Model-based speech enhancement for hearing aids

Kavalekalam, Mathew Shaji

DOI (link to publication from Publisher):
10.5278/vbn.phd.tech.00052

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Kavalekalam, M. S. (2018). *Model-based speech enhancement for hearing aids*. Aalborg Universitetsforlag.
https://doi.org/10.5278/vbn.phd.tech.00052

MODEL-BASED SPEECH ENHANCEMENT FOR HEARING AIDS

BY
MATHEW SHAJI KAVALEKALAM

DISSERTATION SUBMITTED 2018

AALBORG UNIVERSITY
DENMARK

Model-based speech enhancement for hearing aids

Ph.D. Dissertation

Mathew Shaji Kavalekalam

Dissertation submitted December, 2018

Dissertation submitted:	December, 2018
PhD supervisor:	Prof. Mads Græsbøll Christensen Aalborg University
Assistant PhD supervisors:	Jesper Boldt, Research Manager GN Hearing
	Assistant Prof. Jesper Kjær Nielsen Aalborg University
PhD committee:	Associate Professor Thomas Arildsen (Chairman) Aalborg University
	Professor Rainer Martin Ruhr-Universität Bochum
	Senior Research Scientist Emmanuel Vincent Inria Nancy – Grand Est
PhD Series:	Technical Faculty of IT and Design, Aalborg University
Department:	Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-367-9

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2018

Curriculum Vitae

Mathew Shaji Kavalekalam

Mathew was born on October 19, 1989 in Thrissur, India. He received the B.Tech degree in Electronics and Communications Engineering from Amrita University, India and the M. Sc. degree in Communications Engineering from RWTH Aachen University, Aachen, Germany, in 2011 and 2014, respectively. He is currently a PhD student at the Audio Analysis Lab, Department of Architecture, Design and Media Technology, Aalborg University, Aalborg, Denmark. His research interests include speech enhancement for hearing aid applications.

Abstract

According to the World Health Organisation (WHO), around 5% of the world's population suffers from hearing loss. The understanding of speech for a hearing impaired person is severely degraded especially in the presence of interfering speakers such as the cocktail party scenario. A very commonly used approach to overcome hearing loss is to use hearing aid (HA) which performs tasks such as dynamic range compression, feedback cancellation, noise reduction etc. However, the performance of current HA technology in improving the speech understanding in the so called cocktail party scenario has been observed to be limited.

In this thesis, we focus on speech enhancement algorithms capable of improving the speech intelligibility of HA user in such scenarios. In this thesis, we have proposed speech enhancement algorithms that take into account the speech production model. The critical component in this speech enhancement framework is the estimation of the model parameters. In this thesis, this is done using a supervised approach which takes into account a priori information regarding the speech and noise autoregressive (AR) coefficients. The performance of the speech enhancement system is heavily dependent on the estimation of these parameters. Currently, most of the HA users are equipped with HAs at both the ears that can communicate with each other. We have proposed methods to estimate the filter parameters jointly using the information at the two ears and also show the benefit of estimating the filter parameters jointly in comparison to doing it independently for each hear. The proposed algorithms were evaluated using both subjective listening tests and objective measures. Apart from speech enhancement, we have also looked into the problem of noise power spectral density (PSD) estimation which is a critical component in many of the speech enhancement systems. The proposed method was shown to perform better than many of state of the art noise PSD estimators in terms of estimation accuracy and enhancement performance. As the application of these algorithms are only required in the HAs when the speech intelligibility degrades beyond a certain value, it is desirable to measure the intelligibility of the received signal at the HA. To perform this, we have also proposed a non-intrusive method for intelligibility prediction which was shown to have high correlation with the subjective scores.

Lastly, we also investigated on the usage of an external device, e.g. a microphone array along with a HA with the possibility of communicating with each other. We proposed a model based approach to control the beam pattern of the external microphone array based on the look direction of the HA user. The usage of an external device in addition to the HA was found to improve the intelligibility of the received signal.

Resumé

Ifølge Verdenssundhedsorganisationen (WHO) lider omkring 5 % af verdens befolkning af en eller anden form for høretab. Taleforståeligheden for en person med sådan et høretab svækkes i særdeles i situationer, hvor mange personer snakker på én gang som f.eks. til et cocktail-party. Den typiske måde at kompensere for et høretab er ved a give den hørehæmmede et høreapparat, der udfører opgaver som dynamisk kompression og støjundertrykkelse samt modvirker tilbagekoblingsproblemer. Den nuværende teknologi i høreapparater er dog ikke god nok til at kunne forbedre taleforståeligheden i udfordrende situationer som et cocktail-party.

I denne afhandling fokuseres der på taleforbedringsalgoritmer, der er i stand til at forbedre taleforståeligheden for den hørehæmmede i udfordrende situationer som et cocktail-party. I den forbindelse har vi foreslået algoritmer, der tager en model for den menneskelige taleproduktion. Det vigtigste element i udarbejdelsen af disse algoritmer har været estimeringen af modelparametre. Som model for talen har vi brugt autoregressive modeller, hvis parametre er estimeret v.h.a. af en overvåget tilgang baseret på trænede modeller af tale og støj. Kvaliteten af en taleforbedringsalgoritmer afhænger i høj grad af, hvor præcist disse parametre kan estimeres. De fleste høreapparatsbrugere anvender et høreapparat i begge ører, og disse høreapparater kan ofte trådløst kommunikere med hinanden. Vi har udnyttet dette til at foreslå en algoritme, der anvender samtidig information fra begge høreapparater, og vi viser i den forbindelse også hvorfor dette er bedre en at lave dataforarbejdningen uafhængigt af hinanden i de to høreapparater. De foreslåede algoritmer er evalueret med både lyttetest og objektive mål. Udover taleforbedring har vi også undersøgt problemet med at estimere støjspektret, eftersom løsningen til dette problem er meget vigtigt for kvaliteten af en taleforbedringsalgoritme. Den forslåede metode til at estimere støjspektret er demonstreret til at være bedre end mange af de bedste og kendte metoder, når der måles på nøjagtigheden af det estimerede støjspektrum og kvaliteten af taleforbedringsalgoritmer. En forbedring af taleforståeligheden er kun nødvendig i situationer, hvor taleforståeligheden er forringet væsentligt. Derfor er det vigtigt at kunne måle taleforståeligheden med høreapparatet direkte fra et støjfyldt talesignal. I den forbindelse har vi foreslået en metode til netop at gøre dette, og vi har demonstreret, at metodens estimat i høj grad stemmer overens med lytteforsøg.

Et sidste bidrag har været at undersøge om taleforståeligheden for en høreapparatsbruger kunne forbedres, hvis høreapparaterne er forbundet trådløst til et eksternt mikrofonarray. I den forbindelse har vi foreslået en model-baseret metode, der kan styre det eksterne arrays fokussretning v.h.a. at høreapparatbrugerens synsretning. Vi har også vist, at dette kan forbedre taleforståeligheden i sammenligning med kun at bruge et høreapparat.

List of publications

The main body of this thesis consists of the following publications:

[A] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt and M. G. Christensen, "Model-based Speech Enhancement for Intelligibility Improvement in Binaural Hearing aids". *IEEE Trans. on Audio, Speech and Language Processing*, vol. 27, pp. 99-113, 2019.

[B] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen and J. B. Boldt, "Model based Estimation of STP parameters for Binaural Speech Enhancement". *Proc. European Signal Processing Conference*, pp. 2479-2483, 2018.

[C] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen and J. B. Boldt, "Hearing aid-controlled Beamformer for Binaural Speech Enhancement using a Model-based approach". submitted to *IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2019.

[D] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen and J. B. Boldt, "Model-based Noise PSD estimation from Speech in Non-stationary Noise". *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 5424-5428, 2018.

[E] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen and J. B. Boldt, "A study of Noise PSD Estimators for Single channel Speech Enhancement". *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 5464-5468, 2018.

[F] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen and J. B. Boldt, "Online Parametric NMF for Speech Enhancement". *Proc. European Signal Processing Conference*, pp. 2320-2324, 2018.

[G] C. Sorensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt and M. G. Christensen, "Non-intrusive Codebook-based Intelligibility Prediction". *Speech Communication*, vol. 101, pp. 85-93, 2018.

The following additional papers have been published by the author

[1] M. S. Kavalekalam, M. G. Christensen, F. Gran and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook based approach". *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 191-195, 2016.

[2] M. S. Kavalekalam, M. G. Christensen and J. B. Boldt, "Binaural speech enhancement using a codebook based approach". *Proc. International Workshop on Acoustic Signal Enhancement*, 2016.

[3] C. Sorensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt and M. G. Christensen, "Non-intrusive Intelligibility Prediction using a Codebook-based approach". *Proc. European Signal Processing Conference*, pp. 216-220, 2017.

[4] M. S. Kavalekalam, M. G. Christensen and J. B. Boldt, "Model based binaural enhancement of voiced and unvoiced speech". *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 666-670, 2017.

The following patent applications have been filed in relation with the project

[1] M. S. Kavalekalam, M. G. Christensen, F. Gran and J. B. Boldt, "Kalman filtering based speech enhancement using a codebook based approach". *US Patent Application*, US20170265010A1, 2016.

[2] C. Sorensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt and M. G. Christensen, "Hearing Device and Method with Non-Intrusive Speech Intelligibility". *European Patent Application*, 17181107.8 , 2017.

Contents

II Papers 43

Preface

This thesis is submitted to the Technical Faculty of IT and Design at Aalborg University in partial fulfilment of the requirements for the degree of doctor of philosophy. The thesis consists of two parts: the first part is an introduction to the field of speech enhancement and summarises the contributions of the Ph.D. project. The introduction is followed by the scientific papers that have been published through different outlets.

The work was carried out in the period of December 2014 to October 2018 in the Audio Analysis Lab at Department of Architecture, Design and Media Technology at Aalborg University. The project was made possible by a research grant from the Innovation fund Denmark. First of all, I would like to express my sincere gratitude to Mads Græsbøll Christensen for his guidance and motivating me throughout my Ph.D. work. I would like to thank him for his ideas, and giving me the freedom to explore the vast field of speech enhancement. I am also grateful to my co-supervisor Jesper Boldt for his supervision especially in the initial stages of the project and commenting on my papers in spite of his busy schedule. I would also like to thank him for giving me an opportunity to visit GN Hearing. Finally, I would like to thank my second co-supervisor Jesper K Nielsen for his valuable time and guidance during the thesis. I would like to thank him for the long and beneficial discussions which contributed greatly to my project.

I thank my colleagues at the Audio Analysis Lab and friends for making life at Aalborg University a pleasant one and for all the social gatherings we had during the past four years. I would like to thank my parents and my brother for supporting me and standing with me through my ups and downs. Finally, I would like to thank my wife, Maya for her endless love and bearing with my busy schedule especially during the end stages of my Ph.D. work.

Mathew Shaji Kavalekalam
Aalborg University, December 13, 2018

Part I

Introduction

Introduction

1 Speech Communication

Humans have various ways of communicating with each other that is not possessed by animals [1], and one of the dominant forms of communications used by humans is through speech. In speech communication, information is being conveyed between a set of speakers and listeners. Speech produced by humans generally consists of a structured combination of words taken from a lexicon that is uttered by the speaker [2]. To achieve this, our brain first produces a sequence of motor commands that results in the movement of the muscles in the vocal system in a particular manner to produce the desired sound wave. This sound wave travels through a particular channel before it is received by the listener's auditory system (AS). This channel often degrades the speech signal before it is received by the listener's AS. Once it is received by the AS, it is then converted into a set of neurological pulses which is further processed by the brain to interpret what was uttered by the speaker [3].

In an ideal case, there will not be any degradation to the speech that is produced before it reaches the listener's AS. However, in realistic scenarios, there is often some sort of degradation such as additive noise, reverberation etc. that reduces the quality and intelligibility [4] of the received signal. The objective of this thesis is to design algorithms that can alleviate some of these degradations.

The remainder of this chapter will be organised as follows. We will describe the mechanism of the speech production process along with certain important characteristics of these signals in Section 1.1. Following that, we will explain how this signal is perceived by a human ear and also about hearing loss. The degradations experienced by the speech signal which results in significant intelligibility reduction for the hearing impaired (HI) will be explained in Section 1.3. Chapters 2 and 3, respectively, will explain some of the signal models and state of the art methods for performing speech enhancement to mitigate these degradations. An overview of the contributions of this thesis is given in Chapter 4 followed by the conclusion and directions for future research in Chapter 5.

1.1 Speech Production

The human speech production system helps in the conversion of thoughts into speech. A cross section of the human speech production system is shown in Fig. 1. The main components of this system include 1) trachea or the windpipe 2) the vocal chords, 3) the mouth or the vocal tract, and 4) the nasal cavity [3]. A very convenient way to interpret the mechanism of the speech production system is as an acoustic filtering operation that acts on an input signal [3]. The acoustic filter here comprises of a combination of the vocal and nasal tracts whereas the input signal, also termed as the excitation signal is the signal coming out of the vocal chords.

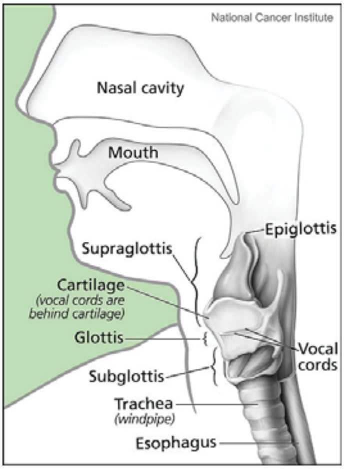


Fig. 1: Human speech production system (source: National Cancer Institute).

The properties of this system can be changed by the excitation signal as well as the shape of the vocal and nasal tracts. Based on the characterestics of the excitation signal, the speech produced by the humans can be broadly categorised into 2 categories namely 1) voiced speech that includes vowels and 2) unvoiced speech which are weaker in amplitude and includes the consonants. In the case of the voiced speech, the air forced out of the lungs travels through periodically vibrating vocal folds to form the excitation signal that is periodic in nature [5], whereas the excitation signal in the case of unvoiced speech is more noise like. The vocal tract in the case of unvoiced speech is constricted in certain areas whereas there is no such constriction for voiced speech [6]. The shape of the vocal tract is associated with a set of resonant frequencies, also referred to as formants which are typically 3 to 5 in humans [3].

As the spectral characterestics of a speech signal vary over time, they are generally analysed using a spectrogram which is a two dimensional representation of the speech signal. A spectrogram plots the power spectra [7] over time and the power of different frequencies is color coded [8]. Fig. 2 shows the spectrogram of a speech signal. The unvoiced speech is characterised by rectangular patches over a wide range of frequencies whereas the voiced speech is characterised by horizontal stripes with dominant energies at the lower frequencies. Moreover, the formants which are represented by peaks in the frequency domain can be seen, e.g, at around 0.4, 1.9 and 2.7 kilohertz at 1.5 seconds.

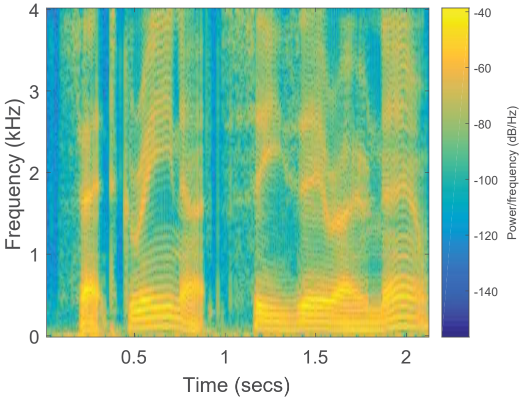


Fig. 2: Spectrogram of the clean signal.

1.2 Speech Perception

The acoustic signal emitted by the speaker is perceived by the listener using the human ear. Fig. 3 shows the structure of the human ear. The acoustic signal arriving at the outer ear travels through the external auditory canal towards the eardrum (tympanic membrane) and vibrates it. As the eardrum vibrates, the ossicular chain consisting of malleus, incus and stapes transfers these vibrations into the inner ear which contains the cochlea [9]. The cochlea consists of the basilar membrane that are planted with hair cells which converts the mechanical vibrations into electric signals that are then sent to the brain. These hair cells have varying degrees of sensitivity to different frequencies, thus acting like spatially distributed band pass filters. The audible range of frequencies in humans is in between 16 Hz and 20 KHz [10].

The amplitude of sound pressure is generally measured in decibels (dB). This is due the large range of amplitude that can be perceived by a healthy human AS (0- 140 dB). These properties vary in the case of a HI person based

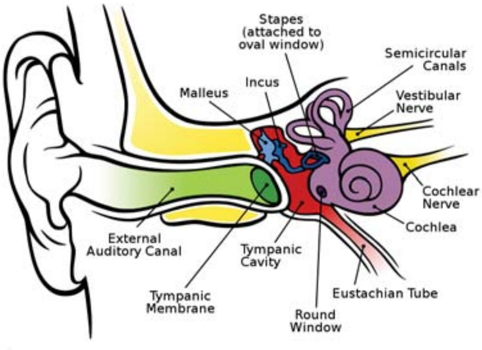


Fig. 3: Structure of the human ear [11].

on the degree of hearing loss. Hearing impairement can occur due to faults in different parts of the ear. Damages to outer or middle ear is termed as the conducting hearing loss whereas damages to the inner ear is termed as the sensorineural hearing loss [12]. Conducting hearing losses, in most cases can be rectified by surgeries but sensorineural hearing losses which are caused due to damage in hair cells are more permanent in nature.

Having two ears enables humans to identify the direction of arrival of the sound source with the help of binaural cues such as the inter-aural time difference (ITD) and the inter-aural level difference (ILD) [10]. As these cues aid the listener to locate the speaker of interest, it is desirable of any processing algorithms, e.g. present in a HA to preserve these cues.

1.3 Speech Degradation

Before the acoustic signal is received by the listener's AS, the speech signal produced by a human often undergoes degradation. Some of the common degradations experienced are reverberation and additive noise. We will not consider here the degradations caused due to clipping, coding etc. For a HI person, such degradations can reduce the speech intelligibility and lead to an increased listening effort. In such scenarios, it is desirable to remove these degradations to some extent before it is perceived by the HI person. We will now explain these degradations in more detail:

Additive noise

The most common type of degradation occurring to a speech signal is due to the presence of additive background noise. The background noise can be due to the presence of interfering speakers as in a cocktail party scenario, traffic, noise from machines etc. Out of these, the most challenging situation for a person with hearing loss is to concentrate on a single speaker in the presence of other competing speakers. This is termed as the cocktail party problem [13]. The AS of a normal hearing person is capable of focusing on one person in such a scenario. However, people with hearing loss lack the ability to do so in an effective manner. This results in a HI person being isolated in such a situation. Fig. 4 shows an example of a cocktail party scenario where the listener (indicated by the blue digits) is having a conversation the speaker (indicated by red(1)) in the presence of multiple interferers (indicated by red (2-9)). The spectrogram of the degraded signal in such a scenario is shown in Fig. 5. It can be seen that the noise by the interferers masks certain portions of the target speech that may be important for speech understanding.

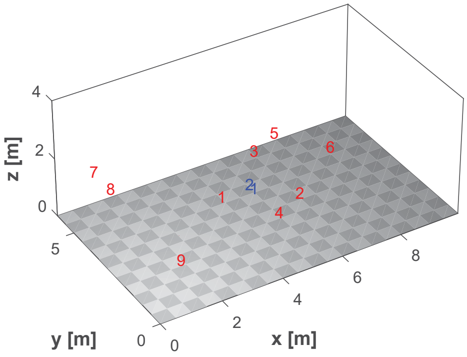


Fig. 4: Simulation setup of a cocktail party scenario.

Reverberation

Reverberation occurs when the speech signal is emittted in an acoustically enclosed space, e.g., a room which has boundaries that reflect sound signals. This leads to sound signals being reflected with certain amount of attenuation due to the absorption charachterestics of the room surface. In such scenarios, apart from the sound signal that travels directly from the speaker to the listener's AS, AS may receive multiple delayed and attenuated versions of the sound signal. This phenomenon is termed as reverberation. The strength of reverberation is generally characterised using reverberation time($T60$). $T60$ is the amount of time required for the reflected sound signal to undergo a

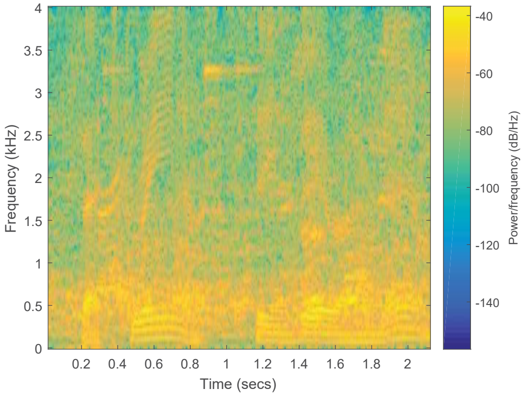


Fig. 5: Spectrogram of the signal corrupted by additive noise.

decay of 60 dB [14, 15]. It is directly proportional to the room size and inversely proportional to the absorption coefficients. An alternative to $T60$ is the direct to reverberation ratio [16]. Fig. 6 shows the spectrogram of signal seen in Fig. 2 with reverberation. It can be seen from the figure that many of the phonemes that are lower in magnitude are masked by preceding louder phonemes. Some of the state of the art methods for performing dereverberation are [16, 17].

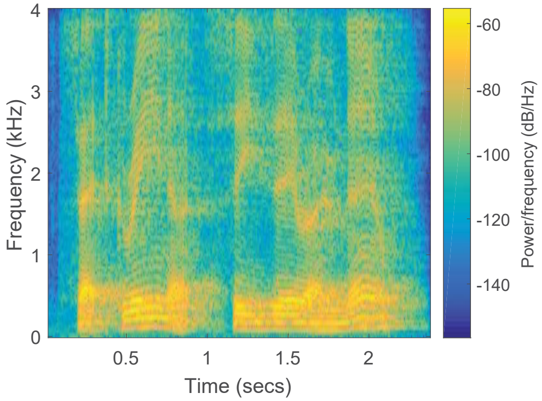


Fig. 6: Spectrogram of the reverberated signal.

2 Speech Modelling

The main motive of this thesis is to remove the degradation present in the speech signal due to additive noise. In such cases, the estimation of clean speech from the noisy signal is non-trivial particularly in case of highly non-stationary background noise. There exists a variety of methods for filtering the noisy speech to obtain the clean speech, and these require the knowledge of speech and noise statistics. Since there is no access to either the clean speech or the noise signal in many of the applications, the statistics have to be estimated directly from the noisy speech which can be modelled in the case of additive noise as,

$$z(n) = s(n) + w(n) \quad \forall n = 0, 1, \ldots \tag{1}$$

where $z(n)$ is the noisy speech, $s(n)$ is the clean speech, and $w(n)$ is the additive background noise. Estimation of the statistics with just access to the noisy signal can prove to be complex. In such scenarios, it is quite useful and convenient to assume a model for the speech and noise. This chapter will provide a few of the signal models that are relevant to this thesis. A model in essence can be represented using a set of parameters which generally has a much smaller dimension than the data to be modelled. In the section that follows this paragraph, we will give a brief introduction into the source-filter model of the speech production process followed by the harmonic model in Section 2.2 which provides a good approximation of the voiced speech. As the processing is generally carried out frame-wise, we will denote the windowed signal containing the f^{th} frame as

$$s_f(m) = s(fM + m)\epsilon(m) \;\; \forall \;\; m = 0, \ldots M - 1 \tag{2}$$

where $\epsilon(m)$ is a window of length M defined in the interval $[0, M-1]$ applied to the data. It is assumed that $s_f(m) = 0$ for $m < 0$ and $m > M - 1$. A frame of clean speech is denoted as $\mathbf{s}_f$ which is defined as

$$\mathbf{s}_f = [s_f(0), s_f(1), \ldots s_f(M-1)]^T.$$

Corresponding counterparts for the noisy and noise signal will be denoted as $\mathbf{z}_f$ and $\mathbf{w}_f$, respectively. The frequency domain counterpart of the speech signal is defined as $S(\omega) = \mathbf{f}^H(\omega)\mathbf{s}_f$ where $\mathbf{f}(\omega) = [\exp(\imath\omega 0), \ldots \exp(\imath\omega(M-1))]^T$, and the corresponding counterparts for the noisy and noise signals are denoted as $Z(\omega)$ and $W(\omega)$, respectively.

2.1 Source-filter model for speech production

This model tries to explain speech signals as a combination of the excitation signal and the vocal tract [18]. As explained previously in Section 1.1, a source-filter model explains the speech signal as the output of a filter (which models the vocal tract) to the excitation signal which is white Gaussian noise (WGN) in case of unvoiced speech and a periodic signal in the case of voiced speech, as shown in Fig. 7. The vocal tract can be represented using an all pole filter which leads to the speech signal being expressed as

$$s(n) = -\sum_{i=1}^{P} a_i s(n-i) + e(n), \tag{3}$$

where $s(n)$ is the speech signal, $e(n)$ is the excitation signal and $A(z) = \frac{1}{1+\sum_{i=1}^{P} a_i z^{-i}}$ is the all pole filter representing the vocal tract. In the case of speech, the characteristics of the vocal tract and the excitation signal produced by the vocal chords are time varying. However, due to the limitations of the human speech production process, the properties associated with them can be assumed to be stationary for frames of approximately 20-25 milliseconds [3]. Given a frame of clean speech , $\{s_f(m)\}_{m=0}^{M-1}$, the autoregressive (AR) coefficients, $\{a_i\}_{i=1}^{P}$, are obtained by the Yule-Walker method [19, 20] or by minimising the mean squared prediction error [21, 22] given as

$$\{\hat{a}_i\}_{i=1}^{P} = \underset{\{a_i\}_{i=1}^{P}}{\arg\min} \sum_{m=0}^{M-1+P} \left| s_f(m) + \sum_{i=1}^{P} a_i s_f(m-i) \right|^2 \tag{4}$$

which leads to the Yule-Walker equations as

$$\begin{bmatrix} R(0) & R(1) & \dots & R(P-1) \\ R(1) & R(0) & \dots & R(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(P-1) & R(P-2) & \dots & R(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_P \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(P) \end{bmatrix} \tag{5}$$

where the autocorrelation coefficients are then estimated using the equation

$$R(p) = \frac{1}{M} \sum_{m=p}^{M-1} s_f(m) s_f(m-p) \quad 0 \leq p \leq P. \tag{6}$$

Methods have been proposed to solve (5) efficiently using the Levinson-Durbin algorithm [23, 24] or the Delsarte-Genin algorithm [25]. This way of estimating the model parameters is very well suited for unvoiced speech as the excitation signal for unvoiced speech can be very well approximated as WGN. However for voiced speech, this model may have its limitations as the voiced speech is characterised by a periodic excitation signal and the periodicity in the excitation signal is perceived as the pitch. Some methods which takes into account the properties of the excitation signal for voiced speech have been proposed in [26, 27]. While having access to clean speech,

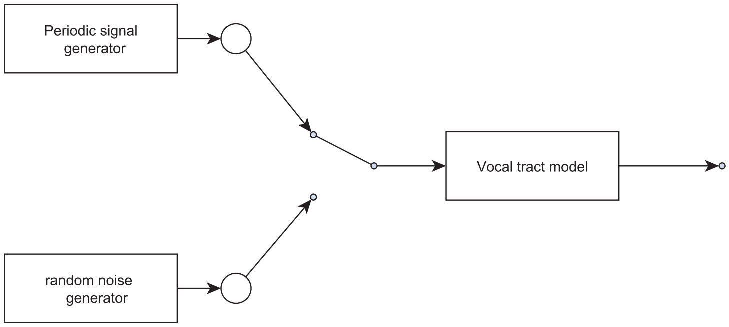


Fig. 7: Schematic for the source-filter model for speech production.

the predictor coefficients as well as the excitation variance can be estimated using the Yule-Walker equations. However, in the presence of noise, the estimation of these parameters is not very trivial and this will be addressed in this thesis.

Apart from the above model, purely voiced speech can be represented by a harmonic model which will be explained in the following section.

2.2 Harmonic model

Another model that is well suited for the representation of the voiced speech is the harmonic model [28]. These models are useful for the estimation of parameters such as the pitch. Apart from speech, many real life signals such as the electrocardiogram (ECG) [29], sound from musical instruments [30] etc. are quasi periodic and can be represented well using the harmonic model. Using this model, the clean speech can be represented as a sum of harmonically related sinusoids as

$$s_f(m) = \textstyle\sum_{l=1}^{L} A_l \cos(l\omega_0 m + \phi_l) \tag{7}$$

$$= \textstyle\sum_{l=1}^{L} a_l e^{jl\omega_0 m} + a_l^* e^{-jl\omega_0 m} \tag{8}$$

where ω_0 is the fundamental frequency, $a_l = \frac{A_l}{2} e^{j\phi_l}$, $A_l > 0$ is the complex amplitude of the l^{th} harmonic, L is the harmonic model order and ϕ_l is the phase of the l^{th} harmonic. Often, it is more convenient to estimate fundamental frequency from analytic signals due to simpler notation. Equivalent analytic or complex signal corresponding to the real signal in (7) can be written as [31, 32]

$$s_f(m) = \sum_{l=1}^{L} \alpha_l e^{jl\omega_0 m} \tag{9}$$

where $\alpha_l = A_l e^{j\phi_l}$. Given a frame of voiced speech, the important parameters to be estimated are the complex amplitudes α_l, the fundamental frequency and the harmonic model order [28]. A erroneous estimate of the model order can lead to errors in the estimates of fundamental frequency (such as frequency halvings and doublings). The model order can be estimated jointly with the fundamental frequency as in [33, 34] or separately as in [35]. Fig. 8 shows the approximation of a frame of voiced speech using the estimated parameters representing a harmonic model. It can be seen that all the properties of the signal are well captured using the harmonic model.

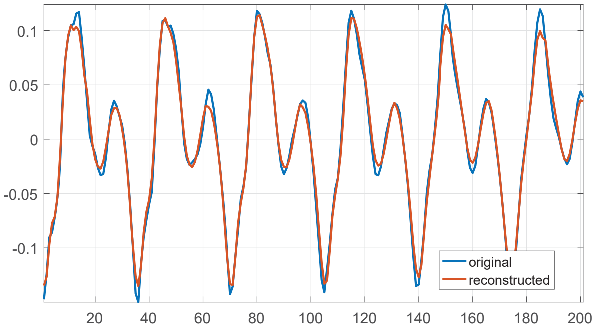


Fig. 8: Reconstruction of voiced speech using the harmonic model.

2.3 Model for noise

It is a common practise to assume a white noise spectrum in many of the signal processing algorithms due to the mathematical tractability and simplicity of the algorithms. In such cases, we need to estimate only a single value corresponding to the intensity of PSD for all frequencies. However, in practical scenarios, the spectral content of noise can have any shape. In such cases, we can model the noise also as an AR process as

$$w(n) = -\sum_{i=1}^{Q} c_i w(n-i) + v(n) \tag{10}$$

where $\{c_i\}_{i=1}^{Q}$ are the AR coefficients corresponding to the noise signal and $v(n)$ is a WGN with variance σ_v^2. Similar to the speech signal, it is a common practise to assume quasi-stationarity for the AR coefficients and the excitation variance. Fig. 9 shows the modelling of noise periodogram of babble noise using a 14 order AR process. The estimation of noise AR coefficients and the excitation variance from the noisy signal will also be addressed in this thesis. As mentioned earlier, certain algorithms assume that the noise is white.

In such cases, pre-whitening [36] of the noisy signal must be performed to obtain satisfactory results.

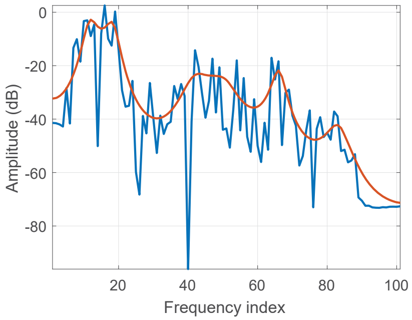


Fig. 9: Approximation of the noise periodogram using a 14 order AR process.

3 Enhancement of Speech

As alluded to in the introduction, degradation caused due to additive noise and reverberation leads to the reduction in the quality and intelligibility of the perceived speech. This degradation is more pronounced among HI people than the normal hearing people. In such scenarios, it is very much desirable to perform enhancement of the noisy speech e.g. in a HA before presenting the speech to the HI people. The field of speech enhancement has been investigated extensively from the early 1960s when Schroeder filed the first patent [37, 38] in this field. A plethora of methods have been proposed since then to perform speech enhancement [39–42]. These methods can be broadly categorised into single and multi channel methods. Multi channel methods which requires access to observation from multiple channels can be effective in exploiting the spatial information to focus on the sound source of interest [43]. Some of the major classes of multi channel enhancement algorithms are 1) fixed beamforming techniques such as delay and sum beamforming and superdirective beamforming [44, 45] 2) adaptive beamforming techniques such as the minimum variance distortionless response (MVDR) beamformer [46] which combines the properties of the fixed beamformers along with adaptive noise reduction and 3) multi channel Wiener filtering techniques [47, 48] that obtains the minimum mean square error (MMSE) estimate of the clean speech component at the reference microphone. In the remainder of this chapter, we will mainly focus on single channel speech enhancement methods. A majority of these methods are based on the Fig. 10. As seen from the figure, the task of speech enhancement can be divided into two blocks. The first step consists of estimating the relevant signal parameters and statistics, and the second step performs the filtering of the noisy signal, $z(n)$, based on the estimated parameters/statistics to obtain the enhanced signal $\hat{s}(n)$. The rest of this chapter elaborates separately on these two steps. In Section 3.3 we will describe how these speech enhancement algorithms can be adapted for HAs.

During the design of speech enhancement algorithms, it is necessary to evaluate the performance of these algorithms. Section 3.4 will elaborate on a few subjective and objective measures we have used in this thesis.

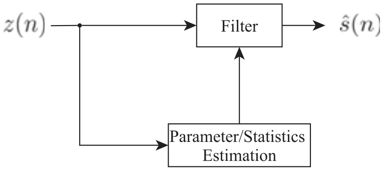


Fig. 10: Basic block diagram of a speech enhancement system.

3.1 Filtering of noisy speech

In this section we will explain a few of the well known methods for performing filtering of the noisy signal to obtain the enhanced signal. These methods require an estimate of the speech/noise statistics or an estimate of the speech/noise model parameters. The performance of the filtering methods depend heavily on these estimates. We will now explain a few of the well known methods for filtering the noisy signal followed by the estimation of the filter parameters in the next section.

Spectral Subtraction- Proposed by Boll in 1979 [49], this was one of the first single channel speech enhancement methods. These methods were motivated by the observation that majority of the background noise is additive in nature and thus estimates the magnitude spectrum of the clean speech by subtracting the estimated noise magnitude spectrum from the noisy magnitude spectrum as

$$|\hat{S}(\omega)| = |Z(\omega)| - |\hat{W}(\omega)| \tag{11}$$

where $|Z(\omega)|$ is the noisy magnitude spectrum, $|\hat{W}(\omega)|$ is the estimated noise magnitude spectrum and $|\hat{S}(\omega)|$ is the magnitude spectrum of the enhanced signal. There can be cases where the estimated magnitude spectrum of speech becomes negative due to the overestimation of the noise magnitude spectrum. This issue is resolved by setting those values to zero or a very small positive value, which results in a non-linear operation. Another possibility to resolve this is by using a frequency dependent threshold. The estimated clean magnitude spectrum , $|\hat{S}(\omega)|$, along with the phase of the noisy signal is used to generate the estimated speech in the time domain. Eventhough the principle behind spectral subtraction is fairly simple, this method has been observed to have certain shortcomings. The most dominant of these shortcomings is that it introduces an artefact termed as the musical noise [50], which is caused due to presence of small isolated randomly located peaks in the spectrum. There have been methods proposed to tackle this issue such as in [50], which subtracts an over estimate of the noise spectrum. Extensions to spectral subtraction have been proposed over the years such as the parametric spectral subtraction in [51], multi band spectral subtraction in [52] and non linear spectral subtraction in [53].

Wiener filtering- In comparison to the spectral subtraction methods that were based on intutive and heuristic principles, Wiener filter is an optimum filter in the sense that it minimises the mean squared error between the desired signal and the estimated signal [54] when both the speech and noise discrete Fourier transform (DFT) coefficients are assumed to be independently distributed complex Gaussian random variables. Wiener filtering of speech can be performed either in the time or frequency domain. The Wiener filter in the frequency domain can be expressed as

$$H(\omega) = \frac{\mathbb{E}[|S(\omega)|^2]}{\mathbb{E}[|S(\omega)|^2] + \mathbb{E}[|W(\omega)|^2]} \approx \frac{\hat{\phi}_{ss}(\omega)}{\hat{\phi}_{zz}(\omega)} \tag{12}$$

where $\hat{\phi}_{ss}(\omega)$ is the estimate of speech power spectral density (PSD) [7] and $\hat{\phi}_{zz}(\omega)$ is the estimate of noisy PSD. Different extensions to the Wiener filter such as the parametric Wiener filter [55] which allows the mechanism to introduce additional noise attenuation have been proposed. Even though the Wiener filter is optimal in the mean square error sense under the assumptions stated above, it might not be optimal from a perceptual perspective point of view. To incorporate the perception into account, constraints which allows us to control the speech distortion and noise attenuation can be imposed during the derivation of the Wiener filter [56]. An extensive study on the trade off between signal distortion and noise reduction while using Wiener filter can be found in [57]. The Wiener filter assumes that the signals being analysed are stationary. To deal with nonstationary speech and noise signals, Kalman filtering based methods can be used. In these methods, the speech as well as the noise signals are characterised by a state space model and the enhanced signal is then computed by estimating the state vector. Kalman filtering for speech enhancement in the presence of white background noise was proposed in [58] and was later extended to deal with colored noise in [59, 60].

Statistical model based methods- This class of methods assumes a certain probability density function model for the speech and noise discrete Fourier transform (DFT) coefficients/magnitudes and use these probability density functions as a basis for deriving estimators for the magnitude or complex spectrum of the speech signal. These estimators can be non linear in nature compared to the Wiener filter (discussed in the previous section) which yielded a linear estimator of the complex spectrum of the signal when the speech and noise DFT coefficients were assumed to be complex Gaussian random variables. One of the first methods belonging to this class was proposed in 1980 where a maximum likelihood estimator of the speech magnitude spectrum was derived by assuming a complex Gaussian model for the noise DFT coefficients while assuming a deterministic model for the speech [61]. This was then combined with a two state model (describing the speech presence or absence) to obtain higher noise reduction in the noise only regions. Another well known approach to perform statistical model based speech enhancement was proposed by Ephraim and Malah in [62], where an MMSE estimator for the magnitude spectrum was derived by modelling the real and imaginary parts of the speech and noise DFT coefficients as Gaussian random variables. The same authors, motivated by the human perception of speech, later proposed the MMSE estimator of the logarithm of the magnitude spectrum [63]. This was found to further improve the noise reduction capabilities. Some of the more recent statistical model based methods are [64, 65]. In [64], a maximum a posteriori (MAP) estimator of the spectral magnitudes was proposed by modelling the magnitude of the speech DFT coefficients using super Gaussian distributions whereas the method in [65] proposed MMSE estimators of speech DFT coefficient magnitudes, where a generalised Gamma prior distribution is used to model the distribution of magnitude of the speech DFT coefficients.

Subspace based methods- These methods are based on the principle that the desired signal lies in a subspace of the full space which contains the signal subspace and noise subspace [66, 67]. The desired signal is then obtained by projecting the observed signal onto the signal subspace. The signal and noise subspace can be identified using orthogonal matrix factorisation techniques from linear algebra as done in [68] using singular value decomposition (SVD). These approaches were first designed for white noise and then later extended for colored noise case in [69–71]. In [72], the basis vectors representing the signal and noise space are computed by performing the joint diagonalisation of the speech and noise correlation matrices. The filter coefficients for performing enhancement was then designed by choosing a weighted linear combination of the basis vectors. The resulting filters had the ability to trade off between speech distortion and noise reduction.

3.2 Statistics and Parameter estimation

In the previous section, we gave a brief overview on the different classes of filtering methods. Most of these algorithms require an estimate of the speech or noise statistics/parameters. In this section we give a brief overview of different methods for estimating the statistics/parameters. These methods can be broadly categorised into supervised and unsupervised methods [73]. Supervised methods use prior training data consisting of speech and noise from a certain database to build models. On the other hand, unsupervised methods generally do not need any prior training data. A majority of these methods assume a statistical model for speech and noise and derive estimators based on these models.

Unsupervised methods

Noise Statistics estimation The estimation of noise statistics forms a very a critical component for the functioning of many speech enhancement methods discussed in the previous section. The noise statistics that are relevant for the speech enhancement is the power spectral density, $\phi_{ww}(\omega)$, which is defined as [7]

$$\phi_{ww}(\omega) = \lim_{M\to\infty} \mathbb{E}\left\{ \left| \frac{1}{M} \sum_{m=0}^{M-1} w_f(m) e^{-\imath\omega m} \right|^2 \right\}, \tag{13}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator. A plethora of methods have been proposed to solve this problem in the past few decades. One of the most basic approaches to solve this problem is by using a voice activity detector which tries to estimate the noise statistics in the absence of speech. However these methods are very difficult to tune and they have been observed to be inefficient in the presence of non-stationary noise [74]. Another popular method for noise PSD estimation is the minimum statistics (MS) proposed by Rainer Martin in [75, 76]. This method is based on the observation that the power of the noisy speech signal goes down to the power of the noise signal often. The noise PSD is then computed by taking the minimum of the recursively smoothed noisy periodogram over a window of variable size. In [77], Doblinger proposed a noise PSD estimator with similar principles. However, the methods based on the MS principle are very sensitive to the size of the window which limits the effectiveness of the method to track highly non-stationary noise. Another class of algorithms is based on the recursive averaging. In [74, 78] the noise PSD was obtained by recursive averaging of the noisy spectral power where the recursive averaging coefficient was a time varying frequency dependent value. This coefficient was adjusted according to the estimated speech presence probability. However, some tracking delay still existed in the presence of rapidly changing noise power. These issues were addressed to some extent in the methods proposed in [79, 80] that were based on the MMSE criterion which modelled the speech and noise DFT coefficients as independently distributed complex Gaussian random variables. The final noise PSD estimate is obtained by recursive averaging of the MMSE estimate of the noise periodogram. Another class of noise PSD estimation algorithms are the histogram based noise estimation algorithms. These algorithms are based on the observation that the maximum number of values in the individual frequency indices correspond to the noise power at that frequency index [81–83]. There have also been approaches to estimate a parametric form of the noise PSD using the expectation maximisation algorithm as done in [59, 60].

The performance of a noise PSD estimation algorithm can be evaluated by measuring its spectral estimation accuracy and the enhancement performance. Such an evaluation of the noise PSD estimators can be found in [84]. In this thesis, we have also carried out an evaluation of some of the state of the art noise PSD estimators along with a method we proposed in paper D.

Fundamental frequency estimation Fundamental frequency is an important characteristic of the voiced speech and it is very critical for the perception of speech [28]. Apart from speech enhancement [85, 86], fundamental frequency is a critical parameter of interest signal coding [87–89], music transcription [90, 91], signal modification [92, 93]. Some of the popular fundamental frequency estimation methods include that of [94, 95]. The basic principle of these methods is to compare the signal under analysis with a delayed version of itself using a certain similarity measure. The fundamental frequency can then be obtained by taking the reciprocal of delay for which the similarity measure is maximised. The methods do not assume any model for the speech signal and are hence referred to as non-parametric methods. While these methods are computationally efficient, they often rely on heuristics and are prone to errors in the presence of noise. Another class of methods are the parametric methods [28] which is derived on the basis of the harmonic model introduced in Section 2.2. These methods are attractive due to their mathematical tractability and robustness to noise for very low signal to noise ratio (SNR), and efficient implementations of these have been proposed in [96].

