CHAT
WORKSHOP

Proceedings of the 1st International Workshop on **Challenges in Hearing Assistive Technology**

CHAT 2017

Stockholm, 19th August 2017

Published by:
Jon Barker
http://spandh.dcs.shef.ac.uk/chat/

ISBN: 978-1-5272-1357-9

Credits:
Editors: Jon Barker, John Culling, John Hansen, Amir Hussain, Peter Nordqvist, Masahiro Sunohara
Cover photo: By Brorsson (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3AStockholms_Old_Town_seen_from_Skeppsholmen.jpg
Cover design: Mia K. Barker
LaTeX editor: Jon Barker
using LaTeX's 'confproc' package, version 0.8 (by V. Verfaille)

The workshop took place in Stockholm, Sweden — 19th August, 2017

Published online at http://spandh.dcs.shef.ac.uk/chat/ — September 1, 2017

Workshop Organisation

Organising Committee

Jon Barker University of Sheffield, UK

John Culling University of Cardiff, UK

John Hansen University of Texas, Dallas, US

Amir Hussain University of Stirling, UK

Peter Nordqvist KTH, Stockholm, Sweden

Masahiro Sunohara Rion Co. Ltd., Japan

Scientific Committee

Kathy Arehart University of Colorado, US

Peter Assmann University of Texas Dallas, US

Stefan Bleeck University of Southampton, UK

Robert Brennan ON Semiconductor, Canada

Jorg Buchholz National Acoustic Lab., Australia

Leslie Collins Duke University, US

Chris Davis Western Sydney University, Australia

Peter Derleth Sonova AG, Switzerland

Bas van Dijk Cochlear, Belgium

Naomi Harte Trinity College Dublin, Ireland

Andrew Hines Dublin Institute of Technology, Ireland

Jesper Jensen Oticon + Aalborg University, Denmark

Yatin Mahajan Western Sydney University, Australia

Bernd T. Meyer Johns Hopkins University, US

Ben Milner University of East Anglia, UK

Brian Moore University of Cambridge, UK

Graham Naylor MRC IHR, Univ.Ñottingham, UK

Peter Nopp MED-EL Ltd, Austria

Niels Pontoppidan Eriksholm Research, Denmark

Bernhard Seeber Tech. Univ. of Munich, Germany

Constantin Spille University of Oldenburg, Germany

Bill Whitmer MRC IHR, Univ. of Nottingham, UK

Tao Zhang Starkey Hearing Technologies, US

Sponsors

MRC Medical Research Council

EPSRC Engineering and Physical Sciences Research Council

Welcome from Conference Chair

The CHAT workshop was a one-day workshop organised as a satelitte event of Interspeech 2017 held at the University of Stockholm, Sweden on the 19th August 2017. It was designed to bring together people from the speech and hearing research communities to explore fresh approaches to hearing assistive technology. The event was sponsored by the International Speech Communication Association (ISCA), the UK Engineering and Physical Sciences Research Council (EPSRC) and the UK Medical Research Council (MRC).

For further details of the event and news of forthcoming CHAT workshops, please visit http://spandh.dcs.shef.ac.uk/chat.

Conference Program

Keynote 1

Oral Session 1

Poster Session 1

Keynote 2

Poster Session 2

Oral Session 2

Space-aware hearing devices - Making hearing aids smarter

Volker Hohmann

Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, Germany

Abstract

In spite of the tremendous advances in hearing device technology since the introduction of the first hearing aid with digital signal processing in 1996, the rehabilitation of acoustic communication in patients with sensorineural hearing loss is still limited. In particular, significant problems are reported in difficult acoustic conditions characterized by high levels of background noise and reverberation. In addition to a reduction in speech intelligibility, other important factors are affected, such as the awareness of the acoustic space and of the spatial configuration and movement of sound sources.

To tackle these problems, recent approaches incorporate knowledge about the principles of human auditory scene analysis to build a representation of the acoustic environment and to decide about the appropriate filtering that makes the attended sound source better audible while keeping the sound features that affect the perception of the acoustic space intact. As an example, a binaural multi-microphone system will be described that estimates the direction of arrival of several sound sources present in the scene, and selects and enhances one of the sources that was identified as the attended (target) source by analyzing the eye movements of the subject.

To develop and test such "space-aware" hearing devices and their underlying signal processing schemes, established lab-based methods are not sufficient, as they make unrealistic assumptions about the acoustic conditions in real life. In particular, different from the stationary spatial configuration of fixed sound sources used in lab-based setups, real-life scenarios are dynamic in the sense that the sound sources constantly move, that the attended source may switch and that the subject is actively listening, i.e., moves in response to visual and auditory input conditioned on its current hearing wish. To incorporate these key factors of acoustic communication in reproducible lab-based measurement setups, virtual audiovisual environments in combination with "subject-in-the-loop" evaluation methods are increasingly used. One study will presented that tested the performance of different classes of hearing aid algorithms in a number of different virtual acoustic environments including scenes with a moving listener. The results confirm previous findings that spatial complexity has a major impact on algorithm benefit and shows that performance measured with established lab-based setups does not predict performance in more complex conditions well. In a second study, the influence of visual cues on motion behavior and involvement of the subject in the listening task was measured. It was found that subjects have different movement strategies when following a conversation. This shows that active listening is individual and requires the hearing devices to accurately represent the acoustic scene and to dynamically detect the attended sound source in order to keep the spatial impression intact while enhancing the target source.

A Simulation Study on Binaural Dereverberation and Noise Reduction based on Diffuse Power Spectral Density Estimators

Ina Kodrasi, Daniel Marquardt, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All, Oldenburg, Germany
{ina.kodrasi,daniel.marquardt,simon.doclo}@uni-oldenburg.de

Abstract

Enhancement techniques in binaural hearing aids are crucial to improve speech understanding for hearing impaired persons in reverberation and noise. Since reverberation and noise can be commonly modeled as diffuse sound fields, many state-of-the-art techniques require an estimate of the diffuse power spectral density (PSD). In this paper we evaluate the performance of binaural dereverberation and noise reduction techniques using several diffuse PSD estimators in realistic acoustic scenarios. Two state-of-the-art techniques are considered, i.e., the binaural multi-channel Wiener filter and the binaural minimum variance distortionless response beamformer with partial noise estimation followed by a postfilter. The considered diffuse PSD estimators are blocking matrix-based and eigenvalue decomposition-based estimators. A least-squares generalization of dual-channel blocking matrix-based estimators to the multi-channel case is also presented, yielding the same diffuse PSD estimate as a recently proposed maximum likelihood estimator. Simulation results show the applicability of diffuse PSD estimators for binaural dereverberation and noise reduction, with the eigenvalue decomposition-based estimators always yielding the best performance.

Index Terms: hearing aids, binaural cues, blocking matrix, eigenvalue decomposition

1. Introduction

Dereverberation and noise reduction techniques in binaural hearing aids are crucial to improve speech intelligibility for hearing impaired persons [1]. In addition to reducing the interference, i.e., reverberation and noise, another important objective of such techniques is the preservation of the listener's impression of the acoustical scene by preserving the binaural cues of the speech source and of the interference [2, 3].

In [2] the binaural multi-channel Wiener filter (MWF) has been presented, which can be decomposed into a binaural minimum variance distortionless response (MVDR) beamformer and a single-channel Wiener postfilter. The binaural MWF and MVDR beamformer preserve the binaural cues of the desired speech source, but distort the cues of the interference such that both the speech source and the residual interference are perceived as coming from the same direction [4]. In order to also (partially) preserve the binaural cues of the residual interference, the binaural MWF with partial noise estimation (MWF-N) [4, 5] and the binaural MVDR beamformer with partial noise estimation (MVDR-N) [3] have been proposed, where a trade-off parameter controls the trade-off between interference reduction and cue preservation. The trade-off parameter yielding a desired cue preservation level can be analytically computed only for the binaural MVDR-N beamformer [3], making it a computationally advantageous technique in comparison to the MWF-N. To further increase the interference reduction performance, a single-channel Wiener postfilter can be applied at the output of the MVDR-N beamformer [3]. Since reverberation is commonly modeled as a diffuse sound field [6–11] and since diffuse background noise is commonly encountered in binaural applications, these binaural speech enhancement techniques require (among other parameters) an estimate of the diffuse power spectral density (PSD).

