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. p. 369–372.

FreeSound. 2014. "https://www.freesound.org/".

Freyman, R.L, R.K Clifton, and R.Y Litovsky. 1991. "Dynamic processes in the precedence effect.". *The Journal of the Acoustical Society of America*, vol. 90, n° 2 Pt 1, p. 874–84.

Glasberg, B. and B. Moore. 1990. "Derivation of auditory filter shapes from notched-noise data". *Hearing research*, vol. 47, n° 1, p. 103–138.

Goldberg, D.E., 1989. *Genetic algorithms in search, optimization, and machine learning*.

Grancharov, V and W B Kleijn. 2008. Speech quality assessment. *Springer Handbook of Speech Processing*, p. 83–99.

Haas, H. 1972. "The influence of a single echo on the audibility of speech". *Journal of the Audio Engineering Society*, vol. 20, n° 2, p. 146–159.

Hasan, T. and M.K. Hasan. 2009. "Suppression of residual noise from speech signals using empirical mode decomposition". *IEEE Signal Processing Letters*, vol. 16, n° 1, p. 2–5.

Hermansky, H. and N. Morgan. 1994. "RASTA processing of speech". *IEEE Transactions on Speech and Audio Processing*, vol. 2, n° 4, p. 578–589.

Hirsch, H. and D. Pearce. 2000. "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions". In *ISCA ITRW ASR2000 "Automatic Speech Recognition : Challenges for the Next Millennium"*.

Hong, O., D. Samo, R. Hulea, and B. Eakin. 2008. "Perception and attitudes of firefighters on noise exposure and hearing loss.". *Journal of occupational and environmental hygiene*, vol. 5, n° 3, p. 210–215.

Hotvet, D.A. 1996. "Directional ear device with adaptive bandwidth and gain control".

Hsu, C.C, T. Lin, J.H Chen, and T.S Chi. 2013. "Voice activity detection based on frequency modulation of harmonics". In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. p. 6679–6683.

Hu, Y. and P. Loizou. jul 2007a. "Subjective comparison and evaluation of speech enhancement algorithms.". *Speech communication*, vol. 49, n° 7, p. 588–601.

Hu, Y and P Loizou. 2007b. "A comparative intelligibility study of signle microphone noise reduction algorithms". *Journal of Acoustical Society of America*, vol. 22, n° 3, p. 1777–1786.

Hu, Yi and P.C Loizou. 2008. "Evaluation of objective quality measures for speech enhancement". *IEEE Transactions on audio, speech, and language processing*, vol. 16, n° 1, p. 229–238.

Inoue, T., H. Saruwatari, K. Shikano, and K. Kondo. 2011. "Theoretical analysis of musical noise in Wiener filtering family via higher-order statistics". In *International Conference on AcousSpeech and Signal Processing (ICASSP)*. p. 5076–5079.

ITU T. 1996. *Annex B: a silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70*. Technical report.

ITU-T. 2003a. *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm.* Technical report.

ITU-T. 2003b. *Transmission systems and media, digital systems and networks: one-way transmission time*. Technical report.

ITU-T P.862. 2001. *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. Technical report.

ITU-T P.862. 2011. *Methods for Objective and Subjectçve Assessment of Speech Quality: Perceptual Objective Listening Quality Assessment*. Technical report.

Jabloun, F. and A.E..ngin Çetin. 1999. "Teager energy based feature parameters for speech recognition in car noise". *IEEE Signal Processing Letters*, vol. 6, n° 10, p. 259–261.

Janssen, J., D.D. Vleeschauwer, R. Windey, G.H Petit, and J.M Leroy. 2002. "Delay bounds for voice over IP calls transported over satellite access networks". *Journal on Mobile Networks and Applications (MONET), special issue on Satellite-Based Information Services*, vol. 7, n° 1, p. 79–89.

Kamath, D.S and P.C Loizou. 2002. "A Multi-band spectral subtraction method for enhancing speech corrupted by colored noise". *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, , p. 2–5.

Kondo, K. 2012. Speech quality. *Subjective quality measurement of speech: its evaluation, estimation and applications*.

Kuk, F., C. Ludvigsen, and C. Paludan-müller. 2002. "Improving hearing aid performance in noise : challenges and strategies". *The hearing Journal*, vol. 55, p. 34–46.

Kuo, S.M., B.H. Lee, and W. Tian, 2014. *Real-time digital signal processing: fundamentals, implementations and applications*. ed. Third.

Kvaloy, O., O. Kristen, and O. Pettersen. 2007. "A combined hearing protection device and communication terminal". In *First European Forum on Efficient Solutions for Managing Occupational Noise Risks*. p. 1361–1365.

Lamm, J.G, A.K Berg, C.M Künzler, B. Kuenzle, and C.G Glück. 2011. "Procedure for the steady-state verification of modulation-based noise reduction systems in hearing instruments". *EURASIP Journal on Advances in Signal Processing*, p. 1–20.