Supervised methods

The unsupervised methods for the estimation of the noise PSD have their limitations in the presence of non-stationary background noise. In comparisons to unsupervised methods, supervised methods are able to deal better with such scenarios. In this section, we will give a brief overview on the supervised methods for estimating the speech and noise statistics. Contrary to unsupervised methods, supervised methods make use of a priori information about the noise type or speech in the form of training data. In the case of prior knowledge about the noise environment or about the speaker of interest, supervised algorithms can be trained to that particular scenario using the training data corresponding to that scenario. However, in absence of such information, the training data can consist of speech and noise from databases containing different speakers and common background noises. The training data is then used to build models of speech and noise [97]. Subsequently, a certain interaction model is defined between the speech and noise models that generates the noisy data. The estimation of the statistics is then carried out by the means of optimisation algorithms on the basis of the above mentioned models and the observed data. Some of the popular supervised methods are the non-negative matrix factorisation (NMF) based methods [97–99], hidden markov model (HMM) based methods [100–102] and codebook based methods [103–105]. We will now briefly explain the NMF and codebook based methods as we have used these in our thesis. NMF techniques allow us to approximate non-negative data using a set of basis vectors and activation coefficients. The basic objective in NMF is to approximate a non-negative matrix V with dimensions $F \times N$ using two non-negative matrices as

$$V \approx WH \tag{14}$$

where W and H are of dimensions $F \times K$ and $K \times N$, respectively, and $FK + KN \ll FN$. The matrix V contains the K F-dimensional observation vectors arranged as the columns. Using (14), each column of V can be represented as a linear combination of the columns of W weighted by the elements in the corresponding column in H. Therefore, W can be regarded as the basis matrix containing the basis vectors along its columns which is used to linearly approximate the data in V. As the elements in H are used to activate the basis vectors present in W, H is referred to as the activation matrix. Fig. 11 shows an example of performing NMF on the spectrogram of twelve seconds piano snippet into 2 basis vectors and the corresponding activation coefficients. As can be seen in the figure, the note *C5* is played in the first segment followed by the note *F4* in the second segment and both the notes are played together in the third segment. This can also be seen from the right side of the figure where the activation coefficient corresponding to first spectral basis vector (representing *C5*) and the second spectral basis vector (representing *F4*) is active in the first and second segments, respectively, whereas both the spectral basis vectors are active in the third segment as both the notes are played together.

To use NMF from a speech enhancement perspective, the observation matrix V contains the magnitude/power spectrogram of the noisy signal where each column corresponds to the magnitude/power spectrum of the noisy signal at a certain frame index. For the supervised NMF based speech enhance-

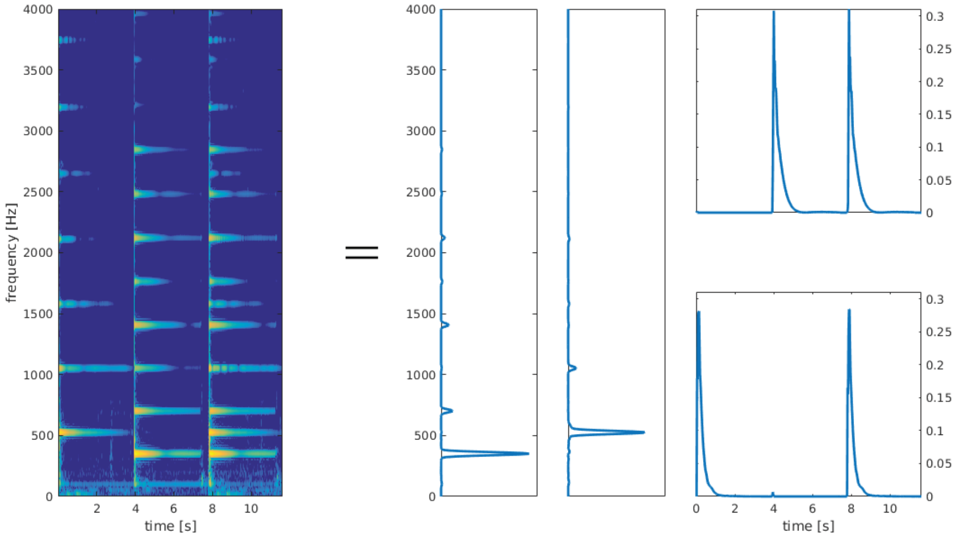


Fig. 11: This figure shows NMF being performed on the spectrogram a piano segment into two basis vectors and the corresponding activation coefficients.

ment, the basis matrix W consists of separate spectral basis vectors trained for speech and noise arranged as shown in (15) and (16) as,

$$V \approx \underbrace{\begin{bmatrix} d_1(1) & \dots & d_K(1) \\ \vdots & \ddots & \vdots \\ d_1(F) & \dots & d_K(F) \end{bmatrix}}_{W} \underbrace{\begin{bmatrix} \sigma_1^2(1) & \dots & \sigma_1^2(N) \\ \vdots & \ddots & \vdots \\ \sigma_K^2(1) & \dots & \sigma_K^2(N) \end{bmatrix}}_{H} . \tag{15}$$

$$V \approx [W_s \, W_w] \begin{bmatrix} H_s \\ H_w \end{bmatrix} = WH, \tag{16}$$

and a particular column of the activation matrix contains the activation coefficients corresponding to the different basis vectors for a particular frame of the signal being analysed. In supervised NMF approaches, the objective is to estimate the activation coefficients such that the observation matrix V is well approximated by WH in terms of certain divergence criterion. The type of divergence criterion is often chosen based on the type of data to be approximated. In the case of speech, some of the typically used divergence criterions are the Itakura-Saito (IS) divergence [106] and the Kullback-Leibler (KL) divergence [107]. It has been shown in [99], that NMF with IS divergence is equivalent to maximum likelihood (ML) estimation of activation coefficients when the observation is modeled as a sum of Gaussian components. Similary, NMF with KL divergence is shown to be optimal when the observation is modelled as sum of Poisson components [108]. The optimisation problem is then solved using iterative algorithms such as expectation-maximisation [109] or multiplicative update rules [98]. The estimated activation coefficients for the speech along with the spectral basis vectors for speech can be used to obtain the speech PSD estimate and similarly for the noise. Several methods have been proposed to enhance the performance of conventional NMF by enforcing sparsity [99, 110, 111] and temporal continuity [112–114] of the activation coefficients.

Another class of supervised methods proposed for the estimation of speech and noise statistics are the codebook-based approaches [103–105]. These approaches model the a priori information regarding the speech and noise spectral envelopes in the form AR coefficients and a parametric representation of the speech and noise statistics is estimated. During the training phase, AR coefficients extracted from frames of speech and noise training data, which are converted into line spectral frequency (LSF) coefficients [115] are passed as input to a vector quantiser to obtain speech and noise codebooks of size N_S and N_W, respectively. During the estimation stage, parameters describing the speech and noise statistics consisting of the AR coefficients and the excitation variances are estimated. In comparison to NMF based approaches which model the noisy observation using a single model containing the weighted linear combination of spectral basis vectors, codebook based approaches model the noisy signal using $N_S N_W$ models. These models are later averaged based on how well they describe the noisy observation. The estimated model averaged parameters can then be used to parametrically represent the PSD of the speech and noise as

$$\phi_{ss}(\omega) = \frac{\sigma_u^2}{|A_s(\omega)|^2} \qquad \phi_{ww}(\omega) = \frac{\sigma_v^2}{|A_w(\omega)|^2}$$

where σ_u^2 and σ_v^2 represents the speech and noise excitation variance, respectively, and $A_s(\omega) = 1 + \sum_{p=1}^{P} a_p e^{-\imath p\omega}$, $A_w(\omega) = 1 + \sum_{q=1}^{Q} c_q e^{-\imath q\omega}$ where $\{a_p\}_{p=1\ldots P}$ and $\{c_q\}_{q=1\ldots Q}$ are the AR coefficients corresponding to speech and noise, respectively. The number of parameters to be trained in the codebook based approach is smaller in comparison to the NMF based approaches as the AR coefficients are typically in the order of 14. In paper F, we have proposed to parametrically represent the basis vectors used within the NMF framework using AR coefficients. We show in this paper that if we represent a frame of noisy signal in the time domain as a sum of autoregressive processes, maximising the likelihood corresponds to performing NMF of the observed data using the IS divergence as the optimisation criterion. The spectral basis matrix in this case would contain the spectral envelopes corresponding to the AR coefficients and the activation coefficients can be interpreted as the variance of the excitation noise that excites the AR filter.

Another class of supervised methods for the enhancement of speech which has gained attention recently are the deep neural network (DNN) based methods [116, 117]. This class of methods formulates speech enhancement as a learning problem where the system uses the training data to learn discriminative patterns of speech and noise [116]. A majority of these algorithms can be divided into three components: 1) learning machines which includes the different DNNs (such as multi-layer perceptrons [118], convolutional neural networks [119], recurrent neural networks [120], generative adversarial networks [121] etc.) 2) training targets that specifies the target that is to be achieved and 3) acoustic features which are fed as an input in to the neural network. Apart form DNN based methods for speech enhancement in noise, they have also been successfully used in source separation in single channel [122] and multi channel [123] scenarios, and dereverberation [124, 125].

3.3 Application to hearing aids

The main motive of this thesis is to develop speech enhancement algorithms for HA applications. Often, a hearing impaired person is equipped with two HAs, at the left and right ear [126, 127]. The two HAs can work independently or together. Recent developments in HA technology makes it possible for the HAs to communicate with each other and exchange information. This enables the use of binaural processing algorithms. Binaural processing is advantageous due to the usage of spatial information [128]. Fig. 13 shows a possible schematic of the signal processing system in such cases. The noisy signals received at the left and right ears are denoted by $z_l(n)$ and $z_r(n)$, respectively, as shown in Fig. 12. The noisy signals at both the ears are used jointly to estimate the parameters/statistics of the speech/noise signal. The estimated parameters/statistics are consequently used at left/right ears to filter the noisy signals $z_l(n)$ and $z_r(n)$ to obtain $\hat{s}_l(n)$ and $\hat{s}_r(n)$ respectively. There are cases where the HA at each ear have multiple microphones. Some of the the algorithms proposed for this case includes [43, 129, 130]. In [43] a multi-channel Wiener filter is adapted for usage in binaural hearing aids, and this method was shown to preserve the binaural cues from the desired source while distorting the binaural cues of the interfering sources. However, in this thesis we will restrict ourselves to the case where we have a single obervation per HA. This situation can be observed in in-the-ear HAs as the the limited space restricts the number of microphones that can be used. One of the first algorithms to propose such a two-input two-output system was in 1996 by Ernst and Dorbecker [131]. The enhancement system here consisted of two stages, the first stage performed a two channel spectral subtraction and a post filter based on the Wiener filtering techniques is applied in the second stage. The spectral subtraction performed here was performed using a common noise PSD estimate that was obtained using the observation at both the left and right channels. The binaural cues, ITD and ILD, were not preserved in this case as a result of different gains being applied on different channels. The estimation of noise PSD in this case was based on the uncorrelatedness assumption. This was later extended to deal with arbitary coherence in [132]. Some of the more recent approaches to perform dual channel speech enhancement includes the work done in [133–135]. In [134], the noisy signals at the left and right channels are fed into a superdirective beamformer to obtain an enhanced signal. This signal is then used as the clean reference signal to estimate the gain applied onto the noisy signal. The two stage enhancement system in [135] uses equalisation cancellation theory to estimate the interference signals in the first stage and applies a Wiener filter in the second stage. The methods in [134, 135] used a common gain on both the channels to preserve the inter-aural cues. In this thesis, papers A and B are based on the schematic shown in Fig. 13. In paper A, the noisy signals at the left and right ears were filtered using a Kalman filter and filter parameters required for the functioning of the Kalman filter were estimated jointly using the information on the left and right channels. The filter parameters here consisted of the AR coefficients and the excitation variance corresponding to the speech and noise as well as the pitch parameters corresponding to the clean speech. Using the pitch information was found to improve the performance as it could better explain the voiced portions of speech. In paper A, we have proposed an optimal way to estimate these parameters under the assumption that the speaker of interest is located in the nose direction of the listener and have also demonstrated the benefit of using the left and right channels jointly to estimate the filter parameters. In paper B, we extended this to take into account the cases where the speaker of interest is not constrained to be located in the nose direction of the listener. Taking into account this information, leads to better modelling and enhancement of the noisy signal when the speaker of interest is not located in the nose direction of the listener.

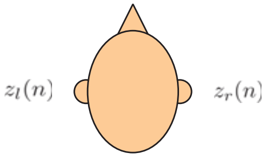


Fig. 12: Binaural noisy signals.

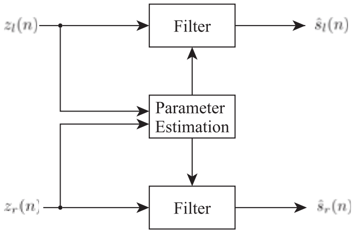


Fig. 13: Basic block diagram of dual channnel enhancement framework.

3.4 Evaluation of processed and degraded speech

During the design of speech enhancement algorithms, it is desirable to test the quality/intelligibilty of the processed and degraded signal using certain measures. These measures can be broadly categorised into subjective [136] and objective measures [137]. In this section we will introduce few of the subjective and objective measures that are commonly used in the field of speech enhancement, and also in this thesis. Finally, we will give a brief introduction on non-intrusive speech intelligibility prediction.

Subjective measures

The end product of most of the speech enhancement algorithms is to be heard by a human e.g. in HAs. In these scenarios, it is preferred for the algorithm to be tested by a set of humans. The processed audio files are tested on basis of different aspects. Two main aspects used commonly for the speech enhancement algorithms are quality [138] and intelligibility [139]. One way to measure the quality is using the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) tests [140] which has been used in the field of speech coding. For speech enhancement, MUSHRA tests can be conducted by presenting the test subjects with different processed signals along with the reference signal which is the clean signal. Fig. A.6 shows an example of the graphical user interface (GUI) [141] that is presented to the test subjects. The listeners are then asked to give the score to different signals after hearing the reference signal. Another quantity that is of particular interest to the HA industry is to measure the intelligibilty of the processed signal. Obtaining an improvement in intelligibility has been found to be more challenging in comparison to improving the quality [142]. While measuring the intelligibility, the test subjects are asked to listen to sentences with a certain syntax that contain information [143, 144] and asked to identify the information. The sentences generally consists of the noisy signals and the processed signals played pack in random order. One example of such a sentence would be: Bin Blue by S 5 please. The listeners are then asked to identify the color, alphabet and the number in the sentence using a GUI as shown in Fig. 15. Based on the answers given by the subject, one can create a mean intelligibility curve giving an indication whether the processing algorithm increases/decreases intelligibility relative to the unprocessed signal. Even though, the subjective tests give us a very good indication of the processing capabilities of the enhancement algorithm, a disadvantage of the subjective tests is the need of resources such as time and enough participants. Therefore, for practical purposes, it is handy to have objective measures that can be run on a computer.

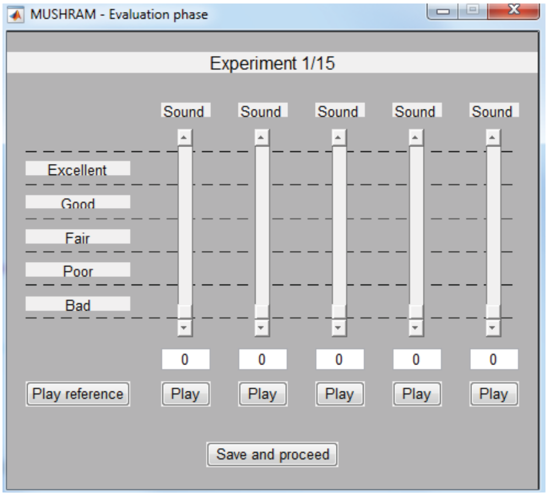


Fig. 14: Graphical user interface for conducting MUSHRA tests [141].

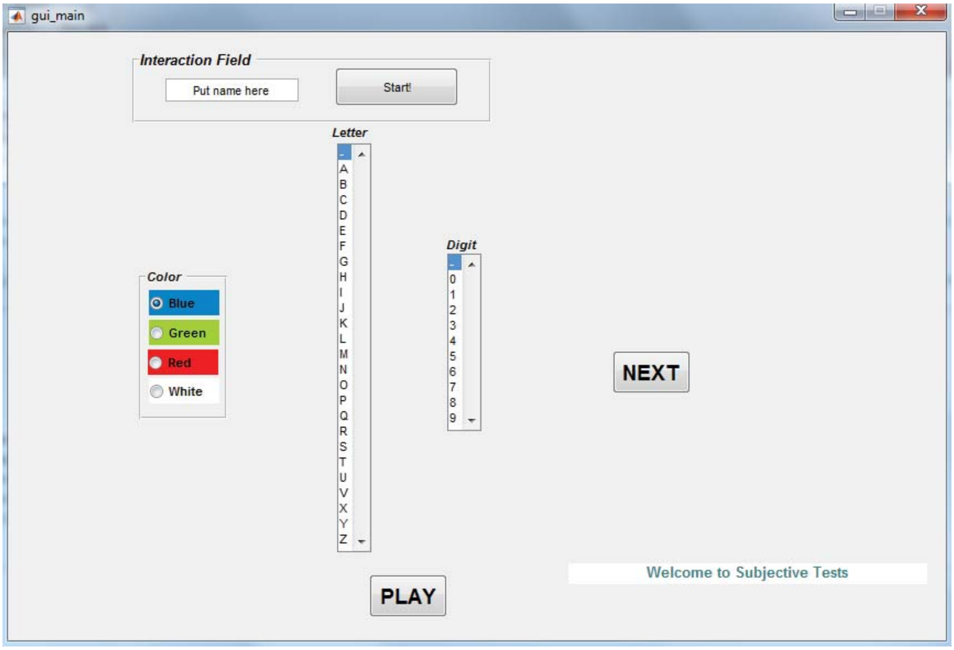


Fig. 15: Graphical user interface for conducting the intelligibility tests.

Objective measures

A commonly used objective measure for measuring the speech enhancement performance is the segmental signal to noise ratio which is defined as [80]

$$\text{segSNR} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{m=1}^{M} s^2(lM+m)}{\sum_{m=1}^{M}(s(lM+m)-\hat{s}(lM+m))^2} \tag{17}$$

where L denotes the number of frames. This measure indicates how close the enhanced signal is to the clean signal in a MSE sense. Along with the segmental signal to noise ratio, some of the other commonly used objective measures are the Segmental speech SNR (spSNR) and Segmental noise reduction (segNR) which are defined as [80]

$$\text{spSNR} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{m=1}^{M} s^2(lM+m)}{\sum_{m=1}^{M}(s(lM+m)-\tilde{s}(lM+m))^2} \tag{18}$$

$$\text{segNR} = \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\frac{\sum_{m=1}^{M} e(lM+m)^2}{\sum_{m=1}^{M}\tilde{e}(lM+m)^2} \tag{19}$$

where $\tilde{s}(n)$ is the filtered clean speech and $\tilde{e}(n)$ is the filtered noise signal. Eventhough these measures are useful, they might not really correlate with our hearing mechanism. A widely used measure that correlates with our hearing is the Itakura-Saito divergence [106] which is defined as

$$d_{\text{IS}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{\phi_{ss}(\omega)}{\phi_{\hat{s}\hat{s}}(\omega)} - \log\frac{\phi_{ss}(\omega)}{\phi_{\hat{s}\hat{s}}(\omega)} - 1\right)\, d\omega. \tag{20}$$

Other objective measures that have been designed to correlate with the human hearing are short time objective intelligibilty (STOI) and perceptual evaluation of subjective quality (PESQ) [145]. STOI gives an indication regarding the intelligibility of the processed signal and is computed by decomposing the reference and the processed signals into one-third octave bands and taking the correlation between short time temporal envelope segments of the corresponding signals. PESQ, which has been used in speech coding applications, gives a measure of the quality. As the objective measures are useful in giving us an indication regarding the quality or intelligibility of the received signal, they can be used as an indicator in the speech enhancement system within the HA when the intelligibility goes below a particular threshold. Both these measures, PESQ and STOI, are intrusive, meaning that they need access to the clean signal. However, in situations e.g. in HAs, we have at our disposal only the noisy signals. In such cases, it is desirable to predict the intelligibility just using the noisy signal.

Non intrusive Speech Intelligibility Prediction

The objective measures that we have explained in the previous section requires the access to the clean speech or noise signal. However, in practical scenarios, e.g., in a HA this is seldom true. In such cases, it might be required to measure the quality/intelligibility of the degraded signal with access only to the degraded signal and the methods capable of doing so are termed as non-intrusive. Non-intrusive methods have been proposed for measuring quality [146] and intelligibilty [147, 148]. In [148] the intelligibility prediction is done using modulation spectral representation of the degraded signal. These methods were mainly designed for signals degraded by reverberation but were also found to work in the presence of noise. Non-intrusive speech intelligibility prediction algorithms dedicated for signals degraded by noise have been proposed in [149, 150]. In this thesis, we have proposed a non-intrusive method for intelligibility prediction in paper G which is based on estimating the spectral features of speech and noise which is later given as an input to the intrusive STOI measure.

4 Contributions

This thesis mainly deals with the enhancement of speech in presence of noise with applications to hearing aids. In this thesis, we have proposed both single channel and dual channel enhancement algorithms, whose schematics are given by Fig. 10 and 13, respectively. The main body of this thesis is constituted by papers A-G. The papers D and F deal with single channel speech enhancement algorithms whereas papers A and B deal with binaural enhancement of speech using the schematic shown in Fig. 13. In addition to these, we have carried out an evaluation of state of the art noise PSD estimators for single channel speech enhancement in paper E. In paper G, we propose a method for non-intrusive speech intelligibility prediction which gives us an indication of when to apply the speech enhancement algorithms. Finally, in paper C, we investigate on how an external device, e.g , a microphone array can be used along with a HA to better enhance the speech signal. We now give a more detailed description regarding the contributions of the individual papers.

Paper A The first paper in this thesis deals with binaural enhancement of speech in the presence of babble noise. In this paper, we proposed a dual channel speech enhancement framework which takes into account the unvoiced speech as well as voiced speech. The method proposed in this paper was based on Kalman filtering of speech where the filter parameters (consisting of the AR coefficients and excitation variances of speech and noise) were estimated jointly from the noisy signals at the left and right ears. The proposed method was found to considerably improve the objective measures such as STOI and PESQ. Moreover, we conducted subjective listening tests to measure the performance of the proposed enhancement framework in terms of intelligibility and quality. Subjective tests indicated an improvement of upto 15% in intelligibility.

Paper B The dual channel enhancement method in Paper A was proposed based on the assumption that the speaker of interest is located in front of the listener. This assumption may not always hold true and it was shown in [151, 152], that the hearing impaired users in some cases tend to orient their head away from the listener. Thus, in this paper we have proposed a model based method for estimating the filter parameters in such a scenario. In this paper, we model the speech signals at the left and right ears using an AR process having identical AR coefficients but different excitation variances. A codebook based method is proposed to estimate these parameters, and the excitation variances are estimated using the the multiplicative update method. Taking into account the position of the target speech using the proposed method led to improvements in modelling and enhancement of the noisy signal.

Paper C As the number of microphones present in a HA are limited due to the space and power constraints present in the HA, the beamforming capabilities to select a particular source of interest may be limited. In this paper, we investigated on how using an external device, e.g., a microphone array can benefit the HA user. This situation can be encountered when the HA user is participating in a meeting with colleagues or sitting at a dinner table with family. In this paper, we proposed a model-based approach using the model proposed in paper F to control the beam pattern of this external device based on the look direction of the HA user. It was shown by simulations, the robustness of the proposed method at very low SNRs in a reverberant scenario. Moreover, experiments were conducted to show the benefits of using this framework for binaural/monaural enhancement of speech.

Paper D In this paper, we looked into the problem of noise PSD estimation which is an important block in many speech enhancement algorithms. A model based noise PSD estimator that allows us to include prior information regarding the spectral envelope of speech and noise is proposed. A variational Bayesian framework was used to estimate the posterior density of the noise component whose second order moment was calculated to estimate the noise PSD. This method of estimating the noise PSD was shown to obtain better spectral estimation accuracy than state of the art methods for noise PSD estimation while also having zero tracking delay.

Paper E As we have mentioned earlier, noise PSD estimation forms a critical component in many of the speech enhancement systems. In this paper, we evaluated state of the art noise PSD estimation algorithms along with the model based approach we proposed in paper D, in terms of enhancement performance and spectral estimation accuracy. The model based approach proposed in paper D was shown to outperform the state of the art noise PSD estimators in terms of the spectral estimation accuracy and enhancement performance.

Paper F In this paper, we have investigated on a parametric representation of the spectral basis vectors for NMF based speech enhancement. The spectral basis vectors are here parametrically represented using AR coefficients. This parametrisation was motivated by the source-filter model for speech production. In this work we show that if we model a frame of the noisy signal in the time domain as a sum of AR processes, the maximum likelihood estimation of the activation coefficients corresponds to performing NMF of the observed noisy periodogram into a basis matrix and activation coefficient vector using IS divergence as the optimisation criterion. Using objective measures, we show the benefit of parametric representation of basis vectors.

Paper G In this paper, we looked into the problem of non intrusive speech intelligibility prediction. The method proposed in this paper was based on STOI which measures the similarity between short time temporal envelope segments of the clean and degraded speech which are decomposed into DFT - based one third octave bands. In this paper, we propose to estimate the spectral components of the clean and noise signal using a codebook based approach. The estimated components are then used to construct the reference and degraded spectrum used in the traditional STOI measure. It was demonstrated through experiments that the proposed method was able to predict well the scores obtained by STOI using oracle information. Listening tests were also conducted to validate the performance of the method for measuring the intelligibility over a wide range of SNRs.

5 Conclusion and directions for future research

The main outcome of this thesis was the proposal of speech enhancement methods that takes into account the speech production model, and the estimation of the parameters explaining this model. The a priori information used to build the models here consisted of the AR coefficients explaining the signal. The methods proposed in this thesis can be categorised into dual channel and single channel methods. We have shown by means of objective experiments, the benefit of taking into account more channels for the estimation of the filter parameters and how they can be used for binaural speech enhancement in HAs. Apart from these, we have also proposed a method for noise PSD estimation and non-intrusive speech intelligibility prediction based on the above model. The proposed system when trained to the noise type or speaker of interest was, found to give better results in comparison to the case when the testing data consisted of speech and noise type that was not included for training. One area of further research would be to investigate the possibilities of training these models online when we have speech only or noise only scenarios. In paper C, we had investigated on how an external device could be used to benefit the HA user by controlling it based on the look direction of the HA user. We believe that this framework helps us in exploring many future possibilities such as providing an enhanced signal for estimating the clean speech statistics and providing us data for training the models for clean speech and noise. Moreover, we would also like to mention about the possibility of taking into account the human perception while performing training of the speech and noise codebooks. We believe that this needs to be further investigated as the current method of training the speech and noise AR coefficients does not exploit the difference in how the speech and noise is perceived by humans. Finally, we would like to remark that the computational complexity of these algorithms have to be analysed and further optimised.

References

[1] M. Tomasello, *Origins of human communication*. MIT press, 2010.

[2] W. J. Levelt, "Models of word production," *Trends in cognitive sciences*, vol. 3, no. 6, pp. 223–232, 1999.

[3] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. Ieee New York, NY, USA:, 2000.

[4] C. Sorin and C. Thouin-Daniel, "Effects of auditory fatigue on speech intelligibility and lexical decision in noise," *The Journal of the Acoustical Society of America*, vol. 74, no. 2, pp. 456–466, 1983.

[5] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001, vol. 1.

[6] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[7] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005, vol. 452.

[8] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.

[9] R. Drake, A. W. Vogl, and A. W. Mitchell, *Gray's Anatomy for Students E-Book*. Elsevier Health Sciences, 2009.

[10] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[11] L. Chittka and A. Brockmann, "Perception space—the final frontier," *PLoS biology*, vol. 3, no. 4, p. e137, 2005.

[12] C. J. Plack, *The sense of hearing*. Routledge, 2018.

[13] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[14] M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[15] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[16] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," 2007.

[17] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2008, pp. 85–88.

[18] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Walter de Gruyter, 2012, vol. 2.

[19] G. Udny Yule, "On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers," *Philosophical Transactions of the Royal Society of London Series A*, vol. 226, pp. 267–298, 1927.

[20] G. T. Walker, "On periodicity in series of related terms," *Proc. R. Soc. Lond. A*, vol. 131, no. 818, pp. 518–532, 1931.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[22] P. Vaidyanathan, "The theory of linear prediction," *Synthesis lectures on signal processing*, vol. 2, no. 1, pp. 1–184, 2007.

[23] N. Levinson, “The wiener (root mean square) error criterion in filter design and prediction,” *Journal of Mathematics and Physics*, vol. 25, no. 1-4, pp. 261–278, 1946.

[24] J. Durbin, “The fitting of time-series models,” *Revue de l'Institut International de Statistique*, pp. 233–244, 1960.

[25] P. Delsarte and Y. Genin, “The split levinson algorithm,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 3, pp. 470–478, 1986.

[26] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, “Sparse linear prediction and its applications to speech processing,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 5, p. 1644, 2012.

[27] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, “Stabilised weighted linear prediction,” *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[28] M. G. Christensen and A. Jakobsson, “Multi-pitch estimation,” *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[29] V. K. Murthy, L. J. Haywood, J. Richardson, R. Kalaba, S. Salzberg, G. Harvey, and D. Vereeke, “Analysis of power spectral densities of electrocardiograms,” *Mathematical Biosciences*, vol. 12, no. 1-2, pp. 41–51, 1971.

[30] J. W. Beauchamp, “Time-variant spectra of violin tones,” *The Journal of the Acoustical Society of America*, vol. 56, no. 3, pp. 995–1004, 1974.

[31] S. L. Hahn, *Hilbert transforms in signal processing*. Artech House, 1996.

[32] L. Marple, “Computing the discrete-time" analytic" signal via fft,” *IEEE Trans. on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[33] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, “An approximate bayesian fundamental frequency estimator.” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2012, pp. 4617–4620.

[34] M. G. Christensen, A. Jakobsson, and S. H. Jensen, “Joint high-resolution fundamental frequency and order estimation,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1635–1644, 2007.

[35] P. Stoica and Y. Selen, “Model-order selection: a review of information criterion rules,” *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[36] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 092953, 2007.

[37] M. R. Schroeder, "Apparatus for suppressing noise and distortion in communication signals," Apr. 27 1965, uS Patent 3,180,936.

[38] ——, "Processing of communications signals to reduce effects of noise," Sep. 24 1968, uS Patent 3,403,224.

[39] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[40] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Springer Science & Business Media, 2009, vol. 2.

[41] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[42] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[43] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

[44] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.

[45] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Trans. on Signal Processing*, vol. 51, no. 10, pp. 2511–2526, 2003.

[46] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[47] S. Doclo and M. Moonen, "Gsvd-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[48] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, pp. 487–503, 2005.

[49] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on speech and audio processing*, vol. 27, no. 2, pp. 113–120, 1979.

[50] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4. IEEE, 1979, pp. 208–211.

[51] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, 1998.

[52] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise." in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. 44 164–44 164.

[53] P. Lockwood and J. Boudy, "Experiments with a non-linear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars," in *Second European Conference on Speech Communication and Technology*, 1991.

[54] N. Wiener, "Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications," 1949.

[55] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[56] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. on Speech and Audio processing*, vol. 12, no. 1, pp. 59–67, 2004.

[57] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[58] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1987.

[59] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[60] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Process.*, vol. 39, no. 8, pp. 1732–1742, 1991.

[61] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[62] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[63] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[64] T. Lotter and P. Vary, “Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model,” *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, p. 354850, 2005.

[65] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[66] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective.* Elsevier, 2014.

[67] Y. Ephraim and H. L. V. Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Audio and Speech Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[68] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[69] H. Lev-Ari and Y. Ephraim, “Extension of the signal subspace speech enhancement approach to colored noise,” *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 104–106, 2003.

[70] U. Mittal and N. Phamdo, “Signal/noise klt based approach for enhancing speech degraded by colored noise,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 159–167, 2000.

[71] A. Rezayee and S. Gazor, “An adaptive klt approach for speech enhancement,” *IEEE Trans. on Speech and Audio Processing*, p. 1, 1999.

[72] J. R. Jensen, J. Benesty, and M. G. Christensen, “Noise reduction with optimal variable span linear filters,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 631–644, 2016.

[73] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning.* Springer series in statistics New York, NY, USA:, 2001, vol. 1, no. 10.

[74] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[75] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. European Signal Processing Conference*, vol. 6, p. 8, 1994.

[76] ——, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[77] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[78] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, vol. 9, no. 1, pp. 12–15, 2002.

[79] R. C. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2010, pp. 4266–4269.

[80] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[81] H.-G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 153–156.

[82] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust asr," *Speech Communication*, vol. 34, no. 1-2, pp. 141–158, 2001.

[83] B. Ahmed and P. Holmes, "A voice activity detector using the chi-square test," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, pp. I–625.

[84] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2011, pp. 4640–4643.

[85] J. Jensen and J. H. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001.

[86] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1948–1963, 2012.

[87] M. G. Christensen, *Estimation and modeling problems in parametric audio coding*. Department of Communication Technology, Aalborg University, 2005.

[88] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. Elsevier Science Inc., 1995.

[89] H. Purnhagen and N. Meine, "Hiln-the mpeg-4 parametric audio coding tools," in *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*, vol. 3. IEEE, 2000, pp. 201–204.

[90] A. T. Cemgil, *Bayesian music transcription*. [Sl: sn], 2004.

[91] A. Klapuri and M. Davy, *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[92] E. B. George and M. J. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 389–406, 1997.

[93] T. Quatieri and R. McAulay, "Speech transformations based on a sinusoidal representation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 10. IEEE, 1985, pp. 489–492.

[94] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.

[95] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[96] J. K. Nielsen, T. Lindstr, J. R. Jensen, M. G. Christensen *et al.*, "Fast and statistically efficient fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2016, pp. 86–90.

[97] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[98] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in neural information processing systems*, 2001, pp. 556–562.

[99] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[100] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, “Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. on Speech and Audio processing*, vol. 6, no. 5, pp. 445–455, 1998.

[101] D. Y. Zhao and W. B. Kleijn, “Hmm-based gain modeling for enhancement of speech in noise,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, 2007.

[102] H. Veisi and H. Sameti, “Speech enhancement using hidden markov models in mel-frequency domain,” *Speech Communication*, vol. 55, no. 2, pp. 205–220, 2013.

[103] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook-based Bayesian speech enhancement for nonstationary environments,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[104] ——, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[105] Q. He, F. Bao, and C. Bao, “Multiplicative update of auto-regressive gains for codebook-based speech enhancement,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 457–468, 2017.

[106] F. Itakura, “Analysis synthesis telephony based on the maximum likelihood method,” in *The 6th international congress on acoustics, 1968*, 1968, pp. 280–292.

[107] J. R. Hershey and P. A. Olsen, “Approximating the kullback leibler divergence between gaussian mixture models,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4. IEEE, 2007, pp. IV–317.

[108] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational intelligence and neuroscience*, vol. 2009, 2009.

[109] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[110] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[111] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-euclidean sparse nmf via bilevel optimization with applications to speech enhancement," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 2014, pp. 11–15.

[112] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[113] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2011, pp. 17–20.

[114] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using bayesian nmf with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2012, pp. 4561–4564.

[115] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[116] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. on Audio, Speech, and Language Processing*, 2018.

[117] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[118] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[119] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[120] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.

[121] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[122] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. Wang, “An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4230–4239, 2017.

[123] A. A. Nugraha, A. Liutkus, and E. Vincent, “Multichannel audio source separation with deep neural networks.” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[124] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[125] B. Wu, K. Li, M. Yang, and C.-H. Lee, “A reverberation-time-aware approach to speech dereverberation based on deep neural networks,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[126] E. Lindemann and J. L. Melanson, “Binaural hearing aid,” Dec. 26 1995, uS Patent 5,479,522.

[127] H. Dillon, *Hearing aids*. Hodder Arnold, 2008.

[128] P. M. Zurek, “Binaural advantages and directional effects in speech intelligibility,” *Acoustical factors affecting hearing aid performance*, vol. 2, pp. 255–275, 1993.

[129] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, “Theoretical analysis of binaural multimicrophone noise reduction techniques,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[130] T. V. D. Bogaert, S. Doclo, J. Wouters, and M. Moonen, “Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids,” *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[131] M. Dorbecker and S. Ernst, “Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation,” in *Proc. European Signal Processing Conference*. IEEE, 1996, pp. 1–4.

[132] M. Jeub, C. Nelke, H. Kruger, C. Beaugeant, and P. Vary, “Robust dual-channel noise power spectral density estimation,” in *Proc. European Signal Processing Conference*. IEEE, 2011, pp. 2304–2308.

[133] A. H. Kamkar-Parsi and M. Bouchard, “Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 521–533, 2009.

[134] T. Lotter and P. Vary, “Dual-channel speech enhancement by superdirective beamforming,” *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, p. 063297, 2006.

[135] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, “Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication,” *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[136] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[137] ——, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[138] S. Bech and N. Zacharov, *Perceptual audio evaluation-Theory, method and application*. John Wiley & Sons, 2007.

[139] G. A. Miller, “The masking of speech.” *Psychological bulletin*, vol. 44, no. 2, p. 105, 1947.

[140] I. Recommendation, “1534-1: Method for the subjective assessment of intermediate quality level of coding systems,” *International Telecommunication Union*, 2003.

[141] E. Vincent, “Mushram: A matlab interface for mushra listening tests,” *Online] http://www. elec. qmul. ac. uk/people/emmanuelv/mushram*, 2005.

[142] S. Kochkin, “10-year customer satisfaction trends in the US hearing instrument market,” *Hearing Review*, vol. 9, no. 10, pp. 14–25, 2002.

[143] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[144] C. Elberling, C. Ludvigsen, and P. Lyregaard, "Dantale: a new danish speech material," *Scandinavian Audiology*, vol. 18, no. 3, pp. 169–175, 1989.

[145] "Perceptual evaluation of speech quality, an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation*, p. 862, 2001.

[146] T. H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*. Citeseer, 2008.

[147] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical signal processing and control*, vol. 8, no. 3, pp. 311–314, 2013.

[148] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[149] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2017, pp. 5085–5089.

[150] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2017, pp. 386–390.

[151] J. A. Grange and J. F. Culling, "The benefit of head orientation to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 703–712, 2016.

[152] ——, "Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4061–4072, 2016.

Part II

Papers

Paper A

Model-based Speech Enhancement for Intelligibilty Improvement in Binaural Hearing Aids

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Jesper B. Boldt and Mads Græsbøll Christensen

The paper has been accepted for publication in
IEEE Trans. on Audio, Speech and Language Processing, 2019

Abstract

Speech intelligibility is often severely degraded among hearing impaired individuals in situations such as the cocktail party scenario. The performance of the current hearing aid technology has been observed to be limited in these scenarios. In this paper, we propose a binaural speech enhancement framework that takes into consideration the speech production model. The enhancement framework proposed here is based on the Kalman filter that allows us to take the speech production dynamics into account during the enhancement process. The usage of a Kalman filter requires the estimation of clean speech and noise short term predictor (STP) parameters, and the clean speech pitch parameters. In this work, a binaural codebook-based method is proposed for estimating the STP parameters, and a directional pitch estimator based on the harmonic model and maximum likelihood principle is used to estimate the pitch parameters. The proposed method for estimating the STP and pitch parameters jointly uses the information from left and right ears, leading to a more robust estimation of the filter parameters. Objective measures such as PESQ and STOI have been used to evaluate the enhancement framework in different acoustic scenarios representative of the cocktail party scenario. We have also conducted subjective listening tests on a set of nine normal hearing subjects, to evaluate the performance in terms of intelligibility and quality improvement. The listening tests show that the proposed algorithm, even with access to only a single channel noisy observation, significantly improves the overall speech quality, and the speech intelligibility by up to 15%.

1 Introduction

Normal hearing (NH) individuals have the ability to concentrate on a single speaker even in the presence of multiple interfering speakers. This phenomenon is termed as the cocktail party effect. However, hearing impaired individuals lack this ability to separate out a single speaker in the presence of multiple competing speakers. This leads to listener fatigue and isolation of the hearing aid (HA) user. Mimicking the cocktail party effect in a digital HA is very much desired in such scenarios [1]. Thus, to help the HA user to focus on a particular speaker, speech enhancement has to be performed to reduce the effect of the interfering speakers. The primary objectives of a speech enhancement system in HA are to improve the intelligibility and quality of the degraded speech. Often, a hearing impaired person is fitted with HAs at both ears. Modern HAs have the technology to wirelessly communicate with each other making it possible to share information between the HAs. Such a property in HAs enables the use of binaural speech enhancement algorithms. The binaural processing of noisy signals has shown to be more effective than processing the noisy signal independently at each ear due to the utilization of spatial information [2]. Apart from a better noise reduction performance, binaural algorithms make it possible to preserve the binaural cues which contribute to spatial release from masking [3]. Often, HAs are fitted with multiple microphones at both ears. Some binaural speech enhancement algorithms developed for such cases are [4, 5]. In [4], a multichannel Wiener filter for HA applications is proposed which results in a minimum mean squared error (MMSE) estimation of the target speech. These methods were shown to distort the binaural cues of the interfering noise while maintaining the binaural cues of the target. Consequently, a method was proposed in [6] that introduced a parameter to trade off between the noise reduction and cue preservation. The above mentioned algorithms have reported improvements in speech intelligibility.

We are here mainly concerned with the binaural enhancement of speech with access to only one microphone per HA [7–9]. More specifically, this paper is concerned with a two-input two-output system. This situation is encountered in in-the-ear (ITE) HAs, where the space constraints limit the number of microphones per HA. Moreover, in the case where we have multiple microphones per HA, beamforming can be applied individually on each HA to form the two inputs, which can then be processed further by the proposed dual channel enhancement framework. One of the first approaches to perform dual channel speech enhancement was that of [7] where a two channel spectral subtraction was combined with an adaptive Wiener post-filter. This led to a distortion of the binaural cues, as different gains were applied to the left and right channels. Another approach to performing dual channel speech enhancement was proposed in [8] and this solution consisted of two stages. The first stage dealt with the estimation of interference signals using an equalisation-cancellation theory, and the second stage was an adaptive Wiener filter. The intelligibility improvements corresponding to the algorithms stated above have not been studied well. These algorithms perform the enhancement in the frequency domain by assuming that the speech and noise components are uncorrelated, and do not take into account the nature of the speech production process. In this paper, we propose a binaural speech enhancement framework that takes the speech production model into account. The model used here is based on the source-filter model, where the filter corresponds to the vocal tract and the source corresponds to the excitation signal produced by the vocal chords. Using a physically meaningful model gives us a sufficiently accurate way for explaining how the signals were generated, but also helps in reducing the number of parameters to be estimated. One way to exploit this speech production model for the enhancement process is to use a Kalman filter, as the speech production dynamics can be modelled within the Kalman filter using the state space equations while also accounting for the background noise. Kalman filtering for single channel speech enhancement in the presence of white background noise was first proposed in [10]. This work was later extended to deal with coloured noise in [11, 12].

One of the main limitations of Kalman filtering based enhancement is that the state space parameters required for the formulation of the state space equations need to be known or estimated. The estimation of the state space parameters is a difficult problem due to the non-stationary nature of speech and the presence of noise. The state space parameters are the autoregressive (AR) coefficients and the excitation variances for the speech and noise respectively. Henceforth, AR coefficients along with the excitation variances will be denoted as the short term predictor (STP) parameters. In [11, 12] these STP parameters were estimated using an approximated expectation-maximisation algorithm. However, the performance of these algorithms were noted to be unsatisfactory in non-stationary noise environments. Moreover, these algorithms assumed the excitation signal in the source-filter model to be white Gaussian noise. Even though this assumption is appropriate for modelling unvoiced speech, it is not very suitable for modelling voiced speech. This issue was handled in [13] by using a modified model for the excitation signal capable of modelling both voiced and unvoiced speech. The usage of this model for the enhancement process required the estimation of the pitch parameters in addition to the STP parameters. This modification of the excitation signal was found to improve the performance in voiced speech regions, but the performance of the algorithm in the presence of non-stationary background noise was still observed to be unsatisfactory. This was primarily due to the poor estimation of the model parameters in non-stationary background noise. The noise STP parameters were estimated in [13] by assuming that the first 100 milli seconds of the speech segment contained only noise and the parameters were then assumed to be constant.

In this work, we introduce a binaural model-based speech enhancement framework which addresses the poor estimation of the parameters explained above. We here propose a binaural codebook-based method for estimating the STP parameters, and a directional pitch estimator based on the harmonic model for estimating the pitch parameters. The estimated parameters are subsequently used in a binaural speech enhancement framework that is based on the signal model used in [13]. Codebook-based approaches for estimating STP parameters in the single channel case have been previously proposed in [14], and has been used to estimate the filter parameters required for the Kalman filter for single channel speech enhancement in [15]. In this work we extend this to the dual channel case, where we assume that there is a wireless link between the HAs. The estimation of STP and pitch parameters using the information on both the left and right channels leads to a more robust estimation of these parameters. Thus, in this work, we propose a binaural speech enhancement method that is model-based in several ways as 1) the state space equations involved in the Kalman filter takes into account the dynamics of the speech production model; 2) the estimation of STP parameters utilised in the Kalman filter is based on trained spectral models of speech and noise; and

3) the pitch parameters used within the Kalman filter are estimated based on the harmonic model which is a good model for voiced speech. We remark that this paper is an extension of previous conference papers [16, 17]. In comparison to [16, 17], we have used an improved method for estimating the excitation variances. Moreover, the proposed enhancement framework has been evaluated in more realistic scenarios and subjective listening tests have been conducted to validate the results obtained using objective measures.