Several multi-channel diffuse PSD estimators have been proposed, such as blocking matrix-based estimators [6, 8, 12–15] and eigenvalue decomposition-based estimators [10, 11]. Blocking matrix-based estimators estimate the diffuse PSD by blocking the target signal using knowledge of the direction of arrival (DOA) of the speech source [6, 8, 15], blind source separation methods [13], or blind system identification methods [14]. The multi-channel blocking matrix-based estimator in [6] uses a maximum likelihood formulation to estimate the diffuse PSD from multiple reference signals, whereas the dual-channel blocking matrix-based estimators in [13–15] estimate the diffuse PSD by solving an equation based on a single reference signal. Eigenvalue decomposition-based estimators on the other hand do not require a blocking matrix and directly estimate the diffuse PSD using the eigenvalues of the prewhitened input PSD matrix.

The objective of this paper is to evaluate the performance of the binaural MVDR beamformer followed by a postfilter (i.e., the binaural MWF) and the binaural MVDR-N beamformer followed by a postfilter using blocking matrix-based and eigenvalue decomposition-based diffuse PSD estimators. In addition, a least-squares generalization of the dual-channel blocking matrix-based estimators from [13–15] to the multi-channel case is presented, which happens to be equivalent to the multi-channel estimator from [6]. The blocking matrix is constructed based on the DOA of the speech source, which is estimated using the binaural DOA estimator proposed in [15]. Simulation results show that the performance of all considered diffuse PSD estimators is high, with the eigenvalue decomposition-based PSD estimators resulting in the best performance. In addition, it is shown that the performance of the blocking matrix-based dual-channel estimator from [15] is very similar to the performance of the blocking matrix-based multi-channel estimator from [6], suggesting that increasing the number of microphones within the blocking matrix-based framework does not increase the diffuse PSD estimation accuracy.

2. Configuration and Notation

We consider a binaural hearing aid configuration consisting of $M = M_{\text{L}} + M_{\text{R}}$ microphones, with M_{L} denoting the number of microphones of the left hearing aid and M_{R} denoting the number of microphones of the right hearing aid. In the short-time Fourier transform domain, the M-dimensional vector of the re-

This work was supported in part by the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG), and the joint Lower Saxony-Israeli Project ATHENA, funded by the State of Lower Saxony.

ceived microphone signals at frequency index k and frame index l can be written as

$$\begin{aligned}\mathbf{y}(k,l) = [&Y_{\mathrm{L},1}(k,l)\ \ldots\ Y_{\mathrm{L},M_{\mathrm{L}}}(k,l) \\ &Y_{\mathrm{R},1}(k,l)\ \ldots\ Y_{\mathrm{R},M_{\mathrm{R}}}(k,l)]^T,\end{aligned} \tag{1}$$

with $Y_{\{\mathrm{L,R}\},m}(k,l)$ the m-th microphone signal of the left and right hearing aid. In a reverberant and noisy acoustic scenario, $\mathbf{y}(k,l)$ is given by

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{d}(k,l) + \mathbf{v}(k,l), \tag{2}$$

with $\mathbf{x}(k,l)$ the direct and early reverberation speech component, $\mathbf{d}(k,l)$ the diffuse sound component, and $\mathbf{v}(k,l)$ the noise component. The diffuse sound component $\mathbf{d}(k,l)$ models the late reverberation [6–11] as well as any noise which can be well approximated by a diffuse sound field, such as background noise in large crowded rooms. The noise component $\mathbf{v}(k,l)$ represents any remaining noise which cannot be modeled by a diffuse sound field, such as uncorrelated sensor noise. For conciseness, the frequency index k will be omitted in the remainder of this paper.

For a single-source scenario, the direct and early reverberation speech component $\mathbf{x}(l)$ can be expressed in terms of the target signals $S_{\mathrm{L}}(l)$ and $S_{\mathrm{R}}(l)$ (i.e., direct and early reverberation speech components) in the reference microphones of the left and right hearing aids as

$$\mathbf{x}(l) = S_{\{\mathrm{L,R}\}}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l), \tag{3}$$

with $\mathbf{a}_{\mathrm{L}}(l)$ and $\mathbf{a}_{\mathrm{R}}(l)$ the M-dimensional vectors of relative early transfer functions (RETFs) of the target signals from the reference microphones to all M microphones. The target signal $S_{\{\mathrm{L,R}\}}(l)$ is often defined as the direct speech component only [6–11], such that the vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$ can be constructed based on a DOA estimate and head models or measurements of anechoic acoustic transfer functions (ATFs). The PSD matrix of the microphone signals is defined as

$$\mathbf{\Phi}_{\mathbf{y}}(l) = \mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\}, \tag{4}$$

with $\mathcal{E}\{\cdot\}$ the expected value operator. As in many speech enhancement techniques, in the following it is assumed that the components in (2) are mutually uncorrelated, such that $\mathbf{\Phi}_{\mathbf{y}}(l)$ can be written as

$$\mathbf{\Phi}_{\mathbf{y}}(l) = \mathbf{\Phi}_{\mathbf{x}}(l) + \mathbf{\Phi}_{\mathbf{d}}(l) + \mathbf{\Phi}_{\mathbf{v}}(l), \tag{5}$$

with $\mathbf{\Phi}_{\mathbf{x}}(l)$, $\mathbf{\Phi}_{\mathbf{d}}(l)$, and $\mathbf{\Phi}_{\mathbf{v}}(l)$ denoting the PSD matrices of $\mathbf{x}(l)$, $\mathbf{d}(l)$, and $\mathbf{v}(l)$, respectively. Using (3), $\mathbf{\Phi}_{\mathbf{y}}(l)$ can be expressed as

$$\mathbf{\Phi}_{\mathbf{y}}(l) = \underbrace{\Phi_{S_{\{\mathrm{L,R}\}}}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l)\mathbf{a}^H_{\{\mathrm{L,R}\}}(l)}_{\mathbf{\Phi}_{\mathbf{x}}(l)} + \underbrace{\Phi_{\mathrm{d}}(l)\mathbf{\Gamma}}_{\mathbf{\Phi}_{\mathbf{d}}(l)} + \mathbf{\Phi}_{\mathbf{v}}(l), \tag{6}$$

with $\Phi_{S_{\{\mathrm{L,R}\}}}(l)$ the time-varying PSD of the target signal, i.e., $\Phi_{S_{\{\mathrm{L,R}\}}}(l) = \mathcal{E}\{|S_{\{\mathrm{L,R}\}}(l)|^2\}$, $\Phi_{\mathrm{d}}(l)$ the time-varying PSD of the diffuse sound component, and $\mathbf{\Gamma}$ the time-invariant spatial coherence matrix of the diffuse sound field. The spatial coherence matrix $\mathbf{\Gamma}$ is assumed to be known, since it can be constructed based on head models [16] or measurements of anechoic ATFs [9, 17]. In order to simplify the notation, in the following we define the interference component $\mathbf{u}(l) = \mathbf{d}(l) + \mathbf{v}(l)$ and the interference PSD matrix

$$\mathbf{\Phi}_{\mathbf{u}}(l) = \Phi_{\mathrm{d}}(l)\mathbf{\Gamma} + \mathbf{\Phi}_{\mathbf{v}}(l). \tag{7}$$

The objective of binaural speech enhancement techniques is to suppress the interference and obtain estimates of the target signals $\hat{S}_{\mathrm{L}}(l)$ and $\hat{S}_{\mathrm{R}}(l)$ by applying M-dimensional filter vectors $\mathbf{w}_{\mathrm{L}}(l)$ and $\mathbf{w}_{\mathrm{R}}(l)$ to all microphone signals (cf. Section 3), i.e.,

$$\hat{S}_{\{\mathrm{L,R}\}}(l) = \mathbf{w}^H_{\{\mathrm{L,R}\}}(l)\mathbf{y}(l). \tag{8}$$