Lamothe, J., C. Gascon, M. Lariviere, and C. Laroche. 2002. "Standarisation of the Hearing In Noise Test (HINT) for bilingual francophone population and anglophone population". *Revue d'orthophonie et d'audiologie2*, vol. 26, n° 2, p. pp. 81–89.

Lee, H. and D. Yook. 2009. "Space-time voice activity detection". *IEEE Transactions on Consumer Electronics*, vol. 55, n° 3, p. 1471–1476.

Lee, H., S. Chang, D. Yook, and Y. Kim. 2009. "A voice trigger system using keyword and speaker recognition for mobile devices". *IEEE Transactions on Consumer Electronics*, vol. 55, n° 4, p. 2377–2384.

Lei, J., J. Yang, J. Wang, and Z. Yang. 2009. "A robust voice activity detection algorithm in nonstationary noise". In *International Conference on Industrial and Information Systems,*. p. 195–198.

Leitner, C. and F. Pernkopf. 2012. "Musical noise suppression for speech enhancement using pre-image iteration". In *International Conference on Systems, Signals and Image Processing (IWSSIP)*. p. 464 – 467.

Levitt, H. 2001. "Noise reduction in hearing aids: a review.". *Journal of Rehabilitation Research and Development*, vol. 38, n° 1, p. 111–121.

Lezzoum, N., G. Gagnon, and J. Voix. 2013. "A low-complexity voice activity detector for smart hearing protection of hyperacusic persons". In *INTERSPEECH*. p. 723–727.

Lezzoum, N, G Gagnon, and J Voix. 2014a. "Voice activity detection system for smart earphones". *IEEE Transactions on Consumer Electronics*, vol. 60, n° 4, p. 737–744.

Lezzoum, N., G. Gagnon, and J. Voix. 2014b. "A Demonstration of a Single Channel Blind Noise Reduction Algorithm with Live Recordings". *International Conference on Acoustics Speech and Signal Processing, Show and Tell*.

Lezzoum, N., G. Gagnon, and J. Voix. 2015. "Audio samples: http://critias.etsmtl.ca/S-HPD".

Lim, J. 1978. "Evaluation of a correlation subtraction method for enhancing speech degraded by additive noise". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-37, n° 6, p. 471–472.

Litovsky, R.Y, H.s Colburn, W. Yost, and S.J Guzman. 1999. "The precedence effect.". *Journal of Acoustical Society of America*, vol. 106, n° 4 Pt 1, p. 1633–54.

Liu, X., Y. Liang, Y. Lou, H. Li, and B. Shan. 2010. "Noise-robust voice activity detector based on hidden semi-markov models". In *International Conference on Pattern Recognition*. p. 81–84. Ieee.

Loizou, P.C., 2007. *Speech enhancement: theory and practice*.

Ma, J., Y. Hu, and P.C Loizou. 2009. "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions.". *Journal of the Acoustical Society of America*, vol. 125, n° 5, p. 3387–405.

Macdonald, E.N, A. Behar, W. Wong, and H. Kunov. 2008. "Noise exposure of opera musicians". *Canadian Acoustics*, vol. 36, n° 4, p. 11–16.

Mazur, K and J Voix. 2013. "Implementing 24-hour in-ear dosimetry with recovery". In *international Conference on Acoustics*.

Metz, C.E. 1978. "Basic principles of ROC analysis". *Seminars in Nuclear Medicine*, vol. 8, n° 4, p. 283–298.

Ming, J, R. Srinivasan, and D. Crookes. 2011. "A corpus-based approach to speech enhancement from nonstationary noise". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n° 4, p. 822–836.

Moattar, M.H and M.M Homayounpour. 2009. "A simple but efficient real-time voice activity detection algorithm". In *17th European Signal Processing Conference*. p. 2549–2553.

Mueller, G. and T. Ricketts. 2005. "Digital noise reduction : much ado about something?". *The Hearing Journal*, vol. 58, n° 1, p. 10–17.

Mueller, H.G., J. Weber, and B.W.Y Hornsby. 2006. "The effects of digital noise reduction on the acceptance of background noise.". *Trends in amplification*, vol. 10, n° 2, p. 83–93.

Nadon, V., A. Bockstael, D. Botteldooren, J.M 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.

NIDCD. 2015. "Noise Induced Hearing Loss". <http://www.nidcd.nih.gov/health/hearing/pages/noise.aspx>.

Nilson, M., Soli S.D., and J.A. Sullivan. 1994. "Development of the hearing in noise test for measurement of speech reception thresholds in quiet and in noise". *Journal of Acoustical Society of America*, vol. 2, n° 95, p. 1085–1099.

NIOSH. 1998. "Occupational noise exposure :U.S department of health and human services".

ON Semiconductors. 2009. "Introduction to audio processing using the WOLA filterbank coprocessor".

OSHA. 1981. "Occupational Noise Exposure; hearing conservation amendment".

OSHA. 1983. "Occupational Noise Exposure; hearing conservation amendment; final rule".

O'Shaughnessy, D., 2000. *Speech communication: human and machine*.

O'Shaughnessy, D. 2008. Formant estimation and tracking. *Springer Handbook of Speech Processing*, p. 213–227.