2 Problem formulation

In this section, we formulate the problem and state the assumptions that have been used in this work. The noisy signals at the left/right ears at time index n are denoted by

$$z_{l/r}(n) = s_{l/r}(n) + w_{l/r}(n) \qquad \forall n = 0, 1, 2 \ldots, \tag{A.1}$$

where $z_{l/r}$, $s_{l/r}$ and $w_{l/r}$ denote the noisy, clean and noise components at the left/right ears, respectively. It is assumed that the clean speech component is statistically independent with the noise component. Our objective here is to obtain estimates of the clean speech signals denoted as $\hat{s}_{l/r}(n)$, from the noisy signals. The processing of the noisy speech using a speech enhancement system to estimate the clean speech signal requires the knowledge of the speech and noise statistics. To obtain this, it is convenient to assume a statsitical model for the speech and noise components, making it easier to estimate the statistics from the noisy signal. In this work, we model the clean speech as an AR process, which is a common model used to represent the speech production process [18].

We also assume that the speech source is in the nose direction of the listener, so that the clean speech component at the left and right ears can be represented by AR processes having the same parameters,

$$s_{l/r}(n) = \sum_{i=1}^{P} a_i s_{l/r}(n-i) + u(n), \tag{A.2}$$

where $\mathbf{a} = [-a_1, \ldots, -a_P]^T$ is the set of speech AR coefficients, P is the order of the speech AR process and $u(n)$ is the excitation signal corresponding to the speech signal. Often, $u(n)$ is modelled as white Gaussian noise with variance σ_u^2 and this will be referred to as the unvoiced (UV) model [11]. It should be noted that we do not model the reverberation here. Similar to the speech, the noise components are represented by AR processes as,

$$w_{l/r}(n) = \sum_{i=1}^{Q} c_i w_{l/r}(n-i) + v(n), \tag{A.3}$$

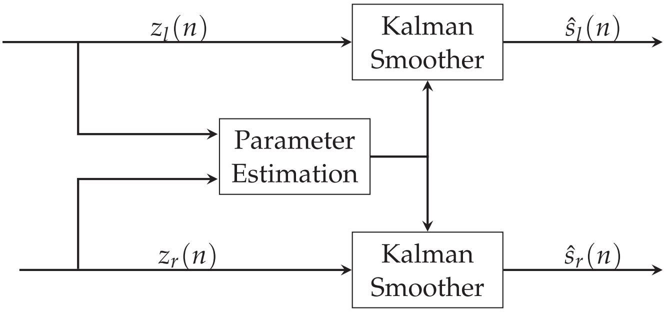


Fig. A.1: Basic block diagram of the binaural enhancement framework.

where $\mathbf{c} = [-c_1, \ldots, -c_Q]^T$ is the set of noise AR coefficients, Q is the order of the noise AR process and $v(n)$ is white Gaussian noise with variance σ_v^2.

As we have seen previously, the excitation signal, $u(n)$, in (A.2) was modelled as a white Gaussian noise. Although this assumption is suitable for representing unvoiced speech, it is not appropriate for modelling voiced speech. Thus, inspired by [13], the enhancement framework here models $u(n)$ as

$$u(n) = b(p)u(n-p) + d(n), \qquad \text{(A.4)}$$

where $d(n)$ is white Gaussian noise with variance σ_d^2, p is the pitch period and $b(p) \in (0,1)$ is the degree of voicing. In portions containing predominantly voiced speech, $b(p)$ is assumed to be close to 1 and the variance of $d(n)$ is assumed to be small, whereas in portions of unvoiced speech, $b(p)$ is assumed to be close to zero so that (A.2) simplifies into the conventional unvoiced AR model. The excitation model in (A.4) when used together with (A.2) is referred to as the voiced-unvoiced (V-UV) model. This model can be easily incorporated into the speech enhancement framework by modifying the state space equations. The incorporation of the V-UV model into the enhancement framework requires the pitch parameters, p and $b(p)$, in addition to the STP parameters to be estimated from the noisy signal. We would like to remark here that these parameters are usually time varying in the case of speech and noise signals. Herein, these parameters are assumed to be quasi-stationary, and are estimated for every frame index $f_n = \lfloor \frac{n}{M} \rfloor + 1$, where M is the frame length. The estimation of these parameters will be explained in the subsequent section.

3 Proposed Enhancement framework

3.1 Overview

The enhancement framework proposed here assumes that there is a communication link between the two HAs that makes it possible to exchange information. Fig. A.1 shows the basic block diagram of the proposed enhancement framework. The noisy signals at the left and right ears are enhanced using a fixed lag Kalman smoother (FLKS), which requires the estimation of STP and pitch parameters. These parameters are estimated jointly using the information in the left and right channels. The usage of identical filter parameters at both the ears leads to the preservation of binaural cues. In this paper, the details regarding the proposed binaural framework will be explained and the performance of the binaural framework will be compared with that of the bilateral framework, where it is assumed that there is no communication link between the two HAs which leads to the filter parameters being estimated independently at each ear. We will now explain the different components of the proposed enhancement framework in detail.

3.2 FLKS for speech enhancement

As alluded to in the introduction, a Kalman filter allows us to take into account the speech production dynamics in the form of state space equations while also accounting for the observation noise. In this work, we use FLKS which is a variant of the Kalman filter. A FLKS gives a better performance than a Kalman filter, but has a higher delay. In this section, we will explain the functioning of FLKS for both the UV and V-UV models that we have introduced in Section 2. We assume here that the model parameters are known. For the UV model, the usage of a FLKS (with a smoother delay of $d_s \geq P$) from a speech enhancement perspective requires the AR signal model in (A.2) to be written as a state space form as shown below

$$\bar{\mathbf{s}}_{l/r}(n) = \mathbf{A}(f_n)\bar{\mathbf{s}}_{l/r}(n-1) + \mathbf{\Gamma}_1 u(n), \tag{A.5}$$

where $\bar{\mathbf{s}}_{l/r}(n) = [s_{l/r}(n), s_{l/r}(n-1), \ldots, s_{l/r}(n-d_s)]^T$ is the state vector containing the $d_s + 1$ recent speech samples, $\mathbf{\Gamma}_1 = [1, 0, \ldots, 0]^T$ is a $(d_s+1) \times 1$ vector, $u(n) = d(n)$ and $\mathbf{A}(f_n)$ is the $(d_s+1) \times (d_s+1)$ speech state transition matrix written as

$$\mathbf{A}(f_n) = \begin{bmatrix} -\mathbf{a}(f_n)^T & \mathbf{0}^T & 0 \\ \mathbf{I}_P & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d_s-P} & \mathbf{0} \end{bmatrix}. \tag{A.6}$$

The state space equation for the noise signal in (B.3) is similarly written as

$$\bar{\mathbf{w}}_{l/r}(n) = \mathbf{C}(f_n)\bar{\mathbf{w}}_{l/r}(n-1) + \mathbf{\Gamma}_2 v(n), \tag{A.7}$$

where $\bar{\mathbf{w}}_{l/r}(n) = [w_{l/r}(n), w_{l/r}(n-1), \ldots, w_{l/r}(n-Q+1)]^T$, $\mathbf{\Gamma}_2 = [1, 0, \ldots, 0]^T$ is a $Q \times 1$ vector and

$$\mathbf{C}(f_n) = \begin{bmatrix} [c_1(f_n), \ldots, c_{Q-1}(f_n)] & c_Q(f_n) \\ \mathbf{I}_{Q-1} & \mathbf{0} \end{bmatrix} \tag{A.8}$$

is a $Q \times Q$ matrix. The state space equations in (A.5) and (A.7) are combined to form a concatenated state space equation for the UV model as

$$\begin{bmatrix} \bar{\mathbf{s}}_{l/r}(n) \\ \bar{\mathbf{w}}_{l/r}(n) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(f_n) & \mathbf{0} \\ \mathbf{0} & \mathbf{C}(f_n) \end{bmatrix} \begin{bmatrix} \bar{\mathbf{s}}_{l/r}(n-1) \\ \bar{\mathbf{w}}_{l/r}(n-1) \end{bmatrix} + \begin{bmatrix} \mathbf{\Gamma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2 \end{bmatrix} \begin{bmatrix} d(n) \\ v(n) \end{bmatrix}$$

which can be rewritten as

$$\bar{\mathbf{x}}^{\text{UV}}_{l/r}(n) \triangleq \mathbf{F}^{\text{UV}}(f_n)\mathbf{x}(n-1) + \mathbf{\Gamma}_3\mathbf{y}(n), \tag{A.9}$$

where $\bar{\mathbf{x}}^{\text{UV}}_{l/r}(n) = \left[\bar{\mathbf{s}}_{l/r}(n)^T \; \bar{\mathbf{w}}_{l/r}(n)^T\right]^T$ is the concatenated state space vector and $\mathbf{F}^{\text{UV}}(f_n)$ is the concatenated state transition matrix for the UV model. The observation equation to obtain the noisy signal is then written as

$$z_{l/r}(n) = \mathbf{\Gamma}^{\text{UV}^T}\bar{\mathbf{x}}^{\text{UV}}_{l/r}(n), \tag{A.10}$$

where $\mathbf{\Gamma}^{\text{UV}} = \left[\mathbf{\Gamma}_1^T \, \mathbf{\Gamma}_2^T\right]^T$. The state space equation (A.9) and the observation equation (A.10) can then be used to formulate the prediction and correction stages of the FLKS for the UV model. We will now explain the formulation of the state space equations for the V-UV model. The state space equation for the V-UV model of speech is written as

$$\bar{\mathbf{s}}_{l/r}(n) = \mathbf{A}(f_n)\bar{\mathbf{s}}_{l/r}(n-1) + \mathbf{\Gamma}_1 u(n), \tag{A.11}$$

where the excitation signal in (A.4) is also modelled as a state space equation as

$$\bar{\mathbf{u}}(n) = \mathbf{B}(f_n)\bar{\mathbf{u}}(n-1) + \mathbf{\Gamma}_4 d(n), \tag{A.12}$$

where $\bar{\mathbf{u}}(n) = [u(n), u(n-1), \ldots, u(n-p_{\max}+1)]^T$, $p_{\max}$ is the maximum pitch period in integer samples, $\mathbf{\Gamma}_4 = [1, 0 \ldots 0]^T$ is a $(p_{\max}) \times 1$ vector and

$$\mathbf{B}(f_n) = \begin{bmatrix} [b(1), \ldots, b(p_{\max}-1)] & b(p_{\max}) \\ \mathbf{I}_{p_{\max}-1} & \mathbf{0} \end{bmatrix} \tag{A.13}$$

is a $p_{\max} \times p_{\max}$ matrix where $b(i) = 0 \; \forall i \neq p(f_n)$. The concatenated state space equation for the V-UV model is

$$\begin{bmatrix} \bar{\mathbf{s}}_{l/r}(n) \\ \mathbf{u}(n+1) \\ \bar{\mathbf{w}}_{l/r}(n) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(f_n) & \mathbf{\Gamma}_1\mathbf{\Gamma}_2^T & \mathbf{0} \\ \mathbf{0} & \mathbf{B}(f_n) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}(f_n) \end{bmatrix} \begin{bmatrix} \bar{\mathbf{s}}_{l/r}(n-1) \\ \bar{\mathbf{u}}(n) \\ \bar{\mathbf{w}}_{l/r}(n-1) \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{\Gamma}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_2 \end{bmatrix} \begin{bmatrix} d(n+1) \\ v(n) \end{bmatrix},$$

which can also be written as

$$\bar{\mathbf{x}}_{l/r}^{\text{V-UV}}(n+1) \triangleq \mathbf{F}^{\text{V-UV}}(f_n)\bar{\mathbf{x}}_{l/r}^{\text{V-UV}}(n) + \boldsymbol{\Gamma}_5\mathbf{g}(n+1), \tag{A.14}$$

where $\bar{\mathbf{x}}_{l/r}^{\text{V-UV}}(n+1) = [\bar{\mathbf{s}}_{l/r}(n)^T\, \bar{\mathbf{u}}(n+1)^T\, \bar{\mathbf{w}}_{l/r}(n)^T]^T$ is the concatenated state space vector, $\mathbf{g}(n+1) = [d(n+1)\, v(n)]^T$ and $\mathbf{F}^{\text{V-UV}}(f_n)$ is the concatenated state transition matrix for the V-UV model. The observation equation to obtain the noisy signal is written as

$$z_{l/r}(n) = \boldsymbol{\Gamma}^{\text{V-UV}^T}\bar{\mathbf{x}}_{l/r}^{\text{V-UV}}(n+1), \tag{A.15}$$

where $\boldsymbol{\Gamma}^{\text{V-UV}} = \left[\boldsymbol{\Gamma}_1^T\, \mathbf{0}^T\, \boldsymbol{\Gamma}_2^T\right]^T$. The state space equation (A.14) and the observation equation (A.15) can then be used to formulate the prediction and correction stages of the FLKS for the V-UV model (see Appendix 7.1). It can be seen that the formulation of the prediction and correction stages of the FLKS requires the knowledge of the speech and noise STP parameters, and the clean speech pitch parameters. The estimation of these model parameters are explained in the subsequent sections.

3.3 Codebook-based binaural estimation of STP parameters

As mentioned in the introduction, the estimation of the speech and noise STP parameters forms a very critical part of the proposed enhancement framework. These parameters are here estimated using a codebook-based approach. The estimation of STP parameters using a codebook-based approach, when having access to a single channel noisy signal has been previously proposed in [14, 19]. Here, we extend this to the case when we have access to binaural noisy signals. Codebook-based estimation of STP parameters uses the a priori information about speech and noise spectral shapes stored in trained speech and noise codebooks in the form of speech and noise AR coefficients, respectively. The codebooks offer us an elegant way of including prior information about the speech and noise spectral models e.g. if the enhancement system present in the HA has to operate in a particular noisy environment, or mainly process speech from a particular set of speakers, the codebooks can be trained accordingly. Contrarily, if we do not have any specific information regarding the speaker or the noisy environment, we can still train general codebooks from a large database consisting of different speakers and noise types. We would like to remark here that we assume the UV model of speech for the estimation of STP parameters.

A Bayesian framework is utilised to estimate the parameters for every frame index. Thus, the random variables (r.v.) corresponding to the parameters to be estimated for the f_n^{th} frame are concatenated to form a single vector $\boldsymbol{\theta}(f_n) = [\boldsymbol{\theta}_s(f_n)^T\, \boldsymbol{\theta}_w(f_n)^T]^T = [\mathbf{a}(f_n)^T\, \sigma_d^2(f_n)\, \mathbf{c}(f_n)^T\, \sigma_v^2(f_n)]^T$, where $\mathbf{a}(f_n)$ and $\mathbf{c}(f_n)$ are r.v. representing the speech and noise AR coefficients, and $\sigma_d^2(f_n)$ and $\sigma_v^2(f_n)$ are r.v. representing the speech and noise excitation variances. The MMSE estimate of the parameter vector is

$$\hat{\boldsymbol{\theta}}(f_n) = \mathbb{E}(\boldsymbol{\theta}(f_n)|\mathbf{z}_l(f_nM), \mathbf{z}_r(f_nM)), \tag{A.16}$$

where $\mathbb{E}(\cdot)$ is the expectation operator and $\mathbf{z}_{l/r}(f_nM) = [z_{l/r}(f_nM), \ldots, z_{l/r}(f_nM+m), \ldots, z_{l/r}(f_nM+M-1)]^T$ denotes the f_n^{th} frame of noisy speech at the left/right ears. The frame index, f_n, will be left out for the remainder of the section for notational convenience. Equation (A.16) is then rewritten as

$$\hat{\boldsymbol{\theta}} = \int_{\Theta} \boldsymbol{\theta} \frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta})\ p(\boldsymbol{\theta})}{p(\mathbf{z}_l, \mathbf{z}_r)} d\boldsymbol{\theta}, \tag{A.17}$$

where Θ denotes the combined support space of the parameters to be estimated. Since we assumed that the speech and noise are independent (see Section 2), it follows that $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_s)p(\boldsymbol{\theta}_w)$ where $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_w$ speech and noise STP parameters respectively. Furthermore, the speech and noise AR coefficients are assumed to be independent with the excitation variances leading to $p(\boldsymbol{\theta}_s) = p(\mathbf{a})p(\sigma_d^2)$ and $p(\boldsymbol{\theta}_w) = p(\mathbf{c})p(\sigma_v^2)$. Using the aforementioned assumptions, (A.17) is rewritten as

$$\hat{\boldsymbol{\theta}} = \int_{\Theta} \boldsymbol{\theta} \frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta})\ p(\mathbf{a})p(\sigma_d^2)p(\mathbf{c})p(\sigma_v^2)}{p(\mathbf{z}_l, \mathbf{z}_r)} d\boldsymbol{\theta}. \tag{A.18}$$

The probability density of the AR coefficients is here modelled as a sum of Dirac delta functions centered around each codebook entry as $p(\mathbf{a}) = \frac{1}{N_s}\sum_{i=1}^{N_s}\delta(\mathbf{a}-\mathbf{a}_i)$ and $p(\mathbf{c}) = \frac{1}{N_w}\sum_{j=1}^{N_w}\delta(\mathbf{c}-\mathbf{c}_j)$, where $\mathbf{a}_i$ is the i^{th} entry of the speech codebook (of size N_s), $\mathbf{c}_j$ is the j^{th} entry of the noise codebook (of size N_w) . Defining $\boldsymbol{\theta}_{ij} \triangleq [\mathbf{a}_i^T\ \sigma_d^2\ \mathbf{c}_j^T\ \sigma_v^2]^T$, (A.18) can be rewritten as

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_sN_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w}\int_{\sigma_d^2}\int_{\sigma_v^2}\boldsymbol{\theta}_{ij}\frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij})\ p(\sigma_d^2)p(\sigma_v^2)}{p(\mathbf{z}_l, \mathbf{z}_r)}d\sigma_d^2 d\sigma_v^2. \tag{A.19}$$

For a particular set of speech and noise AR coefficients, $\mathbf{a}_i$ and $\mathbf{c}_j$, it can be shown that the likelihood, $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij})$, decays rapidly from its maximum value when there is a small deviation in the excitation variances from its true value [14] (see Appendix 7.2). If we then approximate the true values of the excitation variances with the corresponding maximum likelihood (ML) estimates denoted as $\sigma_{d,ij}^2$ and $\sigma_{v,ij}^2$, the likelihood term $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij})$ can be approximated as $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij})\delta(\sigma_d^2-\sigma_{d,ij}^2)\delta(\sigma_v^2-\sigma_{v,ij}^2)$. Defining $\boldsymbol{\theta}_{ij}^{\text{ML}} \triangleq [\mathbf{a}_i^T\ \sigma_{d,ij}^2\ \mathbf{c}_j^T\ \sigma_{v,ij}^2]^T$, and using the above approximation and the property, $\int_x f(x)\delta(x-x_0)dx = f(x_0)$, we can rewrite (A.19) as

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_sN_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w}\boldsymbol{\theta}_{ij}^{\text{ML}}\frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}})p(\sigma_{d,ij}^2)p(\sigma_{v,ij}^2)}{p(\mathbf{z}_l, \mathbf{z}_r)}, \tag{A.20}$$

where

$$p(\mathbf{z}_l, \mathbf{z}_r) = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}_{ij}^{\mathrm{ML}}) p(\sigma_{d,ij}^2) p(\sigma_{v,ij}^2).$$

Details regarding the prior distributions used for the excitation variances is given in Appendix 7.3. It can be seen from (A.20) that the final estimate of the parameter vector is a weighted linear combination of $\boldsymbol{\theta}_{ij}^{\mathrm{ML}}$ with weights proportional to $p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}_{ij}^{\mathrm{ML}}) p(\sigma_{d,ij}^2) p(\sigma_{v,ij}^2)$. To compute this, we need to first obtain the ML estimates of the excitation variances for a given set of speech and noise AR coefficients, $\mathbf{a}_i$ and $\mathbf{c}_j$, as

$$\{\sigma_{d,ij}^2, \sigma_{v,ij}^2\} = \underset{\sigma_d^2, \sigma_v^2 \geq 0}{\arg\max} \quad p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}_{ij}). \tag{A.21}$$

For the models we have assumed previously in Section 2, we can show that $\mathbf{z}_l$ and $\mathbf{z}_r$ are statistically independent given $\boldsymbol{\theta}_{ij}$ [20, Sec 8.2.2], which results in

$$p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}_{ij}) = p(\mathbf{z}_l | \boldsymbol{\theta}_{ij}) p(\mathbf{z}_r | \boldsymbol{\theta}_{ij}).$$

We first derive the likelihood for the left channel, $p(\mathbf{z}_l | \boldsymbol{\theta}_{ij})$, using the assumptions we have introduced previously in Section 2. Using these assumptions, frame of speech and noise component associated with the noisy frame $\mathbf{z}_l$ denoted by $\mathbf{s}_l$ and $\mathbf{w}_l$ respectively can be expressed as

$$\begin{aligned} p(\mathbf{s}_l | \sigma_d^2, \mathbf{a}_i) &\sim \mathcal{N}(\mathbf{0}, \sigma_d^2 \mathbf{R}_s(\mathbf{a}_i)) \\ p(\mathbf{w}_l | \sigma_v^2, \mathbf{c}_j) &\sim \mathcal{N}(\mathbf{0}, \sigma_v^2 \mathbf{R}_w(\mathbf{c}_j)), \end{aligned}$$

where $\mathbf{R}_s(\mathbf{a}_i)$ is the normalised speech covariance matrix and $\mathbf{R}_w(\mathbf{c}_j)$ is the normalised noise covariance matrix. These matrices can be asymptotically approximated as circulant matrices which can be diagonalised using the Fourier transform as [14, 21],

$$\mathbf{R}_s(\mathbf{a}_i) = \mathbf{F} \mathbf{D}_{s_i} \mathbf{F}^H \text{ and } \mathbf{R}_w(\mathbf{c}_j) = \mathbf{F} \mathbf{D}_{w_j} \mathbf{F}^H,$$

where $\mathbf{F}$ is the discrete Fourier transform (DFT) matrix defined as $[\mathbf{F}]_{m,k} = \frac{1}{\sqrt{M}} \exp(\frac{\imath 2\pi mk}{M}), \ \forall m, k = 0, \ldots M-1$ where k represents the frequency index and

$$\mathbf{D}_{s_i} = (\boldsymbol{\Lambda}_{s_i}^H \boldsymbol{\Lambda}_{s_i})^{-1}, \quad \boldsymbol{\Lambda}_{s_i} = \operatorname{diag}\left(\sqrt{M} \mathbf{F}^H \begin{bmatrix} 1 \\ \mathbf{a}_i \\ \mathbf{0} \end{bmatrix} \right),$$

$$\mathbf{D}_{w_j} = (\boldsymbol{\Lambda}_{w_j}^H \boldsymbol{\Lambda}_{w_j})^{-1}, \quad \boldsymbol{\Lambda}_{w_j} = \operatorname{diag}\left(\sqrt{M} \mathbf{F}^H \begin{bmatrix} 1 \\ \mathbf{c}_j \\ \mathbf{0} \end{bmatrix} \right).$$

Thus we obtain the likelihood for the left channel as,

$$p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}) \sim \mathcal{N}(\mathbf{0}, \sigma_d^2\mathbf{F}\mathbf{D}_{s_i}\mathbf{F}^H + \sigma_v^2\mathbf{F}\mathbf{D}_{w_j}\mathbf{F}^H).$$

The log-likelihood $\ln p(\mathbf{z}_l|\boldsymbol{\theta}_{ij})$ is then given by

$$\begin{aligned}\ln p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}) \overset{c}{=}& \ln\left|\sigma_d^2\mathbf{F}\mathbf{D}_{s_i}\mathbf{F}^H + \sigma_v^2\mathbf{F}\mathbf{D}_{w_j}\mathbf{F}^H\right|^{-\frac{1}{2}} \\ &-\frac{1}{2}\mathbf{z}_l^T\left[\sigma_d^2\mathbf{F}\mathbf{D}_{s_i}\mathbf{F}^H + \sigma_v^2\mathbf{F}\mathbf{D}_{w_j}\mathbf{F}^H\right]^{-1}\mathbf{z}_l,\end{aligned} \tag{A.22}$$

where $\overset{c}{=}$ denotes equality up to a constant and $|\cdot|$ denotes the matrix determinant operator. Denoting $\frac{1}{A_s^i(k)}$ as the k^{th} diagonal element of $\mathbf{D}_{s_i}$ and $\frac{1}{A_w^i(k)}$ as the k^{th} diagonal element of $\mathbf{D}_{w_j}$, (A.22) can be rewritten as

$$\begin{aligned}\ln p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}) \overset{c}{=}& \ln\prod_{k=0}^{K-1}\left(\frac{\sigma_d^2}{A_s^i(k)} + \frac{\sigma_v^2}{A_w^j(k)}\right)^{-\frac{1}{2}} \\ -\frac{1}{2}\mathbf{z}_l^T\mathbf{F}&\begin{bmatrix}\frac{\sigma_d^2}{A_s^i(0)} + \frac{\sigma_v^2}{A_w^j(0)} & \mathbf{0} & 0 \\ \mathbf{0} & \ddots & \mathbf{0} \\ 0 & \mathbf{0} & \frac{\sigma_d^2}{A_s^i(K-1)} + \frac{\sigma_v^2}{A_w^j(K-1)}\end{bmatrix}^{-1}\mathbf{F}^H\mathbf{z}_l.\end{aligned} \tag{A.23}$$

Defining the modelled spectrum as $\hat{P}_{z_{ij}}(k) \triangleq \frac{\sigma_d^2}{A_s^i(k)} + \frac{\sigma_v^2}{A_w^j(k)}$, (A.23) can be written as

$$\ln p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}) \overset{c}{=} \ln\prod_{k=0}^{K-1}\left(\hat{P}_{z_{ij}}(k)\right)^{-\frac{1}{2}} - \frac{1}{2}\sum_{k=0}^{K-1}\frac{P_{z_l}(k)}{\hat{P}_{z_{ij}}(k)}, \tag{A.24}$$

where $P_{z_l}(k)$ is the squared magnitude of the k^{th} element of the vector $\mathbf{F}^H\mathbf{z}_l$. Thus,

$$\ln p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}) \overset{c}{=} -\frac{1}{2}\sum_{k=0}^{K-1}\left(\frac{P_{z_l}(k)}{\hat{P}_{z_{ij}}(k)} + \ln\hat{P}_{z_{ij}}(k)\right). \tag{A.25}$$

We can then see that the log-likelihood is equal, up to a constant, to the Itakura-Saito (IS) divergence between $\mathbf{P}_{z_l}$ and $\hat{\mathbf{P}}_{z_{ij}}$ which is defined as [22]

$$d_{\text{IS}}(\mathbf{P}_{z_l}, \hat{\mathbf{P}}_{z_{ij}}) = \frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{P_{z_l}(k)}{\hat{P}_{z_{ij}}(k)} - \ln\frac{P_{z_l}(k)}{\hat{P}_{z_{ij}}(k)} - 1\right),$$

where $\mathbf{P}_{z_l} = \left[P_{z_l}(0), \ldots, P_{z_l}(K-1)\right]^T$ and $\hat{\mathbf{P}}_{z_{ij}} = \left[\hat{P}_{z_{ij}}(0), \ldots, \hat{P}_{z_{ij}}(K-1)\right]^T$. Using the same result for the right ear, the optimisation problem in (A.21), under the aforementioned conditions can be equivalently written as

$$\{\sigma^2_{d,ij}, \sigma^2_{v,ij}\} = \underset{\sigma^2_d, \sigma^2_v \geq 0}{\arg\min} \left[d_{\text{IS}}(\mathbf{P}_{z_l}, \hat{\mathbf{P}}_{z_{ij}}) + d_{\text{IS}}(\mathbf{P}_{z_r}, \hat{\mathbf{P}}_{z_{ij}}) \right]. \tag{A.26}$$

Unfortunately, it is not possible to get a closed form expression for the excitation variances by minimising (A.26). Instead, this is solved iteratively using the multiplicative update (MU) method [23]. For notational convenience, $\hat{\mathbf{P}}_{z_{ij}}$ can be written as $\hat{\mathbf{P}}_{z_{ij}} = \mathbf{P}_{s,i}\sigma^2_d + \mathbf{P}_{w,j}\sigma^2_v$, where

$$\mathbf{P}_{s,i} = \left[\frac{1}{A^i_s(0)}, \ldots, \frac{1}{A^i_s(K-1)} \right]^T, \quad \mathbf{P}_{w,j} = \left[\frac{1}{A^j_w(0)}, \ldots, \frac{1}{A^j_w(K-1)} \right]^T.$$

Defining $\mathbf{P}_{ij} = \left[\mathbf{P}_{s,i} \;\; \mathbf{P}_{w,j} \right]$, and $\mathbf{\Sigma}^{(l)}_{ij} = [\sigma^{2(l)}_{d,ij} \;\; \sigma^{2(l)}_{v,ij}]^T$ where $\sigma^{2(l)}_{d,ij}$ and $\sigma^{2(l)}_{v,ij}$ represents the ML estimates of the excitation variances at the l^{th} MU iteration, the values for the excitation variances using the MU method are computed iteratively as [24],

$$\sigma^{2(l+1)}_{d,ij} \leftarrow \sigma^{2(l)}_{d,ij} \frac{\mathbf{P}^T_{s,i} \left[(\mathbf{P}_{ij}\mathbf{\Sigma}^{(l)}_{ij})^{-2} \cdot (\mathbf{P}_{z_l} + \mathbf{P}_{z_r}) \right]}{2\mathbf{P}^T_{s,i}(\mathbf{P}_{ij}\mathbf{\Sigma}^{(l)}_{ij})^{-1}}, \tag{A.27}$$

$$\sigma^{2(l+1)}_{v,ij} \leftarrow \sigma^{2(l)}_{v,ij} \frac{\mathbf{P}^T_{w,j} \left[(\mathbf{P}_{ij}\mathbf{\Sigma}^{(l)}_{ij})^{-2} \cdot (\mathbf{P}_{z_l} + \mathbf{P}_{z_r}) \right]}{2\mathbf{P}^T_{w,j}(\mathbf{P}_{ij}\mathbf{\Sigma}^{(l)}_{ij})^{-1}}, \tag{A.28}$$

where $(\cdot)$ denotes the element wise multiplication operator and $(\cdot)^{-2}$ denotes element-wise inverse squared operator. The excitation variances estimated using (A.27) and (A.28) lead to the minimisation of the cost function in (A.26). Using these results, $p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}^{\text{ML}}_{ij})$ can be written as

$$p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}^{\text{ML}}_{ij}) = Ce^{\left(-\frac{M}{2} \left[d_{\text{IS}}(\mathbf{P}_{z_l}, \hat{\mathbf{P}}^{\text{ML}}_{z_{ij}}) + d_{\text{IS}}(\mathbf{P}_{z_r}, \hat{\mathbf{P}}^{\text{ML}}_{z_{ij}}) \right] \right)}, \tag{A.29}$$

where C is a normalisation constant, and $\hat{\mathbf{P}}^{\text{ML}}_{z_{ij}} = [\hat{P}^{\text{ML}}_{z_{ij}}(0), \ldots, \hat{P}^{\text{ML}}_{z_{ij}}(K-1)]^T$ and

$$\hat{P}^{\text{ML}}_{z_{ij}}(k) = \frac{\sigma^2_{d,ij}}{A^i_s(k)} + \frac{\sigma^2_{v,ij}}{A^j_w(k)}. \tag{A.30}$$

Once the likelihoods are calculated using (A.29), they are substituted into (A.20) to get the final estimate of the speech and noise STP parameters. Some other practicalities involved in the estimation procedure of the STP parameters are explained next.

Adaptive noise codebook

The noise codebook used for the estimation of the STP parameters is usually generated by using a training sample consisting of the noise type of interest. However, there might be scenarios where the noise type is not known a priori. In such scenarios, to make the enhancement system more robust, the noise codebook can be appended with an entry corresponding to the noise power spectral density (PSD) estimated using another dual channel method. Here, we utilise such a dual channel method for estimating the noise PSD [7], which requires the transmission of noisy signals between the HAs. The estimated dual channel noise PSD, $\hat{P}_w^{\mathrm{DC}}(k)$, is then used to find the AR coefficients and the variance representing the noise spectral envelope. At first, the autocorrelation coefficients corresponding to the noise PSD estimate are computed using the Wiener-Khinchin theorem as

$$r_{ww}(q) = \sum_{k=0}^{K-1} \hat{P}_w^{\mathrm{DC}}(k) \exp\left(\imath 2\pi \frac{qk}{K}\right), \;\; 0 \leq q \leq Q.$$

Subsequently, the AR coefficients denoted by $\hat{\mathbf{c}}^{\mathrm{DC}} = [1, \hat{c}_1^{\mathrm{DC}}, \ldots, \hat{c}_Q^{\mathrm{DC}}]^T$, and the excitation variance corresponding to the dual channel noise PSD estimate are estimated by Levinson-Durbin recursive algorithm [25, p. 100]. The estimated AR coefficient vector, $\hat{\mathbf{c}}^{\mathrm{DC}}$, is then appended to the noise codebook. The final estimate of the noise excitation variance can be taken as a mean of the variance obtained from the dual channel noise PSD estimate and the variance obtained from (A.20). It should be noted that, in the case a noise codebook is not available a priori, the speech codebook can be used in conjunction with dual channel noise PSD estimate alone. This leads to a reduction in the computational complexity [16]. Some other dual channel noise PSD estimation algorithms present in the literature are [26, 27], and these can in principle also be included in the noise codebook.

3.4 Directional pitch estimator

As we have seen previously, the formulation of the state transition matrix in (A.12) requires the estimation of pitch parameters. In this paper, we propose a parametric method to estimate the pitch parameters of clean speech present in noise. The babble noise generally encountered in a cocktail party scenario is spectrally coloured. As the pitch estimator proposed here is optimal only for white Gaussian noise signals, pre-whitening is first performed on the noisy signal to whiten the noise component. Pre-whitening is performed using the estimated noise AR coefficients as

$$\tilde{z}_{l/r}(n) = z_{l/r}(n) + \sum_{i=1}^{Q} \hat{c}_i(f_n) z_{l/r}(n-i). \tag{A.31}$$

The method proposed here operates on signal vectors $\tilde{\mathbf{z}}_{l/r_c}(f_nM) \in \mathbb{C}^M$ defined as $\tilde{\mathbf{z}}_{l/r_c}(f_nM) = [\tilde{z}_{l/r_c}(f_nM), \ldots, \tilde{z}_{l/r_c}(f_nM + M - 1)]^T$ where $\tilde{z}_{l/r_c}(n)$ is the complex signal corresponding to $\tilde{z}_{l/r}(n)$, which is obtained using the Hilbert transform. This method uses the harmonic model to represent the clean speech as a sum of L harmonically related complex sinusoids. Using the harmonic model, the noisy signal at the left ear in vector of Gaussian noise $\tilde{\mathbf{w}}_{l_c}(f_nM)$, with covariance matrix, $\mathbf{Q}_l(f_n)$, is represented as

$$\tilde{\mathbf{z}}_{l_c}(f_nM) = \mathbf{V}(f_n)\mathbf{D}_l\mathbf{q}(f_n) + \tilde{\mathbf{w}}_{l_c}(f_nM) \tag{A.32}$$

where $\mathbf{q}(f_n)$ is a vector of complex amplitudes, $\mathbf{V}(f_n)$ is the Vandermonde matrix defined as $\mathbf{V}(f_n) = [\mathbf{v}_1(f_n) \ldots \mathbf{v}_L(f_n)]$, where $[\mathbf{v}_p(f_n)]_m = e^{\imath\omega_0 p(f_nM+m-1)}$ with ω_0 being the fundamental frequency and $\mathbf{D}_l$ being the directivity matrix from the source to the left ear. The directivity matrix contains a frequency and angle dependent delay and magnitude term along the diagonal, designed using the method in [28, eq. 3]. Similarly, the noisy signal at the right ear is written as

$$\tilde{\mathbf{z}}_{r_c}(f_nM) = \mathbf{V}(f_n)\mathbf{D}_r\mathbf{q}(f_n) + \tilde{\mathbf{w}}_{r_c}(f_nM). \tag{A.33}$$

The frame index f_n will be omitted for the remainder of the section for notational convenience. Assuming independence between the channels, the likelihood, due to Gaussianity can be expressed as

$$p(\tilde{\mathbf{z}}_{l_c}, \tilde{\mathbf{z}}_{r_c}|\boldsymbol{\epsilon}) = \mathcal{CN}(\tilde{\mathbf{z}}_{l_c}; \mathbf{V}\mathbf{D}_l\mathbf{q}, \mathbf{Q}_l)\, \mathcal{CN}(\tilde{\mathbf{z}}_{r_c}; \mathbf{V}\mathbf{D}_r\mathbf{q}, \mathbf{Q}_r) \tag{A.34}$$

where $\boldsymbol{\epsilon}$ is the parameter set containing ω_0, the complex amplitudes, the directivity matrices and the noise covariance matrices. Assuming that the noise is white in both the channels, the likelihood is rewritten as

$$p(\tilde{\mathbf{z}}_{l_c}, \tilde{\mathbf{z}}_{r_c}|\boldsymbol{\epsilon}) = \frac{e^{-\left(\frac{||\tilde{\mathbf{z}}_{l_c} - \mathbf{V}\mathbf{D}_l\mathbf{q}||^2}{\sigma_l^2} + \frac{||\tilde{\mathbf{z}}_{r_c} - \mathbf{V}\mathbf{D}_r\mathbf{q}||^2}{\sigma_r^2}\right)}}{(\pi\sigma_l\sigma_r)^{2M}} \tag{A.35}$$

and the log-likelihood is then

$$\begin{aligned}\ln p(\tilde{\mathbf{z}}_{l_c},\tilde{\mathbf{z}}_{r_c}|\boldsymbol{\epsilon}) &= -M(\ln \pi\sigma_l^2 + \ln \pi\sigma_r^2) \\ &\quad - \left(\frac{||\tilde{\mathbf{z}}_{l_c} - \mathbf{V}\mathbf{D}_l\mathbf{q}||^2}{\sigma_l^2} + \frac{||\tilde{\mathbf{z}}_{r_c} - \mathbf{V}\mathbf{D}_r\mathbf{q}||^2}{\sigma_r^2}\right).\end{aligned} \tag{A.36}$$

Assuming the fundamental frequency to be known, the ML estimate of the amplitudes is obtained as

$$\hat{\mathbf{q}} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{y}, \tag{A.37}$$

where $\mathbf{H} = \left[(\mathbf{V}\mathbf{D}_l)^T \;\; (\mathbf{V}\mathbf{D}_r)^T\right]^T$ and $\mathbf{y} = [\tilde{\mathbf{z}}_{l_c}^T \tilde{\mathbf{z}}_{r_c}^T]^T$. These amplitude estimates are further used to estimate the noise variances as

$$\hat{\sigma}_{l/r}^2 = \frac{1}{M}||\hat{\tilde{\mathbf{w}}}_{l/r_c}||^2 = \frac{1}{M}||\tilde{\mathbf{z}}_{l/r_c} - \mathbf{V}\mathbf{D}_{l/r}\hat{\mathbf{q}}||^2. \tag{A.38}$$

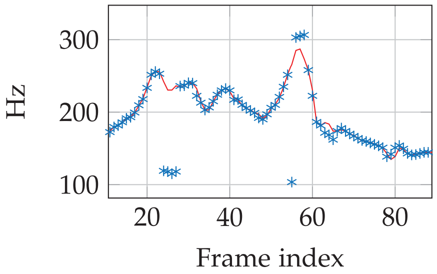


Fig. A.2: Fundamental frequency estimates using the proposed method (SNR = 3 dB). The red line indicates the true fundamental frequency and the blue aterisk denotes the estimated fundamental frequency.

Substituting these into (A.36), we obtain the log-likelihood as

$$\ln p(\tilde{\mathbf{z}}_{l_c}, \tilde{\mathbf{z}}_{r_c} | \boldsymbol{\epsilon}) \overset{c}{=} -M(\ln \hat{\sigma}_l^2 + \ln \hat{\sigma}_r^2). \tag{A.39}$$

The ML estimate of the fundamental frequency is then

$$\hat{\omega}_0 = \underset{\omega_0 \in \Omega_0}{\arg\min} \, (\ln \hat{\sigma}_l^2 + \ln \hat{\sigma}_r^2), \tag{A.40}$$

where Ω_0 is the set of candidate fundamental frequencies. This leads to (A.40) being evaluated on grid of candidate fundamental frequencies. The pitch is then obtained by rounding the reciprocal of the estimated fundamental frequency in Hz. We remark that the model order L is estimated here using the maximum a posteriori (MAP) rule [29, p. 38]. The degree of voicing is calculated by taking the ratio between the energy (calculated as the square of the l^2-norm) present at integer multiples of the fundamental frequency and the total energy present in the signal. This is motivated by the observation that, in case of highly voiced regions, the energy of the signal will be concentrated at the harmonics. Figures A.2 and A.3 show the pitch estimation plot from the binaural noisy signal (SNR = 3 dB) for the proposed method (which uses information from the two channels), and a single channel pitch estimation method which uses only the left channel, respectively. The red line denotes the true fundamental frequency and the blue asterisk denotes the estimated fundamental frequency. It can be seen that the use of the two channels leads to a more robust pitch estimation.

The main steps involved in the proposed enhancement framework for the V-UV model are shown in Algorithm 1. The enhancement framework for the UV model differs from the V-UV model in that it does not require estimation of the pitch parameters, and that the FLKS equations would be derived based on (A.9) and (A.10) instead of (A.14) and (A.15).

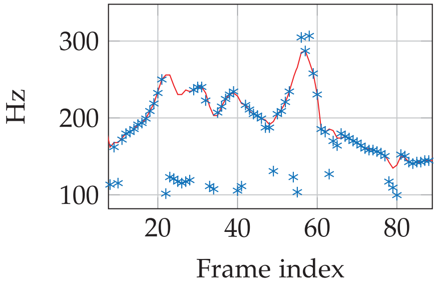


Fig. A.3: Fundamental frequency estimates using the corresponding single channel method [29] (SNR = 3 dB).

4 Simulation Results

In this section, we will present the experiments that have been carried out to evaluate the proposed enhancement framework.

4.1 Implementation details

The test audio files used for the experiments consisted of speech from the GRID database [30] re-sampled to 8 kHz. The noisy signals were generated using the simulation set-up explained in Section 4.2. The speech and noise STP parameters required for the enhancement process were estimated every 25 milliseconds using the codebook-based approach, as explained in Section 3.3. The speech codebook and noise codebook used for the estimation of the STP parameters are obtained by the generalised Lloyd algorithm [31]. During the training process, AR coefficients (converted into line spectral frequency coefficients) are extracted from windowed frames, obtained from the training signal and passed as an input to the vector quantiser. Working in the line spectral frequency domain is guaranteed to result in stable inverse filters [32]. Codebook vectors are then obtained as an output from the vector quantiser depending on the size of the codebook. For our experiments, we have used both a speaker-specific codebook and a general speech codebook. A speaker-specific codebook of 64 entries was generated using head related impulse response (HRIR) convolved speech from the specific speaker of interest. A general speech codebook of 256 entries was generated from a training sample of 30 minutes of HRIR convolved speech from 30 different speakers. Using a speaker-specific codebook instead of a general speech codebook leads to an improvement in the performance, and a comparison between the two was made in [15]. It should be noted that the sentences used for training the codebook were not included in the test sequence. The noise codebook consisting of only 8 entries, was generated using thirty seconds of noise signal [33]. The

Algorithm 1 Main steps involved in the binaural enhancement framework

```
while new time-frames are available do
  Estimate the dual channel noise PSD and append the noise codebook
  with the AR coefficients corresponding to the estimated noise PSD $\hat{P}_w^{\mathrm{DC}}$
  (see Section 3.3).
  for $\forall i \in N_s$ do
    for $\forall j \in N_w$ do
      compute the ML estimates of excitation noise variances
      ($\sigma^2_{d,ij}$ and $\sigma^2_{v,ij}$) using (A.27) and (A.28).
      compute the modelled spectrum $\hat{P}^{\mathrm{ML}}_{z_{ij}}$ using (A.30).
      compute the likelihood values $p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}^{\mathrm{ML}}_{ij})$ using (A.29).
    end for
  end for
  Get the final estimates of STP parameters using (A.20).
  Estimate the pitch parameters using the algorithm explained in Section
  3.4.
  Use the estimated STP parameters and the pitch parameters in the
  FLKS equations (see Appendix 7.1) to get the enhanced signal.
end while
```

AR model order for both the speech and noise signal was empirically chosen to be 14. The pitch period and degree of voicing was estimated as explained in Section 3.4 where the cost function in (A.40) was evaluated on a 0.5 Hz grid for fundamental frequencies in the range $80 - 400$ Hz.