The time-varying input interaural coherence (IC) of the interference is defined as

$$\mathrm{IC}_{\mathrm{in}}(l) = \frac{\mathbf{e}_{\mathrm{L}}^T\mathbf{\Phi}_{\mathbf{u}}(l)\mathbf{e}_{\mathrm{R}}}{\sqrt{\mathbf{e}_{\mathrm{L}}^T\mathbf{\Phi}_{\mathbf{u}}(l)\mathbf{e}_{\mathrm{L}}\mathbf{e}_{\mathrm{R}}^T\mathbf{\Phi}_{\mathbf{u}}(l)\mathbf{e}_{\mathrm{R}}}}, \tag{9}$$

with $\mathbf{e}_{\{\mathrm{L,R}\}}$ an M-dimensional selector vector with one element equal to 1 and all other elements equal to 0 such that $\mathbf{e}^T_{\{\mathrm{L,R}\}}\mathbf{a}_{\{\mathrm{L,R}\}}(l) = 1$. The time-varying output IC of the interference is defined as

$$\mathrm{IC}_{\mathrm{out}}(l) = \frac{\mathbf{w}_{\mathrm{L}}^H(l)\mathbf{\Phi}_{\mathbf{u}}(l)\mathbf{w}_{\mathrm{R}}(l)}{\sqrt{\mathbf{w}_{\mathrm{L}}^H(l)\mathbf{\Phi}_{\mathbf{u}}(l)\mathbf{w}_{\mathrm{L}}(l)\mathbf{w}_{\mathrm{R}}^H(l)\mathbf{\Phi}_{\mathbf{u}}(l)\mathbf{w}_{\mathrm{R}}(l)}}. \tag{10}$$

Since the IC is complex-valued, binaural speech enhancement techniques typically aim at preserving the real-valued magnitude-squared coherence (MSC) of the interference, defined as

$$\mathrm{MSC}(l) = |\mathrm{IC}(l)|^2. \tag{11}$$

3. Binaural Speech Enhancement

In this section the derivation of the filter $\mathbf{w}_{\{\mathrm{L,R}\}}(l)$ based on the binaural MVDR and MVDR-N beamformers followed by a Wiener postfilter is briefly discussed.

3.1. Binaural MVDR and MVDR-N beamformers

The binaural MVDR beamformer [2] aims at minimizing the output PSD of the interference while preserving the target signal in the left and right reference microphones. The binaural MVDR beamformer can be computed as

$$\mathbf{w}^{\mathrm{MVDR}}_{\{\mathrm{L,R}\}}(l) = \frac{\mathbf{\Phi}_{\mathbf{u}}^{-1}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l)}{\mathbf{a}^H_{\{\mathrm{L,R}\}}(l)\mathbf{\Phi}_{\mathbf{u}}^{-1}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l)}. \tag{12}$$

As shown in [4], the beamformer in (12) preserves the binaural cues of the speech source but distorts the output MSC of the interference such that both the speech source and the residual interference are perceived as coming from the same direction. In order to better preserve the interference output MSC, and hence, the impression of the acoustical scene, the binaural MVDR-N beamformer has been proposed [3]. Aiming at preserving both the target signal as well as a scaled version of the interference in the left and right reference microphones, the binaural MVDR-N beamformer can be computed as

$$\mathbf{w}^{\mathrm{MVDR\text{-}N}}_{\{\mathrm{L,R}\}}(l) = [1-\eta(l)]\mathbf{w}^{\mathrm{MVDR}}_{\{\mathrm{L,R}\}}(l) + \eta(l)\mathbf{e}_{\{\mathrm{L,R}\}}, \tag{13}$$

where $\eta(l)$ denotes a (real-valued) scaling parameter between 0 and 1 which provides a trade-off between interference reduction and MSC preservation. The value of the parameter $\eta(l)$ yielding a desired user-defined interference output MSC can be computed analytically [3].

3.2. Wiener postfilter

In order to further increase the interference reduction performance, a single-channel Wiener postfilter can be applied at the output of the MVDR and MVDR-N beamformers [3, 18], i.e.,

$$G_{\{\mathrm{L,R}\}}(l) = \frac{\xi_{\{\mathrm{L,R}\}}(l)}{1+\xi_{\{\mathrm{L,R}\}}(l)}, \tag{14}$$

with $\xi_{\{\mathrm{L,R}\}}(l)$ the a-priori signal-to-interference ratio (SIR) at the beamformer output in the left and right hearing aid. The a-priori SIR can be estimated using the decision-directed approach based on an estimate of the interference PSD at the beamformer output [19]. The interference PSD at the beamformer output can be computed as

$$\Phi^{\mathrm{out}}_{\{\mathrm{L,R}\},\mathrm{u}}(l) = \mathbf{w}^H_{\{\mathrm{L,R}\}}(l)\mathbf{\Phi_u}(l)\mathbf{w}_{\{\mathrm{L,R}\}}(l), \tag{15}$$

with $\mathbf{w}_{\{\mathrm{L,R}\}}(l)$ the MVDR beamformer in (12) or the MVDR-N beamformer in (13). In order to preserve the binaural cues of the speech source and interference, a common postfilter $G(l)$ is applied to both hearing aids, with

$$G(l) = \sqrt{G_{\mathrm{L}}(l)G_{\mathrm{R}}(l)}. \tag{16}$$

In summary, in Section 5 we consider two different methods for computing the filter $\mathbf{w}_{\{\mathrm{L,R}\}}(l)$, i.e.,

1. using an MVDR beamformer and a Wiener postfilter:

$$\mathbf{w}_{\{\mathrm{L,R}\}}(l) = \mathbf{w}^{\mathrm{MVDR}}_{\{\mathrm{L,R}\}}(l)G(l) \tag{17}$$

2. using an MVDR-N beamformer and a Wiener postfilter:

$$\mathbf{w}_{\{\mathrm{L,R}\}}(l) = \mathbf{w}^{\mathrm{MVDR\text{-}N}}_{\{\mathrm{L,R}\}}(l)G(l) \tag{18}$$

Computing the filters in (17) and (18) requires estimates of the diffuse PSD $\Phi_{\mathrm{d}}(l)$, noise PSD matrix $\mathbf{\Phi_v}(l)$, and RETF vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$.

4. Diffuse PSD Estimators

In this section it is assumed that estimates of the noise PSD matrix $\mathbf{\Phi_v}(l)$ and RETF vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$ are available, such that only the diffuse PSD $\Phi_{\mathrm{d}}(l)$ needs to be estimated. The noise PSD matrix $\mathbf{\Phi_v}(l)$ can in practice be estimated from the microphone signals using e.g. a multi-channel speech presence probability estimator [20]. The RETF vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$ can in practice be estimated as in Section 5, i.e., using a DOA estimator and measurements of anechoic ATFs [15]. To estimate the diffuse PSD $\Phi_{\mathrm{d}}(l)$, we consider blocking matrix-based and eigenvalue decomposition-based estimators.

4.1. Blocking matrix-based estimators

In [13–15] dual-channel (i.e., $M = 2$) diffuse PSD estimators using a single reference signal at the output of a blocking matrix have been proposed. In the following, a least-squares generalization of these estimators for $M > 2$ is presented.