Paliwal, K., K. Wójcicki, and B. Schwerin. 2010. "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain". *Speech Communication*, vol. 52, n° 5, p. 450–475.

Parikh, D., S. Ravindran, and D. Anderson. 2009. "Gain adaptation based on signal to noise ratio for noise suppression.". In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. p. 185–188.

Parikh, G. and P.C. Loizou. 2005. "The influence of noise on vowel and consonant cues". *Journal of the Acoustical Society of America*, vol. 118, n° 6, p. 3874–3888.

Pickett, J.M. 1958. "Limits of Direct Speech Communication in Noise". *Journal of the Acoustical Society of America*, vol. 30, n° 4, p. 278–281.

PureData. 2014. "http://puredata.info/".

Quené, H. 2007. "On the just noticeable difference for tempo in speech". *Journal of Phonetics*, vol. 35, n° 3, p. 353–362.

Ramírez, J., J.C Segura, C. Benítez, L. García, and A. Rubio. 2005. "Statistical voice activity detection using multiple observation likelihood ratio test". *IEEE Signal Processing Letters*, vol. 12, n° 10, p. 689–692.

Saberi, K. and J.V. Antonio. 2003. "Precedence-effect thresholds for a population of untrained listeners as a function of stimulus intensity and interclick interval". *Journal of the Acoustical Society of America*, vol. 114, n° 1, p. 420.

Sandlin, R.E., 2002. *Text book of hearing aid amplification*.

Saon, G., S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury. 2013. "The IBM speech activity detection system for the DARPA RATS program". In *INTERSPEECH*. p. 3497–3501.

Scalart, P. and J. Fiho. 1996. "Speech enhancement based on a priori signal to noise estimation.". In *IEEE Int. Conf. Acoust. , Speech, Signal Processing (ICASSP)*. p. 629–632.

Segbroeck, M.V, A. Tsiartas, and S.S. Narayanan. 2013. "A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice". In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. p. 704–708.

Semiconductor, ON. 2012. "BELASIGNA 250 Programmable Audio Processing System".

Shahid, M., R. Ishaq, B. Sällberg, N. Grbic, B. Lövström, and C. Ingvar. 2011. "Modulation domain adaptive gain equalizer for speech Enhancement". In *International Conference on Signal and Image Processing and Applications*.

Shen, J., J. Hung, and L. Lee. 1998. "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments". In *5th International conference ICSLP*. p. 4.

Sohn, J., N.S. Kim, and W. Sung. 1999. "A statistical model-based voice activity detection". *IEEE Signal Processing Letters*, vol. 6, n° 1, p. 1998–2000.

Stedmon, A. W. 1997. "Stressors, speech production and automatic speech recognition (ASR)". In *International Conference on Cognitive systems*.

Stegmann, J. and G. Schroder. 1997. "Robust Voice-Activity Detection Based on the Wavelet Transform". In *IEEE Workshop on Speech Coding For Telecommunications Proceeding,*. p. 99–100.

Tucker, R. 1992. "Voice activity detection using a periodicity measure". In *IEE Proceedings Communications, Speech and Vision*. p. 377–380.

Voix, J. 2014. "Did you say "Bionic Ear"?". *Canadian Acoustics*, vol. 42, n° 3, p. 68–69.

Voix, J. and F. Laville. 2005. "Problematiques associées au développement d'un bouchon d'oreille "intelligent"". *Perspectives interdisciplinaires sur le travail (PISTES)*, vol. 7, n° 2.

Voix, J. and F. Laville. 2009. "Prediction of the attenuation of a filtered custom earplug". *Applied Acoustics*, vol. 70, n° 7, p. 935–944.

Voix, J., J.N Laperle, J. Mazur, and A. Bernier. 2014. "Advanced communication earpiece device and method".

Wei, C.W., C.C. Tsai, T.S. Chang, and S.J. Jou. 2010. "Perceptual multiband spectral subtraction for noise reduction in hearing aids". In *IEEE Asia-Pacific Conference on Circuits and Systems, Proceedings, APCCAS*. p. 692–695.

Westerlund, N., M. Dahl, and I. Claesson. 2004. "Speech enhancement using an adaptive gain equalizer with frequency dependent parameter settings". *IEEE Vehicular Technology Conference*.

Wu, J. and X. Zhang. 2011. "Efficient multiple kernel support vector machine based voice activity detection". *IEEE Signal Processing Letters*, vol. 18, n° 8, p. 466–469.

Yang, X. and D.W Grantham. 1997. "Echo suppression and discrimination suppression aspects of the precedence effect.". *Perception & psychophysics*, vol. 59, n° 7, p. 1108–17.

Younkin, A.C. and P.J. Corriveau. 2008. "Determining the amount of audio-video synchronization errors perceptible to the average end-user". *IEEE Transactions on Broadcasting*, vol. 54, n° 3, p. 623–627.

Zue, V., S. Sennef, and J. Glass. 1990. "Speech database development: TIMIT and beyond". *Speech Communication*, vol. 9, n° 4, p. 351–356.