4.2 Simulation set-up

In this paper we have considered two simulation set-ups representative of the cocktail party scenario. The details regarding the two set-ups are given below:

Set-up 1

The clean signals were at first convolved with an anechoic binaural HRIR corresponding to the nose direction, taken from a database [34]. Noisy signals are then generated by adding binaurally recorded babble noise taken from the ETSI database [33].

Set-up 2

The noisy signals were generated using the McRoomSim acoustic simulation software [35]. Fig. A.4 shows the geometry of the room along with the

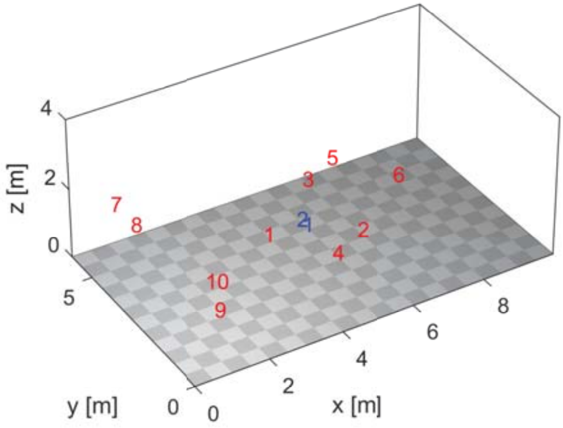


Fig. A.4: Set-up 2 showing the cocktail scenario where 1 (red) indicates the speaker of interest and 2-10 (red) are the interferers and 1,2 (blue) are the microphones on the left ear and right ear respectively.

speaker, listener and the interferers. This denotes a typical cocktail party scenario, where 1 (red) indicates the speaker of interest, 2-10 (red) are the interferers, and 1, 2 (blue) are the microphones on the left, right ears respectively. The dimensions of the room in this case is $10 \times 6 \times 4$ m. The reverberation time of the room was chosen to be 0.4 s.

4.3 Evaluated enhancement frameworks

In this section we will give an overview about the binaural and bilateral enhancement frameworks that have been evaluated in this paper using the objective and subjective scores.

Binaural enhancement framework

In the binaural enhancement framework, we assume that there is a wireless link between the HAs. Thus, the filter parameters are estimated jointly using the information at the left and right channels.

> *Proposed methods* : The binaural enhancement framework utilising the V-UV model, when used in conjunction with a general speech codebook is denoted as Bin-S(V-UV), whereas Bin-Spkr(V-UV) denotes the case where we use a speaker-specific codebook. The binaural enhancement framework utilising the UV model, when used in conjunction with a general speech codebook is denoted as Bin-S(UV), whereas Bin-Spkr(UV) denotes the case where we use a speaker-specific codebook.

> *Reference methods* : For comparison, we have used the methods proposed in [7] and [8] which we denote as TwoChSS and TS-WF respectively. We chose these methods for comparison, as TwoChSS was one of the first methods designed for a two-input two-output configuration and TS-WF is one of the state of the art methods belonging to this class.

Bilateral enhancement framework

In the bilateral enhancement framework, single channel speech enhancement techniques are performed independently on each ear.

> *Proposed methods* : The bilateral enhancement framework utilising the V-UV model, when used in conjunction with a general speech codebook is denoted as Bil-S(V-UV), whereas Bil-Spkr(V-UV) denotes the case where we use a speaker-specific codebook. The bilateral enhancement framework utilising the UV model, when used in conjunction with a general speech codebook is denoted as Bil-S(UV), whereas Bil-Spkr(UV) denotes the case where we use a speaker-specific codebook. The difference of the bilateral case in comparison to the binaural case is in the estimation of the filter parameters. In the bilateral case, the filter parameters are estimated independently for each ear which leads to different filter parameters for each ear, e.g., the STP parameters are estimated using the method in [19] independently for each ear.

> *Reference methods* : For comparison, we have used the methods proposed in [36] and [37] which we denote as MMSE-GGP and PMBE respectively.

4.4 Objective measures

The objective measures, STOI [38] and PESQ [39] have been used to evaluate the intelligibility and quality of different enhancement frameworks. We have evaluated the performance of the algorithms, separately for the 2 different simulation set-ups explained in Section 4.2. Table A.1 and A.2 show the objective measures obtained for the binaural and bilateral enhancement frameworks, respectively, when evaluated in the set-up 1. The test signals that have been used for the binaural and bilateral enhancement frameworks are identical. The scores shown in the tables are the averaged scores across the left and right channels. In comparison to the reference methods which reduce the STOI scores, it can be seen that all of the proposed methods improve the STOI scores. It can be seen from Tables A.1 and A.2 that the Bin-Spkr(V-UV) performs the best in terms of STOI scores. In addition to preserving the binaural cues, it is evident from the scores that the binaural frameworks perform in general better than the bilateral frameworks, and the improvement of binaural framework over bilateral framework is more pronounced at low SNRs. It can also be seen that the V-UV model which takes into account the pitch information performs better than the UV model. Tables A.3 and A.4 show the objective measures obtained for the different binaural and bilateral enhancement frameworks, respectively, when evaluated in the simulation set-up 2. The results obtained for set-up 2 shows similar trends to the results obtained for set-up 1. We would also like to remark here that in the range of 0.6-0.8, an increase in 0.05 in STOI score corresponds to approximately 16 percentage points increase in subjective intelligibility [40].

4.5 Inter-aural errors

We now evaluate the proposed algorithm in terms of binaural cue preservation. This was evaluated objectively using inter-aural time difference (ITD) and inter-aural level difference (ILD) also used in [8]. ITD is calculated as

$$\mathrm{ITD} = \frac{|\angle C_{\mathrm{enh}} - \angle C_{\mathrm{clean}}|}{\pi}, \tag{A.41}$$

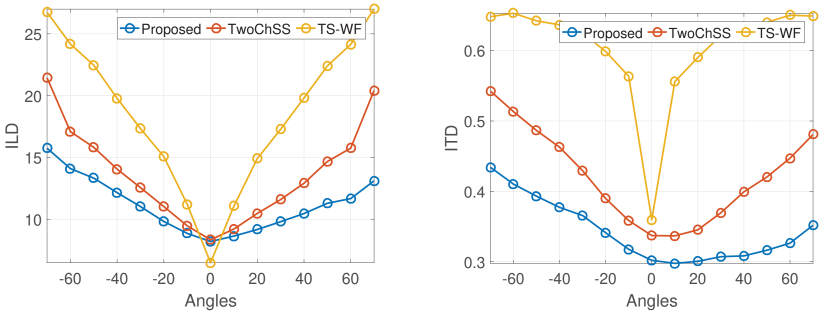
where $\angle C_{\mathrm{enh}}$ and $\angle C_{\mathrm{clean}}$ denotes the phases of the cross PSD of the enhanced and clean signal respectively, given by $C_{\mathrm{enh}} = \mathbb{E}\{\hat{S}_l \hat{S}_r\}$ and $C_{\mathrm{clean}} = \mathbb{E}\{S_l S_r\}$, where $\hat{S}_{l/r}$ denotes the spectrum of enhanced signal at the left/right ear and $S_{l/r}$ denotes the spectrum of the clean signal at the left/right ear. The expectation is calculated by taking the average value over all frames and frequency indices (which has been omitted here for notational convenience). ILD is calculated as

$$\mathrm{ILD} = \left| 10\log_{10} \frac{I_{\mathrm{enh}}}{I_{\mathrm{clean}}} \right|, \tag{A.42}$$

where $I_{\mathrm{enh}} = \frac{\mathbb{E}\{|\hat{S}_l|^2\}}{\mathbb{E}\{|\hat{S}_r|^2\}}$ and $I_{\mathrm{clean}} = \frac{\mathbb{E}\{|S_l|^2\}}{\mathbb{E}\{|S_r|^2\}}$. Fig. A.5 shows the ILD and ITD cues for the proposed method, Bin-Spkr(V-UV), TwoChSS and TS-WF for different angles of arrivals. It can be seen that the proposed method has a lower ITD and ILD in comparison to TwoChSS and TS-WF. It should be noted that the proposed method and TwoChSS do not use the angle of arrival and assume that the speaker of interest is in the nose direction of the listener. TS-WF, on the other hand requires the a priori knowledge of the angle of arrival. Thus, to make a fair comparison we have included here the inter-aural cues for TS-WF when the speaker of interest is assumed to be in the nose direction.

4.6 Listening tests

We have conducted listening tests to measure the performance of the proposed algorithm in terms of quality and intelligibility improvements. The



(a) ILD

(b) ITD

Fig. A.5: Inter-aural cues for different speaker positions.

Table A.1: This table shows the comparison of objective measures (PESQ & STOI) for the different **BINAURAL** enhancement frameworks for 4 different signal to noise ratios. Noisy signals used for the evaluation here is generated using the simulation set-up 1.

		Bin-Spkr(UV)	Bin-Spkr(V-UV)	Bin-S(UV)	Bin-S(V-UV)	TS-WF	TwoChSS	Noisy
STOI	0 dB	0.71	**0.75**	0.68	0.72	0.62	0.64	0.67
	3 dB	0.80	**0.82**	0.77	0.79	0.69	0.72	0.73
	5 dB	0.84	**0.85**	0.81	0.83	0.74	0.77	0.78
	10 dB	0.91	**0.91**	0.90	0.90	0.85	0.86	0.87
PESQ	0 dB	1.43	**1.53**	1.37	1.45	1.40	1.49	1.33
	3 dB	1.67	**1.72**	1.58	1.68	1.55	1.66	1.43
	5 dB	1.80	**1.85**	1.73	1.78	1.68	1.79	1.50
	10dB	**2.24**	2.22	2.13	2.14	2.13	2.20	1.70

tests were conducted on a set of nine NH subjects. These tests were performed in a silent room using a set of Beyerdynamic DT 990 pro headphones. The speech enhancement method that we have evaluated in the listening tests is Bil-Spkr(V-UV) for a single channel. We chose this case for the tests as we wanted to test the simpler, but more challenging case of intelligibility and quality improvement when we have access to only a single channel. Moreover, as the tests were conducted with NH subjects, we also wanted to eliminate any bias in the results that can be caused due to the binaural cues [41], as the benefit of using binaural cues is higher for a NH person than for a hearing impaired person.

Quality tests

Quality performance of the proposed algorithms were evaluated using MUSHRA experiments [42]. The test subjects were asked to evaluate the quality of the processed audio-files using a MUSHRA set-up. The subjects were presented with the clean, processed and the noisy signals. The processing algorithms

Table A.2: This table shows the comparison of objective measures (PESQ & STOI) for the different **BILATERAL** enhancement frameworks for 4 different signal to noise ratios. Noisy signals used for the evaluation here is generated using the simulation set-up 1.

		Bil-Spkr(UV)	Bil-Spkr(V-UV)	Bil-S(UV)	Bil-S(V-UV)	MMSE-GGP	PMBE	Noisy
STOI	0 dB	0.68	**0.72**	0.66	0.70	0.66	0.66	0.67
	3 dB	0.77	**0.79**	0.75	0.78	0.73	0.73	0.73
	5 dB	0.81	**0.83**	0.80	0.82	0.78	0.78	0.78
	10 dB	0.90	**0.90**	0.89	0.90	0.87	0.87	0.87
PESQ	0 dB	1.37	**1.45**	1.34	1.40	1.26	1.30	1.33
	3 dB	1.58	**1.65**	1.53	1.60	1.43	1.43	1.43
	5 dB	1.72	**1.76**	1.66	1.72	1.50	1.56	1.50
	10 dB	**2.12**	2.10	2.04	2.05	1.73	1.79	1.70

Table A.3: This table shows the comparison of STOI scores for the different **BINAURAL** enhancement frameworks for 4 different signal to noise ratios. Noisy signals used for the evaluation here is generated using the simulation set-up 2.

		Bin-Spkr(UV)	Bin-Spkr(V-UV)	Bin-S(UV)	Bin-S(V-UV)	TS-WF	TwoChSS	Noisy
STOI	0 dB	0.63	**0.68**	0.61	0.66	0.62	0.58	0.60
	3 dB	0.73	**0.75**	0.71	0.74	0.69	0.67	0.68
	5 dB	0.78	**0.80**	0.76	0.79	0.73	0.72	0.73
	10 dB	0.88	**0.89**	0.87	0.88	0.81	0.83	0.84

Table A.4: This table shows the comparison of STOI scores for the different **BILATERAL** enhancement frameworks for 4 different signal to noise ratios. Noisy signals used for the evaluation here is generated using the simulation set-up 2.

		Bil-Spkr(UV)	Bil-Spkr(V-UV)	Bil-S(UV)	Bil-S(V-UV)	MMSE-GGP	PMBE	Noisy
STOI	0 dB	0.61	**0.65**	0.60	0.64	0.58	0.60	0.60
	3 dB	0.71	**0.74**	0.69	0.73	0.66	0.68	0.68
	5 dB	0.76	**0.79**	0.75	0.78	0.72	0.73	0.73
	10 dB	0.87	**0.88**	0.86	0.88	0.83	0.84	0.84

considered here are Bil-Spkr(V-UV) and MMSE-GGP. The SNR of the noisy signal considered here was 10 dB. The subjects were then asked to rate the presented signals in a score range of $0 - 100$. Fig. A.6 shows the mean scores along with 95% confidence intervals that were obtained for the different methods. It can be seen from the figure that the proposed method performs significantly better than the reference method.

Intelligibility tests

Intelligibility tests were conducted using sentences from the GRID database [30]. The GRID database contains sentences spoken by 34 different speakers (18 males and 16 females). The sentences are of the following syntax: Bin Blue (Color) by S (Letter) 5 (Digit) please. Table G.1 shows the syntax of all the possible sentences. subjects are asked to identify the color, letter

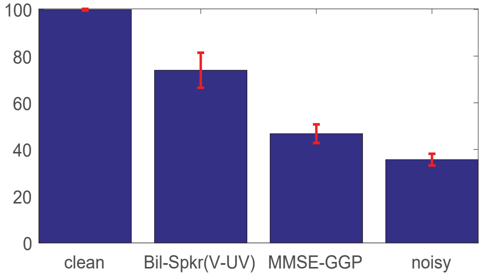


Fig. A.6: Figure showing the mean scores and the 95% confidence intervals obtained in the MUSHRA test for the different methods.

and number after listening to the sentence. The sentences are played back in the SNR range -8 to 0 dB for different algorithms. This SNR range is chosen as all the subjects were NH which led to the intelligibility of the unprocessed signal above 2 dB to be close to 100%. A total of nine test subjects were used for the experiments and the average time taken for carrying out the listening test for a particular person was approximately two hours. The noise signal that we have used for the tests is the babble signal from the AURORA database [43]. The test subjects evaluated the noisy signals (unp) and two versions of the processed signal, nr_{100} and nr_{85}. The first version, nr_{100}, refers to the completely enhanced signal and the second version, nr_{85}, refers to a mixture of the enhanced signal and the noisy signal with 85% of the enhanced signal and 15% of the noisy signal. This mixing combination was empirically chosen [44]. Figures A.7, A.8 and A.9 show the intelligibility percentage along with 90% probability intervals obtained for digit, color and the letter field respectively as a function of SNR, for the different methods. It can be seen that nr_{85} performs the best consistently followed by nr_{100} and the unp. Fig. A.10 shows the mean accuracy over all the 3 fields. It can be seen from the figure that nr_{85} gives up to 15% improvement in intelligibility at -8 dB SNR. We have also computed the probabilities that a particular method is better than the unprocessed signal in terms of intelligibility. For the computation of these probabilities, the posterior probability of success for each method is modelled using a beta distribution. Table A.6 shows these probabilities at different SNRs for the 3 different fields. $P(nr_{85} > unp)$ denotes the probability that nr_{85} is better than unp. It can be seen from the table that nr_{85} consistently has a very high probability of being better than unp for all the SNRs, whereas nr_{100} has a high probability of decreasing the intelligibility for the color field at -2 dB and the letter field at 0 dB. This can also be seen from Figures A.8 and A.9. In terms of the mean intelligibility across all fields, it can be seen that the probability that nr_{85} performs better than unp is 1 for all the SNRs. Similarly, the probability that nr_{100} also performs better than

Table A.5: Sentence syntax of the GRID database.

Sentence structure					
command	color	preposition	letter	digit	adverb
bin	blue	at	A-Z (no W)	0-9	again
lay	green	by			now
place	red	in			please
set	white	with			soon

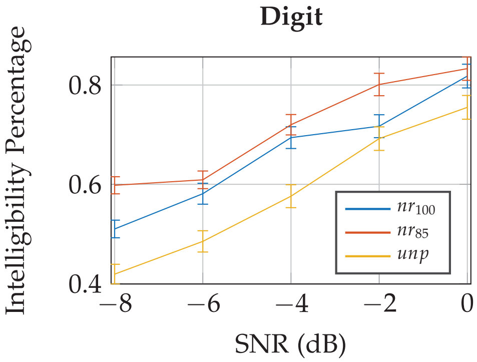


Fig. A.7: Mean percentage of correct answers given by participants for the digit field as function of SNR for different methods. (unp) refers to the noisy signal, (nr_{100}) refers to the completely enhanced signal and (nr_{85}) refers to a mixture of the enhanced signal and the noisy signal with 85% of the enhanced signal and 15% of the noisy signal.

unp is very high across all SNRs.

5 Discussion

The noise reduction capabilities of a HA are limited especially in situations such as the cocktail party scenario. Single channel speech enhancement algorithms which do not use any prior information regarding the speech and noise type have not been able to show much improvements in speech intelligibility [45]. A class of algorithms that has received significant attention recently have been the deep neural network (DNN) based speech enhancement systems. These algorithms use a priori information about speech and noise types to learn the structure of the mapping function between noisy and clean speech features. These methods were able to show improvements in speech intelligibility when trained to very specific scenarios. Recently, the performance of a general DNN based enhancement system was investigated in terms of objective measures and intelligibility tests [46]. Even though the

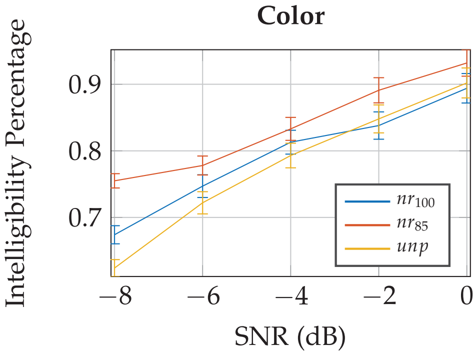


Fig. A.8: Mean percentage of correct answers given by participants for color field as function of SNR for different methods.

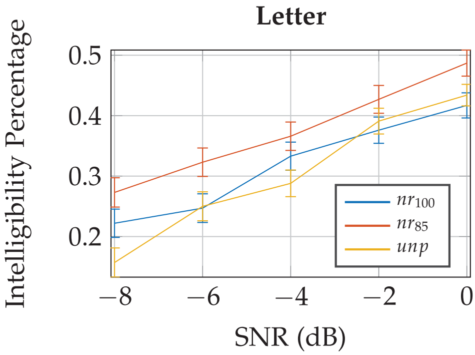


Fig. A.9: Mean percentage of correct answers given by participants for letter field as function of SNR for different methods.

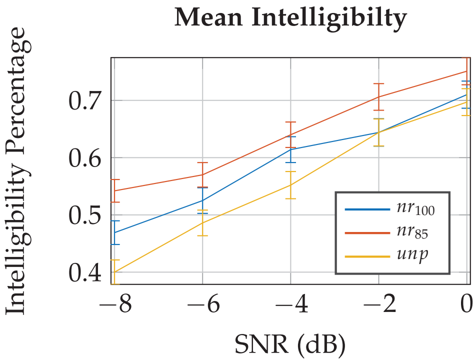


Fig. A.10: Mean percentage of correct answers given by participants for all the fields as function of SNR for different methods.

Table A.6: This table shows the probabilities that a particular method is better than the unprocessed signal.

		SNR (dB)				
		-8	-6	-4	-2	0
Digit	$P(nr_{85} > unp)$	1	1	1	1	1
	$P(nr_{100} > unp)$	1	1	1	0.91	0.99
Color	$P(nr_{85} > unp)$	1	0.99	0.99	0.99	0.99
	$P(nr_{100} > unp)$	0.98	0.91	0.89	0.24	0.27
Letter	$P(nr_{85} > unp)$	1	1	1	0.96	0.99
	$P(nr_{100} > unp)$	1	0.44	0.99	0.22	0.19
Mean	$P(nr_{85} > unp)$	1	1	1	1	1
	$P(nr_{100} > unp)$	1	0.99	1	0.50	0.87

general system showed improvements in the objective measures, the intelligibility tests failed to show consistent improvements across the SNR range. In this paper we have proposed a model-based speech enhancement framework that takes into account the speech production model, characterised by the vocal tract and the excitation signal. The proposed framework uses a priori information regarding the speech spectral envelopes (which is used for modelling the characteristics of the vocal tract) and noise spectral envelopes. In comparison to DNN based algorithms the training data required by the proposed algorithm, and the parameters to be trained for the proposed algorithm is significantly less. The parameters to be trained in the proposed algorithm includes the AR coefficients corresponding to the speech and noise spectral shapes which is considerably less compared to the weights present in a DNN. As the amount of parameters to be trained is much smaller, it should also be possible to train these parameters on-line in case of noise only scenarios or speech only scenarios. The proposed framework was able to show consistent improvements in the intelligibility tests even for the single channel case as shown in section 4.6. Moreover, we have shown the benefit of using multiple channels for enhancement by the means of objective experiments. We would like to remark that the enhancement algorithm proposed in this paper is computationally more complex when compared to conventional speech enhancement algorithms such as [36]. However, there exists some methods in the literature which can reduce the computational complexity of the proposed algorithm. The pitch estimation algorithm can be sped up using the principles proposed in [47]. There also exists efficient ways of performing Kalman filtering due to the structured and sparse matrices involved in the operation of a Kalman filter [13].

6 Conclusion

In this paper, we have proposed a model-based method for performing binaural/bilateral speech enhancement in HAs. The proposed enhancement framework takes into account the speech production dynamics by using a FLKS for the enhancement process. The filter parameters required for the functioning of the FLKS are estimated jointly using the information at the left and right microphones. The filter parameters considered here are the speech and noise STP parameters and the speech pitch parameters. The estimation of these parameters in not trivial due to the highly non-stationary nature of speech and the noise in a cocktail party scenario. In this work, we have proposed a binaural codebook-based method, trained on spectral models of speech and noise, for estimating the speech and noise STP parameters, and a pitch estimator based on the harmonic model is proposed to estimate the pitch parameters. We then evaluated the proposed enhancement framework in two experimental set-ups representative of the cocktail party scenario. The objective measures, STOI and PESQ, were used for evaluating the proposed enhancement framework. The proposed method showed considerable improvement in STOI and PESQ scores, in comparison to a number of reference methods. Subjective listening tests when having access to single channel noisy observation also showed improvement in terms of intelligibility and quality. In the case of intelligibility tests, a mean improvement of about 15 % was observed at -8 dB SNR.

7 Appendix

7.1 Prediction and Correction stages of the FLKS

This section gives the prediction and correction stages involved in the FLKS for the V-UV model. The same equations apply for the UV model, except that the state vector and the state transition matrices will be different. The prediction stage of the FLKS, which computes the a priori estimates of the state vector ($\hat{\mathbf{x}}_{l/r}^{\text{V-UV}}(n|n-1)$) and error covariance matrix ($\mathbf{M}(n|n-1)$) is given by

$$\begin{aligned}\hat{\mathbf{x}}_{l/r}^{\text{V-UV}}(n|n-1) &= \mathbf{F}^{\text{V-UV}}(f_n)\hat{\mathbf{x}}_{l/r}^{\text{V-UV}}(n-1|n-1)\\ \mathbf{M}(n|n-1) &= \mathbf{F}^{\text{V-UV}}(f_n)\mathbf{M}(n-1|n-1)\mathbf{F}^{\text{V-UV}}(f_n)^T +\\ &\qquad\qquad \mathbf{\Gamma}_5\begin{bmatrix}\sigma_d^2(f_n) & 0\\ 0 & \sigma_v^2(f_n)\end{bmatrix}\mathbf{\Gamma}_5^T.\end{aligned}$$

The Kalman gain is computed as

$$\mathbf{K}(n) = \frac{\mathbf{M}(n|n-1)\mathbf{\Gamma}^{\text{V-UV}}}{[\mathbf{\Gamma}^{\text{V-UV}^T}\mathbf{M}(n|n-1)\mathbf{\Gamma}^{\text{V-UV}}]}. \tag{A.43}$$

The correction stage of the FLKS, which computes the a posteriori estimates of the state vector and error covariance matrix is given by

$$\hat{\mathbf{x}}^{\text{V-UV}}_{l/r}(n|n) = \hat{\mathbf{x}}^{\text{V-UV}}_{l/r}(n|n-1) + \mathbf{K}(n)[z_{l/r}(n) - \boldsymbol{\Gamma}^{\text{V-UV}^T}\hat{\mathbf{x}}^{\text{V-UV}}_{l/r}(n|n-1)]$$

$$\mathbf{M}(n|n) = (\mathbf{I} - \mathbf{K}(n)\boldsymbol{\Gamma}^{\text{V-UV}^T})\mathbf{M}(n|n-1).$$

Finally, the enhanced signal at time index $n - (d_s + 1)$ is obtained by taking the $(d_s + 1)^{th}$ entry of the a posteriori estimate of the state vector as

$$\hat{s}_{l/r}(n - (d_s + 1)) = \left[\hat{\mathbf{x}}^{\text{V-UV}}_{l/r}(n|n)\right]_{d_s+1}. \tag{A.44}$$

7.2 Behaviour of the likelihood function

For a given set of speech and noise AR coefficients, we show the behaviour of the likelihood $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta})$ as a function of the speech and noise excitation variance. For the experiments, we have set the excitation variances to be 10^{-3}. Fig. A.11 plots the likelihood as a function of the speech and noise excitation variance. It can be seen from the figure that likelihood is the maximum at the true values and decays rapidly as it deviates form its true value. This behaviour motivates the approximation in Section 3.3.

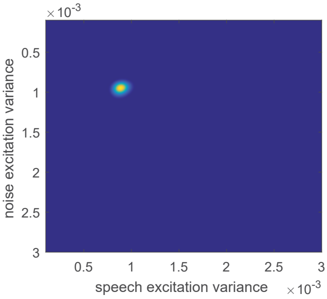


Fig. A.11: Likelihood shown as a function of the speech and noise excitation variance.

7.3 A priori information on the distribution of the excitation variances

It can be seen from (A.20) that the prior distributions of the excitation variances are used in the estimation of STP parameters. In the case of no a priori knowledge regarding the excitation variances, a uniform distribution can be used as done in [14], but a priori knowledge regarding the distribution of the noise excitation variance can be beneficial. Fig. A.12 shows the histogram of the noise excitation variance plotted for a minute of babble noise [43]. It can be observed from the figure that the histogram approximately follows a

Gamma distribution. Thus, we here use a Gamma distribution to model the a priori information about the noise excitation variance, which is modelled using two parameters (shape parameter κ and the scale parameter ζ) as

$$p(\sigma_v^2) = \frac{1}{\Gamma(\kappa)\zeta^k} \sigma_v^{2^{\kappa-1}} e^{-\frac{\sigma_v^2}{\zeta}}, \tag{A.45}$$

where $\Gamma(\cdot)$ is the Gamma function. The parameters ζ and κ can be learned from the training data.

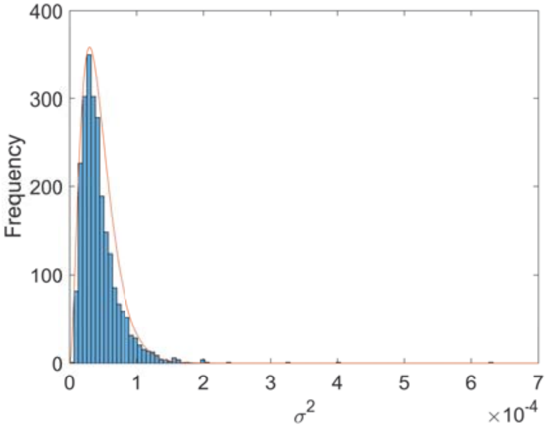


Fig. A.12: Plot showing the histogram fitting for noise excitation variance. Curve (red) is obtained by fitting the histogram with a Gamma distribution with two parameters.

References

[1] S. Kochkin, "10-year customer satisfaction trends in the US hearing instrument market," *Hearing Review*, vol. 9, no. 10, pp. 14–25, 2002.

[2] T. V. D. Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[3] A. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

[4] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

[5] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[6] T. J. Klasen, T. V. D. Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. on Signal Process.*, vol. 55, no. 4, pp. 1579–1585, 2007.

[7] M. Dorbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," in *Signal Processing Conference, 1996 European*. IEEE, 1996, pp. 1–4.

[8] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[9] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–14, 2006.

[10] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1987.

[11] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Process.*, vol. 39, no. 8, pp. 1732–1742, 1991.

[12] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 6, no. 4, pp. 373–385, 1998.

[13] Z. Goh, K. C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 7, no. 5, pp. 510–524, 1999.

[14] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[15] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook based approach," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 2016.

[16] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Binaural speech enhancement using a codebook based approach," *Proc. Int. Workshop on Acoustic Signal Enhancement*, 2016.

[17] ——, "Model based binaural enhancement of voiced and unvoiced speech," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 2017.

[18] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[19] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 457–468, 2017.

[20] M. B. Christopher, *Pattern recognition and machine learning*. Springer-Verlag New York, 2006.

[21] R. M. Gray *et al.*, "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[22] F. Itakura, "Analysis synthesis telephony based on the maximum likelihood method," in *The 6th international congress on acoustics, 1968*, 1968, pp. 280–292.

[23] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[24] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[25] P. Stoica, R. L. Moses *et al.*, *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005, vol. 452.

[26] A. H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 521–533, 2009.

[27] M. Jeub, C. Nelke, H. Kruger, C. Beaugeant, and P. Vary, "Robust dual-channel noise power spectral density estimation," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 2304–2308.

[28] P. C. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 6, no. 5, pp. 476–488, 1998.

[29] M. G. Christensen and A. Jakobsson, “Multi-pitch estimation,” *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[31] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[32] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, 1976.

[33] ETSI202396-1, “Speech and multimedia transmission quality; part 1: Background noise simulation technique and background noise database.” 2009.

[34] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–10, 2009.

[35] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, “Room acoustics simulation for multichannel microphone arrays,” in *Proceedings of the International Symposium on Room Acoustics*, 2010, pp. 1–6.

[36] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[37] P. C. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 13, no. 5, pp. 857–869, 2005.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[39] “Perceptual evaluation of speech quality, an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” *ITU-T Recommendation*, p. 862, 2001.

[40] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, “Objective quality and intelligibility prediction for

users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[41] A. W. Bronkhorst and R. Plomp, "A clinical test for the assessment of binaural speech perception in noise," *Audiology*, vol. 29, no. 5, pp. 275–285, 1990.

[42] I. Recommendation, "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[43] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[44] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and hearing*, vol. 27, no. 5, p. 480, 2006.

[45] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 47–56, 2011.

[46] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[47] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Processing*, vol. 135, pp. 188–197, 2017.

Paper B

Model based Estimation of STP parameters for Binaural Speech Enhancement

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Mads Græsbøll Christensen and Jesper B. Boldt

The paper has been published in the
Proc. European Signal Processing Conference, 2018

Abstract

This paper deals with the estimation of the short-term predictor (STP) parameters of speech and noise in a binaural framework. A binaural model based approach is proposed for estimating the power spectral density (PSD) of speech and noise at the individual ears for an arbitrary position of the speech source. The estimated PSDs can be subsequently used for enhancement in a binaural framework. The experimental results show that taking into account the position of the speech source using the proposed method leads to improved modelling and enhancement of the noisy speech.

1 Introduction

Understanding of speech in difficult listening situations like cocktail party scenarios is a major issue for the hearing impaired. Speech enhancement capabilities of a hearing aid (HA) in such scenarios have been observed to be limited. Generally, a hearing impaired person is fitted with HAs at both ears. With the recent developments in HA technology, HAs are able to communicate with each other through a wireless link and share information. This enables binaural processing of signals. Binaural processing of noisy signals has shown to be more effective than processing the noisy signal independently at each ear [1]. Some binaural speech enhancement algorithms with multiple microphones present in each hearing aid have been previously proposed in [2, 3].

However, in this work we are concerned with binaural speech enhancement algorithms with access to only one microphone per HA. This is obseved in in-the-ear (ITE) HAs, where the space constraints limit the number of microphones per HA. Some of the existing algorithms with a single microphone present in each hearing aid are [4–6]. These algorithms perform the enhancement in the frequency domain by assuming that the speech and noise components are uncorrelated, and do not take into consideration the dynamics of speech production process. It was recently proposed in [7, 8] to perform binaural enhancement of speech while taking into account the speech production model. The filter parameters here consists of the STP parameters of speech and noise. STP parameters constitute of the autoregressive (AR) parameters representing the spectral envelope and the excitation variance corresponding to the gain of the envelope. These parameters can be used to parametrically model the speech and noise PSDs at the individual ears. The estimation of these filter parameters in [7, 8] assumed that the speaker source is in the nose direction of the listener. Due to this assumption, the speech PSDs at the two ears were modelled in [7, 8] using the same set of STP parameters. This type of modelling might not be appropriate if the speaker is not in the nose direction. This scenario is of interest, as it has been observed in [9, 10], that the Speech Reception Threshold (SRT) is not always the minimum when the speaker is in the nose direction. It was noticed that the listeners often tend to orient their head away from the speech source for an improvement in the SRT. Thus, in this paper, we propose a method to take the position of the speaker into account while estimating the speech and noise PSDs at the two ears. This leads to the estimation of individual speech PSDs for the two ears. A codebook based approach, which takes into account the a priori information regarding the speech and noise AR spectral envelopes is proposed to estimate the STP parameters. The method proposed in this paper uses a multiplicative update method [11] commonly used in non-negative matrix factorisation (NMF) applications [12] to estimate the gain parameters corresponding to the speech and noise AR processes.

The remainder of the paper is structured as follows. Section 2 motivates the problem and also introduces the signal model used in the paper. Section 3 explains the proposed method of estimating the speech and noise STP parameters in detail. Experiments and results are presented in Section 4 followed by conclusion in Section 5.

2 Motivation

In this section we introduce the signal model and motivate this work. The binaural noisy signals at the left/right ear, denoted by $z_{l/r}(n)$ is written as

$$z_{l/r}(n) = s_{l/r}(n) + w_{l/r}(n) \qquad \forall n = 0, 1, 2 \ldots, \tag{B.1}$$

where $s_{l/r}(n)$ is the clean speech component and $w_{l/r}(n)$ is the noise component. A very popular way to represent the clean speech component is in the form of an AR process. In [7, 8], it was assumed that the target speaker is located in the nose direction of the listener. Due to this assumption, the clean speech component at both ears were represented using AR processes having the same set of STP parameters. This modelling is reasonable as long as the speaker is in the nose direction of the listener. However, it might not be an appropriate model for the case when speaker is not present in the nose direction. Here, we have conducted a few simulations to show the properties of the parameters corresponding to the speech component present at the left and right microphones. The speaker position is set to be 40 degree right of the listener at a distance of 80 cm. Fig. B.1 shows a snapshot of the gain normalised spectral envelopes for the left and right channel. It can be seen that the gain normalised spectral envelopes at the left and right channels have approximately the same content. In comparison to the AR spectral envelopes, it can be seen from Fig. B.2, that there is considerable difference in the excitation variances between the left and right channels. This can be explained due to the head shadowing effect, which leads to an attenuation of the intensity at

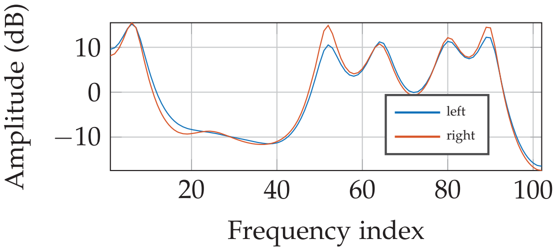


Fig. B.1: Gain normalised spectral envelopes for the left and right channel

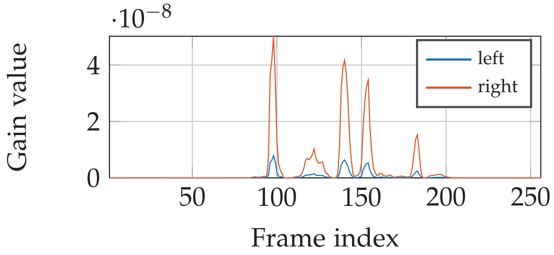


Fig. B.2: Plot of the excitation variances for the left and right channel

the ear on the far side (left ear in this case). Motivated by these observations in figures B.1 and B.2, we model the speech component at the left and right ears using the same spectral envelope but different excitation variances as

$$s_{l/r}(n) = \left(\sum_{i=1}^{P} a_i s_{l/r}(n-i) \right) + u_{l/r}(n), \tag{B.2}$$

where $\{a_i\}_{i=1}^{P}$ is the set of speech AR parameters and $u_{l/r}(n)$ is white Gaussian noise (WGN) with zero mean and excitation variance $\sigma^2_{u_{l/r}}(n)$. It is also assumed that the noise component at both ears have similar spectral shape. This is due to the diffuse noise field assumption. The noise components can be similarly expressed as an AR process of order Q as follows,

$$w_{l/r}(n) = \left(\sum_{i=1}^{Q} b_i w_{l/r}(n-i) \right) + v(n). \tag{B.3}$$

where $\{b_i\}_{i=1}^{Q}$ is the set of noise AR parameters and $v(n)$ is white Gaussian noise (WGN) with zero mean and excitation variance $\sigma^2_v(n)$. STP parameters corresponding to speech and noise are considered to be constant over a duration of 25ms.

3 Model based estimation of STP parameters

The speech and noise STP parameters required for the enhancement are estimated frame-wise using a codebook based approach [7, 13]. The estimation of these parameters uses a priori information about the speech and noise spectral envelopes present in trained codebooks in the form of Linear Prediction Coefficients (LPC). These trained parameters offers us an elegant way to take into account prior information regarding the noise type and speaker of interest. Here, we use a Bayesian framework for estimating the STP parameters. The random variables (r.v) corresponding to the parameters to be estimated are represented as $\boldsymbol{\theta} = [\boldsymbol{\theta}_s \ \boldsymbol{\theta}_w] = [\mathbf{a}; \sigma_u^2; \mathbf{b}; \sigma_v^2; c]$, where $\mathbf{a}, \mathbf{b}$ corresponds to r.v representing the speech and noise AR parameters, σ_u^2, σ_v^2 representing the speech and noise excitation variances and c corresponds to the scale parameter that relates to the excitation variance between the left and right ear *i.e.* $\sigma_{u_l}^2 = \sigma_u^2$ and $\sigma_{u_r}^2 = c \times \sigma_u^2$. In this work, scale parameter is considered time varying, to take into account the changes in speaker position. Fig. B.3 shows a basic block diagram of the enhancement framework, where it can be seen that the STP parameters are estimated jointly using the information at the left and right channels. Thus, the MMSE estimate of the parameter vector

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{z}_l, \mathbf{z}_r) = \int_{\Theta} \boldsymbol{\theta} \frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{z}_l, \mathbf{z}_r)} d\boldsymbol{\theta}, \tag{B.4}$$

where $\mathbf{z}_l$ and $\mathbf{z}_r$ is a frame of length N of noisy speech at the left and right ears respectively. Let us define $\boldsymbol{\theta}_{ij}^{\mathrm{ML}} = [\mathbf{a}_i; \sigma_{u,ij}^{2,\mathrm{ML}}; \mathbf{b}_j; \sigma_{v,ij}^{2,\mathrm{ML}}; c_{ij}^{\mathrm{ML}}]$ where $\mathbf{a}_i$ is the i^{th} entry of speech codebook (of size N_s), $\mathbf{b}_j$ is the j^{th} entry of the noise codebook (of size N_w) and $\sigma_{u,ij}^{2,\mathrm{ML}}, \sigma_{v,ij}^{2,\mathrm{ML}}$ and c_{ij}^{ML}represents the maximum likelihood (ML) estimates of the excitation variances and the scale parameter re-

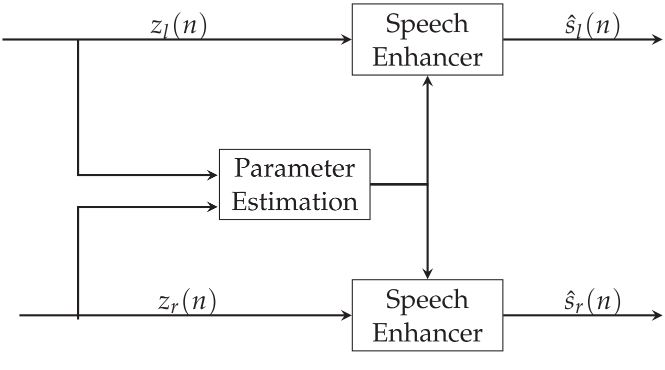


Fig. B.3: Basic block diagram of the binaural enhancement framework

spectively for the ij th combination of the codebook entries. Using the above definition, (B.4) is approximated as [13]

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} \boldsymbol{\theta}_{ij}^{\text{ML}} \frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}}) p(\boldsymbol{\theta}_{ij}^{\text{ML}})}{p(\mathbf{z}_l, \mathbf{z}_r)}, \tag{B.5}$$

where the MMSE estimate is expressed as a weighted linear combination of $\boldsymbol{\theta}_{ij}^{\text{ML}}$ with weights proportional to $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}})$. It is assumed that the left and right noisy signal are conditionally independent given $\boldsymbol{\theta}_{ij}^{\text{ML}}$, which leads to $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}})$ being written as,

$$p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}}) = p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\text{ML}}) p(\mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}}). \tag{B.6}$$

As the scale term is not used for modelling the spectrum at the left ear the likelihood $p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\text{ML}})$ is expressed as

$$p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\text{ML}}) = p(\mathbf{z}_l|[\mathbf{a}_i; \sigma_{u,ij}^{2,\text{ML}}; \mathbf{b}_j; \sigma_{v,ij}^{2,\text{ML}}]). \tag{B.7}$$

Similarly $p(\mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}})$ is expressed as

$$p(\mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}}) = p(\mathbf{z}_r|[\mathbf{a}_i; c_{ij}^{\text{ML}} \times \sigma_{u,ij}^{2,\text{ML}}; \mathbf{b}_j; \sigma_{v,ij}^{2,\text{ML}}]) \tag{B.8}$$

Logarithm of the likelihood $p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\text{ML}})$ can be written as being proportional to the negative of Itakura-Saito (IS) divergence between the noisy periodogram at the left ear $P_{z_l}(k)$ and the modelled noisy spectral envelope $\hat{P}_{z_{l,ij}}^{\text{ML}}(k)$, where k corresponds to the frequency index [13]. Using the same result for the right ear, $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\text{ML}})$ can be written as

$$p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij}^{\text{ML}}) = K \exp\Bigg(-\frac{N}{2}\Big(d_{\text{IS}}\big[P_{z_l}(k), \hat{P}_{z_{l,ij}}^{\text{ML}}(k)\big] + d_{\text{IS}}\big[P_{z_r}(k), \hat{P}_{z_{r,ij}}^{\text{ML}}(k)\big] \Big)\Bigg) \tag{B.9}$$

where $\hat{P}_{z_{l,ij}}^{\text{ML}}(k)$ and $\hat{P}_{z_{r,ij}}^{\text{ML}}(k)$ are denoted as

$$\hat{P}_{z_{l,ij}}^{\text{ML}}(k) = \frac{\sigma_{u,ij}^{2,\text{ML}}}{|A_s^i(k)|^2} + \frac{\sigma_{v,ij}^{2,\text{ML}}}{|A_w^j(k)|^2}, \tag{B.10}$$

$$\hat{P}_{z_{r,ij}}^{\text{ML}}(k) = \frac{c_{ij}^{\text{ML}} \sigma_{u,ij}^{2,\text{ML}}}{|A_s^i(k)|^2} + \frac{\sigma_{v,ij}^{2,\text{ML}}}{|A_w^j(k)|^2} \tag{B.11}$$

and $1/|A_s^i(k)|^2$ is the spectral envelope corresponding to the i^{th} entry of the speech codebook, $1/|A_w^j(k)|^2$ is the spectral envelope corresponding to the j^{th} entry of the noise codebook. For a particular combination of the speech and noise codebook entries, the ML estimates of the excitation variances are estimated by maximising $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij})$. This is equivalent to minimising the total IS distortion as seen in (B.9) given by

$$T_{\text{IS}} = d_{\text{IS}}[P_{z_l}(k), \hat{P}_{z_{l,ij}}(k)] + d_{\text{IS}}[P_{z_r}(k), \hat{P}_{z_{r,ij}}(k)], \qquad \text{(B.12)}$$

where $\hat{P}_{z_{l,ij}}$ and $\hat{P}_{z_{r,ij}}$ has the same form as in (B.10) and (B.11). Here, we use a multiplicative method to estimate the excitation variances and scale term that leads to a minimisation of the cost function in (B.12). In the multiplicative update method, the value of the variable at $(l+1)^{\text{th}}$ iteration is computed by multiplying the value of the variable at l^{th} iteration with the ratio between the negative component of the the gradient and the positive component of the gradient, which is mathematically written as [12], $\phi^{l+1} \leftarrow \phi^{l} \frac{\nabla f(\phi^l)_-}{\nabla f(\phi^l)_+}$, where ϕ is the variable of interest. Taking the derivative of (B.12) with respect to speech and noise excitation variances, and the scaling term c, we get

$$\frac{\partial T_{\text{IS}}}{\partial \sigma^2_{u,ij}} = \frac{1}{N} \sum_{k=1}^{N} \frac{\frac{1}{|A_s^i(k)|^2}}{\hat{P}_{z_{l,ij}}(k)} - \frac{\frac{P_{z_l}(k)}{|A_s^i(k)|^2}}{\hat{P}_{z_{l,ij}}(k)^2} + \frac{\frac{c_{ij}}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)} - \frac{\frac{c_{ij}P_{z_r}(k)}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)^2} \qquad \text{(B.13)}$$

$$\frac{\partial T_{\text{IS}}}{\partial \sigma^2_{v,ij}} = \frac{1}{N} \sum_{k=1}^{N} \frac{\frac{1}{|A_w^j(k)|^2}}{\hat{P}_{z_{l,ij}}(k)} - \frac{\frac{P_{z_l}(k)}{|A_w^j(k)|^2}}{\hat{P}_{z_{l,ij}}(k)^2} + \frac{\frac{1}{|A_w^j(k)|^2}}{\hat{P}_{z_{r,ij}}(k)} - \frac{\frac{P_{z_r}(k)}{|A_w^j(k)|^2}}{\hat{P}_{z_{r,ij}}(k)^2} \qquad \text{(B.14)}$$

$$\frac{\partial T_{\text{IS}}}{\partial c_{ij}} = \frac{1}{N} \sum_{k=1}^{N} \frac{\frac{\sigma^2_{u,ij}}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)} - \frac{\frac{\sigma^2_{u,ij}P_{z_r}(k)}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)^2} \qquad \text{(B.15)}$$

Using the multiplicative update rule, the values for the excitation noise variances are computed iteratively as shown below

$$\sigma^{2(l+1)}_{u,ij} \leftarrow \sigma^{2(l)}_{u,ij} \frac{\sum\limits_{k=1}^{N} \frac{P_{z_l}(k)}{|A_s^i(k)|^2 \hat{P}_{z_{l,ij}}(k)^2} + \frac{c^{(l)}_{ij} P_{z_r}(k)}{|A_s^i(k)|^2 \hat{P}_{z_{r,ij}}(k)^2}}{\sum\limits_{k=1}^{N} \frac{1}{|A_s^i(k)|^2 \hat{P}_{z_{l,ij}}(k)} + \frac{c^{(l)}_{ij}}{|A_s^i(k)|^2 \hat{P}_{z_{r,ij}}(k)}} \qquad \text{(B.16)}$$

$$\sigma^{2(l+1)}_{v,ij} \leftarrow \sigma^{2(l)}_{v,ij} \frac{\sum\limits_{k=1}^{N} \frac{P_{z_l}(k)}{|A_w^j(k)|^2 \hat{P}_{z_{l,ij}}(k)^2} + \frac{P_{z_r}(k)}{|A_w^j(k)|^2 \hat{P}_{z_{r,ij}}(k)^2}}{\sum\limits_{k=1}^{N} \frac{1}{|A_w^j(k)|^2 \hat{P}_{z_{l,ij}}(k)} + \frac{1}{|A_w^j(k)|^2 \hat{P}_{z_{r,ij}}(k)}} \qquad \text{(B.17)}$$

$$c^{(l+1)}_{ij} \leftarrow c^{(l)}_{ij} \frac{\sum\limits_{k=1}^{N} \frac{\sigma^{2(l)}_{u,ij} P_{z_r}(k)}{|A_s^i(k)|^2 \hat{P}_{z_{r,ij}}(k)^2}}{\sum\limits_{k=1}^{N} \frac{\sigma^{2(l)}_{u,ij}}{|A_s^i(k)|^2 \hat{P}_{z_{r,ij}}(k)}}. \qquad \text{(B.18)}$$

It should be noted that $\hat{P}_{z_{l,ij}}(k)$ and $\hat{P}_{z_{r,ij}}(k)$ used in (B.16), (B.17) and (B.18) in the l^{th} iteration is computed using excitation variances and the scale parameter from the $(l-1)^{\text{th}}$ iteration. We have summarised the proposed algorithm for estimating the speech and noise STP parameters in Algorithm 2.