In order to estimate the diffuse PSD, an $M \times (M-1)$-dimensional blocking matrix $\mathbf{B}(l)$ can be used to generate a set of $M-1$ reference signals containing only the interference component, i.e.,

$$\tilde{\mathbf{u}}(l) = \mathbf{B}^H(l)\mathbf{y}(l), \tag{19}$$

with $\mathbf{B}(l)$ such that $\mathbf{B}^H(l)\mathbf{a}_{\mathrm{L}}(l) = \mathbf{0}$ or $\mathbf{B}^H(l)\mathbf{a}_{\mathrm{R}}(l) = \mathbf{0}$. Using $\mathbf{a}_{\mathrm{L}}(l)$, a blocking matrix can be computed from the first $M-1$ columns of the matrix $\mathbf{T}(l)$ defined as

$$\mathbf{T}(l) = \mathbf{I} - \frac{\mathbf{a}_{\mathrm{L}}(l)\mathbf{a}^H_{\mathrm{L}}(l)}{\|\mathbf{a}_{\mathrm{L}}(l)\|^2}, \tag{20}$$

where $\mathbf{I}$ denotes the $M \times M$-dimensional identity matrix. It should be noted that many blocking matrices exist and one can also be computed using $\mathbf{a}_{\mathrm{R}}(l)$ instead of $\mathbf{a}_{\mathrm{L}}(l)$ in (20). Based on (6), the PSD matrix of the $M-1$ reference signals at the blocking matrix output is equal to

$$\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l) = \Phi_{\mathrm{d}}(l)\underbrace{\mathbf{B}^H(l)\mathbf{\Gamma}\mathbf{B}(l)}_{\tilde{\mathbf{\Gamma}}(l)} + \underbrace{\mathbf{B}^H(l)\mathbf{\Phi_v}(l)\mathbf{B}(l)}_{\mathbf{\Phi}_{\tilde{\mathbf{v}}}(l)}. \tag{21}$$

The PSD matrix $\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l)$ can be directly estimated from $\tilde{\mathbf{u}}(l)$, whereas the matrices $\tilde{\mathbf{\Gamma}}(l)$ and $\mathbf{\Phi}_{\tilde{\mathbf{v}}}(l)$ can be computed using the available diffuse coherence matrix $\mathbf{\Gamma}$ and the available noise PSD matrix $\mathbf{\Phi_v}(l)$. Since the only unknown quantity is the diffuse PSD $\Phi_{\mathrm{d}}(l)$, the system of equations in (21) represents an overdetermined system of equations. A least-squares estimate of the diffuse PSD can be obtained by minimizing the cost function

$$J(l) = \|\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l) - \mathbf{\Phi}_{\tilde{\mathbf{v}}}(l) - \Phi_{\mathrm{d}}(l)\tilde{\mathbf{\Gamma}}(l)\|^2_F, \tag{22}$$

where $\|\cdot\|_F$ denotes the matrix Frobenious norm. Setting the derivative of (22) with respect to $\Phi_{\mathrm{d}}(l)$ equal to 0, the least-squares estimate of the diffuse PSD can be computed as

$$\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d}}(l) = \frac{\mathrm{trace}\{[\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l) - \mathbf{\Phi}_{\tilde{\mathbf{v}}}(l)]^H\tilde{\mathbf{\Gamma}}(l)\}}{\mathrm{trace}\{\tilde{\mathbf{\Gamma}}^H(l)\tilde{\mathbf{\Gamma}}(l)\}}, \tag{23}$$

where $\mathrm{trace}\{\cdot\}$ denotes the trace operator. For $M = 2$, $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d}}(l)$ is equal to the PSD estimate derived in [13–15]. Interestingly, for $M > 2$, $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d}}(l)$ is equal to the maximum likelihood PSD estimate derived in [6].

4.2. Eigenvalue decomposition-based estimators

While the estimator in Section 4.1 requires knowledge of the RETF vector, an RETF-independent eigenvalue decomposition-based PSD estimator is proposed in [10, 11]. This estimator requires knowledge of the PSD matrix $\mathbf{\Phi_c}(l) = \mathbf{\Phi_x}(l) + \mathbf{\Phi_d}(l)$, which can be computed as

$$\mathbf{\Phi_c}(l) = \mathbf{\Phi_y}(l) - \mathbf{\Phi_v}(l), \tag{24}$$

with $\mathbf{\Phi_y}(l)$ directly estimated from the microphone signals. Based on (6), the prewhitened PSD matrix $\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)$ is equal to the sum of a rank-1 matrix and a scaled identity matrix, i.e.,

$$\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l) = \Phi_{S_{\{\mathrm{L,R}\}}}(l)\mathbf{\Gamma}^{-1}\mathbf{a}_{\{\mathrm{L,R}\}}(l)\mathbf{a}^H_{\{\mathrm{L,R}\}}(l) + \Phi_{\mathrm{d}}(l)\mathbf{I}. \tag{25}$$

As a result, the eigenvalues of $\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)$ are equal to

$$\lambda_1\{\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)\} = \sigma(l) + \Phi_{\mathrm{d}}(l), \tag{26}$$

$$\lambda_j\{\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)\} = \Phi_{\mathrm{d}}(l), \;\; j = 2, \,\ldots,\, M, \tag{27}$$

with $\sigma(l)$ the only non-zero eigenvalue of the rank-1 term in (25). In [11] it is proposed to estimate the diffuse PSD using any of the last $M-1$ eigenvalues $\lambda_j\{\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)\}$, $j = 2, \,\ldots,\, M$. Due to signal model violations and estimation errors in $\mathbf{\Phi_c}(l)$, the last $M-1$ eigenvalues of $\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)$ are not equal in practice. In this paper we consider two alternative eigenvalue decomposition-based PSD estimates $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}(l)$ and $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}(l)$, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}(l)$ computed as the mean of the last $M-1$ eigenvalues and $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}(l)$ computed as the second eigenvalue, i.e.,

$$\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}(l) = \frac{\mathrm{trace}\{\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)\} - \lambda_1\{\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)\}}{M-1}, \tag{28}$$

$$\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}(l) = \lambda_2\{\mathbf{\Gamma}^{-1}\mathbf{\Phi_c}(l)\}. \tag{29}$$

Using any diffuse PSD estimate in (23), (28), or (29), the available coherence matrix $\mathbf{\Gamma}$, and the available noise PSD matrix $\mathbf{\Phi_v}(l)$, an estimate of the interference PSD matrix $\mathbf{\Phi_u}(l)$ in (7) can now be computed.

5. Experimental Results

In this section the dereverberation and noise reduction performance using the filters in (17) and (18) is investigated for different reverberation times and signal-to-noise ratios (SNRs). In addition, the performance is investigated for a stationary speaker as well as for a moving speaker. In order to focus on the diffuse sound suppression, in the following it is assumed that the microphone signals consist only of a direct and early reverberation speech component and a diffuse sound component (i.e., late reverberation and diffuse background noise), i.e., $\mathbf{v}(l) = \mathbf{0}$ and $\mathbf{\Phi_u}(l) = \Phi_\mathrm{d}(l)\mathbf{\Gamma}$. For $\mathbf{\Phi_u}(l) = \Phi_\mathrm{d}(l)\mathbf{\Gamma}$, the MVDR and MVDR-N beamformers in (12) and (13) can be constructed using only the diffuse spatial coherence matrix $\mathbf{\Gamma}$ (i.e., the scalar $\Phi_\mathrm{d}(l)$ cancels out).

5.1. Setup

Signals were recorded in a laboratory with variable acoustics at the University of Oldenburg using two 2-channel behind-the-ear hearing aid dummies placed on the ears of a head-and-torso simulator (HATS), i.e., $M_\mathrm{L} = 2$, $M_\mathrm{R} = 2$, and $M = 4$. The stationary speaker was simulated by playing back clean speech from a loudspeaker placed at a distance of 2 m from the center of the head. Two stationary speaker scenarios were generated by placing the loudspeaker at two different angles θ_1 and θ_2, with $\theta_1 = 35^\circ$ and $\theta_2 = -35^\circ$. The considered reverberation times for the stationary speaker scenarios were $T_{60} \in \{0.5\text{ s}, 0.75\text{ s}, 1\text{ s}\}$. The moving speaker was a human speaker naturally walking in the frontal hemisphere of the HATS. The considered reverberation time for the moving speaker scenario was $T_{60} \approx 1$ s. To simulate a diffuse noise field, the background noise was generated by placing four loudspeakers facing the corners of the laboratory playing back uncorrelated multi-talker noise. It should be noted that although this background noise was not perfectly diffuse, its MSC was rather similar to the MSC of a diffuse noise field. The speech and the noise signals were recorded separately such that we were able to mix them at different input SNRs (iSNRs) afterward. The considered iSNRs are $\text{iSNR} \in \{0\text{ dB}, 5\text{ dB}, \ldots, 20\text{ dB}\}$.