4 Experiments

This section will elaborate on the experiments used to evaluate the proposed algorithm. The test audio files used for the experiments consisted of speech from GRID database [14] resampled to 8 k Hz. The noise signal used is a binaural babble recording from the ETSI database [15], which was recorded with two microphones placed on a dummy head. Binaural noisy signals were generated by convolving the clean speech signal with binaural anechoic head related impulse responses (HRIR) corresponding to ITE HAs obtained from [16] and adding the binaural noise signals . The experiments were performed for different positions of the speakers (the position of the speaker is defined as in Fig. B.4). The speech and noise STP parameters required for the enhancement are estimated every 25 milliseconds, as explained in Section 4. For our experiments, we have used a speech codebook of 64 entries, which was generated using the generalised Lloyd algorithm [17] on a training sample of around 30 minutes of HRIR convolved speech from the specific speaker of interest. Using a speaker specific codebook instead of a general speech codebook leads to improvement in performance, and a comparison between the two has been made in [18]. The HRIR used for convolving the training signal corresponded to zero degrees, whereas the test signals consisted of speech coming form different directions. The noise codebook which consists of only 8 entries, is generated using thirty seconds of noise signal. The audio samples used for training the noise signal was different from audio samples used for testing. The AR order for the speech and noise signal is chosen to be 14. The codebooks as well as MATLAB code for generating the codebooks will be available at `https://tinyurl.com/mskcreatevbn`. We have evaluated the proposed method in terms of the accuracy in the estimation of STP parameters as well as the enhancement performance.

4.1 Accuracy in the estimation of STP parameters

This section evaluates the proposed algorithm in terms of the accuracy in the estimation of STP parameters. Fig. B.5 shows the plots of the true and estimated speech excitation variances (for the left and right channels) for speaker position at 30 degree to the left of the listener at a distance of 80 cm, for a particular test signal. It can be seen that the proposed method captures the difference in speech excitation variances between the two channels. We

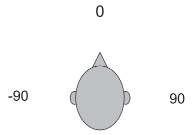


Fig. B.4: Figure showing the top view of the listener. Position of the speaker has been varied for the experiments

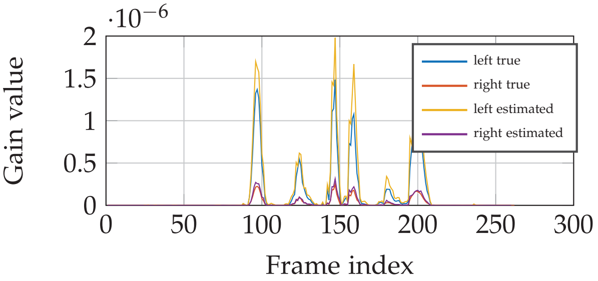


Fig. B.5: Plot of the true and estimated speech excitation variances

now evaluate the ability of the proposed algorithm to deal with changes in the speaker position. For the experiments, the position of the speaker has been varied from -15 degree to 0 degree at frame index 149 and from 0 degree to 10 degree at frame index 285 at a distance of 80 cm. Fig. B.6 shows the estimated value of the scale parameter along the frame index for different speaker positions. It can be seen from Fig. B.6 that the $\hat{c}$ has a value of approximately 0.2 until frame index 141 and then changes to approximately 1 from frame index 149 until 282, and finally changes to 2 from then onwards. The $\hat{c}$ for the first one third portion has a value less than 1 as the speaker is located to the left of the listener. In this case, the level of the signal at the right ear is attenuated in comparison to the level at the left ear, due to the head shadowing effect. For the second portion $\hat{c}$ is approximately 1 as the speaker is located in front of the listener. As the speaker position is changed to 10 degree right of listener, $\hat{c}$ has a value of around 2, as the speaker is closer to the right ear. The position of the speaker can be easily tracked without any delay using the proposed method, as the scale parameter is estimated for every frame index. Moreover, the proposed method does not require the knowledge of the speaker position at any stage to initialise the value of the scale parameter. It should be noted that the scale parameter is relevant only in the speech active regions. Thus, the aberrations present in Fig. B.6 can be explained by the speech being absent in certain time frames.

Next, we compute the total IS divergence between the observed noisy periodograms and the modelled spectrums for test signals taken form the GRID

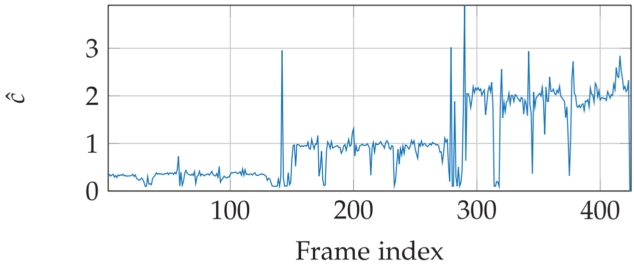


Fig. B.6: Plot of the estimated scale parameter ($\hat{c}$)

database. This measure shows the ability of the estimated parameters to fit the observed noisy spectrum. For this experiment, the position of the speaker is varied around the listener for two different distances at SNR = 5 dB. Table B.1 shows the computed IS divergences for different speaker positions for the proposed method and the method in [7] which we denote as BSTP. It should be noted that the excitation gains in [7] were calculated by minimising an approximate cost function as opposed to here. Thus, to make a fair comparison, we have used the multiplicative update method [11] for computing the excitation variances as used here for [7]. It can be seen that the estimation of the STP parameters using the proposed method leads to a reduced IS divergence between the modelled and the observed spectrums.

Table B.1: Table showing total IS divergence between the modelled noisy spectrum and the observed noisy periodograms (left + right channels) for different speaker positions

		Angle of the speaker			
	Distance (cm)	-85	-75	-65	-55
Proposed	80	**3.61**	**3.75**	**3.73**	**3.65**
	300	**3.62**	**3.73**	**3.72**	**3.62**
BSTP [7]	80	3.98	4.30	4.35	4.20
	300	3.85	4.16	4.25	4.08

4.2 Enhancement performance

We now evaluate the benefit of incorporating the speaker position for enhancement. The frame work that we have used for the experiments is similar to [7] where a fixed lag Kalman smoother is used for enhancement on each channel. Fig. B.7 shows the short-term objective intelligibility (STOI) [19] scores obtained for the two methods when the speaker is at a position of -50 degree at 300 cm. The STOI score shown in the Fig. B.7 corresponds to the score obtained for the better ear. We have compared the propsed method to

BSTP and dual channel speech enhancement method proposed in [4] which we denote here as TwoChSS. It can be seen that taking into account the position of the speaker using the proposed method leads to improvement in the STOI scores especially in low SNR region. It can be seen that TwoChSS degrades the performance of the signal in terms of STOI. This is mainly due to the assumption in TwoChSS that the speaker is in the nose direction of the listener. It should also be noted that the performance of the proposed method and BSTP is similar when the speaker is in the nose direction as $\hat{c} \approx 1$.

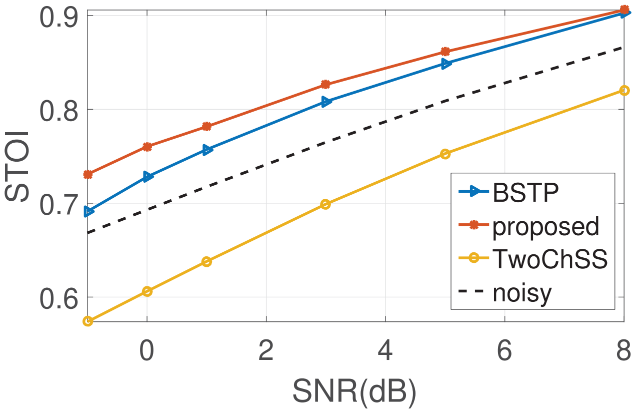


Fig. B.7: Comparison of the STOI scores when the speaker is 50 degrees to the left of the speaker

5 Conclusion

This paper proposed a model based approach for estimating the STP parameters of speech and noise in a binaural framework. The proposed method is able to take into account the position of the speaker while estimating the parameters which leads to an improved modelling of the observed spectrum in comparison to a previous method proposed in [7]. The estimated parameters are subsequently used for enhancement of speech in a binaural framework.

References

[1] T. V. D. Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

[3] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[4] M. Dorbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," in *European Signal Processing Conference, 1996. EUSIPCO 1996. 8th*. IEEE, 1996, pp. 1–4.

[5] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[6] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–14, 2006.

[7] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Binaural speech enhancement using a codebook based approach," *Proc. Int. Workshop on Acoustic Signal Enhancement*, 2016.

[8] ——, "Model based binaural enhancement of voiced and unvoiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2017.

[9] J. A. Grange and J. F. Culling, "The benefit of head orientation to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 703–712, 2016.

[10] ——, "Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4061–4072, 2016.

[11] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 457–468, 2017.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[13] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[15] ETSI202396-1, "Speech and multimedia transmission quality; part 1: Background noise simulation technique and background noise database." 2009.

[16] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–10, 2009.

[17] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[18] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook based approach," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 2016.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

Algorithm 2 Summary of the estimation framework

```
while new time-frames are available do
    for $\forall i \in N_s$ do
        for $\forall j \in N_w$ do
            compute the ML estimates of excitation noise variances and the scale term
            $(\sigma^{2,\mathrm{ML}}_{u,ij}, \sigma^{2,\mathrm{ML}}_{v,ij}, c^{\mathrm{ML}}_{ij})$ using (B.16), (B.17) and (B.18)
            compute the modelled spectrum for left channel $\hat{P}^{\mathrm{ML}}_{z_{l,ij}}$ and right channel
            $\hat{P}^{\mathrm{ML}}_{z_{r,ij}}$ using (B.10) and (B.11) respectively
            compute the likelihood values $p(\mathbf{z}_l, \mathbf{z}_r | \boldsymbol{\theta}^{\mathrm{ML}}_{ij})$ using (B.9)
        end for
    end for
    Get the estimates of STP parameters $(\hat{\sigma}^2_u, \{\hat{a}_i\}^P_{i=1}, \hat{\sigma}^2_v, \{\hat{b}_i\}^Q_{i=1}, \hat{c})$ using (B.5)
end while
```

Paper C

Hearing Aid-Controlled Beamformer for Binaural Speech Enhancement using a Model-based Approach

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Mads Græsbøll Christensen and Jesper B. Boldt

The paper has been submitted to
IEEE Int. Conf. Acoust., Speech, and Signal Processing, 2019

In peer review

The layout has been revised.

Abstract

The understanding of speech from a particular speaker in the presence of other interfering speakers can be severely degraded for a hearing impaired person. Beamforming techniques have been proven to be effective to improve the speech understanding in such scenarios. However, the number of microphones in a hearing aid (HA) is limited due to the space and power constraints present in the HA. In this paper, we propose to use an external device e.g., a microphone array, that can communicate with the HA to overcome this limitation. We propose a method to control this external device based on the look direction of the HA user. We show, by means of simulations, the robustness of the proposed method at very low SNRs in a reverberant scenario. Moreover, we have also conducted experiments that show the benefit of using this framework for binaural and monaural enhancement.

1 Introduction

The ability of a hearing impaired person to understand speech is severely degraded in situations such as the cocktail party scenario, and this can subsequently lead to social isolation of the hearing impaired person. Therefore, it is crucial in such scenarios to perform speech enhancement. The speech enhancement algorithms can be broadly categorised into single and multichannel methods [1]. In comparison to single channel algorithms which can exploit the temporal and spectral information, multi-channel algorithms are more effective as they can also exploit the spatial information. This property is useful in the cases where the speech and noise sources are spatially separated [2]. Beamforming, which forms a class of multi-channel enhancement algorithms, has been proven to be useful as it can exploit the spatial information to selectively attenuate the interferers in comparison to the speaker of interest [2]. The state of the art HAs are equipped with multiple microphones present at each ear which enables the use of beamforming algorithms. However, the space and power constraints on the HAs limit the number of microphones that can be used within a HA which consequently limits the performance of the beamformer within the HA to focus on the speaker of interest and attenuate the competing speakers. These limitations can be overcome by using an external device which can communicate wirelessly with the HA. In [3], it was investigated on how the speech intelligibility can be improved when the target speaker wears a microphone which picks up the speech signal uttered by the target speaker and transmits it wirelessly to the HA. The transmitted signal can then be binaurally saptialised according to the target speaker's location [4–6]. However, this solution has the constraint that the speaker of interest wears the microphone and that the listener is interested only in that speaker. In this paper, we try to relax this constraint by using a microphone array as the external device. Fig. C.1 shows the scenario that we are interested in where the HA user is perhaps interested in listening to any of the speakers located at the table. As the external device is equipped with more microphones, it may be used to better exploit the spatial information for focusing on the speaker the HA user is listening to. However, a problem that can be encountered in this setup is that the external microphone array needs to know the direction of arrival of the source that the HA user is interested in. In this work, we propose a model-based method to estimate the direction of arrival via a collaboration between the HA and the external device. We use a model that we proposed in a previous paper [7] to represent the signal received at the HA as well as the external device. The estimated model parameters are subsequently used to estimate the direction of arrival by measuring the similarity between the model parameters. Using a model to represent the signal facilitates a low dimensional representation of the signal, which leads to less information being transmitted from the HA to the external device which is critical as power is a limiting factor in the HAs. To the best of author's knowledge, such an approach to control an external device based on the look direction of the HA user has not been done before.

The remainder of the paper is organised as follows. Section 2 introduces the setup, the signal model and the problem mathematically. The solution to the defined problem is then explained in Section 3 followed by the results and conclusion in Sections 4 and 5, respectively.

2 Problem Formulation

2.1 Scenario of interest

Fig. C.1 shows an example of the scenario of interest. This situation can be encountered when the HA user is participating in a meeting with colleagues or sitting at a dinner table with family or friends. From the figure, it can be seen that the HA user is listening to speaker A. It will be assumed in this work that the HA user is looking at the source of interest. From the perspective of the HA user, speaker A is the target whereas speakers B and C are interferers. The objective is to focus the beamformer present in the external device (uniform circular array (UCA) in this case) towards speaker A as the HA user is looking towards speaker A. This requires a communication link between the HA and the external device. To compute the data to be transmitted from the HA to the external device, a conventional beamformer focusing on the nose direction is used on the HA to form a preliminary enhanced signal which we denote as $x_{\text{HA}}(n)$. To estimate the source of interest, the UCA focuses its beam towards I different candidate directions ($I = 8$ in the figure) which are uniformly spaced along the azimuthal plane as shown

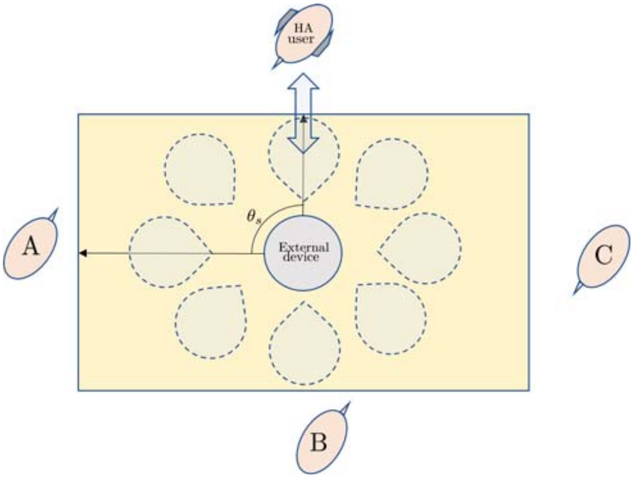


Fig. C.1: Scenario of interest. HA user is listening to speaker A while speakers B and C are interferers.

in the figure. The set of beamformed signals from the different candidate directions are denoted as $\{x_{\mathrm{CA}(\theta_i)}(n)\}_{i=1}^{I}$. Our objective here is to estimate θ_s (see Fig. C.1) such that the UCA focuses on the source of interest, using the beamformed signal at the HA, $x_{\mathrm{HA}}(n)$, and the set of beamformed signals from the candidate directions, $\{x_{\mathrm{CA}(\theta_i)}(n)\}_{i=1}^{I}$.

2.2 Signal Model

We now introduce the signal model [7] that is used to represent the signals $x_{\mathrm{HA}}(n)$ and $\{x_{\mathrm{CA}(\theta_i)}(n)\}_{i=1}^{I}$. A frame of the signal of interest in the time domain denoted as $\mathbf{x} = [x(0) \ldots x(N-1)]^T$ is modelled as a sum of U autoregressive (AR) processes as

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{c}_u \ . \tag{C.1}$$

Each of the AR process can be expressed as a multivariate Gaussian [8, 9], i.e. ,

$$\mathbf{c}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{Q}_u), \tag{C.2}$$

where σ_u^2 is the excitation variance and $\mathbf{Q}_u$ is the gain normalised covariance matrix. $\mathbf{Q}_u$ can be asymptotically approximated as a circulant matrix which can be further diagonalised using by the Fourier transform as [10]

$$\mathbf{Q}_u = \mathbf{F}\mathbf{D}_u\mathbf{F}^H \tag{C.3}$$

where $\mathbf{F}$ is the DFT matrix defined as $[\mathbf{F}]_{k,n} = \frac{1}{\sqrt{N}}\exp(j2\pi nk/N)\ n,k = 0 \ldots N-1$ and

$$\mathbf{D}_u = (\mathbf{\Lambda}_u^H \mathbf{\Lambda}_u)^{-1}, \qquad \mathbf{\Lambda}_u = \text{diag}(\sqrt{N}\mathbf{F}^H \begin{bmatrix} \mathbf{a}_u \\ \mathbf{0} \end{bmatrix}) \tag{C.4}$$

where $\mathbf{a}_u = [1, a_u(1), \ldots a_u(P)]^T$ represents the vector of AR coefficients corresponding to the u^{th} AR process and P is the AR order. The diagonal entries of the matrix $\mathbf{D}_u$ contains the eigenvalues of the matrix $\mathbf{Q}_u$ and these correspond to the power spectral density (PSD) of the u^{th} gain normalised AR process. The set of U PSDs can be arranged as the columns of a spectral basis matrix $\mathbf{D}$ as

$$\mathbf{D} = [\mathbf{d}_1 \ldots \mathbf{d}_u \ldots \mathbf{d}_U] \tag{C.5}$$

where $\mathbf{d}_u = [d_u(1) \ldots d_u(k) \ldots d_u(K)]^T$ and $d_u(k)$ is the k^{th} diagonal element of $\mathbf{D}_u$. Using the above model explained by (C.1) and (C.2), a frame of the beamformed signal at the HA denoted as $\mathbf{x}_{\text{HA}} = [x_{\text{HA}}(0) \ldots x_{\text{HA}}(N-1)]^T$ is expressed as

$$\mathbf{x}_{\text{HA}} = \sum_{u=1}^{U} \mathbf{c}_{\text{HA}_u} \tag{C.6}$$

where $\mathbf{c}_{\text{HA}_u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{HA}_u}\mathbf{Q}_u)$. Denoting $\boldsymbol{\sigma}_{\text{HA}} = [\sigma^2_{\text{HA}_1} \ldots \sigma^2_{\text{HA}_U}]^T$, the PSD of the modelled signal at the HA can be represented as $\mathbf{D}\boldsymbol{\sigma}_{\text{HA}}$. Similarly a frame of the beamformed signal at the UCA for the ith candidate direction denoted as $\mathbf{x}_{\text{CA}(\theta_i)} = [x_{\text{CA}(\theta_i)}(0), \ldots x_{\text{CA}(\theta_i)}(N-1)]^T$ can be modelled as

$$\mathbf{x}_{\text{CA}(\theta_i)} = \sum_{u=1}^{U} \mathbf{c}_{\text{CA}(\theta_i)_u} \tag{C.7}$$

where $\mathbf{c}_{\text{CA}(\theta_i)_u} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{CA}(\theta_i)_u}\mathbf{Q}_u)$. Denoting $\boldsymbol{\sigma}_{\text{CA}(\theta_i)} = [\sigma^2_{\text{CA}(\theta_i)_1} \ldots \sigma^2_{\text{CA}(\theta_i)_U}]^T$, the PSD of the modelled signal at the UCA for the ith candidate direction is obtained as $\mathbf{D}\boldsymbol{\sigma}_{\text{CA}(\theta_i)}$. In the case of observing $V > 1$ frames, the PSD of the modelled signal for each frame can be arranged as columns to form a matrix $\mathbf{D}\mathbf{\Sigma}_{\text{HA}}$ or $\mathbf{D}\mathbf{\Sigma}_{\text{CA}(\theta_i)}$ where $\mathbf{\Sigma}_{\text{HA}} = [\boldsymbol{\sigma}_{\text{HA}}(1) \ldots \boldsymbol{\sigma}_{\text{HA}}(V)]$ and $\mathbf{\Sigma}_{\text{CA}(\theta_i)} = [\boldsymbol{\sigma}_{\text{CA}(\theta_i)}(1) \ldots \boldsymbol{\sigma}_{\text{CA}(\theta_i)}(V)]$. $\mathbf{\Sigma}_{\text{HA}}$ and $\mathbf{\Sigma}_{\text{CA}(\theta_i)}$ will be henceforth denoted as activation coefficients.

2.3 Mathematical problem

Our objective here is to estimate the direction of arrival of the speaker talking to the HA user relative to the UCA. In this paper, we propose to solve this problem by measuring the similarity between the beamformed signal received at the HA and the beamformed signal at the UCA for different candidate directions. This can be done in different ways. The first method, denoted as IS based method, is by using the spectral similarity to estimate θ_s as

$$\hat{\theta}_s = \underset{\theta_i \in \{\theta_i\}_{i=1}^{I}}{\arg\min} \quad d_{\text{IS}}(\mathbf{D}\mathbf{\Sigma}_{\text{HA}}|\mathbf{D}\mathbf{\Sigma}_{\text{CA}(\theta_i)}), \tag{C.8}$$

where $d_{\text{IS}}(\cdot,\cdot)$ is the Itakura-Saito divergence [11]. The second approach proposed here is to use the correlation between the estimated activation coefficients at the HA and UCA to estimate θ_s as

$$\hat{\theta}_s = \underset{\theta_i \in \{\theta_i\}_{i=1}^{I}}{\arg\max} \quad \text{corr}(\mathbf{\Sigma}_{\text{HA}}|\mathbf{\Sigma}_{\text{CA}(\theta_i)}) \tag{C.9}$$

where

$$\begin{aligned}\text{corr}(\mathbf{\Sigma}_{\text{HA}}|\mathbf{\Sigma}_{\text{CA}(\theta_i)}) = \frac{1}{UV}\sum_{v=1}^{V}\sum_{u=1}^{U}&\left(\frac{\mathbf{\Sigma}_{\text{HA}}(u,v)-\mu_{\text{HA}}}{\epsilon_{\text{HA}}}\right)\\ &\times\left(\frac{\mathbf{\Sigma}_{\text{CA}(\theta_i)}(u,v)-\mu_{\text{CA}(\theta_i)}}{\epsilon_{\text{CA}(\theta_i)}}\right)\end{aligned} \tag{C.10}$$

where $\mu_{\text{HA/CA}(\theta_i)}$ and $\epsilon_{\text{HA/CA}(\theta_i)}$ are the sample mean and standard deviation, respectively.

3 Estimation of the model parameters

As explained previously in Section 2.2, a frame of the signal is modelled as a sum of U AR processes with AR coefficients $\mathbf{a}_u$. In this work, the set of U AR coefficients are trained a priori using a standard vector quantisation technique used in speech coding applications. During the training stage, a speech codebook is first computed using the generalised Lloyd algorithm (GLA) [8, 12]. The speech codebook contains AR coefficients corresponding to the spectral envelopes of speech. In the training process, linear prediction coefficients which are transformed into line spectral frequency coefficients are extracted from windowed frames, from the training signal and subsequently passed as the input data to the vector quantiser. Once the speech codebook is created, the spectral envelopes corresponding to the AR coefficients ($\{\mathbf{a}_u\}_{u=1}^{U}$) are computed and arranged as columns of the spectral basis matrix $\mathbf{D}$ as explained by (C.4) and (C.5). Given the observed data and the spectral basis matrix $\mathbf{D}$, it has been shown in [7] that the maximum likelihood estimation of the activation coefficients corresponds to minimising the IS divergence between the periodogram of the observed signal and the modelled PSD. Since there is no closed form solution for this, it is generally estimated iteratively using the multiplicative update (MU) rule [13] as

$$\hat{\mathbf{\Sigma}}_{\text{HA/CA}(\theta_i)} \leftarrow \hat{\mathbf{\Sigma}}_{\text{HA/CA}(\theta_i)} \odot \frac{\mathbf{D}^T\big((\mathbf{D}\hat{\mathbf{\Sigma}}_{\text{HA/CA}(\theta_i)})^{[-2]} \odot \mathbf{\Phi}_{\text{HA/CA}(\theta_i)}\big)}{\mathbf{D}^T(\mathbf{D}\hat{\mathbf{\Sigma}}_{\text{HA/CA}(\theta_i)})^{[-1]}}, \tag{C.11}$$

Algorithm 3 Main steps involved in the proposed framework

```
while new time-frames are available do
    Apply beamforming in the HA as well as the UCA for different candidate directions to obtain $\mathbf{x}_{\text{HA}}$ and $\mathbf{x}_{\text{CA}(\theta_i)}$
    Assuming the spectral basis matrix $\mathbf{D}$ is trained a priori, estimate $\mathbf{\Sigma}_{\text{HA}}$ and $\mathbf{\Sigma}_{\text{CA}(\theta_i)}$ using (C.11)
    Transmit the estimated activation coefficients $\hat{\mathbf{\Sigma}}_{\text{HA}}$ from the HA to external device
    Estimate $\theta_s$ using (C.8) or (C.9)
    Use the beamformed signal from the UCA as a reference signal for performing binaural enhancement
end while
```

where $\mathbf{\Phi}_{\text{HA}/\text{CA}(\theta_i)}$ contains the periodogram of frames of the signals arranged as columns, $\odot$ is the element-wise product and the division is an element-wise division. The spectral basis matrix along with the estimated activation coefficients can be utilised as shown in (C.8) and (C.9) to estimate the direction of arrival to control the beam pattern of the UCA. It should be noted that the method proposed here to estimate θ_s requires only the transmission of $\hat{\mathbf{\Sigma}}_{\text{HA}}$ from the HA to the external device. This is generally much less than the amount of signal samples. The proposed algorithm is summarised in Algorithm 3.

4 Experiments

4.1 Experimental setup

This section describes the experimental results obtained for the proposed method. The setup used for carrying out the experiments will be explained in this section followed by the results. Fig. C.2 shows a portion of the experimental setup in a room of dimensions $12 \times 6 \times 6$ m with a reverberation time of 0.4 seconds. The room impulse responses were generated using [14]. In this figure, the HA user is trying to listen to the target speaker denoted by a green dot. Along with the speaker of interest, there are 3 other interferers located around the table as shown in the figure. The HA user is simulated with a ULA of 5 microphones with a span of 0.24 m which is a typical ear to ear distance of a human head. The external device that we have considered here is a UCA of radius of 0.13 m with 10 microphones. The signals used for testing consisted of speech signals spoken by two males and two females taken from the CHIME database [15]. A codebook of size 64 entries was generated using the GLA using speech from the EUROM database [16]. We have used

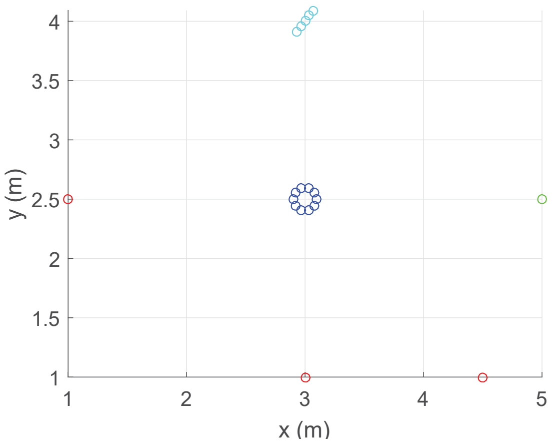


Fig. C.2: Figure shows the experimental setup where the green dot indicates the source of interest and the red dot indicates the interferers.

different databases for testing and training to test the robustness of the proposed method against the mismatches that maybe encountered in a practical scenario. The parameters used for the experiments have been summarised in Table C.1. The beamformed signals at the HA and the external device can be obtained using any of the conventional beamforming algorithms. For the experiments conducted in this section, we have used the robust Capon beamforming (RCB) [17, 18], as this method has been shown to be robust to reverberation and uncertainty in the steering vector [18]. The number of candidate directions I has been chosen to be 8 in the experiments. It should be noted that speakers are not constrained to be at the candidate positions as the RCB takes into account uncertainties in the steering vector using the parameter ϵ [17, eq. (14)] which was chosen to be 3.5 in our experiments. The experiments we have conducted to validate the robustness of the proposed method is shown in the following section.

4.2 Experimental Results

In this section we evaluate the accuracy of the proposed method in the experimental setup explained above. In addition to the case shown in Fig. C.2, we have varied the source of interest, so the results shown in this section are averaged over all the speakers (4 in this case). In addition to the interferers, we also add spherically isotropic babble ambient noise generated using the implementation in [19]. Figures C.3 and C.4 show the average accuracy obtained over all the speakers for 10 iterations per speaker as a function of SNR and the memory for correlation based and IS based methods, respectively.

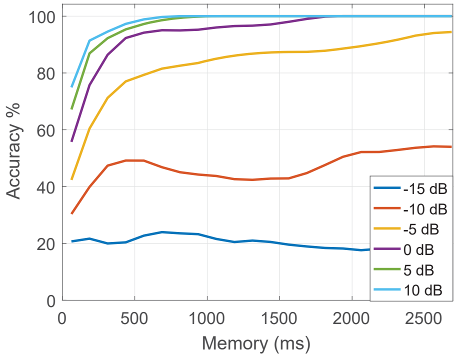


Fig. C.3: Percentage of correctly detected direction for the correlation based method as a function of the SNR and memory.

Memory here is related to the number of frames V (used for the computation of the model parameters) that is needed to estimate θ_s. It can be seen from the figure that the correlation based method converges to 100 % accuracy for SNRs 0 dB and above when a memory greater than 1700 ms is used for computation of the model parameters, whereas the IS based method converges to 100 % accuracy for SNRs 5 dB and above when a memory of 2200 ms is used. It should be noted that the experiments conducted in this section, assumed the positions of the HA user, the target speaker and the interferers to be stationary. However, in practical scenarios it may be useful to update the result at much finer time scale, as the HA user may continously change the look direction. The influence of memory has also been investigated in figures C.3 and C.4 and it can be seen that as the SNR increases the memory required for the proposed method to obtain certain accuracy decreases, e.g., to reach 80 % accuracy, it requires a memory of approximately 100, 240 and 600 ms for SNRs 10, 0 and -5 dB, respectively for the correlation based method.

Table C.1: Parameters used for the experiments

Parameters	
sampling frequency	8000 Hz
Frame size (N)	200
Frame overlap	50%
AR order (P)	14
U	64
MU iterations	50
I	8

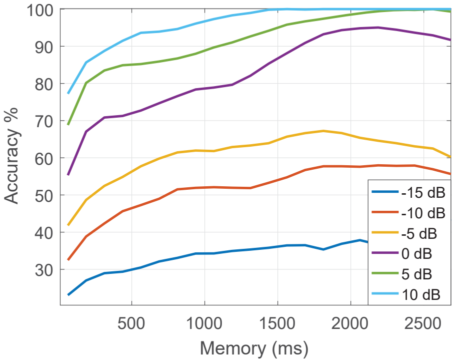


Fig. C.4: Percentage of correctly detected direction for the IS based method as a function of the SNR and memory.

Enhancement performance

In this section, we show an example of how this setup can be used for monaural/binaural enhancement in HAs. One option is to wirelessly transmit the beamformed signal at the external device to the HA and play back that signal at both the ears. Playing back the monaural signal, however, may lead to distortion of binaural cues. Thus in this paper, we also perform binaural speech enhancement, where we consider the signals received at the left and rightmost microphones as the binaural noisy signal. To perform the enhancement we consider the binaural enhancement framework proposed in [20], which is based on the MMSE criterion. This method applies a common gain on the left and right channels which leads to the preservation of the binaural cues. The common gain applied in this case requires the estimation of the speech/noise statistics [20, eq. (17)]. In this work, we propose to use the beamformed signal at the UCA from the estimated direction to be used as the reference signal to estimate the clean speech statistics. Fig. C.5 shows the averaged short time objective intelligibilty (STOI) [21] scores for the left and right channels obtained for the different configurations. The beamformed signals at the HA and UCA are denoted as mono-HA and mono-CA, respectively. The binaural enhancement method where we use the beamformed signals at the HA and UCA to estimate the clean speech statistics is denoted as Bin-HA and Bin-CA, respectively. It can be seen that using the beamformed signal at the UCA shows an improvement in both the binaural and monaural configurations.

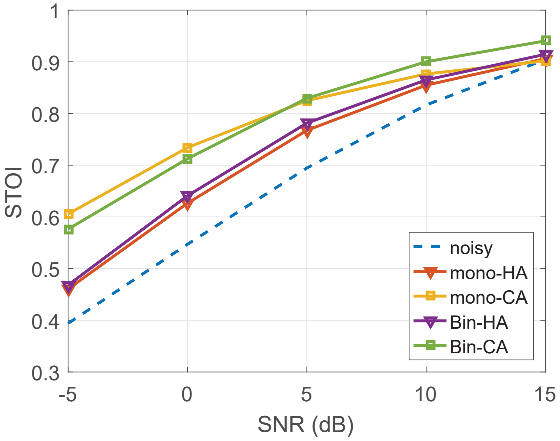


Fig. C.5: Average STOI score obtained for the left and right channels for binaural and monaural configurations.

5 Conclusion

In this paper we have proposed a framework for improving the speech understanding for HA users in the presence of multiple interferers. The proposed system consisted of using an external microphone array whose beam-pattern is controlled by the look direction of the HA user using a model-based approach. The robustness of the proposed method at very low SNRs in a reverberant scenario has been shown by the means of simulations. Moreover, the benefits of using the external device in addition to the HA for performing binaural enhancement has been shown using an objective measure for intelligibility.

References

[1] J Benesty, S Makino, and J Chen, "Speech enhancement, ser. signals and communication technology," 2005.

[2] S. Doclo, S. Gannot, M. Moonen, and A Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

[3] M. S. Lewis, C. C. Crandell, M. Valente, and J. E. Horn, "Speech perception in noise: Directional microphones versus frequency modulation (fm) systems," *Journal of the American Academy of Audiology*, vol. 15, no. 6, pp. 426–439, 2004.

[4] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

[5] J. M. Kates, K. H. Arehart, R. K. Muralimanohar, and K. Sommerfeldt, "Externalization of remote microphone signals using a structural binaural model of the head and pinna," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 2666–2677, 2018.

[6] M. Farmani, M. S. Pedersen, and J. Jensen, "Sound source localization for hearing aid applications using wireless microphones," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. IEEE, 2018, pp. 455–459.

[7] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric nmf for speech enhancement," in *Proceedings of the European Signal Processing Conference*, 2018.

[8] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[9] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model-based noise psd estimation from speech in non-stationary noise," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2018.

[10] R. M Gray et al., "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[11] F. Itakura, "Analysis synthesis telephony based on the maximum likelihood method," in *The 6th international congress on acoustics, 1968*, 1968, pp. 280–292.

[12] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[14] E. A. P. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[15] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ’chime’ speech separation and recognition challenge: Dataset, task and baselines,” *IEEE 2015 Automatic Speech Recognition and Understanding Workshop*, 2015.

[16] D. Chan, A. Fourcin, D. Gibbon, B Granstrom, et al., “Eurom-a spoken language resource for the eu,” in *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech’95*, 1995, pp. 867–880.

[17] J. Li, P. Stoica, and Z. Wang, “On robust capon beamforming and diagonal loading,” *IEEE transactions on signal processing*, vol. 51, no. 7, pp. 1702–1715, 2003.

[18] Y. Zhao, J. R. Jensen, M. G. Christensen, S. Doclo, and J. Chen, “Experimental study of robust beamforming techniques for acoustic applications,” in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 86–90.

[19] E. A. P. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[20] G. Enzner, M. Azarpour, and J. Siska, “Cue-preserving mmse filter for binaural speech enhancement,” in *Acoustic Signal Enhancement (IWAENC), 2016 IEEE International Workshop on*. IEEE, 2016, pp. 1–5.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

Paper D

Model-based Noise PSD Estimation from Speech in Non-stationary Noise

Jesper Kjær Nielsen, Mathew Shaji Kavalekalam,
Mads Græsbøll Christensen and Jesper B. Boldt

The paper has been published in the
Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, 2018

Abstract

Most speech enhancement algorithms need an estimate of the noise power spectral density (PSD) to work. In this paper, we introduce a model-based framework for doing noise PSD estimation. The proposed framework allows us to include prior spectral information about the speech and noise sources, can be configured to have zero tracking delay, and does not depend on estimated speech presence probabilities. This is in contrast to other noise PSD estimators which often have a too large tracking delay to give good results in non-stationary situations and offer no consistent way of including prior information about the speech or the noise type. The results show that the proposed method outperforms state-of-the-art noise PSD estimators in terms of tracking speed and estimation accuracy.

1 Introduction

The healthy human auditory system has a remarkable ability to extract the desirable information from a noisy speech signal. Even in situations such as a cocktail party where the background noise is non-stationary and the signal-to-noise ratio (SNR) is very low, normal hearing people are not only able to cope with the situation, but able to enjoy it. For people with a hearing defect, however, noisy situations such as a cocktail party are often mentally fatiguing and very challenging to deal with. These hearing impaired people often rely on a hearing aid for the speech enhancement, but the performance of the current hearing aid technology is far from enabling its users to thrive in difficult situations such as a cocktail party. Speech enhancement is not only important to the hearing impaired person in a cocktail party situation, but in any situation where the desired speech is observed in noise. Moreover, not only humans benefit from speech enhancement since, e.g., speaker identification and speech recognition algorithms are often designed for a clean speech signal [1].

Any speech enhancement algorithm must incorporate some prior knowledge in order to successfully separate the desired speech from the unwanted background noise. For example, the popular Wiener filter and many other speech enhancement algorithms such as maximum SNR, MVDR, and LCMV [1] (see also [2] for a comparison) assume that the second-order statistics of the speech and/or noise are known somehow. In practice, however, the statistics is often unknown and time-varying. Therefore, the prior knowledge must be represented in an alternative way so that the statistics can be estimated directly from the noisy speech. In this paper, we make contributions to the solution of exactly this problem.

Many people have been analysing the problem of estimating the noise power spectral density (PSD) or, equivalently, the second-order noise statistics, from a noisy speech signal. The most basic approach to estimating the noise PSD has been to use a voice activity detector (VAD) to inform the estimation algorithm about when speech is absent so that the noise PSD can be estimated. Unfortunately, such VADs are often difficult to tune in low SNRs, and they do not work well when the noise is non-stationary [3, 4]. Moreover, they are inefficient since they typically disable the noise PSD estimator across the entire frequency range, even if speech is only present in a few frequency bands. This has motivated the use of a soft VAD in each frequency band. A prominent example of this is the minimum statistics (MS) method [3, 5]. The algorithm is built on the assumption that the noise PSD is slowly varying with time and that the power of the noisy signal frequently goes down to the noise power level. Although the MS principle is simple, a lot of heuristics go into estimating a very important smoothing parameter and to correct the negative bias of the estimator. In fact, a full journal paper has been published on the latter issue [6]. Other problems with the principle are that the variance of the estimated noise PSD is bigger than for other methods [3, 4] and that very long tracking delays can occur, in particular when the noise power is increasing. Precisely these two issues were addresses in the MCRA [7–9] and later in the improved MCRA (IMCRA) [4] methods. Unfortunately, however, there might still be a considerable tracking delay in IMCRA if the noise power is increasing [10] and a lot of hand-crafting is still involved in tuning the algorithm and in doing bias correction. In [10, 11], the MS principle was abandoned in favour of MMSE estimators. These MMSE estimators were demonstrated to have a much better tracking speed than the MS and IMCRA methods and can be considered to be the best noise tracker currently [12]. One of the disadvantages of the MMSE estimators is that the first five time frames are assumed to be noise only to initialise the tracker. Another disadvantage is that it is not clear what prior information is actually built into the MMSE estimators about the speech and the noise, besides that the speech and noise spectral coefficients are modelled as independent and normally distributed random variables. This model assumption is very common in noise PSD estimation, but does not by itself enable us to separate a mixture into its components. Additional prior information is, therefore, necessary to find a unique solution to the problem, but the current noise trackers often rely on heuristic tricks for making the problem solvable rather than explicitly stating the model assumptions. Approaches based on, e.g., vector Taylor series [13] or nonnegative matrix factorisations (NMF) [14] give such model based estimates of the noise statistics via a separate training step. The clear advantage of these approaches is that it is much easier to understand the applicability and limitations of the model and, consequently, the noise PSD estimator. Moreover, we do not have to compensate for artefacts such as an unwanted bias, and we can change the built-in prior information via the model. For example, a hearing aid user often communicate with the same people, but such information cannot be built into current noise PSD trackers.