The signals are processed using a weighted overlap-add framework with a frame size of 512 samples and an overlap of 50% at a sampling frequency $f_s = 16$ kHz. The first microphone of each hearing aid is arbitrarily selected as the reference microphone. The DOA of the speech source is estimated using the binaural DOA estimator in [15]. It should be noted that the DOA estimate obtained in all considered reverberant and noisy scenarios is highly accurate. Using the estimated DOA, the RETF vector $\mathbf{a}_{\mathrm{L,R}}(l)$ is computed from anechoic ATFs measured on the same dummy head [21]. The diffuse coherence matrix $\mathbf{\Gamma}$ is calculated using spatially averaged auto- and cross-correlations of the anechoic ATFs measured for angles ranging between -180° to 175°. To compute the parameter $\eta(l)$ for the MVDR-N beamformer, the desired interference output MSC is defined based on the frequency-dependent values proposed in [17], which are psychoacoustically motivated [22] and do not alter the listener's impression of a diffuse sound field. The PSD matrices $\mathbf{\Phi_y}(l)$ and $\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l)$ are estimated using recursive averaging with a time constant of 40 ms. The minimum gain of the Wiener postfilter $G(l)$ is -20 dB.

The dereverberation and noise reduction performance is evaluated in terms of the improvement in PESQ (ΔPESQ) [23] and frequency-weighted segmental SNR (ΔfSSNR) [24] between the output signal and the reference microphone signal for each hearing aid. The PESQ and fSSNR measures are intrusive measures comparing the signal being evaluated to a desired signal. The desired signal for each hearing aid is generated by convolving the clean speech signal with the measured anechoic ATFs corresponding to the true DOAs. The clean speech signal for the moving speaker scenario is assumed to be the signal recorded with a close-talk microphone. The ΔPESQ and ΔfSSNR presented in the following are average improvements between the left and right hearing aids.

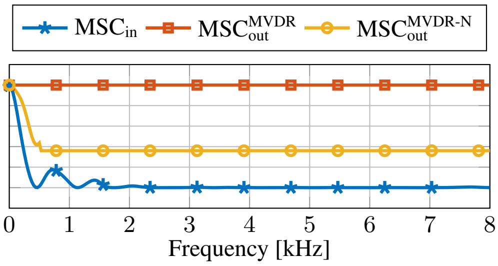


Figure 1: *MSC at the input and output of the MVDR and MVDR-N beamformers.*

The postfilters in (17) and (18) are computed using the blocking matrix-based and eigenvalue decomposition-based diffuse PSD estimators. Two alternative estimates will be investigated for the blocking matrix-based estimator, i.e., $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},2}(l)$ denoting the PSD estimate obtained using only the reference microphones on the left and right hearing aids (corresponding to the dual-channel PSD estimator in [15]) and $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},4}(l)$ denoting the PSD estimate obtained using all 4 microphones (corresponding to the maximum likelihood PSD estimator in [6]).

5.2. MSC preservation

Since the common Wiener postfilter does not change the binaural cues, to evaluate the interference MSC preservation performance of the considered techniques the MSC is computed at the input and output of the MVDR and MVDR-N beamformers using (9), (10), and (11). Fig. 1 presents the obtained MSC values. Since the interference PSD matrix is modeled by a scaled diffuse coherence matrix, the input MSC is time-invariant and equal to the MSC of a diffuse sound field. Furthermore, the MSC at the output of the MVDR and MVDR-N beamformers is also time-invariant, with the MVDR beamformer always distorting the output MSC and the MVDR-N beamformer always yielding the desired user-defined output MSC. Note that since the late reverberation and the noise are not perfectly diffuse, the interference PSD matrix is not equal to a scaled diffuse coherence matrix in practice. Computing the MSC directly from the signals would yield different results from the ones presented in Fig. 1. However, the presented MSC values do illustrate that in all simulations, the MVDR beamformer distorts the cues of the residual interference whereas the MVDR-N beamformer better preserves them.

5.3. Dereverberation performance for a stationary speaker

In this section the dereverberation performance is investigated for several stationary speaker scenarios with different reverberation times and speaker positions. The presented ΔPESQ and ΔfSSNR are averaged between the considered speaker positions. Fig. 2(a) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR beamformer and a Wiener postfilter with different diffuse PSD estimators. It can be observed that in terms of ΔPESQ, using any diffuse PSD estimator yields a similar improvement, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}$ resulting in a slightly higher ΔPESQ than other PSD estimators. In terms of ΔfSSNR, it can be observed that the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based es-

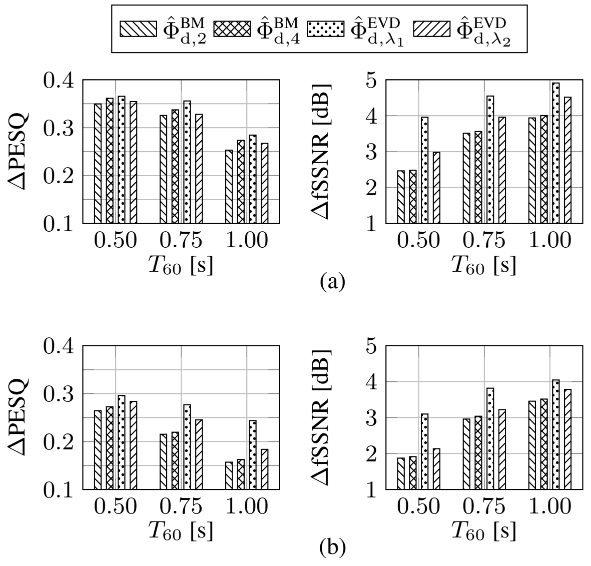


Figure 2: *Dereverberation performance for a stationary speaker using a beamformer and a Wiener postfilter: (a) MVDR and (b) MVDR-N.*

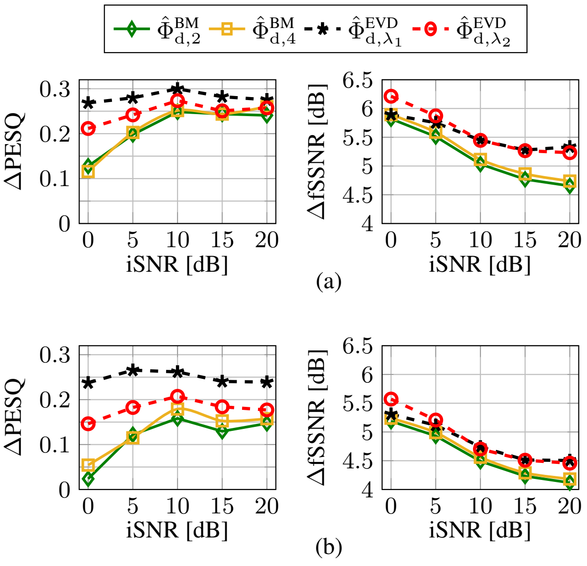


Figure 3: *Dereverberation and noise reduction performance for a stationary speaker using a beamformer and a Wiener postfilter: (a) MVDR and (b) MVDR-N ($T_{60} \approx 1$ s).*

timators, with $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_1}$ resulting in the best performance. In addition, in terms of both performance measures it appears that the performance obtained using $\hat{\Phi}^{\text{BM}}_{\text{d},2}$ and $\hat{\Phi}^{\text{BM}}_{\text{d},4}$ is very similar, suggesting that increasing the number of microphones in the blocking matrix-based framework does not increase the diffuse PSD estimation accuracy. Fig. 2(b) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR-N beamformer and a Wiener postfilter with different diffuse PSD estimators. Overall it can be observed that the performance improvement obtained for all considered reverberation times and diffuse PSD estimators is smaller than in Fig. 2(a). This is to be expected, since the MVDR-N beamformer also (partly) preserves the MSC of the residual interference component (cf. Fig. 1). In terms of both performance measures, it can be observed that the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based estimators, with $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_1}$ resulting in the best performance. In addition, similarly to before, the performance obtained using $\hat{\Phi}^{\text{BM}}_{\text{d},2}$ and $\hat{\Phi}^{\text{BM}}_{\text{d},4}$ is very similar in terms of both performance measures.