In this paper, we propose a new noise PSD estimator which has some resemblance to both the NMF approach and the MMSE estimators. However, we derive our estimator directly in a flexible statistical framework which can be used in situations where we have specific prior information, but also in situations where we do not. By virtue of being model-based, we can in principle also use the proposed framework for noise PSD estimation with no tracking delay, even if speech is continuously present.

2 The Estimation Problem and the Model

We assume that we observe N samples from the noisy speech signal

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{e} \tag{D.1}$$

where $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{s} \in \mathbb{R}^{N\times 1}$, and $\boldsymbol{e} \in \mathbb{R}^{N\times 1}$ are the noisy speech, the clean speech, and the noise, respectively. Given $\boldsymbol{y}$, we seek to estimate the noise PSD which is typically defined as [15, p. 7]

$$\phi_{\mathrm{e}}(\omega) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[|E(\omega)|^2|\boldsymbol{y}\right] \tag{D.2}$$

where $\mathbb{E}$ is the expectation operator and $E(\omega) = \boldsymbol{f}^H(\omega)\boldsymbol{e}$ is the DFT of the noise with $\boldsymbol{f}(\omega) = \{\exp(j\omega n)\}_{n=0,\ldots,N-1}$. The conditional expectation in (D.2) is the second moment of the density $p(E(\omega)|\boldsymbol{y})$. However, it can also be written in terms of the density $p(\boldsymbol{e}|\boldsymbol{y})$ as

$$\mathbb{E}\left[|E(\omega)|^2|\boldsymbol{y}\right] = \boldsymbol{f}^H(\omega)\left[\int_{\mathbb{R}^{N\times 1}} \boldsymbol{e}\boldsymbol{e}^T p(\boldsymbol{e}|\boldsymbol{y})d\boldsymbol{e}\right]\boldsymbol{f}(\omega)\ . \tag{D.3}$$

The problem of estimating the noise PSD is, therefore, essentially that of computing the second moment of the posterior $p(\boldsymbol{e}|\boldsymbol{y})$. This might seem counter-intuitive since noise PSD estimation is usually pitched in the context of speech enhancement, but the above form of the PSD reveals that we actually have to decompose the noisy speech into its components before the noise PSD can be estimated. Nevertheless, speech enhancement is not only a mathematical problem of minimising some objective function, but often the art of improving highly subjective measures such as speech intelligibility and speech quality, so noise PSD estimation is still useful in many applications.

To compute the posterior $p(\boldsymbol{e}|\boldsymbol{y})$, we elicit several statistical models $\{\mathcal{M}_k\}_{k=1}^K$ for how the data vector $\boldsymbol{y}$ was generated. Such models can easily be included in (D.3) as

$$\mathbb{E}\left[|E(\omega)|^2|\boldsymbol{y}\right] = \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{y})\,\mathbb{E}\left[|E(\omega)|^2|\boldsymbol{y},\mathcal{M}_k\right]\ . \tag{D.4}$$

Thus, we obtain a model averaged noise PSD estimator if we insert (D.4) in (D.2). The model probabilities $\{p(\mathcal{M}_k|\boldsymbol{y})\}_{k=1}^{K}$ ensure that those models which explain the data well will contribute with a larger weight than those models which do not explain the data well. In principle, there are no limits on which models can be used. From a practical perspective, however, it is advantageous to use models that lead to tractable algorithms while still being a sufficiently accurate representation of how the speech and the noise were generated. In this paper, we will use autoregressive processes to model the speech and the noise, i.e.,

$$p(\boldsymbol{s}|\sigma^2_{\mathrm{s},k}, \mathcal{M}_k) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\mathrm{s},k}\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)) \tag{D.5}$$

$$p(\boldsymbol{e}|\sigma^2_{\mathrm{e},k}, \mathcal{M}_k) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\mathrm{e},k}\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)) \tag{D.6}$$

where $\sigma^2_{\mathrm{s},k}$, $\sigma^2_{\mathrm{e},k}$, $\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)$, $\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)$, $\boldsymbol{a}_k$, and $\boldsymbol{b}_k$ are the excitation noise variances, the normalised covariance matrices, and the AR-parameters of the speech and the noise, respectively. We assume that the AR-processes are periodic in N since the normalised covariance matrices then are diagonalised by the DFT matrix $\boldsymbol{F}$. That is,

$$\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k) = N^{-1}\boldsymbol{F}\boldsymbol{D}_{\mathrm{s}}(\boldsymbol{a}_k)\boldsymbol{F}^H \tag{D.7}$$

$$[\boldsymbol{F}]_{nl} = \exp(j2\pi(n-1)(l-1)/N)\ , \quad n,l = 1,\ldots,N \tag{D.8}$$

$$\boldsymbol{D}_{\mathrm{s}}(\boldsymbol{a}_k) = \left(\boldsymbol{\Lambda}_{\mathrm{s}}^H(\boldsymbol{a}_k)\boldsymbol{\Lambda}_{\mathrm{s}}(\boldsymbol{a}_k)\right)^{-1} \tag{D.9}$$

$$\boldsymbol{\Lambda}_{\mathrm{s}}(\boldsymbol{a}_k) = \mathrm{diag}(\boldsymbol{F}^H \begin{bmatrix}\boldsymbol{a}_k^T & \boldsymbol{0}\end{bmatrix}^T) \tag{D.10}$$

with similar definitions for $\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)$. Although it might seem unfounded to assume periodicity in N, this assumption is actually implicitly made when using the asymptotic covariance matrix of an AR-process for finite length signals as in [16] or when interpreting the Itakura-Saito (IS) distortion measure [17, 18] as the maximum likelihood estimator of short-time speech spectra. Precisely the IS distortion measure has been very popular in the speech community for decades, partly due to it also being a perceptually meaningful distortion measure [19], and has lately also been used successfully as a distortion measure for nonnegative matrix factorisation (NMF) [20]. Moreover, the above model actually has the signal model used in [10, 11] as a special case. Specifically, if we select $K = 1$ and set the AR-orders to $N-1$, then the speech and noise spectral coefficients are modelled as independent and normally distributed random variables and the noise PSD estimator in (D.2) is the foundation of the MMSE-estimators in [10, 11]. As discussed in the introduction, however, this frequency domain model does not by itself allow us to separate the noisy mixture into its components. , so we have to introduce additional prior information into the model to separate the two components.

2.1 Prior Information

Inspired by the work in [16, 21], the AR-parameters are here assumed known for a given model. Thus, a model in our framework corresponds to one combination of so-called codebook entries in the framework of [16, 21]. That is, if we have a speech and a noise codebook consisting of K_s and K_e trained AR-vectors, respectively, we have a total of $K = K_\mathrm{s}K_\mathrm{e}$ models[1]. At first glance, it might seem a disadvantage that these codebooks have to be trained, but they actually offer an excellent way of including prior spectral information. For example, if the noise PSD estimator has to operate in a particular noise environment such as a car cabin or mainly process speech from a single person such as in mobile telephony, we can use codebooks with typical normalised AR-spectra for these sources. Conversely, in the absence of any specific information about the speaker(s) and the noise environment(s), we can use classified codebooks [16] where we first classify the speaker/noise type and then use the corresponding codebooks which have been trained on different speakers and noise types. Moreover, the noise PSD estimate from any noise tracker can also be included as a noise codebook vector. This also means that the proposed framework can be used to combine existing noise trackers in a consistent fashion. A potential problem of the model-based approach is that the number of models grows with the product of the codebook sizes, and this might lead to an intractable computational complexity. This is also one of the reasons why we use models whose covariance matrices can be diagonalised by the DFT matrix.

The excitation noise variances are not pre-trained, but are treated as unknown random variables with the prior

$$p(\sigma^2_{\mathrm{s},k}|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{s},k}, \beta_{\mathrm{s},k}) \tag{D.11}$$

$$p(\sigma^2_{\mathrm{e},k}|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{e},k}, \beta_{\mathrm{e},k}) \tag{D.12}$$

where Inv-$\mathcal{G}(\cdot,\cdot)$ denotes the inverse Gamma density. Similarly, we also have a prior mass function $p(\mathcal{M}_k)$ for the models. Speech is normally processed frame-by-frame, often with some overlap. Consequently, values for the excitation noise variances and models that work well in one frame, should also work reasonably well in the next frame, and the priors are an excellent tool for using previous information in the current frame. In a completely stationary environment, for example, the posterior distribution of one frame should be the prior distribution in the next frame. The more non-stationary the signals are, the broader the prior of the current frame should be com-

[1] Note that a codebook is not restricted to only include AR-spectra, but can in principle include any type of spectrum as in NMF. We here focus on a parametric representation of the spectra in terms of AR-parameters since this leads to codebooks with a small memory footprint, can be used for short segment sizes, and allows us to train the codebooks using standard vector quantisation techniques developed for speech coding [22].

pared to the posterior of the previous frame. In the limit, no information is carried over from one frame to the next, and we use uninformative priors with $\alpha_{\cdot,k} = \beta_{\cdot,k} \to 0$ and $p(\mathcal{M}_k) = K^{-1}$. In this paper, we focus on exactly this limiting case in the simulations. Besides not having enough space here to give a complete description of a general frame transition model, this choice is motivated by that 1) babble noise is typically very non-stationary, and 2) we wish to demonstrate that the proposed model-based approach works well, even without any smoothing between frames. This is in contrast to current state-of-the-art noise trackers which at best have a tracking delay of a few hundred milliseconds [10]. Before going to the simulations, however, we first describe how the noise PSD is estimated from the model and the data.

3 Inference

To estimate the noise PSD, we have to compute the posterior model probabilities $p(\mathcal{M}_k|\boldsymbol{y})$ as well as the second moment of the posterior $p(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ (see (D.4)) by combining the information in the data with the prior information. Unfortunately, neither of these posteriors exist in closed-form, and we, therefore, have to content ourselves with either analytical or stochastic approximations. For our inference problem, the variational Bayesian (BS) framework [23, 24] produces a simple analytical approximation if we assume that the full joint posterior factorises as

$$p(\boldsymbol{e}, \sigma^2_{\text{s},k}, \sigma^2_{\text{e},k}|\boldsymbol{y}, \mathcal{M}_k)p(\mathcal{M}_k|\boldsymbol{y}) \approx \\ q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)q(\sigma^2_{\text{s},k}, \sigma^2_{\text{e},k}|\boldsymbol{y}, \mathcal{M}_k)q(\mathcal{M}_k|\boldsymbol{y}) \ . \quad \text{(D.13)}$$

Unfortunately, the derivation of the three factors in the approximation is lengthy so we only state the results here and refer the interested reader to a supplementary document for a detailed derivation (available at `http://tinyurl.com/jknvbn`). From the derivation, we obtain that the posterior factor $q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ is given by

$$q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k) = \mathcal{N}(\hat{\boldsymbol{e}}_k, \hat{\boldsymbol{\Sigma}}_k) \quad \text{(D.14)}$$

where

$$\hat{\boldsymbol{\Sigma}}_k = \left[\frac{a_{\text{s},k}}{b_{\text{s},k}}\boldsymbol{R}_\text{s}^{-1}(\boldsymbol{a}_k) + \frac{a_{\text{e},k}}{b_{\text{e},k}}\boldsymbol{R}_\text{e}^{-1}(\boldsymbol{b}_k)\right]^{-1} \quad \text{(D.15)}$$

$$\hat{\boldsymbol{e}}_k = \frac{a_{\text{s},k}}{b_{\text{s},k}}\hat{\boldsymbol{\Sigma}}_k\boldsymbol{R}_\text{s}^{-1}(\boldsymbol{a}_k)\boldsymbol{y} \ . \quad \text{(D.16)}$$

The scalars $a_{\text{s},k}$, $b_{\text{s},k}$, $a_{\text{e},k}$, and $b_{\text{e},k}$ are obtained from the factor $q(\sigma^2_{\text{s},k}, \sigma^2_{\text{e},k}|\boldsymbol{y}, \mathcal{M}_k)$ which is given by

$$q(\sigma^2_{\text{s},k}, \sigma^2_{\text{e},k}|\boldsymbol{y}, \mathcal{M}_k) = \text{Inv-}\mathcal{G}(a_{\text{s},k}, b_{\text{s},k})\text{Inv-}\mathcal{G}(a_{\text{e},k}, b_{\text{e},k}) \quad \text{(D.17)}$$

where

$$a_{\mathrm{s},k} = \alpha_{\mathrm{s},k} + N/2 \tag{D.18}$$

$$b_{\mathrm{s},k} = \beta_{\mathrm{s},k} + \left[\hat{\boldsymbol{s}}_k^T \boldsymbol{R}_\mathrm{s}^{-1}(\boldsymbol{a}_k)\hat{\boldsymbol{s}}_k + \mathrm{tr}\left(\boldsymbol{R}_\mathrm{s}^{-1}(\boldsymbol{a}_k)\hat{\boldsymbol{\Sigma}}_k\right)\right]/2 \tag{D.19}$$

$$a_{\mathrm{e},k} = \alpha_{\mathrm{e},k} + N/2 \tag{D.20}$$

$$b_{\mathrm{e},k} = \beta_{\mathrm{e},k} + \left[\hat{\boldsymbol{e}}_k^T \boldsymbol{R}_\mathrm{e}^{-1}(\boldsymbol{b}_k)\hat{\boldsymbol{e}}_k + \mathrm{tr}\left(\boldsymbol{R}_\mathrm{e}^{-1}(\boldsymbol{b}_k)\hat{\boldsymbol{\Sigma}}_k\right)\right]/2 \tag{D.21}$$

$$\hat{\boldsymbol{s}}_k = \boldsymbol{y} - \hat{\boldsymbol{e}}_k \, . \tag{D.22}$$

The above solution is not a closed-form solution for the parameters of the posterior factors. Instead, these are computed iteratively, and the VB framework guarantees that the algorithm converges to a mode. Since the normalised covariance matrices are diagonalised with the DFT matrix, we can easily evaluate the matrix inverses and the traces above. An interesting observation is that the VB algorithm essentially performs Wiener filtering in (D.16). Convergence of the VB algorithm can be monitored via the variational lower bound $\mathcal{L}_k$ which is related to the posterior model factor as

$$q(\mathcal{M}_k|\boldsymbol{y}) \propto \exp(\mathcal{L}_k)p(\mathcal{M}_k) \, . \tag{D.23}$$

Unfortunately, the variational lower bound consists of many terms so we refer the interested reader to the supplementary document for the full expression.

Since the posterior factor $q(\boldsymbol{e}|\boldsymbol{y}, \mathcal{M}_k)$ is a normal distribution, its second moment is

$$\mathbb{E}[\boldsymbol{e}\boldsymbol{e}^T|\boldsymbol{y}, \mathcal{M}_k] = \hat{\boldsymbol{e}}_k\hat{\boldsymbol{e}}_k^T + \hat{\boldsymbol{\Sigma}}_k \, . \tag{D.24}$$

Inserting this and the posterior model factor in (D.4) and (D.2) gives

$$\phi_\mathrm{e}(\omega) \approx \frac{1}{N}\sum_{k=1}^{K} q(\mathcal{M}_k|\boldsymbol{y})\left[|\boldsymbol{f}^H(\omega)\hat{\boldsymbol{e}}_k|^2 + \boldsymbol{f}^H(\omega)\hat{\boldsymbol{\Sigma}}_k\boldsymbol{f}(\omega)\right]$$

where we have ignored the limit operator. This PSD estimator is essentially a model-averaged version of the MMSE estimators in [10, 11]. However, the proposed estimator does not depend on threshold parameters to avoid stagnation, on bias compensation, or on unknown parameters which have to be estimated by computing speech presence probabilities. Moreover, the proposed estimator has a consistent way of including prior spectral information in the form of codebooks, and it works for a single data frame, even for uninformative prior distributions on all the excitation noise variances.

4 Evalutation

This paper has focused on motivating and deriving the proposed noise PSD estimator. Therefore, there is only a limited space left to provide evidence

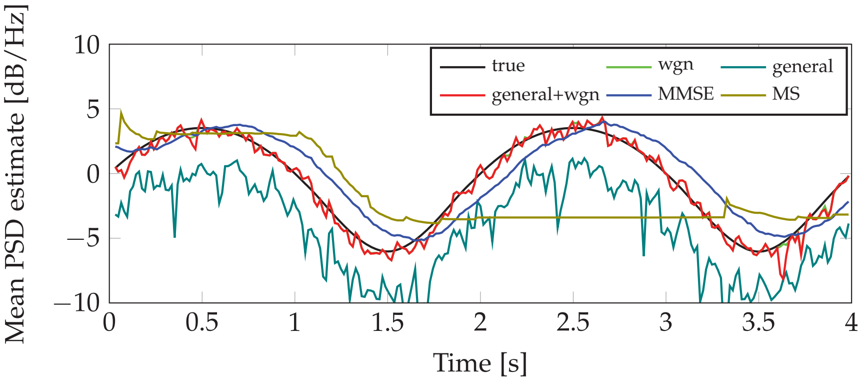


Fig. D.1: Estimates of the noise variance for modulated white Gaussian noise. The displayed results are averaged over frequency.

for that the fundamental principle works, but we have a more thorough evaluation in [25]. Here, we consider two different experiments. First, we demonstrate that the proposed noise PSD estimator works with zero tracking delay. Second, we apply the proposed noise PSD estimator to the difficult problem of estimating the PSD of babble noise from a noisy mixture. In both experiments, the speech codebook consisted of 64 AR vectors of order 14. It was trained using a variation of the LBG-algorithm method [26] on both male and female speech from the EUROM English database [27]. A noise codebook consisting of 12 AR vector of order 14 was trained on different noise types from the NOIZEUS database [28]. These noise types included restaurant, exhibition, street, and station noise. Thus, we did not train the codebook on babble noise which we are using for testing in the second experiment. As alluded to in Sec. 2.1, we used non-informative priors corresponding to no smoothing between frames. The codebooks as well as the MATLAB code for generating the presented results are available at `http://tinyurl.com/jknvbn`.

4.1 Tracking speed

The first experiment assessed the tracking speed of the estimator and is very similar to the first experiment in [10]. Thus, we estimated the noise power of modulated white Gaussian noise where the noise variance was time-varying with a frequency of 2 Hz. We compared the proposed method for three different noise codebooks to the MMSE method [10] and the MS method [3]. For the proposed method, the three different noise codebooks were a) a codebook consisting of only one entry modelling a flat spectrum; b) the noise codebook described above; and c) a combination of a) and b). Fig. D.1 shows the results for the various noise PSD estimates averaged over frequency. As in [10], it is observed that MS tracked the noise variance poorly and that the MMSE

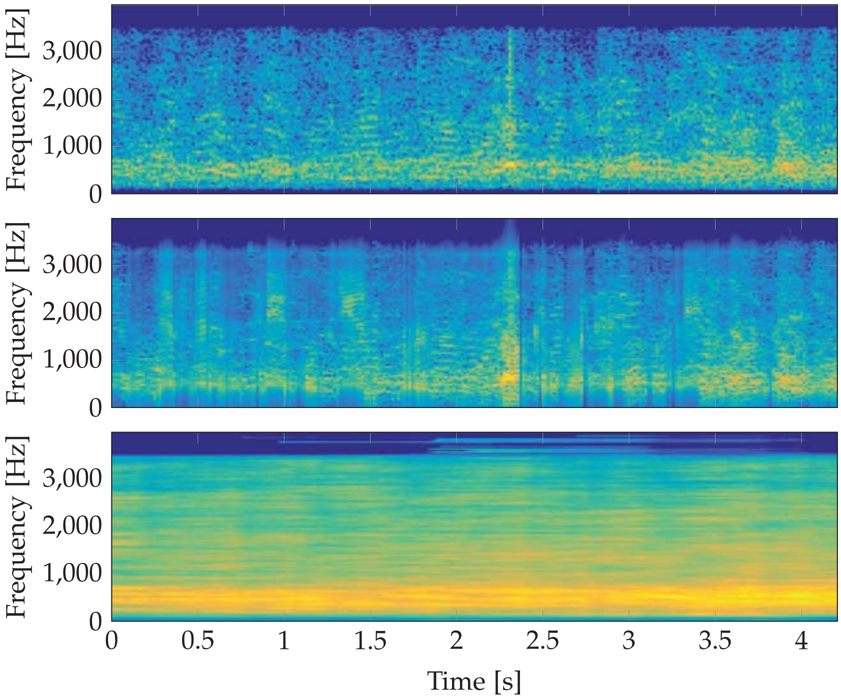


Fig. D.2: The spectrogram of the babble noise PSD (top) compared to the noise PSD estimates of the proposed (middle) and MMSE methods (bottom).

method tracked much better, but with a delay of a few hundred milliseconds. On the other hand, the proposed method with noise codebook a) and c) had no tracking delay and produced visually identical results. The latter observation suggests that the algorithm assigned all weight to the true model and used that for estimating the noise PSD. Finally, the proposed method with noise codebook b) underestimated the noise variance and had a much larger variance. This illustrates that we get a degraded performance if we use incorrect prior information in the codebook.

4.2 Babble noise PSD estimation

In the second experiment, we estimated the babble noise PSD from a mixture of speech and babble noise at different SNRs in steps of 2 dB from -10 dB to 10 dB. The babble noise was taken from the NOIZEUS database [28] and the speech signal was taken from the CHiME database [29]. Thus, neither of these signals were used for training the codebooks. For every SNR, we measured the average Itakura-Saito (IS) distance and the average log-spectral distortion (LSD) between the babble noise spectrogram and the estimated noise PSD for four different methods using the default MMSE method settings of 32

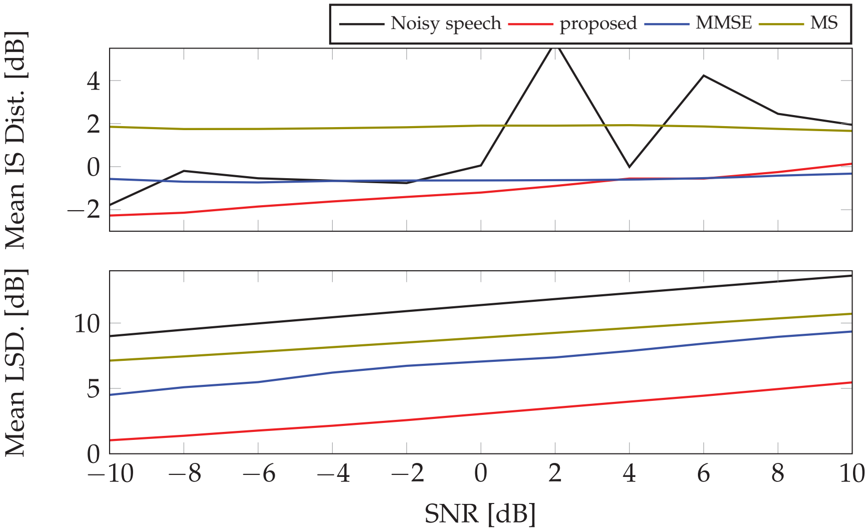


Fig. D.3: The IS distance and LSD between the babble noise spectrogram and estimated noise PSDs for various methods.

ms windows with a 50 % overlap. Aside from the proposed, the MMSE, and the MS methods, we also used the spectrogram of the observed mixture as a reference method. The results are shown in Fig. D.2 and Fig. D.3. In Fig. D.2, we have plotted the babble noise spectrogram (top), the proposed noise PSD estimate (middle), and the MMSE PSD estimate (bottom) for an SNR of 0 dB. Clearly, the proposed PSD estimate contains many more details than the MMSE PSD estimate. For example, there is a short burst in the babble noise at around 2.3 s which was captured by the proposed method, but smoothed out by the MMSE method. In Fig. D.3, the performance of the different estimators are quantified in terms of the IS distance and the LSD. The proposed method outperformed the other methods, except for the IS distance for an SNR above 3 dB where the proposed method and the MMSE method have similar performance.

5 Conclusion

In this paper, we have developed a framework for doing noise PSD estimation using parametric models. These models offer a way of including prior information into the estimator to obtain a better estimation accuracy. More concretely, we proposed a class of models based on pre-training codebooks. These codebooks contained typical spectra for the speech and the noise, but could in principle also include the PSD estimates from other estimators. The developed framework also contained model comparison to ensure that models which explain the data well have a larger weight in the model averaged noise PSD estimate. Via two experiments, we demonstrated the potential applicability and improvements in the tracking speed and estimation accuracy over two state-of-the-art methods.

References

[1] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[2] K. B. Christensen, M. G. Christensen, J. B. Boldt, and F. Gran, "Experimental study of generalized subspace filters for the cocktail party situation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 420–424.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[5] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Processing Conf.*, 1994, vol. 6, pp. 1182–1185.

[6] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Elsevier Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.

[7] I. Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands," in *International Workshop on Hands-Free Speech Communication*, 2001, pp. 95–98.

[8] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Elsevier Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[9] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[10] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[11] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2010, pp. 4266–4269.

[12] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2011, pp. 4640–4643.

[13] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 1996, vol. 2, pp. 733–736.

[14] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[15] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*, Englewood Cliffs, NJ, USA: Prentice Hall, May 2005.

[16] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[17] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. Int. Congr. Acoust.*, 1968, pp. 17–20.

[18] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequency," *Trans. Inst. Electron. Commun. Eng. (Japan)*, vol. 53, no. 1, pp. 36–43, 1970.

[19] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 367–376, 1980.

[20] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[21] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[22] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Dordrecht, The Nederlands: Kluwer Academic Publishers, 1992.

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, Aug. 2006.

[25] M. S. Kavalekalam, J. K Nielsen, M. G. Christensen, and J. B. Boldt, “A study of noise PSD estimators for single channel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018.

[26] Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[27] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, “EUROM-A spoken language resource for the EU,” in *Proc. European Conf. Speech Commun. and Speech Technol.*, 1995, pp. 867–880.

[28] Y. Hu and P. C. Loizou, “Subjective evaluation and comparison of speech enhancement algorithms,” *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[29] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. IEEE Workshop on Autom. Speech Recog. and Underst.*, 2015, pp. 504–511.

Paper E

A Study of Noise PSD Estimators for Single channel Speech Enhancement

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Mads Græsbøll Christensen and Jesper B. Boldt

The paper has been published in the
Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, 2018

Abstract

The estimation of the noise power spectral density (PSD) forms a critical component of several existing single channel speech enhancement systems. In this paper, we evaluate one new and some of the existing and commonly used noise PSD estimation algorithms in terms of the spectral estimation accuracy and the enhancement performance for different commonly encountered background noises, which are stationary and non-stationary in nature. The evaluated algorithms include the Minimum Statistics, MMSE, IMCRA methods and a new model-based method.

1 Introduction

Speech enhancement algorithms have a wide range of applications such as in digital hearing aids, speech recognition systems, mobile communications, etc [1], where the desired speech is degraded by acoustic background noise. These algorithms can be broadly categorised into single and multi channel algorithms. In this paper, we are only concerned with the former class of algorithms. The single channel speech enhancement algorithms must generally incorporate some assumptions to remove the background noise from the desired signal. For example, the Wiener filter assumes the second-order statistics of the speech/noise signal to be known. In practical scenarios, these statistics must be estimated from noisy observations. Thus, a very critical part present in most of the single channel speech enhancement methods is the estimation of the noise PSD [2, 3]. A significant amount of work has been done in the past decades to solve this problem.

In this paper, we evaluate some of the well known noise PSD estimation algorithms along with a new model-based approach [4]. Previously, an evaluation of noise PSD estimators was carried out in [5]. This study compared some of the existing noise PSD estimators in terms of the spectral estimation accuracy. In this study, we also evaluate the noise PSD estimators in terms of its enhancement capabilities in some of the typically encountered background noises. The estimation of noise PSD is not a trivial task especially in the case of non-stationary noises. In such scenarios, the noise PSD estimate has to be updated as rapidly as possible. An under-estimation or over-estimation of the noise PSD can lead to residual noise or speech distortion. In the current study, we evaluate different noise PSD estimation algorithms for different types of commonly encountered background noise, which are stationary and non-stationary in nature. The well-known algorithms that we have evaluated in this paper are Minimum Statistics (MS) method [6], Improved minima controlled recursive averaging (IMCRA) [7] method and minimum mean squared error (MMSE) based estimation [8]. In addition to these algorithms, we also evaluate a new model-based approach for estimating the noise PSD.

A detailed description regarding this method can be found in [4]. Here we focus on evaluating its performance. This method uses a priori information regarding the speech and noise spectral shapes in the form of autoregressive (AR) parameters stored in trained speech and noise codebooks.

The remainder of this paper will be organised as follows. Section 2 gives a brief introduction to the noise PSD estimation problem and an overview of the model-based method for estimating the noise PSD. A brief overview of the compared algorithms is given in Section 3. The experiments used in the evaluation of the algorithms will be explained in section 4 followed by the results and conclusion in Sections 5 and 6 respectively.

2 Model based approach for estimating the noise PSD

This section formulates the noise PSD estimation problem and gives a brief overview of the model-based approach for estimating the noise PSD. We refer the interested readers to a companion paper [4] (for further details). It is assumed here that N samples of noisy signal are observed as

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{e}, \tag{E.1}$$

where $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{s} \in \mathbb{R}^N$, and $\boldsymbol{e} \in \mathbb{R}^N$ are the noisy speech, the clean speech, and the noise, respectively. The basic task here is to estimate the noise PSD which is typically defined as [9]

$$\phi_e(\omega) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[|E(\omega)|^2|\boldsymbol{y}] \tag{E.2}$$

where $\mathbb{E}$ is the expectation operator and $E(\omega) = \boldsymbol{f}^H(\omega)\boldsymbol{e}$ is the DFT of the noise with $\boldsymbol{f}(\omega) = [1 \quad \exp(j\omega) \quad \cdots \quad \exp(j\omega(N-1))]^T$. The conditional expectation in (E.2) is the second moment of the density $p(E(\omega)|\boldsymbol{y})$ which leads to (E.2) be rewritten in terms of $p(\boldsymbol{e}|\boldsymbol{y})$ as

$$\phi_{\mathrm{e}}(\omega) = \lim_{N\to\infty} \frac{1}{N}\boldsymbol{f}^H(\omega) \left[\int_{\mathbb{R}^{N\times 1}} \boldsymbol{e}\boldsymbol{e}^T p(\boldsymbol{e}|\boldsymbol{y})d\boldsymbol{e}\right] \boldsymbol{f}(\omega) \ . \tag{E.3}$$

To compute the posterior $p(\boldsymbol{e}|\boldsymbol{y})$, statistical models denoted as $\{\mathcal{M}_k\}_{k=1}^K$, are used for explaining the generation of data. These models can be incorporated into (E.3) as,

$$\phi_{\mathrm{e}}(\omega) \approx \frac{1}{N}\sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{y}) \\ \times \boldsymbol{f}^H(\omega)\left[\int_{\mathbb{R}^{N\times 1}} \boldsymbol{e}\boldsymbol{e}^T p(\boldsymbol{e}|\boldsymbol{y},\mathcal{M}_k)d\boldsymbol{e}\right]\boldsymbol{f}(\omega) \tag{E.4}$$

$$= \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{y})\phi_{\mathrm{e}}(\omega|\mathcal{M}_k). \tag{E.5}$$

where $\{p(\mathcal{M}_k|\boldsymbol{y})\}_{k=1}^{K}$ denote the model probabilities, which ensure that models explaining the data well are given more weight in comparison to other model. The models that have been used are autoregressive (AR) models for speech and noise denoted by [10, 11]

$$p(\boldsymbol{s}|\sigma^2_{\mathrm{s},k},\mathcal{M}_k) = \mathcal{N}(\boldsymbol{0},\sigma^2_{\mathrm{s},k}\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)) \tag{E.6}$$

$$p(\boldsymbol{e}|\sigma^2_{\mathrm{e},k},\mathcal{M}_k) = \mathcal{N}(\boldsymbol{0},\sigma^2_{\mathrm{e},k}\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)) \tag{E.7}$$

where $\sigma^2_{\mathrm{s},k}$, $\sigma^2_{\mathrm{e},k}$, $\boldsymbol{R}_{\mathrm{s}}(\boldsymbol{a}_k)$, $\boldsymbol{R}_{\mathrm{e}}(\boldsymbol{b}_k)$, $\boldsymbol{a}_k$, and $\boldsymbol{b}_k$ are the excitation noise variance, the normalised covariance matrices, and the AR-parameters of the speech and the noise, respectively. It can be shown under certain assumptions that the normalised covariance matrix corresponding to speech and noise can be diagonalised by the DFT matrix [11, 12]. The excitation variances are treated as unknown random variables with the priors,

$$p(\sigma^2_{\mathrm{s},k}|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{s},k},\beta_{\mathrm{s},k}) \tag{E.8}$$

$$p(\sigma^2_{\mathrm{e},k}|\mathcal{M}_k) = \text{Inv-}\mathcal{G}(\alpha_{\mathrm{e},k},\beta_{\mathrm{e},k}). \tag{E.9}$$

As seen from (E.4) and (E.5), the posteriori model probabilities and the second moment of the posterior needs to be computed to get the final noise PSD estimate. As there is no closed form solution to obtain this, a variational Bayesian framework [13, 14] is used to produce an analytical approximation of the full joint posterior used in (E.4) as

$$p(\boldsymbol{e},\sigma^2_{\mathrm{s},k},\sigma^2_{\mathrm{e},k}|\boldsymbol{y},\mathcal{M}_k)p(\mathcal{M}_k|\boldsymbol{y}) \approx \\ q(\boldsymbol{e}|\boldsymbol{y},\mathcal{M}_k)q(\sigma^2_{\mathrm{s},k},\sigma^2_{\mathrm{e},k}|\boldsymbol{y},\mathcal{M}_k)q(\mathcal{M}_k|\boldsymbol{y})\ . \tag{E.10}$$

Since the posterior factor $q(\boldsymbol{e}|\boldsymbol{y},\mathcal{M}_k)$ is a normal distribution, its second moment and the posterior model probabilities $q(\mathcal{M}_k|\boldsymbol{y})$ is substituted in (E.4) to get the final noise PSD estimate. More details regarding the derivation of this method can be found in `http://tinyurl.com/jknvbn`.

3 Overview of the existing algorithms

In this section, we will give a brief overview of the existing noise PSD estimation algorithms that have been evaluated in this paper.

3.1 Minimum Statistics

This method [6] tracks the minima of the smoothed noisy spectrum for each frequency component. The method is based on the observation that the speech and noise component are statistically independent and that the power of the noisy signal often goes down to the power of the noise signal. The smoothed noisy spectrum is calculated using a recursive smoothing equation. Since this method is based on computing the minimum of the smoothed noisy spectrum over a moving window, the noise PSD estimate is necessarily biased. This is overcome in [6] to some extent by using a bias compensation factor in time and frequency.

3.2 IMCRA

In this method [7], the noise PSD estimate is obtained by a recursive averaging of the noisy spectral values using a time varying frequency dependent smoothing parameter, that is adjusted according to the speech presence probability (SPP) for each frequency component. The a priori SPP are calculated in this method after two iterations of smoothing and minima tracking. The final SPP (used for the recursive averaging) is then computed using the a priori SPP and the estimated a priori SNR.

3.3 MMSE

This method [8] derives an MMSE estimator of the noise PSD coefficients. Here, the speech and noise spectral coefficients are modelled as normally distributed random variables that are independent with each other. The first step involves the computation of the conditional expectation of the noise periodogram given the noisy signal which involves a weighted combination of noise PSD estimate from the previous frame and the noisy periodogram from the current frame. The final noise PSD estimate is then obtained by a recursive averaging of the estimated noise periodogram.

4 Experiments

We will now describe the experiments that have been carried out to evaluate the four noise PSD estimation algorithms. Section 4.1 describes the parameters that have been used for implementing the different noise PSD estimation algorithms. Sections 4.2 and 4.3 explains the experiments done to evaluate the estimation accuracy and the enhancement capabilities of the noise PSD estimation algorithms, respectively.

4.1 Implementation Details

We have evaluated a total of four algorithms: MS, IMCRA, MMSE and the new model based approach. The test signals used for evaluation were taken from the EUROM database [15]. The clean speech signals were then degraded by 5 types of typically encountered background noise: babble, street, station, exhibition and restaurant from the NOIZEUS database [16]. The model based approach for estimating the noise PSD explained in Section 2 requires the speech and noise codebooks to be trained offline. For the experiments we have trained a speech codebook of 64 entries and a noise codebook of 12 entries. The codebooks were trained using a variation of the LBG algorithm [17]. The training data used for creating the speech codebook consisted of audio samples from the EUROM database. It should be noted that we have trained a codebook that is independent of the speaker. The data used for generating the noise codebook consisted of noise files from the NOIZEUS database. Different codebooks were trained for different types of noise, which were then appended together to form a larger codebook. The noise codebook had a size of 16 entries, which consisted of 4 entries each for babble, restaurant and exhibition and 2 entries each for street and station. It should be noted that, while testing for a particular noise scenario, the noise codebook entries corresponding to that scenario is **NOT** used for the estimation of noise PSD. The codebooks as well as MATLAB code for generating the codebooks will be available at `http://tinyurl.com/jknvbn`. The AR order for the speech and noise models were chosen to be 14. All the noise PSD estimation algorithms evaluated here work on a frame size of 32 ms with 50% overlap.

4.2 Estimation Accuracy

We have used the log spectral distortion between the estimated noise PSD and the reference noise PSD to measure the spectral estimation accuracy of the algorithms. The reference PSD in this case is computed by taking the periodogram of the noise only signal. The mean log spectral distortion across the whole signal is given by

$$\text{LogErr} = \frac{10}{NL} \sum_{l=0}^{L-1} \sum_{k=0}^{N-1} \left| \log_{10} \frac{\phi_e(k,l)}{\hat{\phi}_e(k,l)} \right| \tag{E.11}$$

where $\phi_e(k,l)$ is the true noise PSD and $\hat{\phi}_e(k,l)$ is the estimated noise PSD at the k^{th} frequency index of the l^{th} frame. This term can be separated into distortion due to over-estimation and under-estimation of the noise PSD, which can be written as LogErr $=$ $\text{LogErr}_{\text{ov}} + \text{LogErr}_{\text{un}}$, where $\text{LogErr}_{\text{ov}}$ and $\text{LogErr}_{\text{un}}$ are defined as [8]

$$\text{LogErr}_{\text{ov}} = \frac{10}{NL} \sum_{l=0}^{L-1} \sum_{k=0}^{N-1} \left| \min\left(0, \log_{10} \frac{\phi_e(k,l)}{\hat{\phi}_e(k,l)}\right) \right| \tag{E.12}$$

$$\text{LogErr}_{\text{un}} = \frac{10}{NL} \sum_{l=0}^{L-1} \sum_{k=0}^{N-1} \max\left(0, \log_{10} \frac{\phi_e(k,l)}{\hat{\phi}_e(k,l)}\right). \tag{E.13}$$

Overestimation of the noise PSD measured by $\text{LogErr}_{\text{ov}}$ is likely to cause speech distortion during the enhancement stage, whereas $\text{LogErr}_{\text{un}}$ gives a measure of the residual noise present in the enhanced signal. A plot of these measures for different acoustic background noises is shown in Section 5.

4.3 Enhancement performance

The estimated noise PSD is then incorporated in a speech enhancement framework. For this, we first estimate the a priori SNR using the decision directed approach [2]. The estimated a priori SNR is then incorporated in a Wiener filter for speech enhancement. In this work, we have used the Segmental SNR (segSNR), Segmental speech SNR (spSNR) and segmental noise reduction (segNR) which has also been used in [8, 18] to evaluate the enhancement performance. segSNR, spSNR and segNR are denoted as

$$\text{segSNR} = \frac{10}{L} \sum_{l=0}^{L-1} \log_{10} \frac{\sum_{m=1}^{M} s^2(lM+m)}{\sum_{m=1}^{M} (s(lM+m) - \hat{s}(lM+m))^2} \tag{E.14}$$

$$\text{spSNR} = \frac{10}{L} \sum_{l=0}^{L-1} \log_{10} \frac{\sum_{m=1}^{M} s^2(lM+m)}{\sum_{m=1}^{M} (s(lM+m) - \tilde{s}(lM+m))^2} \tag{E.15}$$

$$\text{segNR} = \frac{10}{L} \sum_{l=0}^{L-1} \log_{10} \frac{\sum_{m=1}^{M} e(lM+m)^2}{\sum_{m=1}^{M} \tilde{e}(lM+m)^2} \tag{E.16}$$

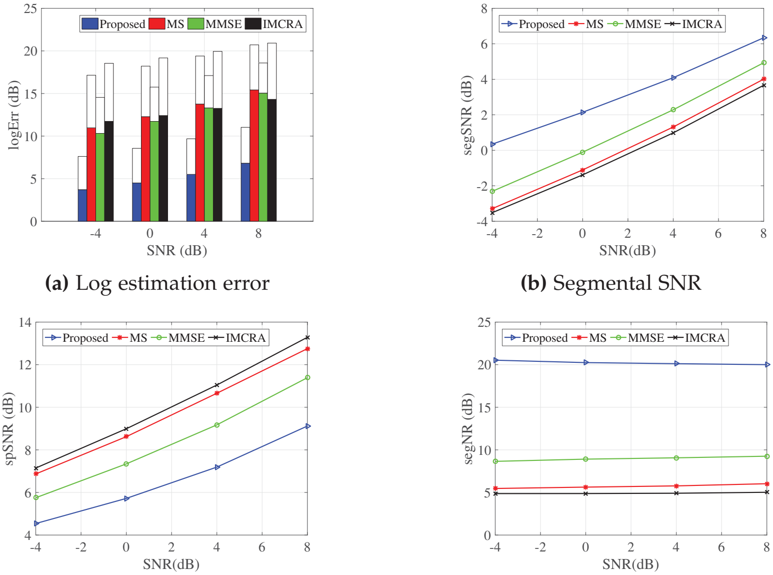
where $s(n)$ denotes the clean signal, $e(n)$ denotes the noise signal, $\hat{s}(n)$ denotes the enhanced signal and M denotes the number of samples in a segment. The term $\tilde{s}(n)$ and $\tilde{e}(n)$ are obtained by the applying the estimated Wiener filter onto $s(n)$ and $e(n)$ respectively. The spSNR measures the speech distortion, where an increase in speech distortion is indicated by a decrease in spSNR. segNR gives a measure of the residual noise present in the signal after enhancement. segSNR improvement takes into account both the speech distortion and noise reduction. A plot of these measures for different acoustic background noise is shown in Section 5.

5 Results

In this section we plot the performance metrics introduced in Sections 4.2 and 4.3 for different background noises. Figure E.1 shows the results obtained for babble noise. Figure E.1a corresponds the log error distortion for different methods as a function of the input SNR. The lower shaded area of the bar plot corresponds to $\text{LogErr}_{\text{ov}}$ caused due to over estimation of the noise PSD and upper part corresponds to $\text{LogErr}_{\text{un}}$ caused due to under estimation of the noise PSD. It can be seen that the model based approach performs the best in terms of log distortion measure followed by MMSE, MS and IMCRA. Figure E.1b shows the segmental SNR for the different methods as a function of the input SNR. Figures E.1c and E.1d show the segmental speech SNR and noise reduction, respectively. It can be seen that even though the model based approach performs the best in terms of segSNR and segNR, it also has the lowest spSNR. This indicates a high noise reduction at the cost of speech distortion. IMCRA which performs the worst in terms of noise reduction performs the best in terms of speech distortion. This is a common trade-off observed in speech enhancement [19]. Figures E.2, E.3, E.4 and E.5 show the obtained results for restaurant, exhibition, street and station noise respectively. These figures also show a similar trend as observed for the babble noise. It should be noted that the benefit of using the model based approach over the other methods is more pronounced in relatively non-stationary noises such as babble and the restaurant noise. This can be explained by the zero tracking delay of the model based approach in comparison to other nethods which atleast have a few hundred milliseconds of tracking delay [4, 8].