5.4. Dereverberation and noise reduction performance for a stationary speaker

In this section the dereverberation and noise reduction performance is investigated for several stationary speaker scenarios with different iSNRs and speaker positions. The considered reverberation time is $T_{60} \approx 1$ s. The presented ΔPESQ and ΔfSSNR are averaged between the considered speaker positions. Fig. 3(a) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR beamformer and a Wiener postfilter with different diffuse PSD estimators. It can be observed that in terms of both performance measures, the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based estimators, with $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_1}$ resulting in the best ΔPESQ and $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_2}$ resulting in the best ΔfSSNR for low iSNRs. In addition, it can be observed that $\hat{\Phi}^{\text{BM}}_{\text{d},2}$ and $\hat{\Phi}^{\text{BM}}_{\text{d},4}$ yield a very similar performance in terms of both performance measures. Fig. 3(b) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR-N beamformer and a Wiener postfilter with different diffuse PSD estimators. Overall it can be observed that as expected, the performance improvement obtained for all considered iSNRs and diffuse PSD estimators is lower than in Fig. 3(a). Furthermore, the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based estimators in terms of both performance measures, with $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_1}$ resulting in the best ΔPESQ and $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_2}$ resulting in the best ΔfSSNR for low iSNRs. Whereas larger differences can be observed in terms of ΔPESQ between the blocking matrix-based and eigenvalue decomposition-based estimators, the obtained ΔfSSNR for all PSD estimators are rather similar. In addition, similarly to before, the performance obtained using $\hat{\Phi}^{\text{BM}}_{\text{d},2}$ and $\hat{\Phi}^{\text{BM}}_{\text{d},4}$ is very similar.

5.5. Dereverberation and noise reduction performance for a moving speaker

In this section the dereverberation and noise reduction performance is investigated for a moving speaker scenario with $T_{60} \approx 1$ s and iSNR $= 10$ dB. Since both ΔPESQ and ΔfSSNR show very similar patterns, Table 1 presents only the ΔfSSNR obtained using the MVDR and MVDR-N beamformers and a Wiener postfilter. It can be observed that using the eigenvalue decomposition-based estimate $\hat{\Phi}^{\text{EVD}}_{\text{d},\lambda_2}$ results in the best performance. However, the performance obtained using the other considered diffuse PSD estimators is also comparable. In addition, it can be observed that as expected, the improvement obtained for all diffuse PSD estimators when using the MVDR-N beamformer is lower than when using the MVDR beamformer. However, the performance loss is rather insignificant, particularly when using the eigenvalue decomposition-based estimators.

In summary, the simulation results presented in this paper show the applicability of diffuse PSD estimators for binaural dereverberation and noise reduction based on beamforming and spectral filtering. Although all PSD estimators yield a high performance, the eigenvalue decomposition-based estimators result in the best performance for all considered techniques and scenarios. It should be noted that although the considered PSD estimators are based on a diffuse sound field model, the late reverberation and background noise considered in these simu-

Table 1: *Dereverberation and noise reduction performance in terms of ΔfSSNR using an MVDR and MVDR-N beamformer and a Wiener postfilter for a moving speaker ($T_{60} \approx 1$ s, iSNR = 10 dB).*

	$\hat{\Phi}^{\rm BM}_{\rm d,2}$	$\hat{\Phi}^{\rm BM}_{\rm d,4}$	$\hat{\Phi}^{\rm EVD}_{\rm d,\lambda_1}$	$\hat{\Phi}^{\rm EVD}_{\rm d,\lambda_2}$
MVDR	7.42	7.55	6.78	**7.86**
MVDR-N	6.83	6.89	6.66	**7.63**

lations were not perfectly diffuse, confirming the applicability of the considered estimators in realistic acoustic environments. Informal listening tests suggest that blocking matrix-based estimators yield a larger interference suppression while causing more signal distortions, whereas eigenvalue decomposition-based estimators yield a smaller interference suppression while introducing less signal distortions. In the future, formal listening tests should be conducted to truly assess the quality of these different late reverberation PSD estimators for binaural dereverberation and noise reduction.

6. Conclusion

In this paper we investigated the dereverberation and noise reduction performance of the binaural MVDR and MVDR-N beamformers followed by a Wiener postfilter when using blocking matrix-based and eigenvalue decomposition-based diffuse PSD estimators. A least-squares generalization of dual-channel blocking matrix-based estimators to the multi-channel case was also presented, yielding the same PSD estimate as a recently proposed multi-channel maximum likelihood estimator. Simulations results show that independently of the technique used, the eigenvalue decomposition-based PSD estimators yield the best performance. Furthermore, it is shown that increasing the number of microphones within the blocking matrix-based framework does not increase the PSD estimation accuracy.

7. References

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Hoboken, USA: John Wiley & Sons, 2010.

[3] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, University of Oldenburg, Oldenburg, Germany, Dec. 2015.

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

[5] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[6] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[7] S. Braun, M. Torcoli, D. Marquardt, E. A. P. Habets, and S. Doclo, "Multichannel dereverberation for hearing aids with interaural coherence preservation," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Juan les Pins, France, Sep. 2014, pp. 124–128.

[8] O. Schwartz, S. Braun, S. Gannot, and E. A. P. Habets, "Maximum likelihood estimation of the late reverberant power spectral density in noisy environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2015.

[9] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1595–1608, Sep. 2016.

[10] I. Kodrasi and S. Doclo, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, San Francisco, USA, Mar. 2017, pp. 116–120.

[11] ——, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, Mar. 2017, pp. 611–615.

[12] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[13] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, May 2013.

[14] M. Azarpour, G. Enzner, and R. Martin, "Binaural noise PSD estimation for binaural speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 7068–7072.

[15] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2017, submitted.

[16] M. Jeub, M. Dorbecker, and P. Vary, "A semi-analytical model for the binaural coherence of noise fields," *IEEE Signal Processing Letters*, vol. 18, no. 3, pp. 197–200, Mar. 2011.

[17] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2162–2176, Dec. 2015.

[18] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

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

[20] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[21] 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, Jul. 2009.

[22] A. Walther and C. Faller, "Interaural correlation discrimination from diffuse field reference correlations," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1496–1502, Mar. 2013.

[23] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[24] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

Low-Latency Real-Time Blind Source Separation with Binaural Directional Hearing Aids

Masahiro Sunohara[1], *Chiho Haruta*[1], *Nobutaka Ono*[2]

[1]RION Co., Ltd., Japan
[2]National Institute of Informatics, Japan

suno@rion.co.jp, c-haruta@rion.co.jp, onono@nii.ac.jp

Abstract

Understanding desired speech in noisy environments is one of the important issues for hearing aid systems, which require a strong real-time property. Recently, the authors have proposed a low-latency algorithm for real-time blind source separation (BSS) based on online auxiliary-function-based independent vector analysis (AuxIVA) by the truncation of non-causal components of time-domain demixing impulse responses, and we evaluated the separation performance using omnidirectional binaural microphones. On the other hand, directional microphones have been widely used for hearing aids to improve the signal-to-noise ratio. In this paper, the effects of the truncation of demixing impulse responses is investigated when using the proposed algorithm with binaural directional microphones. By experimental evaluation using a head and torso simulator in a real environment, the performance of the proposed algorithm with directional microphones in the case of 10 ms latency is 9.0 dB in terms of the signal-to-interference ratio (SIR), which is only a 2.2 dB performance loss from the case of 128 ms latency. Moreover, the performance with directional microphones is about 1.0 dB better than that with omnidirectional microphones.

Index Terms: hearing aids, directional microphone, blind source separation, independent vector analysis, low latency

1. Introduction

Hearing-impaired people have difficulties communicating with others even if they wear hearing aids, especially in a noisy environment such as a party venue or a crowded restaurant. Improving speech communication in such difficult situations is one of the most challenging issues to be solved for hearing aids. As a technique for solving these problems, blind source separation (BSS) may be applicable [1, 2, 3]. BSS is a signal processing method that can extract a desired sound source from a mixture by using multiple microphones without requiring information on the source signals. In the frequency-domain approach for convolutive BSS, independent vector analysis (IVA) has been proposed as a technique that does not require the solution of a permutation ambiguity problem [4, 5, 6]. Furthermore, auxiliary-function-based IVA (AuxIVA) has been proposed as a state-of-the-art approach with rapid convergence and a low calculation cost [7, 8, 9].

To apply BSS as an application of binaural hearing aids, which are real-time systems, it is important to reduce the latency from the input to the output of the system [10, 11]. In addition to computational complexity, an algorithm may require an inherent delay, which is referred to as an algorithmic delay. In the case of frequency-domain BSS, a delay of at least one frame length is necessary for frame analysis [12]. Although several real-time implementations of IVA have been proposed [13, 14], this delay is unavoidable. Such a large delay causes various problems in a hearing aid system such as difficulty in speaking owing to the delayed auditory feedback effect or a sense of discomfort due to the loss of lip synchronization [15].

Recently, the authors have proposed a low-latency algorithm for real-time BSS based on online AuxIVA for hearing aids [16]. This proposed algorithm can significantly shorten the algorithmic delay by the time-domain implementation of demixing matrices as FIR filters and the truncation of part of their non-causal components. Generally, the truncation of the non-causal components should degrade the separation performance. However, if all the non-causal components of the demixing impulse response are originally zero, these components can be truncated without performance degradation. In our previous work [16], we confirmed that the proposed system with an algorithmic delay of within 10 ms worked with little performance degradation by experimental evaluation using binaural behind-the-ear (BTE)-type hearing aids consisting of omnidirectional microphones.

However, bilateral directional microphones have been widely used in actual hearing aids to improve the signal-to-noise ratio of front speech signals in a noisy background [17]. Thus, in this paper, the separation performance of binaural BSS based on the low-latency online AuxIVA algorithm with directional microphones was investigated as a more practical verification.

2. Low-latency real-time BSS

2.1. Overview of online AuxIVA

We assume that K sources are observed by K microphones and that their short-time Fourier transform (STFT) representations are known. Let $\boldsymbol{x}(\omega,\tau) = [x_1(\omega,\tau)\cdots x_K(\omega,\tau)]^t$ be the vector representations of the observation signal in the (ω,τ)th time-frequency bin, where ^{t} denotes the vector transpose. In the frequency domain, the sources are estimated by the following linear demixing process:

$$\boldsymbol{y}(\omega,\tau) = W(\omega;\tau)\boldsymbol{x}(\omega,\tau), \tag{1}$$

where $W(\omega;\tau) = (\boldsymbol{w}_1(\omega;\tau)\cdots\boldsymbol{w}_K(\omega;\tau))^h$ is a demixing matrix, ^{h} denotes the Hermitian transpose, and $\boldsymbol{y}(\omega,\tau) = [y_1(\omega,\tau)\cdots y_K(\omega,\tau)]^t$ represents the estimated sources.

An online AuxIVA algorithm is an effective means of estimating the demixing matrices $W(\omega;\tau)$ in the (ω,τ)th time-frequency bin under dynamic conditions [14]. The algorithm consists of alternate update rules, which update the weighted covariance and the demixing matrix. In this paper, we focus on the case of $K = 2$ for application to hearing aids.

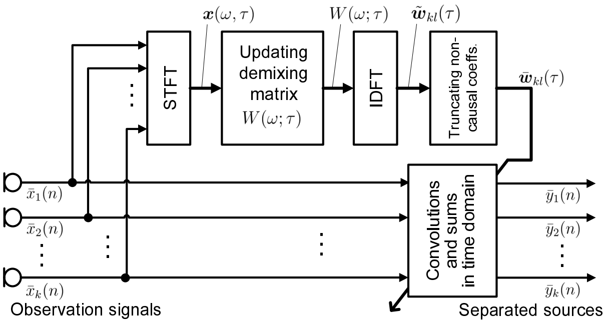

$\boldsymbol{x}(\omega,\tau)$
$W(\omega;\tau)$
$\tilde{\boldsymbol{w}}_{kl}(\tau)$
STFT
Updating demixing matrix $W(\omega;\tau)$
IDFT
Truncating non-causal coeffs.
$\bar{\boldsymbol{w}}_{kl}(\tau)$
Convolutions and sums in time domain
$\bar{x}_1(n)$
$\bar{x}_2(n)$
$\bar{x}_k(n)$
$\bar{y}_1(n)$
$\bar{y}_2(n)$
$\bar{y}_k(n)$
Observation signals
Separated sources

Figure 1: *Signal block diagram of low-latency real-time online independent vector analysis.*

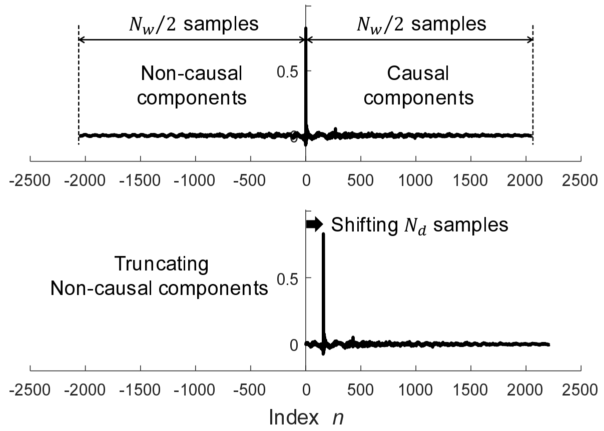


Figure 2: *Time-domain demixing impulse responses. Upper: original response $\tilde{w}_{kl}(n;\tau)$. Lower: shifted and truncated response $\bar{w}_{kl}(n;\tau)$.*

2.2. Realization as quasi-causal FIR filter

Figure 1 shows a signal block diagram of a low-latency version of the online AuxIVA algorithm [16]. A means of shortening the delay is to form two paths, one for updating the demixing matrices in the frequency domain and the other for separating the sources using FIR filters in the time domain. After applying back-projection [18], the frequency-domain demixing matrix $W(\omega;\tau)$ is converted to coefficients of multiple time-domain FIR filters $\tilde{w}_{kl}(n;\tau)$ using the inverse discrete Fourier transform. This structure can shorten the algorithmic delay to half of the frame length ($N_\omega/2$ samples). To further shorten the algorithmic delay, the coefficients of only N_d non-causal components are shifted and the other components are truncated as shown in Fig. 2. After that, the algorithmic delay of the system becomes only N_d samples.

If all the non-causal components of $\tilde{w}_{kl}(\tau)$ are originally zero, the algorithmic delay of the system can theoretically be zero without performance degradation. For the simple model consisting of two sound sources and two observations shown in Fig. 3, a theoretical sufficient condition for the ideal separation filters to be causal is obtained as the following inequality [16]:

$$[\log a(\theta_2) - \log a(\theta_1)] \cdot [\tau(\theta_2) - \tau(\theta_1)] < 0, \tag{2}$$

where $a(\theta_k)$ and $\tau(\theta_k)$ are respectively the amplitude ratio and the time difference of the second channel relative to the first channel for source k with direction θ_k.

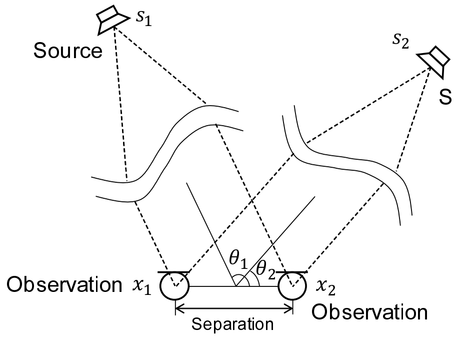


Figure 3: *Simple model consisting of two sound sources and two observations.*

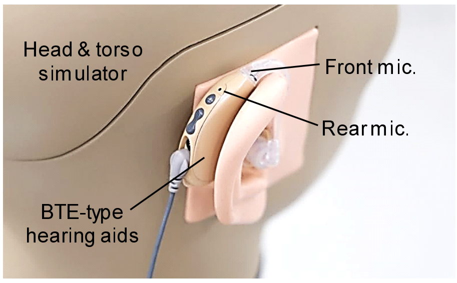


Figure 4: *Locations of the microphones for BTE-type hearing aid with KEMAR dummy head.*