6 Discussion and Conclusion

The estimation of noise PSD is a very critical component of a speech enhancement system. Thus, in this paper, we have evaluated four noise PSD estimators for single channel speech enhancement in some of the typically encountered background noises. The evaluated algorithms consisted of MS, MMSE, IMCRA and a new model based method. It was observed that the model-based method outperformed other algorithms in terms of the spectral estimation accuracy for all the noise types. In terms of the enhancement performance, the model-based approach outperformed the other algorithms for relatively non-stationary noises such as babble and restaurant noise irrespective of the SNR. In the case of more stationary noise types such as station and street noise, the benefit of using the model-based approach is observed only in lower SNRs.



(a) Log estimation error

(b) Segmental SNR

(c) Segmental Speech SNR

(d) Segmental NR

Fig. E.1: Performace measures of the algorithms for babble noise. The lower part of the subfigure E.1a represents the LogErr$_{\text{ov}}$ and the upper part in white represents LogErr$_{\text{un}}$ error due to the underestimation of noise PSD. Subfigures E.1b, E.1c and E.1d represent the segmental SNR, segmental speech SNR and segmental NR respectively

References

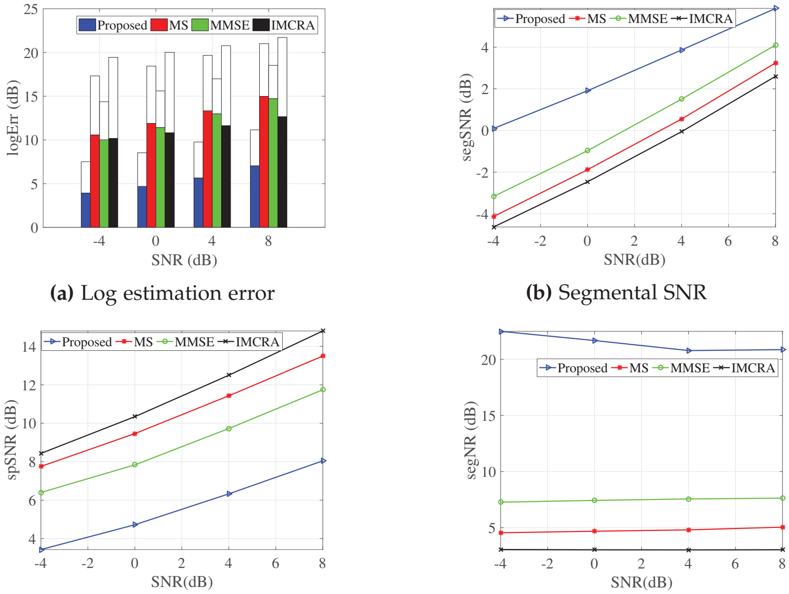
[1] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[4] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Model-based noise psd estimation from speech in non-stationary noise," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2018.

[5] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2011, pp. 4640–4643.



(a) Log estimation error

(b) Segmental SNR

(c) Segmental Speech SNR

(d) Segmental NR

Fig. E.2: Performance measures for different algorithms for restaurant noise

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

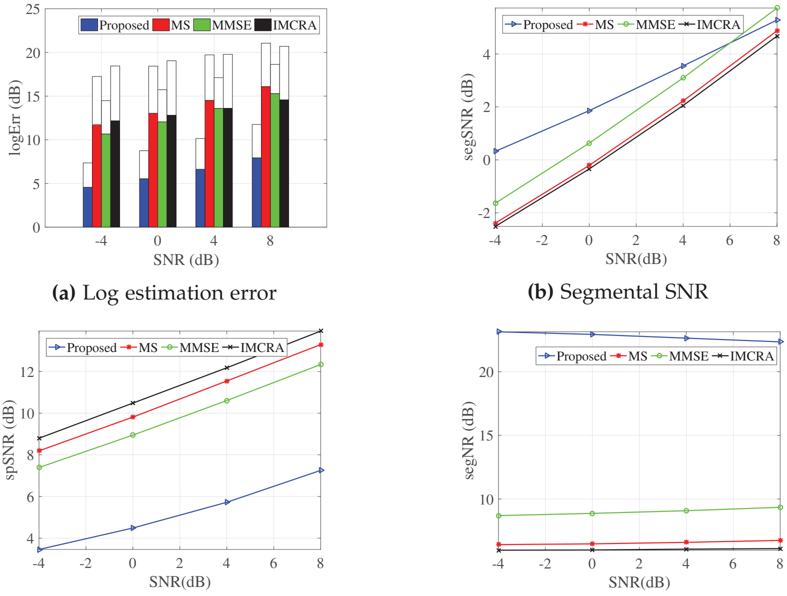
[7] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[8] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[9] P. Stoica, R. L. Moses, et al., *Spectral analysis of signals*, vol. 452, Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[11] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE*



(a) Log estimation error

(b) Segmental SNR

(c) Segmental Speech SNR

(d) Segmental NR

Fig. E.3: Performance measures for different algorithms for exhibition noise

Trans. on Audio, Speech, and Language Processing, vol. 15, no. 2, pp. 441–452, 2007.

[12] R. M. Gray et al., "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

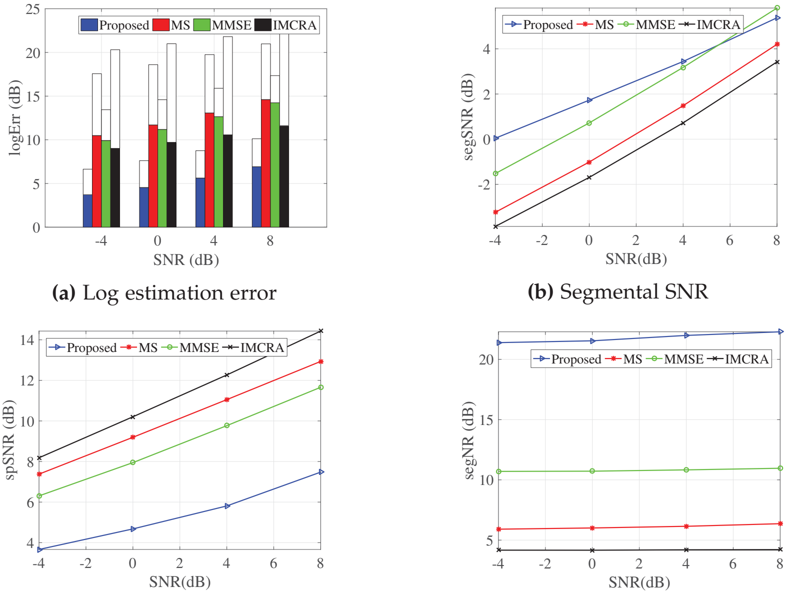
[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[14] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.

[15] D. Chan, A. Fourcin, D. Gibbon, B Granstrom, et al., "Eurom-a spoken language resource for the eu," in *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech'95*, 1995, pp. 867–880.

[16] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2006.

[17] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.



(a) Log estimation error

(b) Segmental SNR

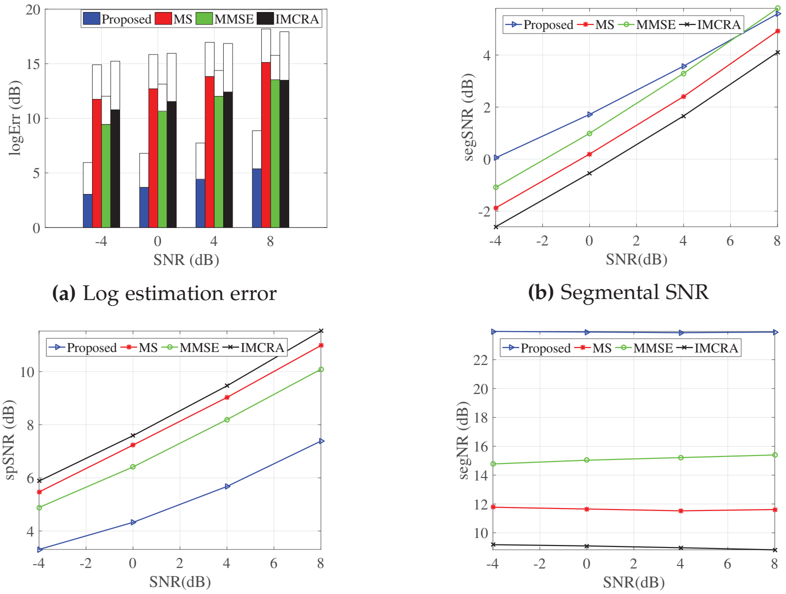
(c) Segmental Speech SNR

(d) Segmental NR

Fig. E.4: Performance measures for different algorithms for street noise

[18] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.

[19] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.



(a) Log estimation error

(b) Segmental SNR

(c) Segmental Speech SNR

(d) Segmental NR

Fig. E.5: Performace measures of the different algorithms for station noise.

Paper F

Online Parametric NMF for Speech Enhancement

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Liming Shi, Mads Græsbøll Christensen and Jesper B. Boldt

The paper has been published in the
Proc. European Signal Processing Conference, 2018

Abstract

In this paper, we propose a speech enhancement method based on non-negative matrix factorization (NMF) techniques. NMF techniques allow us to approximate the power spectral density (PSD) of the noisy signal as a weighted linear combination of trained speech and noise basis vectors arranged as the columns of a matrix. In this work, we propose to use basis vectors that are parameterised by autoregressive (AR) coefficients. Parametric representation of the spectral basis is beneficial as it can encompass the signal characteristics like, e.g. the speech production model. It is observed that the parametric representation of basis vectors is beneficial while performing online speech enhancement in low delay scenarios.

1 Introduction

A healthy human auditory system is capable of focusing on desired signal from a target source while ignoring background noise in a complex noisy environment. In comparison to a healthy auditory system, the auditory system of a hearing impaired person lacks this ability, leading to degradation in speech intelligibility. In such scenarios, a hearing impaired person often relies on speech enhancement algorithms present in a hearing aid. However, the performance of the current hearing aid technology in this aspect is limited [1]. Speech enhancement algorithms that have been developed can be mainly categorised into supervised and unsupervised methods. Some of the existing unsupervised methods are spectral subtraction methods [2], statistical model based methods [3] and subspace based methods [4]. Supervised methods generally use some amount of training data to estimate the model parameters corresponding to speech and noise. The model parameters are subsequently used for enhancement. Examples of supervised enhancement methods include codebook based methods [5, 6], NMF methods [7–9], hidden Markov model based methods [10, 11].

In this paper, we propose a speech enhancement method based on non-negative matrix factorization (NMF) techniques. NMF for source separation and speech enhancement has been previously proposed [7, 8]. NMF techniques allow us to approximate the power spectrum or the magnitude spectrum of the noisy signal as a weighted linear combination of trained speech and noise basis vectors arranged as the columns of a matrix. Generally the basis vectors used in NMF based speech enhancement are not constrained by any parameters. Parameterisation of the basis vectors in the field of music processing has been previously done in [12]. In [12], harmonic combs parametrised by the fundamental frequency was used as the basis vectors. This parametrisation was found to efficiently represent the music signal in comparison to the non parametric counterpart.

In this work, we propose to use basis vectors that are parametrised by autoregressive (AR) coefficients. This parametrisation allows representation of power spectral density (PSD) using a small set of parameters. Parametrisation by AR coefficients is motivated by the source filter model of speech production. This model describes speech components as a combination of a sound source (excitation signal produced by the vocal chords) and an AR filter which models the vocal tract. In this work, we show that if we model the observed data in the time domain as a sum of AR processes, the maximisation of the likelihood corresponds to performing NMF of the observed data into a basis matrix and activation coefficients, using Itakura-Saito (IS) divergence as the optimisation criterion. The IS divergence has been extensively used in speech and music processing due to its similarity to perceptual distance. The basis matrix here consists of AR spectral envelopes parameterised by AR coefficients, and the activation coefficients can be physically interpreted as the excitation variance of the noise that excites the AR filter parametrised by the AR coefficients. A benefit of parametrically representing the spectral basis, is that, it can be represented by a small set of parameters, which means that fewer parameters have to be trained a priori for performing on-line speech enhancement.

The remainder of this paper is organised as follows. Section 2 explains the signal model and formulates the problem mathematically. Training of the speech and noise spectral bases is explained in Section 3. Section 4 explains the on-line estimation of the activation coefficients corresponding to the spectral bases followed by the enhancement procedure using the Wiener filter. Sections 5 and 6 give the experimental results and the conclusion respectively.

2 Mathematical formulation

This section explains the signal model and mathematically formulates the problem. The noisy signal is expressed as

$$x(n) = s(n) + w(n) \tag{F.1}$$

where $s(n)$ is the clean speech and $w(n)$ is the noise signal. The objective of a speech enhancement system is to obtain an estimate of the clean speech signal from the noisy signal. A very popular method for estimating the clean speech signal is by applying a Wiener filter onto the noisy signal. Wiener filtering requires the knowledge of the speech and noise statistics. Since there is no direct access to either speech or noise in practical scenarios, these statistics have to be estimated from the noisy observation. As the speech and noise properties change over time, these statistics are generally time varying. The majority of the speech processing algorithms consider these statistics to be quasi-stationary. Thus, these statistics are assumed to be constant for short segments of time (≈ 25 ms).

We now, explain the signal model used in the estimation of the statistics from a frame of noisy signal. It is assumed that a frame of noisy signal $\mathbf{x} = [x(0), \dots x(N-1)]^T$ can be represented as a sum of $U = U_s + U_w$ AR processes $\mathbf{c}_u$. This is mathematically written as

$$\mathbf{x} = \sum_{u=1}^{U} \mathbf{c}_u \; = \sum_{u=1}^{U_s} \mathbf{c}_u + \sum_{u=U_s+1}^{U} \mathbf{c}_u, \tag{F.2}$$

where the first U_s AR processes correspond to the speech signal and the remaining U_w AR processes correspond to the noise signal. Each of the AR process is expressed as a multivariate Gaussian [6] as shown below

$$\mathbf{c}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{Q}_u). \tag{F.3}$$

The gain normalised covariance matrix, $\mathbf{Q}_u$ can be asymptotically approximated as a circulant matrix which can be diagonalised using the Fourier transform as [13]

$$\mathbf{Q}_u = \mathbf{F}\mathbf{D}_u\mathbf{F}^H \tag{F.4}$$

where $\mathbf{F}$ is the DFT matrix defined as $[\mathbf{F}]_{k,n} = \frac{1}{\sqrt{N}} \exp(j2\pi nk/N), \;\; n,k = 0 \dots N-1$ and

$$\mathbf{D}_u = (\mathbf{\Lambda}_u^H \mathbf{\Lambda}_u)^{-1}, \qquad \mathbf{\Lambda}_u = \text{diag}(\sqrt{N}\mathbf{F}^H \begin{bmatrix} \mathbf{a}_u \\ \mathbf{0} \end{bmatrix}) \tag{F.5}$$

where $\mathbf{a}_u = [1, a_u(1) \dots a_u(P)]^T$ represents the vector of AR coefficients corresponding to u^{th} basis vector and P is the AR order. The likelihood as a function of U excitation variances and AR spectral envelopes are expressed as

$$p(\mathbf{x}|\sigma, \mathbf{D}) \sim \mathcal{N}(\mathbf{0}, \sum_{u=1}^{U} \sigma_u^2 \mathbf{Q}_u) \tag{F.6}$$

where $\boldsymbol{\sigma}$ represents the excitation variances corresponding to the U AR processes and $\mathbf{D}$ represents AR spectral envelopes corresponding to the U AR processes. In this paper, we are interested in the maximum likelihood (ML) estimation of activation coefficients $\boldsymbol{\sigma}$ given the noisy signal $\mathbf{x}$. Since, we are performing supervised enhancement here, we assume that the spectral basis are trained a priori, which is explained in Section 3. Thus, in this work we only estimate the activation coefficients online while the basis vectors are assumed known. This is expressed mathematically as, To solve this, the logarithm of likelihood in (F.6) is written as

$$\begin{aligned}\ln p(\mathbf{x}|\sigma,\ \mathbf{D}) = -\frac{N}{2}\ln 2\pi + \ln\Big|\sum_{u=1}^{U}\sigma_u^2\mathbf{F}\mathbf{D}_u\mathbf{F}^H\Big|^{-\frac{1}{2}} \\ -\frac{1}{2}\mathbf{x}^T[\sum_{u=1}^{U}\sigma_u^2\mathbf{F}\mathbf{D}_u\mathbf{F}^H]^{-1}\mathbf{x}.\end{aligned} \tag{F.7}$$

This is further simplified as

$$\begin{aligned}\ln p(\mathbf{x}|\sigma,\ \mathbf{D}) = -\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\Big(\sum_{u=1}^{U}\sigma_u^2 d_u(k)\Big)^{-\frac{1}{2}} \\ -\frac{1}{2}\mathbf{x}^T\mathbf{F}[\sum_{u=1}^{U}\sigma_u^2\mathbf{D}_u]^{-1}\mathbf{F}^H\mathbf{x}\end{aligned} \tag{F.8}$$

where $d_u(k)$ represents the k^{th} diagonal element of $\mathbf{D}_u$ and number of frequency indices $K = N$. Further simplifying,

$$\begin{aligned}\ln p(\mathbf{x}|\sigma,\ \mathbf{D}) = -\frac{K}{2}\ln 2\pi + \ln\prod_{k=1}^{K}\Big(\sum_{u=1}^{U}\hat{\Phi}_u(k)\Big)^{-\frac{1}{2}} \\ -\frac{1}{2}\sum_{k=1}^{K}\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)}\end{aligned} \tag{F.9}$$

where $\hat{\Phi}_u(k) = \sigma_u^2 d_u(k)$, $\Phi(k) = |X(k)|^2$ and $X(k) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}x(n)\exp(-j2\pi nk/N)$. Log-likelihood is then written as

$$\ln p(\mathbf{x}|\sigma,\mathbf{D}) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\left(\frac{\Phi(k)}{\sum_{u=1}^{U}\hat{\Phi}_u(k)} + \ln\sum_{u=1}^{U}\hat{\Phi}_u(k)\right) \tag{F.10}$$

where

$$\sum_{u=1}^{U}\hat{\Phi}_u(k) = \sum_{u=1}^{U}\sigma_u^2 d_u(k) = \mathbf{d}_k\boldsymbol{\sigma} \tag{F.11}$$

where $\mathbf{d}_k = [d_1(k)\dots d_U(k)]$ and $\boldsymbol{\sigma} = [\sigma_1^2\dots\sigma_U^2]^T$. Thus maximising the likelihood is equivalent to minimising the IS divergence between $\boldsymbol{\phi} = [\Phi(1)\dots\Phi(K)]^T$ and $\mathbf{D}\boldsymbol{\sigma}$ subject to $\Phi(k) > 0\ \forall k$ where $\mathbf{D} = [\mathbf{d}_1^T\dots\mathbf{d}_K^T]^T$. In case we observe $V > 1$ frames, this corresponds to performing NMF of $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1\dots\boldsymbol{\phi}_v\dots\boldsymbol{\phi}_V]$ (where $\boldsymbol{\phi}_v = [\Phi_v(1)\dots\Phi_v(K)]^T$ contains the periodogram of the v^{th} frame) as

$$\boldsymbol{\Phi} \approx \underbrace{\begin{bmatrix} d_1(1) & \dots & d_U(1) \\ \vdots & \ddots & \vdots \\ d_1(K) & \dots & d_U(K)\end{bmatrix}}_{\mathbf{D}}\underbrace{\begin{bmatrix} \sigma_1^2(1) & \dots & \sigma_1^2(V) \\ \vdots & \ddots & \vdots \\ \sigma_U^2(1) & \dots & \sigma_U^2(V)\end{bmatrix}}_{\boldsymbol{\Sigma}}. \tag{F.12}$$

The first U_s columns of $\mathbf{D}$ corresponds to the spectral basis corresponding to the speech and the remaining U_w columns of $\mathbf{D}$ correspond to noise signal. The first U_s rows of $\mathbf{\Sigma}$ correspond to the activation coefficients for speech and the remaining U_w rows of $\mathbf{\Sigma}$ correspond to the activation coefficients corresponding to the noise signal, which leads to (F.12) being rewritten as,

$$\mathbf{\Phi} \approx [\mathbf{D}_s \, \mathbf{D}_w] \begin{bmatrix} \mathbf{\Sigma}_s \\ \mathbf{\Sigma}_w \end{bmatrix} = \mathbf{D\Sigma}. \tag{F.13}$$

3 Training the Spectral Bases

This section explains the training of the basis vectors used for the construction of the basis matrix $\mathbf{D}$. In this work we use a parametric representation of the PSD [14] where the u^{th} spectral basis $\mathbf{d}_u = [d_u(1) ... d_u(k) ... d_u(K)]^T$ is represented as

$$d_u(k) = \frac{1}{\left| 1 + \sum\limits_{p=1}^{P} a_u(p) \exp(\frac{-j2\pi pk}{N}) \right|^2}, \tag{F.14}$$

where $\{a_u(p)\}_{p=1}^{P}$ is the set of AR coefficients corresponding to the u^{th} basis vector. During the training stage, a speech and noise codebook is first computed using the generalised Lloyd algorithm [15] [16] [6]. The speech codebook and noise codebooks contain AR coefficients corresponding to the spectral envelopes of speech and noise. During the training process linear prediction coefficients (converted into line spectral frequency coefficients) are extracted from windowed frames, obtained from the training signal and passed as input to the vector quantiser. Once the speech codebook and noise codebooks are created, the spectral envelopes corresponding to the speech AR coefficients ($\{\mathbf{a}_u\}_{u=1}^{U_s}$) and noise AR coefficients ($\{\mathbf{a}_u\}_{u=U_s+1}^{U}$) are computed using (F.14), and arranged as columns of $\mathbf{D}$. The spectral envelopes generated here are gain normalised, so they do not include the excitation variance. Fig. F.1 shows a few examples of the trained speech and noise spectral envelopes.

4 Enhancement - Multiplicative Update

This section describes the estimation of speech and noise PSDs using the signal model explained in Section 2. Since we are interested in on-line processing of the noisy signal, we here assume that only a frame of noisy signal is available at particular time for enhancement. The method considered here assumes that

$$\boldsymbol{\phi} \approx \mathbf{D}\boldsymbol{\sigma} \tag{F.15}$$

where $\boldsymbol{\phi}$ is a $K \times 1$ vector containing the noisy PSD, $\mathbf{D}$ is $K \times U$ basis matrix and $\boldsymbol{\sigma}$ is $U \times 1$ vector containing the activation coefficients. The objective here, is to estimate $\boldsymbol{\sigma}$ given the noisy periodogram $\boldsymbol{\phi}$ and $\mathbf{D}$. As explained in Section 2, this is done by minimising the IS divergence as

$$\boldsymbol{\sigma}_{est} = [\boldsymbol{\sigma}_{s_{est}}^T \ \boldsymbol{\sigma}_{w_{est}}^T]^T = \underset{\boldsymbol{\sigma} \geq 0}{\arg\min} \quad d_{\text{IS}}(\boldsymbol{\phi}|\mathbf{D}\boldsymbol{\sigma}). \tag{F.16}$$

In this work, a multiplicative update (MU) method is used to estimate the activation coefficients which are calculated as [8, 17]

$$\boldsymbol{\sigma}_{est} \leftarrow \boldsymbol{\sigma}_{est} \frac{\mathbf{D}^T((\mathbf{D}\boldsymbol{\sigma}_{est})^{[-2]}.\boldsymbol{\phi})}{\mathbf{D}^T(\mathbf{D}\boldsymbol{\sigma}_{est})^{[-1]}}. \tag{F.17}$$

Once the gains are estimated, a Wiener filter can be constructed to extract the speech/noise components. The estimated clean speech PSD is obtained as $\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}}$ and the estimated noise PSD is obtained as $\mathbf{D}_w\boldsymbol{\sigma}_{w_{est}}$. The Wiener filter vector constructed to extract the speech component is denoted as

$$\mathbf{g}_{est} = \frac{\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}}}{\mathbf{D}_s\boldsymbol{\sigma}_{s_{est}} + \mathbf{D}_w\boldsymbol{\sigma}_{w_{est}}}, \tag{F.18}$$

where the division is an element wise division.

5 Experiments

5.1 Implementation Details

This section explains the experiments that have been carried out to evaluate the proposed enhancement framework. The test signals used here consist of sentences taken from the GRID database [18]. The speech and noise PSD parameters are estimated (as explained in Section 4) for a segment of 25 ms with 50 percent overlap. The parameters used for the experiments are summarised in table F.1. For our experiments, we have used both a speaker-specific codebook and a general speech codebook. A speaker-specific codebook of 64 entries was trained using a training sample of 5 minutes of speech from the specific speaker of interest. A general speech codebook of 64 entries was trained from a training sample of approximately 150 minutes of speech from 30 different speakers. It should be noted that the sentences used for training the codebook were not included for testing. The proposed enhancement framework was tested on three different types of commonly encountered background noise: babble, restaurant and exhibition noise taken from the NOIZEUS database [19]. We have performed experiments for a noise specific codebook as well as general noise codebook. A noise-specific codebook of 8 entries was trained on the specific noise type of interest. For creating a general noise codebook, a noise codebook of 4 entries was trained for each noise type. While testing for a particular noise scenario, the noise codebook entries corresponding to that scenario are not used for the estimation of noise PSD. For example, while testing in the babble noise scenario, the noise codebook consists a total of 8 entries formed by concatenating the entries trained for restaurant and exhibition scenarios. After obtaining the speech and noise codebooks, the spectral basis matrix is constructed as explained in Section 3. The estimated PSD parameters are then used to create a Wiener filter for speech enhancement. Wiener filter is applied in the frequency domain and time-domain enhanced signal is synthesised using overlap-add.

5.2 Results

We have used the objective measures such as STOI and Segmental SNR to evaluate the proposed algorithm. We will denote the proposed parametric NMF as ParNMF. We have compared the performance of the proposed method to non parametric NMF where there is no parametrisation involved in the creation of the basis vectors. We will denote this method as NonParNMF. It should be noted that we have used the same training set for ParNMF and NonParNMF. We have also used the speech enhancement method proposed in [20] for comparison purposes, which we denote as MMSE-GGP. Traditionally, NMF methods for speech enhancement generally try to approximate the magnitude spectrum than the power spectrum. Even though, this is not theoretically well formulated, this has been observed to give better performance [21]. Thus, here we evaluated the performance of the proposed algorithm for both the cases, which we denote as ParNMF-abs while approximating the magnitude spectrum and ParNMF-pow while approximating the power spectrum. We do the same evaluation in the case of NonParNMF. Figures F.2-F.4 show these measures for different methods in different commonly encountered background noises while using a speaker

Table F.1: Parameters used for the experiments

Parameters	
sampling frequency	8000 Hz
Frame Size	200
Frame Overlap	50%
Speech AR order	14
Noise AR order	14
U_s	64
U_w	8
MU iterations	50

specific codebook and a noise specific codebook. It can be seen that NMF based methods perform better than MMSE-GGP in terms of STOI. When comparing the ParNMF and NonParNMF, it is demonstrated that the former performs better in terms of STOI and Segmental SNR measures. We have also performed experiments when having an access to a general speech codebook and a general noise codebook. Figures F.5-F.7 shows the objective measures obtained for this case. It can be seen that performance in this case degrades in comparison to figures F.2-F.4 due to the mismatch in training and testing conditions. Even though there is a degradation in the performance, the proposed method is able to increase the STOI measure significantly over the conventional method.

6 Conclusion

In this paper, we have proposed an NMF based speech enhancement method where the basis vectors are parametrised using AR coefficients. Parametrisation of the spectral basis vectors helps in encompassing the signal characterestics. We have demonstrated, through objective measures, that the proposed parametric NMF based speech enhancement out performs its non-parametric counterpart in some of the typically encountered background noises.

References

[1] S. Kochkin, “10-year customer satisfaction trends in the US hearing instrument market,” *Hearing Review*, vol. 9, no. 10, pp. 14–25, 2002.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[4] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[6] ——, “Codebook-based bayesian speech enhancement for nonstationary environments,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, 2007.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[8] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[9] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, “Speech enhancement under low snr conditions via noise estimation using sparse and low-rank nmf with kullback–leibler divergence,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, 2015.

[10] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. on Speech and Audio processing*, vol. 6, no. 5, pp. 445–455, 1998.

[11] D. Y. Zhao and W. B. Kleijn, “Hmm-based gain modeling for enhancement of speech in noise,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, 2007.

[12] R. Hennequin, R. Badeau, and B. David, “Time-dependent parametric and harmonic templates in non-negative matrix factorization,” in *Proc. of the 13th International Conference on Digital Audio Effects (DAFx)*, 2010.

[13] R. M. Gray *et al.*, “Toeplitz and circulant matrices: A review,” *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[14] P. Stoica, R. L. Moses *et al.*, *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005, vol. 452.

[15] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[16] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, “Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2016, pp. 191–195.

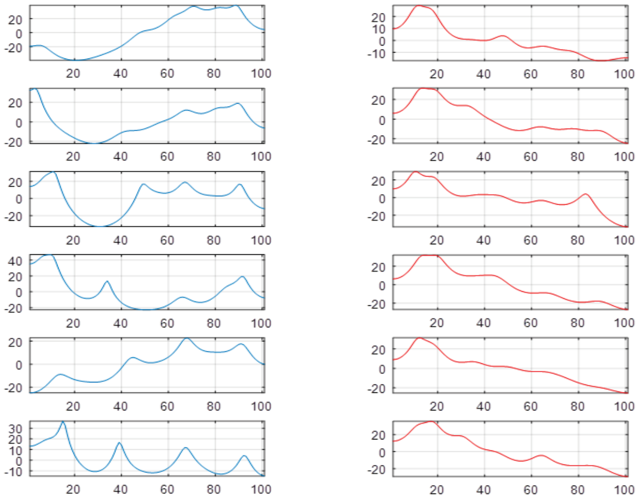
[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," *IEEE 2015 Automatic Speech Recognition and Understanding Workshop*, 2015.

[19] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7, pp. 588–601, 2007.

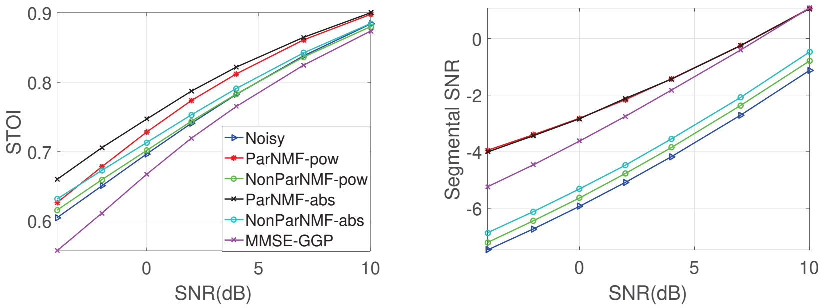
[20] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[21] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*. IEEE, 2015, pp. 266–270.



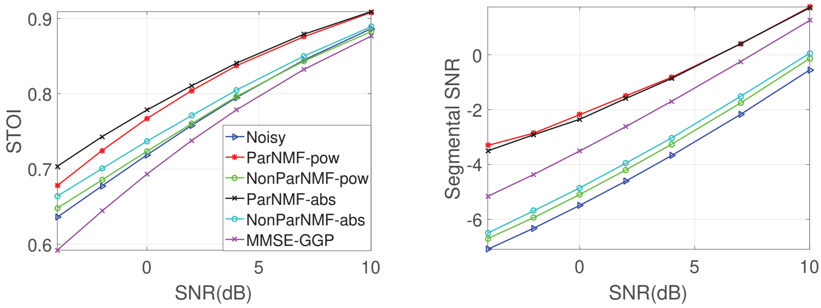
(a) **(b)**

Fig. F.1: Figure showing a set of (a) trained speech spectral envelopes and (b) noise spectral envelopes.



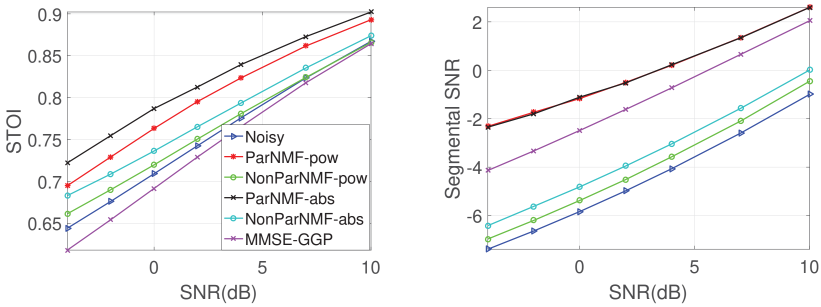
(a) STOI **(b)** Segmental SNR

Fig. F.2: Objective measures for babble noise when using speaker-specific codebook and a noise-specific codebook.



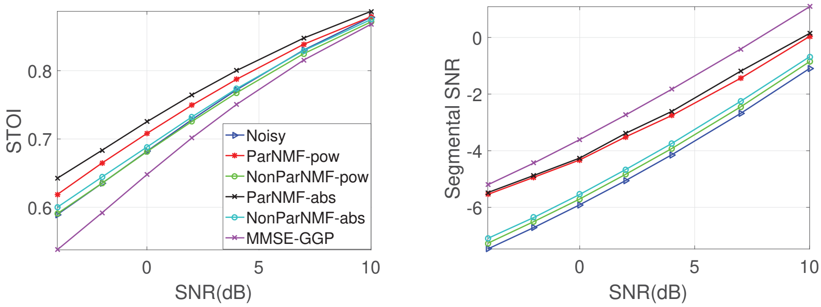
(a) STOI **(b)** Segmental SNR

Fig. F.3: Objective measures for restaurant noise when using speaker-specific codebook and a noise-specific codebook.



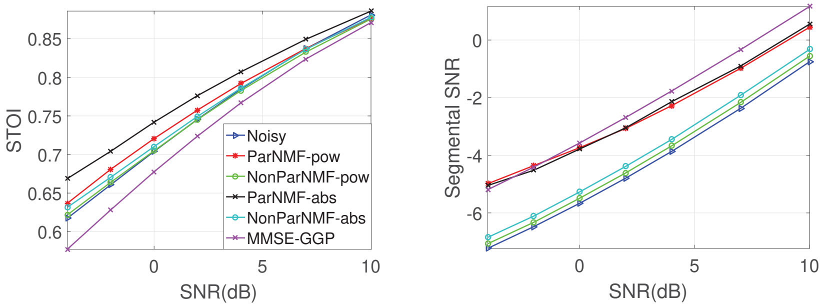
(a) STOI **(b)** Segmental SNR

Fig. F.4: Objective measures for exhibition noise when using speaker-specific codebook and a noise-specific codebook.



(a) STOI **(b)** Segmental SNR

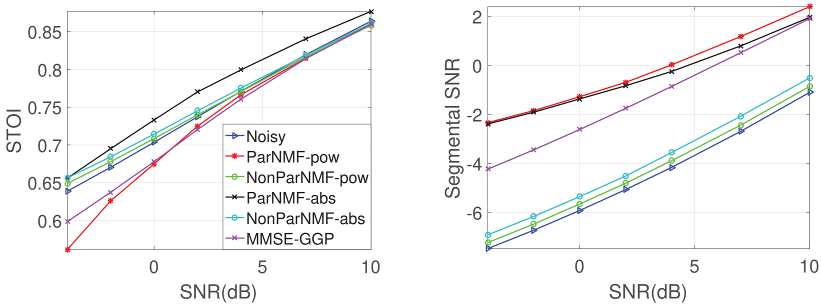
Fig. F.5: Objective measures for babble noise when using general speech codebook and a general noise codebook.



(a) STOI

(b) Segmental SNR

Fig. F.6: Objective measures for restaurant noise when using general speech codebook and a general noise codebook.



(a) STOI

(b) Segmental SNR

Fig. F.7: Objective measures for exhibition noise when using general speech codebook and a general noise codebook.

Paper G

Non-intrusive codebook-based intelligibility prediction

Charlotte Sørensen, Mathew Shaji Kavalekalam, Angeliki Xenaki, Jesper B. Boldt and Mads Græsbøll Christensen

The paper has been published in the
Speech Communication, 2018

Abstract

In recent years, there has been an increasing interest in objective measures of speech intelligibility in the speech processing community. Important progress has been made in intrusive measures of intelligibility, where the Short-Time Objective Intelligibility (STOI) method has become the de facto standard. Online adaptation of signal processing in, for example, hearing aids, in accordance with the listening conditions, requires a non-intrusive measure of intelligibility. Presently, however, no good non-intrusive measures exist for noisy, nonstationary conditions. In this paper, we propose a novel, non-intrusive method for intelligiblity prediction in noisy conditions. The proposed method is based on STOI, which measures long-term correlations in the clean and degraded speech. Here, we propose to estimate the clean speech using a codebook-based approach that jointly models the speech and noisy spectra, parametrized by auto-regressive parameters, using pre-trained codebooks of both speech and noise. In experiments, the proposed method is demonstrated to be capable of accurately predicting the intelligibility scores obtained with STOI from oracle information. Moreover, the results are validated in listening tests that confirm that the proposed method can estimate intelligibility from noisy speech over a range of signal-to-noise ratios.

1 Introduction

Human interaction depends on communication where speech has a central role. Inability to understand speech, e.g., due to hearing impairment, noisy background, or distortion in communication systems, can lead to ineffective communication and social isolation, and the development of speech enhancement methods [1, 2] is, therefore, a key concern in many applications. These include challenging applications such as hearing aids [3], telecommunication systems [4, 5], and architectural acoustics [6]. To assess the listening conditions in which speech processing would be beneficial, but also to evaluate the speech processing algorithms as such, a speech intelligibility measure is required [3, 5, 7].

A natural way of assessing the intelligibility of a degraded, i.e., processed, distorted or noisy speech signal is by performing subjective listening tests. Subjective speech intelligibility scores gives the percentage of correctly identified information from a degraded speech signal. However, subjective speech intelligibility experiments are time-consuming, expensive and cannot be used for real-time applications. Hence, there is a great interest in developing objective measures for speech intelligibility prediction. As opposed to subjective listening tests, objective intelligibility prediction algorithms are faster, cheaper and can be used for real-time processing.

The Articulation Index (AI) [8, 9] and the Speech Intelligibility Index (SII) [10] are some of the earliest metrics for prediction of speech intelligibility scores. The AI and SII use the signal-to-noise ratio (SNR) of speech excerpts in several frequency bands to estimate the intelligibility, hence they require that both the clean speech signal and the noise are available and uncorrelated as well as the noise to be stationary. The Extended SII (ESII) [11] and the Coherence SII (CSII) [12], are variants of SII which account for fluctuating noise and nonlinear distortions from clipping, respectively. The Speech Transmission Index (STI) [4] was introduced to predict the intelligibility of an amplitude modulated signal at the output of a transmission channel based on changes in the modulation depth across frequency of a probe signal. The STI, which requires a probe signal as reference, offers good prediction of speech intelligibility in reverberant and noisy conditions [4], but not for more adverse nonlinear distortions, such as those caused by spectral subtraction [13]. The Short-Time Objective Intelligibility (STOI) metric [14] predicts the intelligibility of a signal by its short-time correlation with its clean counterpart which is required as input. STOI estimates are accurate for time-frequency processed speech [15, 16]. The speech-based Envelope Power Spectrum Model (sEPSM) [17] estimates the SNR in the envelope-frequency domain and uses the noise signal alone as reference. The sEPSM accounts for the effects of additive noise and reverberation and some types of nonlinear processing such as spectral subtraction [17], but fails with other types of nonlinear processing such as ideal binary masks and phase jitter [16]. More recent work includes that of [18], which takes an information theoretical approach to the problem.

All the aforementioned methods are intrusive, i.e., they require either the clean speech signal or the noise interference as reference to estimate the intelligibility of the degraded signal. Access to the clean speech signal is impractical for many real-life applications or real-time processing systems. To overcome this limitation, a number of non-intrusive objective intelligibility measures have been proposed. The Speech to Reverberation Modulation energy Ratio (SRMR) [19] and the average Modulation-spectrum Area (ModA) [20] both provide intelligibility predictions based on the modulation spectrum of the degraded speech signal, i.e., in a non-intrusive manner. Other notable work includes the reduced dynamic range (rDR) based intelligibility measure [21], wherein the intelligibility is predicted directly from the dynamic range of the noisy speech, and the across-band envelope correlation (ABEC) metric [22], which is based on temporal envelope waveforms. Another approach to predict speech intelligibility non-intrusively is to first obtain an estimate of the clean speech signal which is thereafter used as reference to an intrusive method. Machine learning [23, 24], principal component analysis [7] or noise reduction [25, 26] methods have been proposed to reconstruct the clean signal from its degraded version and use it as input to the intrusive STOI metric for objective intelligibility prediction.

The present paper, which is an extension of our prior work [27], proposes a non-intrusive intelligibility metric, which uses the STOI measure non-intrusively by estimating the features of the clean reference signal from its degraded version. The proposed method, however, estimates the reference signal by identifying the entries of pre-trained codebooks of speech and noise spectra which best fit the data, i.e., the noisy speech signal. The resulting new metric is dubbed Non-Intrusive Codebook-based STOI (NIC-STOI). The method is inspired by the work [28, 29] which demonstrates that codebook-based approaches offer effective speech enhancement, even under nonstationary noise such as babble noise. Moreover, the approaches of [28, 29] are based on low-dimensional parametrizations of both the noise and speech spectra, more specifically, via auto-regressive (AR) models, something that engenders both effective training leading to small codebooks and computationally fast implementations. Furthermore, an AR process models the envelope of the signal's spectrum rather than its fine structure. Such models are suitable in this context since it is shown that the spectral envelope of speech is an important cue for intelligibility [30]. Compared to our previous work [26], which can be interpreted as sampling the speech spectrum at high-SNR frequencies based on the pitch, something that is consistent with the glimpsing model of speech perecetion [31], the new method is based on the complete speech spectrum. It should also be noted that we here address the problem of single-channel non-intrusive intelligibility prediction, which is a much more difficult task than the multichannel problem [25, 26], as the latter can use spatial information.

The rest of the paper is organized as follows. First, the principles of intelligibility prediction in the STOI method are described in Section 2. Then, the signal model that the proposed method is based on is detailed in Section 3, and the proposed non-intrusive method is described in in Section 4. The experimental details and results, which include both experiments with objective measures and a listening test, are first described in Section 5 and then discussed in Section 6. Finally, Section 7 concludes on the work.

2 Background

The STOI [14] metric predicts the speech intelligibility based on the correlation between the temporal envelopes of the clean and the degraded speech signal (see Fig. G.1). First, the clean and degraded speech signals are decomposed in time-frequency representations using a discrete Fourier transform. Then, these time-frequency representations are grouped in one-third octave frequency bins and short-time segments (384 ms). The short-time segments are normalized in order to account for global level differences of the input signals. Furthermore, the short-time segments are clipped to prevent time-frequency units that are already completely degraded from excessively

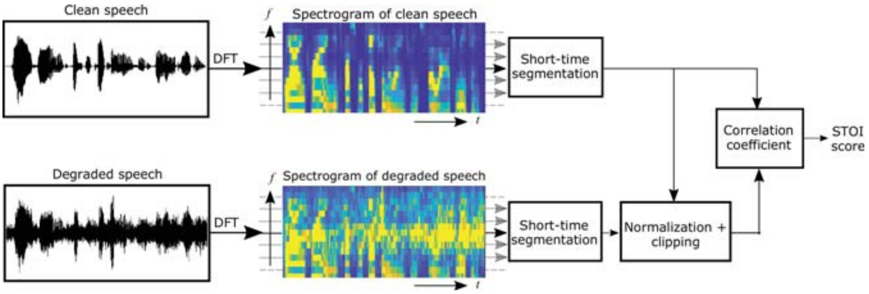


Fig. G.1: Block diagram of the STOI measure [14] that forms the basis for the proposed non-intrusive method. The STOI metric is based on the correlation between temporal envelopes of the clean and degraded speech in short time segments.

influencing the intelligibility score. Finally, the correlation of the signals is calculated over the short-time segments per frequency band. The STOI output is the average of the correlation coefficients across frequency bands and time-segments, i.e., a scalar value in the range 0-1 which relates monotonically to the average speech intelligibility scores.

3 Signal model

Assuming that a speech signal and a noise signal are generated by uncorrelated random processes, the corresponding noisy speech signal, $y(n)$, at time instance n is $y(n) = s(n) + w(n)$. In the proposed method, both the speech and the noise are modeled as stochastic processes, namely AR processes [28, 29]. Using such a stochastic AR model, a segment of the speech signal is expressed as

$$s(n) = -\sum_{i=1}^{P} a_s(i)s(n-i) + u(n), \tag{G.1}$$

which can also be expressed in vector notation as

$$u(n) = \mathbf{a}_s^T \mathbf{s}(n) \tag{G.2}$$

where P is the order of the AR process, $\mathbf{s}(n) = [s(n), s(n-1), \ldots, s(n-P)]^T$ is a vector collecting the P past speech samples, $\mathbf{a}_s = [1, a_s(1), a_s(2), \ldots, a_s(P)]^T$ is a vector containing the speech auto-regressive parameters with $a_s(0) = 1$, and $u(n)$, which here models the excitation, is zero mean white Gaussian noise with excitation variance σ_u^2. Transforming the AR model into the frequency domain, $A_s(\omega)S(\omega) = U(\omega) \Leftrightarrow S(\omega) = U(\omega)/A_s(\omega)$, results in the following power spectrum:

$$P_s(\omega) = |S(\omega)|^2 = \frac{\sigma_u^2}{|A_s(\omega)|^2}, \tag{G.3}$$

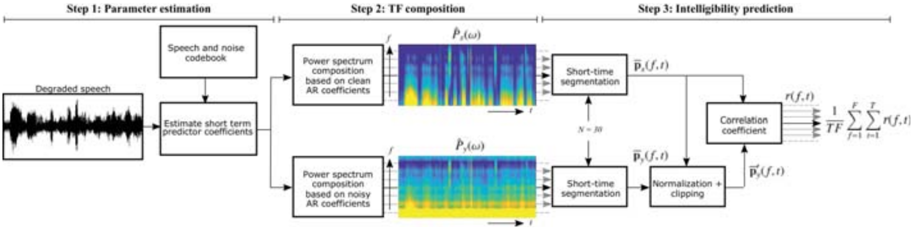

Step 1: Parameter estimation
Step 2: TF composition
Step 3: Intelligibility prediction
Speech and noise codebook
Degraded speech
Estimate short term predictor coefficients
Power spectrum composition based on clean AR coefficients
Power spectrum composition based on noisy AR coefficients
$\hat{P}_s(\omega)$
$\hat{P}_y(\omega)$
f
t
Short-time segmentation
Short-time segmentation
N = 30
$\bar{\mathbf{p}}_s(f,t)$
$\bar{\mathbf{p}}_y(f,t)$
$\bar{\mathbf{p}}'_y(f,t)$
Normalization + clipping
Correlation coefficient
r(f,t)
$\frac{1}{TF}\sum_{f=1}^{F}\sum_{t=1}^{T} r(f,t)$

Fig. G.2: Block diagram depicting the processing scheme of the proposed non-intrusive codebook-based short-time objective intelligibility (NIC-STOI) metric. The relevant features of the clean and degraded speech signals are estimated using a codebook-based approach as time-frequency power spectra, which replace the estimates in the front-end of the STOI method.

where $A_s(\omega) = \sum_{k=0}^{P} a_s(k)e^{-j\omega k}$. Similarly to the speech sginal, the noise signal can be modeled as

$$w(n) = -\sum_{i=1}^{Q} a_w(i)w(n-i) + v(n), \tag{G.4}$$

which can also be expressed as

$$v(n) = \mathbf{a}_w^T \mathbf{w}(n), \tag{G.5}$$

where Q is the order of the AR process, $\mathbf{w}(n) = [w(n), w(n-1), \ldots, w(N-Q)]^T$ is a vector collecting the Q past noise samples, $\mathbf{a}_w = [1, a_w(1), a_w(2) \ldots, a_w(Q)]^T$ where $a_w(0) = 1$, and $v(n)$ is zero mean white Gaussian noise with excitation variance σ_v^2. The noisy power spectrum is likewise given by

$$P_w(\omega) = |W(\omega)|^2 = \frac{\sigma_v^2}{|A_w(\omega)|^2}. \tag{G.6}$$

where $A_w(\omega) = \sum_{k=0}^{Q} a_w(k)e^{-j\omega k}$.

The models of the the speech and noise in (G.2) and (G.5), respectively, can be motivated as follows. The AR model has a long history in speech processing, where one of its uses is in modeling the speech production system (see, e.g., [32]), where it corresponds to a cylinder model of the vocal tract which is excited by a noise signal generated by the lungs. The model is, though, well-known not to be perfect. For example, it does not account for the nasal cavity and the Gaussian model is only a good model for unvoiced speech and less so for voiced speech [33]. Nevertheless, it remains useful for many purposes and here it is used as a low-dimensional representation of the speech spectrum. Regarding the noise, the model is good for many natural noise sources, but, in any case, it can be used for modeling arbitrary, smooth spectra of Gaussian signals [34].

4 The NIC-STOI measure

The proposed method provides an objective measure for speech intelligibility prediction given solely the degraded speech signal, i.e., non-intrusively.

The method is based on the speech and noise being additive and the AR models of the speech (G.2) and noise (G.5) signals. The speech and noise spectra are simultaneously estimated from the degraded speech signal using a Bayesian approach which uses the AR parameters as prior information for inference. The prior information is obtained from trained codebooks (dictionaries) of speech and noise AR parameters. The estimation is performed on short-time frames in order to account for non-stationary noise.

Figure G.2 depicts a block diagram of the NIC-STOI algorithm. The methodology comprises three main steps: 1) estimation of the parameters for the speech and noise AR models, 2) computation of the time-frequency representations for the clean, s, and noisy speech, y, signals from the estimated parameters, 3) prediction of speech intelligibility of the noisy speech signal with the STOI framework from the estimated spectra.

4.1 Step 1: Parameter Estimation

Let the column vector $\boldsymbol{\theta} = [\mathbf{a}_s;\ \mathbf{a}_w;\ \sigma_u^2;\ \sigma_v^2]$ comprise all parameters to be estimated, i.e., the AR coefficients and the excitation variances of the models of both speech and noise.

Bayes' theorem facilitates the computation of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ of the model parameters $\boldsymbol{\theta}$ conditioned on the observation of N noise samples, i.e., $\mathbf{y} = [\ y(0)\ y(1)\ \dots\ y(N-1)\ ]$, from the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, the prior distribution of the model parameters $p(\boldsymbol{\theta})$, and the marginal distribution of the data $p(\mathbf{y})$ [28, 29, 35]:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{G.7}$$

Based on the signal model introduced previously, the data likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, is a multi-variate zero-mean Gaussian distribution with covariance matrix, $\mathbf{R}_Y = \mathbf{R}_s + \mathbf{R}_w$, where $\mathbf{R}_s = \sigma_u^2(\mathbf{G}_s^T\mathbf{G}_s)^{-1}$ and $\mathbf{G}_s$ is a $N \times N$ lower triangular Toeplitz matrix defined by the AR parameters $\mathbf{a}_s$. More specifically, it is given by

$$\mathbf{G}_s = \begin{bmatrix} 1 & 0 & \dots & & 0 \\ a_s(1) & 1 & & & \\ \vdots & a_s(1) & & & \\ a_s(P) & \vdots & \ddots & & \vdots \\ 0 & a_s(P) & \ddots & & \\ \vdots & \vdots & & 1 & \\ 0 & 0 & \dots & a_s(1) & 1 \end{bmatrix} \tag{G.8}$$

while the noise covariance matrix can be expressed as $\mathbf{R}_w = \sigma_v^2(\mathbf{G}_w^T\mathbf{G}_w)^{-1}$ with $\mathbf{G}_w$ being defined in a similar manner as $\mathbf{G}_s$ but from $\mathbf{a}_w$. Then, the minimum mean square error (MMSE) estimate is given by [36]

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MMSE}} &= \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}\left[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})^2\right] = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) \\ &= \int_\Theta \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_\Theta \boldsymbol{\theta} \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} d\boldsymbol{\theta}, \end{aligned} \tag{G.9}$$

where Θ is the support space of the parameters to be estimated. Based on the independence of speech and noise signals, and further assuming that the AR process and excitation variances are independent, the prior distribution of the model parameters can be simplified as

$$p(\boldsymbol{\theta}) = p(\mathbf{a}_s, \sigma_u^2)p(\mathbf{a}_w, \sigma_v^2) \approx p(\mathbf{a}_s)p(\sigma_u^2)p(\mathbf{a}_w)p(\sigma_v^2).$$

Limiting the support of the AR parameter vectors $\mathbf{a}_s$ and $\mathbf{a}_w$ to predefined codebooks of size N_s and N_w, respectively, the corresponding excitation variances are estimated through a maximum likelihood (ML) approach

$$\{\sigma_{u,ij}^{2,\text{ML}}, \sigma_{v,ij}^{2,\text{ML}}\} = \arg\max_{\sigma_u^2,\sigma_v^2} \log p(\mathbf{y}|\mathbf{a}_{s_i}^{\text{CB}};\ \mathbf{a}_{w_j}^{\text{CB}};\ \sigma_u^2;\ \sigma_v^2),$$

where $\mathbf{a}_{s_i}^{\text{CB}}$ is the i^{th} entry of the speech codebook and $\mathbf{a}_{w_j}^{\text{CB}}$ is the j^{th} entry of the noise codebook. The Gaussian likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ can be expressed in the frequency domain in terms of the Itakura-Saito distortion measure between the observed, $P_y(\omega)$, and modeled, $\hat{P}_y^{ij}(\omega)$, noisy data power spectrum, i.e.,

$$p(\mathbf{y}|\mathbf{a}_{s_i}^{\text{CB}};\ \mathbf{a}_{w_j}^{\text{CB}};\ \sigma_{u,ij}^2;\ \sigma_{v,ij}^2) \propto e^{-d_{\text{IS}}(P_y(\omega),\hat{P}_y^{ij}(\omega))}, \tag{G.10}$$

where $d_{\text{IS}}(\cdot,\cdot)$ is the Itakura-Saito divergence, which is given by [29, 37]

$$\begin{aligned} &d_{\text{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) = \\ &\qquad \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)} - \ln\left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)}\right) - 1\right) d\omega. \end{aligned} \tag{G.11}$$

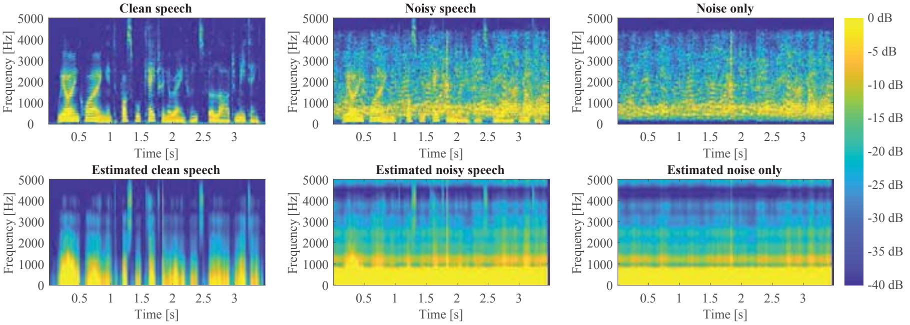


Fig. G.3: The top panel depicts from left to right, respectively, the spectra of the original clean speech signal, the degraded noisy speech signal at 0 dB SNR and noise only. In the bottom panel their corresponding estimated spectra using the codebook-based approach are depicted.

Equation (G.11) makes use of the modeled noisy power spectrum, which is here given by

$$\hat{P}_y^{ij}(\omega) = \frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2}, \tag{G.12}$$

where $A_s^i(\omega) = \sum_{k=0}^{P} a_s^{i,\text{CB}}(k)e^{-j\omega k}$ and $A_w^j(\omega) = \sum_{k=0}^{Q} a_w^{j,\text{CB}}(k)e^{-j\omega k}$ being the spectra of the i^{th} and j^{th} vector from the speech codebook and noise codebook, respectively.

Assuming that the modeling error between $P_y(\omega)$ and $\hat{P}_y^{ij}(\omega)$ is small and by using a second-order Taylor series approximation of $\ln(\cdot)$, the Itakura-Saito divergence can be approximated as [29]

$$d_{\text{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) \approx \frac{1}{2} d_{\text{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right), \tag{G.13}$$

where the log-spectral distortion between the observed and modeled noisy spectrum, $d_{\text{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right)$, which is given by

$$d_{\text{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right) = \tag{G.14}$$
$$\frac{1}{2\pi}\int_0^{2\pi}\left|\ln\left(\frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2}\right) - \ln\left(P_y(\omega)\right)\right|^2 d\omega$$

Finally, the ML estimates of the speech and noise excitation variances, $\sigma_{u,ij}^{2,\text{ML}}$ and $\sigma_{v,ij}^{2,\text{ML}}$ can be obtained by

$$\{\sigma_{u,ij}^{2,\text{ML}}, \sigma_{v,ij}^{2,\text{ML}}\} = \arg\min_{\sigma_u^2, \sigma_v^2} d_{\text{LS}}\left(P_y(\omega), \hat{P}_y^{ij}(\omega)\right), \tag{G.15}$$

which is solved by differentiating (G.14) with respect to σ_u^2 and σ_v^2 and setting the result equal to zero [28, 35]. This results in the following estimate of the excitation variance for the speech:

$$\begin{aligned}\sigma_{u,ij}^{2,\mathrm{ML}} = \frac{1}{\Psi_{ij}}\Bigg(&\sum_{\omega}\frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4}\sum_{\omega}\frac{1}{P_y(\omega)|A_s^i(\omega)|^2}\\ -&\sum_{\omega}\frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2}\sum_{\omega}\frac{1}{P_y(\omega)|A_w^j(\omega)|^2}\Bigg).\end{aligned}$$

Similarly, the estimate of for excitation variance of the noise is given by

$$\begin{aligned}\sigma_{v,ij}^{2,\mathrm{ML}} = \frac{1}{\Psi_{ij}}\Bigg(&\sum_{\omega}\frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4}\sum_{\omega}\frac{1}{P_y(\omega)|A_w^j(\omega)|^2}\\ -&\sum_{\omega}\frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2}\sum_{\omega}\frac{1}{P_y(\omega)|A_s^i(\omega)|^2}\Bigg).\end{aligned}$$

The quantity Ψ_{ij} is given by

$$\begin{aligned}\Psi_{ij} = &\sum_{\omega}\frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4}\sum_{\omega}\frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4}\\ &-\left(\sum_{\omega}\frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2}\right)^2.\end{aligned}\tag{G.16}$$

Finally, based on these estimates, the quantities in (G.9) are estimated from their discrete counterparts, which are given by

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_s N_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w}\boldsymbol{\theta}_{ij}\frac{p(\mathbf{y}|\boldsymbol{\theta}_{ij})}{p(\mathbf{y})}\tag{G.17}$$

and

$$p(\mathbf{y}) = \frac{1}{N_s N_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w}p(\mathbf{y}|\boldsymbol{\theta}_{ij}),\tag{G.18}$$

where $\boldsymbol{\theta}_{ij} = [\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_{u,ij}^{2,\mathrm{ML}};\ \sigma_{v,ij}^{2,\mathrm{ML}}]$ is the resulting parameter vector for the i^{th} entry of the speech codebook and the j^{th} entry of the noise codebook and the final estimates are denoted as $\hat{\boldsymbol{\theta}} = [\hat{\mathbf{a}}_s;\ \hat{\mathbf{a}}_w;\ \hat{\sigma}_u^2;\ \hat{\sigma}_v^2]$. These estimates can be thought of as being obtained from an average over all possible models with each model being weighted by its posterior. We remark that codebook combinations that result in infeasible, negative values for either the speech or noise excitation variances should be neglected. Since all ML estimates of the excitation variances and the predefined codebook entries contribute with equal probability, the prior is non-informative and is omitted in (G.9). It should also be noted that the weighted summation of the AR parameters can be performed in the line spectral frequency (LSF) domain whereby a stable inverse filters is ensured, something that is not always the case when operating directly on the AR parameters [28, 29].

4.2 Step 2: TF composition

The estimated parameters in $\hat{\boldsymbol{\theta}}$, obtained using (G.17), are then used to compute the time-frequency (TF) power spectra of the estimated speech and noise spectra as

$$\hat{P}_s(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2}, \tag{G.19}$$

where $\hat{A}_s(\omega) = \sum_{k=0}^{P} \hat{a}_s(k) e^{-j\omega k}$, and

$$\hat{P}_w(\omega) = \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}, \tag{G.20}$$

where $\hat{A}_w(\omega) = \sum_{k=0}^{Q} \hat{a}_w(k) e^{-j\omega k}$. The AR parameters, i.e., $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_w$, determine the shape of the envelope of the corresponding signals $\hat{S}(\omega)$ and $\hat{W}(\omega)$, respectively. The excitation variances, $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$, determine the overall signal power. Finally, the noisy spectrum is composed as the combined sum of the clean and the noise power spectra:

$$\hat{P}_y(\omega) = \hat{P}_s(\omega) + \hat{P}_w(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2} + \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}. \tag{G.21}$$

These time-frequency spectra replace the discrete Fourier transform of the clean reference signal and the noisy signal in the original STOI measure, respectively.

4.3 Step 3: Intelligibility Prediction

The STOI measure is used for intelligibility prediction with the estimated spectra $\hat{P}_s(\omega)$ (G.19) and $\hat{P}_y(\omega)$ (G.21) as inputs. First, the frequency bins of $\hat{P}_s(\omega)$ and $\hat{P}_y(\omega)$ are grouped into 15 one-third octave bands denoted by $\overline{P}_s(f,t)$ and $\overline{P}_y(f,t)$, respectively, with the lowest center frequency set to 150 Hz and the highest set to 4.3 kHz. The short-time region of the temporal envelopes of the clean speech is defined as $\overline{\mathbf{p}}_s(f,t) = [\overline{P}_s(f,t-N+1), \overline{P}_s(f,t-N+2), \ldots, \overline{P}_s(f,t)]^T$, where N is the length of the short-time regions and is set to 30, resulting in a short-time region of 384 ms as in the original STOI implementation [14]. In the same manner, the short-time region of the degraded speech is given by $\overline{\mathbf{p}}_y(f,t)$. The short-time regions of the degraded speech, $\overline{\mathbf{p}}_y(f,t)$, are further clipped by a normalization procedure in order to de-emphasize the impact of region in which noise dominates the spectrum:

$$\overline{\mathbf{p}}'_y(f,t) = \min\left(\frac{\|\overline{\mathbf{p}}_s(f,t)\|_2}{\|\overline{\mathbf{p}}_y(f,t)\|_2}\overline{\mathbf{p}}_y(f,t), (1+10^{-\beta/20})\overline{\mathbf{p}}_s(f,t)\right)$$

where $\|\cdot\|_2$ denotes the l_2 norm and $\beta = -15$ dB is the lower signal-to-distortion ratio. The local correlation coefficient, $r(f,t)$, between $\overline{\mathbf{p}}'_y(f,t)$ and $\overline{\mathbf{p}}_s(f,t)$ at frequency f and time t, is defined as

$$r(f,t) = \frac{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^T(\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})}{\sqrt{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^2}\sqrt{(\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})^2}},$$

where $\mu(\cdot)$ denotes the sample average of the corresponding vector. Given the local correlation coefficient, the NIC-STOI prediction is given by averaging across all bands and frames as

$$d_{NS} = \frac{1}{TF}\sum_{f=1}^{F}\sum_{t=1}^{T} r(f,t). \tag{G.22}$$

Table G.1: Sentence syntax of the GRID database [38] which is used in the subjective listening test. Each sentence is constructed from (in order) a combination of a command, color, preposition, letter digit, and adverb.

Command	Color	Preposition	Letter	Digit	Adverb
bin lay place set	blue green red white	at by in with	A-Z (no W)	0-9	again now please soon

5 Experimental Details and Results

5.1 Performance Measures

The non-intrusive intelligibility prediction is given by d_{NS}, for the different conditions to be evaluated. Whereas the ground truth, denoted by d_S, for these conditions are given by the intrusive STOI scores. Similarly to the approach in [24], the original true STOI score is expected to be well-correlated with the subjective intelligibility. Thus, the purpose is to predict the intrusive STOI score of a given condition using a non-intrusive method. The performance of the objective intelligibility predictions are evaluated using three performance metrics often used for assessing objective intelligibility predictions [3, 14, 39]:

- The Pearson correlation coefficient (ρ) quantifies the linear relationship between the predicted non-intrusive intelligibility scores and true STOI scores or subjective intelligibility scores, where a higher ρ indicates higher correlation.
- Kendall's Tau (τ) characterizes the ranking capability by describing the monotonic relationship between the predicted intelligibility scores and true STOI scores or subjective intelligibility scores, where a higher τ represents better performance [40]. It is defined as $\tau = 2(n_c - n_d)/N(N-1)$, where n_c is the number of concordant pairs, i.e. ordered in the same way, and n_d is the number of discordant pairs, i.e. ordered differently.
- The standard deviation of the prediction error (σ) is given as a measure of the estimation accuracy of the predicted non-intrusive intelligibility scores, where a lower σ implies better results.

5.2 Experimental Details

The results reported in this paper are based on both objective measurements and subjective listening tests. For the results based on the objective measures, the proposed metric, NIC-STOI, is evaluated on a test set of 100 speech utterances (full sentences), 50 male and 50 female, randomly selected from the EUROM_1 database of the English corpus [41]. The interfering additive noise signal is babble noise from the AURORA database. The babble noise contains many speakers in a reverberant acoustical environment. The sentences and interfering additive noise signal are both resampled to 10 kHz. Segments randomly selected from the additive noise signal are added to the EUROM_1 sentences at different SNR levels in the range of -30 to 30 dB SNR in steps of 10 dB SNR.

For further evaluation of the proposed metric, a subjective listening test has also been carried out to provide a data set for comparing NIC-STOI and SRMR. Stimuli were the fixed-syntax sentences from the GRID corpus database [38] mixed with the babble signal from the AURORA database with an SNR range -8 to 0 dB. The grid corpus consists of sentences spoken by 34 talkers (16 female and 18 male). The sentences are simple, syntactically identical phrases, e.g. "place blue in A 4 again", and the listeners are asked to identify the color, letter, and digit after listening to the stimuli using a user-controlled MATLAB interface. The syntax and words of the GRID corpus are shown in Table G.1. A total of nine subjects were used for the experiment which took around 30 minutes per subject. Intelligibility was defined as the number of keywords correctly identified per stimulus resulting in a fraction of either 0, 1/3, 2/3, or 1 being correct. A total of 220 stimuli were used, approximately 2 s in duration each, with the same stimuli being used for both NIC-STOI and SRMR: 5 SNR levels times 44 different sentences. We remark that to reduce intra- and intersubject variability the condition-averaged results are used for comparison and mapping of the objective results to subjective performance [3, 42]. Measuring intelligibility on a short time-scale (i.e., from short stimuli less than 2 s in duration each) with non-stationary noise types implies a high variance for both subjective and objective evaluations, i.e., precise estimation of intelligibility requires multiple sentences and not only a few words. However, it is difficult to execute subjective listening tests using long sentences or phrases as stimulus for which reason the average of many shorter sentences is here used instead.

The AR parameters and excitation variances of both the speech and noise signal are estimated on frames with a length of 256 samples. The speech and, thus, the estimated parameters are assumed to be stationary over these very short 25.6 ms frames. The frames are windowed using a Hann window with 50 % overlap between adjacent frames. The AR model orders P and Q of the speech and noise, respectively, are set to 14 in accordance with the literature [28, 29, 35]. The speech codebook is trained using the generalized Lloyd algorithm (GLA) on 10 minutes of speech from multiple speakers in the EUROM_1 database in order to ensure a sufficiently general speech model [28, 43]. We stress that the speakers included in the test set are not used for the training of the speech codebook. The noise codebook is trained on 2 minutes of babble talk. The sizes of the speech and noise codebooks are $N_s = 64$ and $N_w = 8$, respectively.

5.3 Experimental Results

An example of the spectrum of a speech signal from the test set is shown in Fig. G.3 . The spectra of the original clean speech signal, the degraded noisy signal at 0 dB SNR and the noisy only are depicted in the top panel from left to right, respectively. The corresponding estimated spectra of the relevant signal features are shown in the bottom panel. The spectra are generated using trained codebooks of speech and noise spectral shapes. The estimated clean spectrum (bottom left panel) and estimated noisy spectrum (bottom middle panel) are used as input to the intrusive STOI framework.

The performance of the NIC-STOI metric is evaluated against the intrusively computed scores of the original STOI metric as ground truth. In Fig. G.4, the estimated NIC-STOI scores have been plotted against the intrusive STOI scores. The plot shows good performance by means of a strong

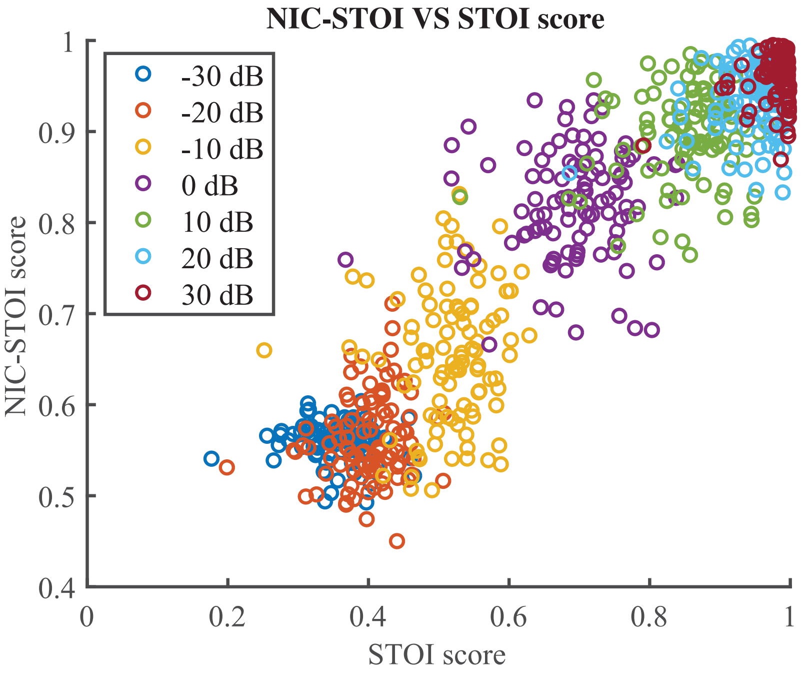


Fig. G.4: Scatter plot of the predicted STOI scores using the non-intrusive codebook-based STOI, NIC-STOI, metric.

monotonic relationship between NIC-STOI and STOI, such that a higher NIC-STOI score also corresponds to a higher STOI score. Furthermore, a strong linear correlation can be observed between the two measures. This observation is also supported by the performance criteria, where NIC-STOI achieves a Pearson's correlation of $\rho = 0.94$, Kendall's Tau of $\tau = 0.70$ and a standard deviation of the prediction error $\sigma = 0.14$ for STOI, implying a high correlation. This indicates that the proposed non-intrusive version of STOI can offer a comparable performance to the original intrusive STOI.

Fig. G.5 depicts the averaged predictions (± standard deviation) of the NIC-STOI scores in the scatter plot in Fig. G.4 for male (blue line), female (red line) and both genders (yellow line), where the performance measures are given in Tab. G.2. As it can be observed, the measure performs equally well whether the method is tested using either a gender specific clean speech codebook or a generic clean speech codebook. This suggests that the method generalizes well and does not capture gender specific effects due to the very generic and smooth structure of the spectra of the auto-regressive processes.

In Fig. G.6 the STOI measure (purple line) and the NIC-STOI measure (male: blue line; female: red line; both genders: yellow line) are depicted as

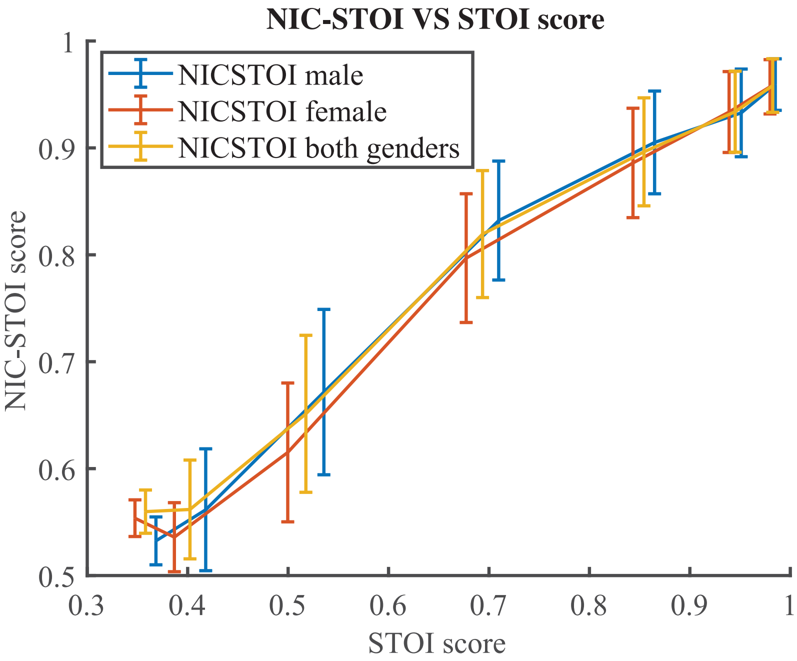


Fig. G.5: Averaged NIC-STOI scores (± standard deviation) against the intrusively computed STOI score.

function of SNR. There is a clear monotonic correspondence between NIC-STOI and STOI, such that a higher STOI measure results in a higher NIC-STOI score. Furthermore, the NIC-STOI scores also increase with increasing SNR.

Subjective results, in terms of intelligibility as a function of SNR, are shown in Fig. G.7 together with objective results obtained using the proposed NIC-STOI and SRMR. The error bars in the Figure are 95 % confidence intervals computed using a normal distribution for the SRMR and NIC-STOI methods and the normal approximation for the binomial confidence interval of the subjective results from the listening test. Note that to map the objective results to subjective intelligibility, a sigmoid function has been fitted to the average data as described in Section 5.2. As can be seen, the proposed method performs well and is capable of predicting the speech intelligibility with similar variance over a range of SNRs. The results do not, however, enable the conclusion that NIC-STOI is superior to SRMR although NIC-STOI has a better alignment with the subjective data, as both metrics have a good performance, even at low SNRs, and the confidence intervals overlap. Concerning the probability intervals, the intervals for both NIC-STOI and SRMR are large, as is to be expected, due to the short sentences in the GRID corpus and the limited number of stimuli for each SNR level. One thing to

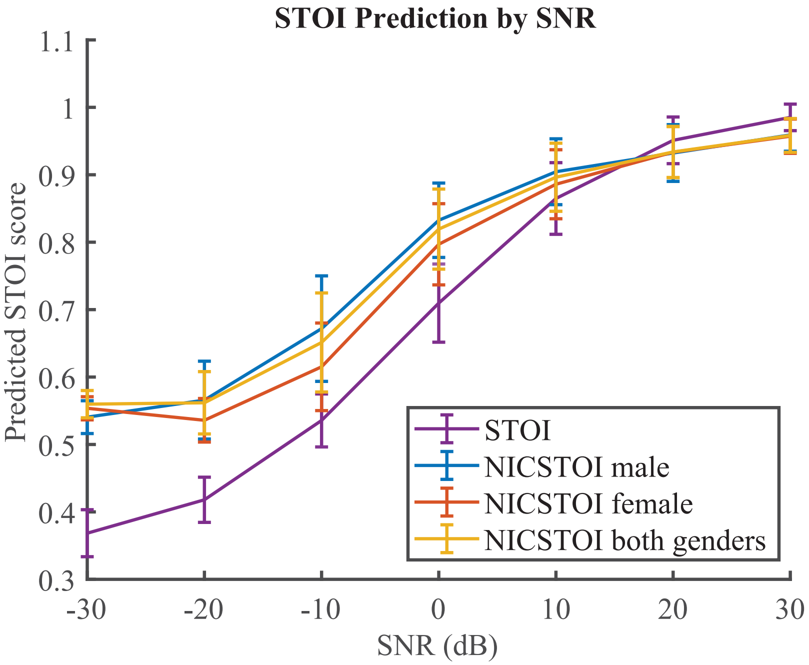


Fig. G.6: Averaged NIC-STOI and STOI scores (± standard deviation) per SNR condition.

note is that the variance for SRMR increases as the SNR decreases, whereas NIC-STOI exhibits a similar variance across SNRs.

6 Discussion

Since the framework of NIC-STOI is based on an AR model, it only captures the overall envelope structure and not the fine structure of the speech signal as illustrated in Fig. G.3 [29, 37]. The envelope of the speech has been shown to be a good predictor for speech intelligibility in previous intrusive intelligibility frameworks, i.e. STI and EPSM [4, 17, 30]. Extensive vocoder simulations also support these findings, where a high speech intelligibility can be obtained in quiet solely from the envelope content in only four spectral bands [30]. As such, only modeling the envelope structure of the clean speech as the essential features in NIC-STOI is assumed to be an appropriate predictor for speech intelligibility. Moreover, the promising results in [28], which show improvements of STOI scores for single channel enhancement over the noise signal, also support that the proposed model captures the essential features of the speech, as the estimated AR parameters and excitation variances are used in a speech production model in [28] to enhance the noisy

Table G.2: Performance of the proposed metric in terms of Pearson's correlation (ρ), and Kendall's tau (τ) and the standard deviation of the prediction error (σ) between NIC-STOI and STOI.

Condition	ρ	τ	σ
Male	0.93	0.70	0.14
Female	0.94	0.71	0.13
Both genders	0.94	0.70	0.14

speech with a Kalman filter.

Both the reported objective and subjective results show that the proposed method works well. The subjective results show that the proposed method can predict the intelligibility of a listening experiment over a range of 10 dB. Although the predicted values exhibit a high variance, as is to be expected of this type of experiment, this variance is similar to the one obtained with SRMR. The objective results indicate that NIC-STOI performs very well for a broad range of SNRs, even down to -30 dB SNR where the noisy speech is expected to be unintelligible. It should be noted that while NIC-STOI appears to deviate from STOI for very low SNRs, this is less important as, according to [3], a STOI score of 0.6 approximately corresponds to zero intelligibility. Even though the absolute value of STOI depends highly on the specific speech material and listening environment, the broad working range of NIC-STOI should cover the range of intelligibility. Hence, any score below this threshold can be simply assumed unintelligible. Here, it is also important to stress that the overall aim of NIC-STOI is to have a monotonic relation with the intrusively computed STOI scores, and not necessarily to predict the absolute STOI scores. However, the offset observed between the predicted NIC-STOI scores and STOI scores in Fig. G.6 can easily be accounted for by the observed linear trend between the two measures depicted in Fig. G.4, such that the absolute STOI score can be predicted by means of the estimated NIC-STOI score.

It should be noted that STOI was among the first intrusive intelligibility metrics with very good performance, but since it was first introduced other intrusive metrics have also been proposed that show good performance. The front-end of NIC-STOI, that forms the basis of the present work, could also quite possibly be used for other intrusive frameworks, provided that they are also based on spectral features of the noisy and clean speech. Regarding this, it is interesting to note that the estimation of the parameters in short-time segments based on the current observation makes the front-end suitable for non-stationary noise conditions. However, STOI does not work well for highly non-stationary interferers due to the analysis window length. Therefore, it could be interesting to investigate using the Extended STOI (ESTOI) as a back-end to NIC-STOI instead, since this method has been developed to

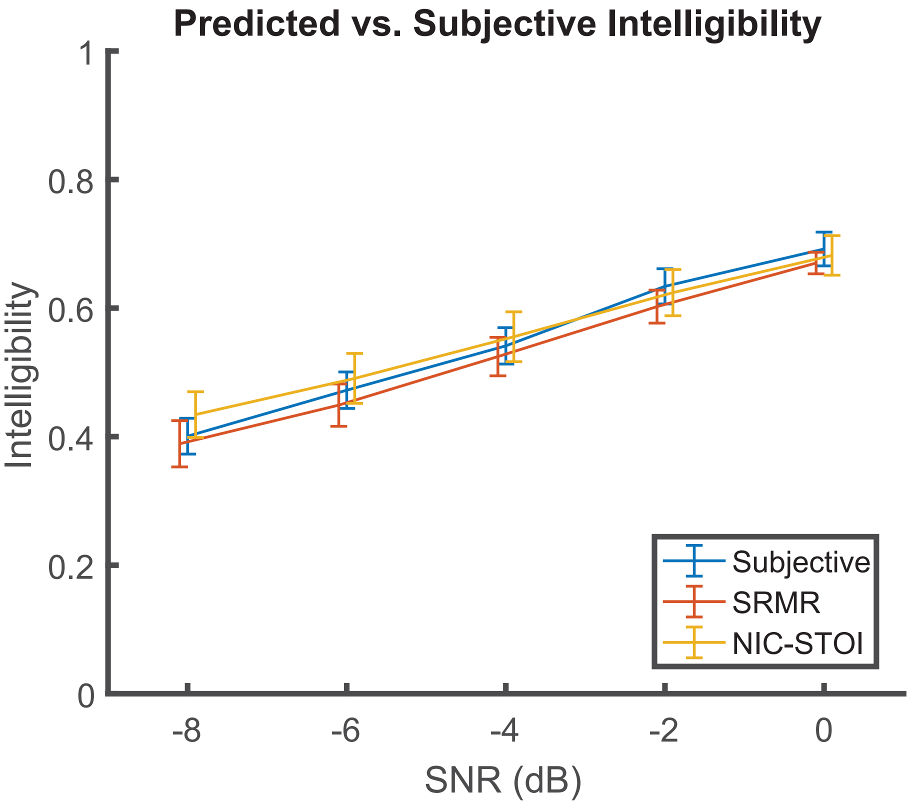


Fig. G.7: Intelligibility as a function of SNR for subjective listening experiments and as predicted by the proposed NIC-STOI and SRMR. Shown are the means and their 95 % confidence intervals.

work well for highly modulated noise sources [44].

Correlation-based metrics including STOI are generally not suitable for predicting the intelligibility of reverberant speech and, thus, it is likely that NIC-STOI will fail in such conditions [14, 45]. Furthermore, the short time frames used in STOI might also have a negative impact on the application of NIC-STOI to reverberant speech, as short time frames cannot capture all the effects of reverberation, such as temporal smearing [14]. Currently, SRMR and ModA are the most well-studied non-intrusive intelligibility metrics. They have both been proposed for predicting the intelligibility of reverberant speech, where they both show good performance [3, 19, 20]. Even though these metrics are aimed for reverberant speech, they have also been tested for noisy and processed speech [3], where they perform reasonably well. However, it would seem that SRMR and ModA are a more suitable choice for reverberant speech, while our proposed method, NIC-STOI, which takes into account the presence of noise, is a more suitable choice for additive degradations, such as background noise and interferences. In this connection, it should also be mentioned that the proposed method is computationally much more demanding than SRMR and ModA, mainly due to the codebook search, although approximate methods for implementation of this exist [46].

In closing, we remark that the proposed method is not expected to account well for non-linear signal processing, since it is based on an additive noise model as well as the codebooks being trained on clean speech signals and noise signals. However, testing the method on the Ideal Time-Frequency Segregation (IFTS) data set from [47], which was used for evaluating the original STOI measure [14], results in a Pearson correlation of 0.70, which is surprisingly good. For comparison, NIC-STOI outperforms the non-intrusive intelligibility metric, SRMR [3, 19], which achieves a Pearson correlation of 0.24 [7], although it should be noted that SRMR, as already mentioned, was designed for reverberant speech. However, the newly proposed Non-Intrusive STOI (NI-STOI) measure [7] achieves a Pearson correlation of 0.71 for the data set [47], which is on par with the results obtained for NIC-STOI. We remark that NI-STOI is not completely non-intrusive, as it is based on the ideal voice activity detector used in the intrusive STOI metric [7].

7 Conclusion

In this paper, a non-intrusive codebook-based short-time objective intelligibility metric, called NIC-STOI, has been proposed. It is based on an intrusive intelligibility metric, STOI, but, unlike STOI, it does not require access to the clean speech signal. Instead, the proposed method estimates the spectrum of the reference signal by identifying the entries of pre-trained spectral codebooks of speech and noise spectra, parametrized by auto-regressive parameters, which best fit the observed signal, i.e., the noisy speech signal. This is done in a statistical framework wherein parameters are estimated by minimizing the Itakura-Saito divergence for combinations of speech and noise models. This is equivalent to maximum likelihood estimation for Gaussian distributed signals. The proposed NIC-STOI metric is shown, in experiments, to be highly correlated with STOI (with a Pearson correlation of 0.94 and a standard deviation of the prediction error of 0.14) and is also validated in a listening experiment assessing speech intelligibility. Hence, it can be used for the assessment of speech intelligibility when a clean reference signal is not available. This could be used, for example, for online optimization of hearing aids.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. Taylor & Francis, 2007.

[2] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Commun.*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[3] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, “Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[4] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech-transmission quality,” *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[5] S. Jørgensen, J. Cubick, and T. Dau, “Speech intelligibility evaluation for mobile phones.” *Acustica United with Acta Acustica*, vol. 101, p. 1016 – 1025, 2015.

[6] T. Houtgast and H. J. M. Steeneken, “A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[7] A. Heidemann Andersen, J. de Haan, Z.-H. Tan, and J. Jensen, “A non-intrusive short-time objective intelligibility measure,” in *ICASSP*, March 2017, pp. 5085–5089.

[8] J. B. Allen, “Harvey Fletcher’s role in the creation of communication acoustics,” *J. Acoust. Soc. Am.*, vol. 99, no. 4, pp. 1825–1839, 1996.

[9] N. R. French and J. C. Steinberg, “Factors governing the intelligibility of speech sounds,” *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[10] ANSI S3.5, 1997, *Methods for the Calculation of the Speech Intelligibility Index, American National Standards Institute, New York, USA*, Std., 1997.

[11] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise,” *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.

[12] J. M. Kates and K. H. Arehart, “Coherence and the speech intelligibility index,” *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[13] C. Ludvigsen, C. Elberling, and G. Keidser, “Evaluation of a noise reduction method comparison between observed scores and scores predicted from STI,” *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.

[16] H. Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, and T. Dau, “Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain,” *J. Acoust. Soc. Am.*, vol. 140, no. 4, p. 2670–2679, 2016.

[17] S. Jørgensen and T. Dau, “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing,” *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[18] S. van Kuyk, W. B. Kleijn, and R. Hendriks, “An instrumental intelligibility metric based on information theory,” in *ICASSP*, 2018.

[19] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[20] F. Chen, O. Hazrati, and P. C. Loizou, “Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure,” *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 311 – 314, 2013.

[21] F. Chen, “Modeling noise influence in speech intelligibility non-intrusively by reduced speech dynamic range,” in *Interspeech*, 2016.

[22] ——, “Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation,” *Biomedical Signal Processing and Control*, vol. 24, pp. 109 – 113, 2016.

[23] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, “Twin-hmm-based non-intrusive speech intelligibility prediction,” in *ICASSP*, March 2016, pp. 624–628.

[24] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, “A data-driven non-intrusive measure of speech quality and intelligibility,” *Speech Commun.*, vol. 80, pp. 84–94, 2016.

[25] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, “Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids,” in *EUSIPCO*, August 2016, pp. 1358–1362.

[26] C. Sørensen, A. Xenaki, J. Boldt, and M. Christensen, “Pitch-based non-intrusive objective intelligibility prediction,” in *ICASSP*, March 2017, pp. 386–390.

[27] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, “Non-intrusive intelligibility prediction using a codebook-based approach,” in *EUSIPCO*, August 2017, pp. 226–230.

[28] M. Kavalekalam, M. Christensen, F. Gran, and J. Boldt, “Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach,” in *ICASSP*, March 2016, pp. 191–195.

[29] S. Srinivasan, J. Samuelsson, and W. Kleijn, “Codebook-based Bayesian speech enhancement for nonstationary environments,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[30] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[31] M. Cooke, “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[32] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*. Prentice-Hall, 1987.

[33] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, “Sparse linear prediction and its applications to speech processing,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20(5), pp. 1644–1657, 2012.

[34] J. Makhoul, “Linear Prediction: A Tutorial Review,” *Proceedings of the IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[35] S. Srinivasan, J. Samuelsson, and W. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[37] K. Paliwal and W. Kleijn, “Quantization of LPC parameters,” in *Speech Coding and Synthesis*. Elsevier Science, 1995, pp. 433–468.

[38] M. Cooke and J. Barker, “An Audio-visual corpus for speech perception and automatic speech recognition (L),” *J. Acoust. Soc. Am.*, vol. 120(5), pp. 2421–2424, Nov. 2006.

[39] A. H. Andersen, J. M. de Hann, Z.-H. Tan, and J. Jensen, “Predicting the intelligibility of noisy and nonlinearly processed binaural speech,” *IEEE Tran. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.

[40] M. G. Kendall, “A new measure of rank correlation,” *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.

[41] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, “EUROM - a spoken language resource for the EU,” in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[42] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, “Speech quality estimation: Models and trends,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[43] Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[44] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov 2016.

[45] R. L. Goldsworthy and J. E. Greenberg, “Analysis of speech-based speech transmission index methods with implications for nonlinear operations,” *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[46] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1993.

[47] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, 2009.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-367-9

AALBORG UNIVERSITY PRESS