3. Directional microphone in hearing aids

Generally, the directivity of the microphone in a hearing aid is produced by a pair of omnidirectional microphone signals. Fig. 4 shows the locations of omnidirectional microphones in a BTE-type hearing aid mounted on an auricle of a KEMAR dummy head. The separation between the front and rear omnidirectional microphones was 1.08 cm in this case. Fig. 5 shows a signal block diagram of the directional microphone as a spatial signal-processing system. The output of the directional microphone system can be expressed in terms of the microphone separation d, the angle of arrival θ, the rear microphone time delay τ_r, and the gain b. Then, the directional microphone response in the frequency domain $X_d(\omega,\theta)$ when $b = 1$ can be approximated by the following equation [17]:

$$|X_d(\omega,\theta)| \approx \omega \left(\frac{d}{c} \cos\theta + \tau_r \right), \tag{3}$$

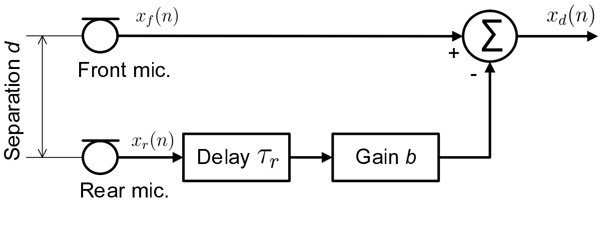


Figure 5: *Block diagram of directional microphone as a spatial signal-processing system.*

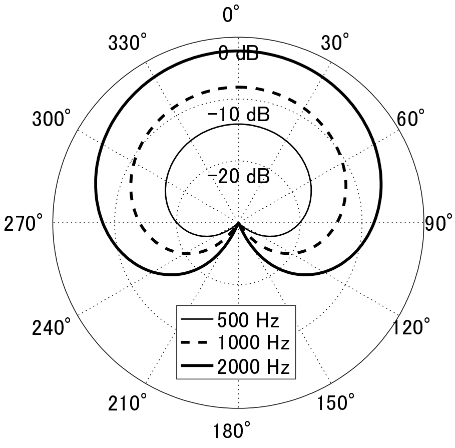


Figure 6: *Microphone directional patterns ($\tau_r = d/c$).*

where c denotes the sound velocity. Fig. 6 shows microphone directional patterns when $\tau_r = d/c$. From the figure, it is found that the sensitivity of the response at lower frequencies is attenuated by 6 dB per octave. These spatial directional responses may affect the extent to which Eq. (2) is satisfied. Therefore, the purpose of this paper is to experimentally evaluate how the proposed low-latency BSS works in the binaural directional system.

4. Evaluation

4.1. Setup

To evaluate the performance of the low-latency online IVA with binaural directional microphones for hearing aids, a PC simulation was carried out using real mixtures of two speeches recorded by four microphones in binaural BTE-type hearing aids with a head and torso simulator (G.R.A.S.: KEMAR type 45BB) in a meeting room. Fig. 7 shows the setup of the loudspeakers and microphones in the evaluation. Two omnidirectional electret condenser microphones (front and rear) were installed in one BTE-type hearing aid. The hearing aids were attached to each ear of the head and torso simulator. The direction of one of the two sources was fixed at 0° and that of the other source was varied from 30° to 180° in steps of 30°. We selected ten speech sources for each direction from the RWCP Japanese News Speech Corpus [19]. The other experimental conditions are summarized in Table 1.

Table 1: *Experimental conditions*

microphone spacing (interaural)	18.0 cm
microphone distance (front and rear)	1.08 cm
reverberation time	650 ms at 500 Hz
signal length	30 s × 10
sampling frequency	16 kHz
frame length	4096
frame shift	1024
window function	Hanning
forgetting factor	0.98

For comparison, we used two different latencies, where the numbers of remaining non-causal components N_d were 160 and 2048 samples, corresponding to algorithmic delays of 10 and 128 ms, respectively. The experiments on the recorded mixtures were performed using MATLAB R2016a on a laptop PC with an Intel Core i7-3770 3.40 GHz processor. The performance was evaluated by the average signal-to-interference ratio (SIR) over all trials with the exception of the first three seconds on each trial, which is defined as the ratio of the signal power of the desired speaker to the signal power from the interfering speaker. The SIR was calculated by bss_eval_images.m in the BSS toolbox [20].

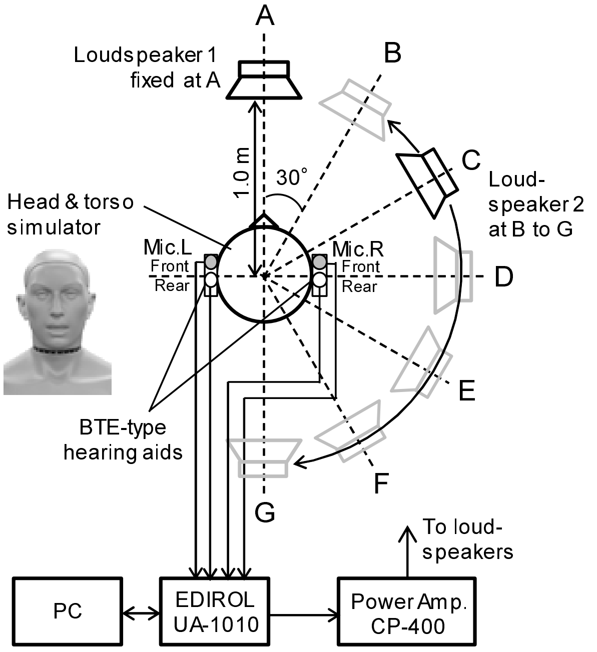


Figure 7: *Setup of loudspeakers and microphones in the evaluation.*

4.2. Method

In this experiment, the separated signals were obtained by applying back-projection to the front channel in the omnidirectional case and to the directional channel (the output of Fig. 5). To compare them fairly, it is necessary to compensate the dif-

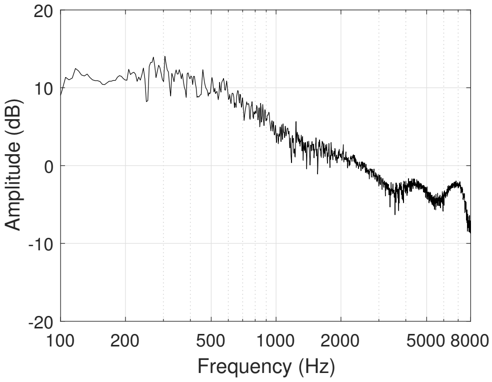


Figure 8: *Example of amplitude response of compensation filter $c_k(\omega)$.*

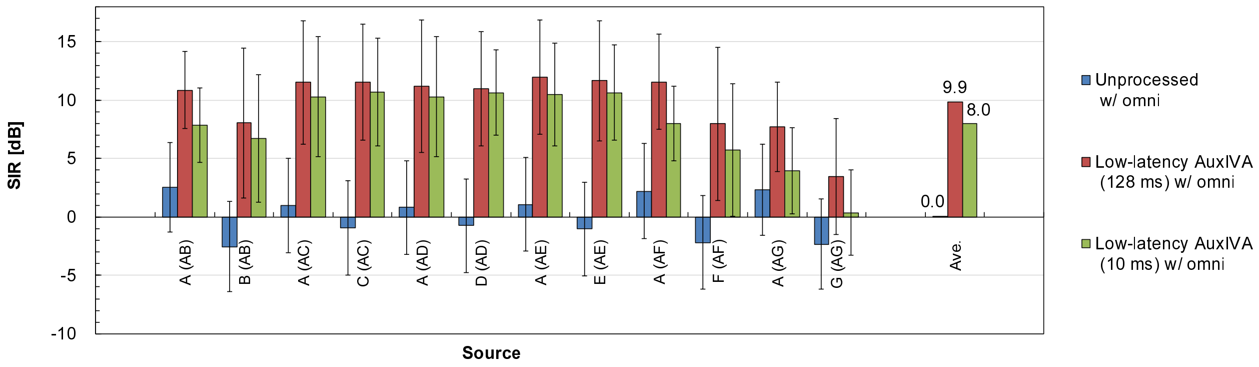


Figure 9: *Separation performance with omnidirectional microphones.*

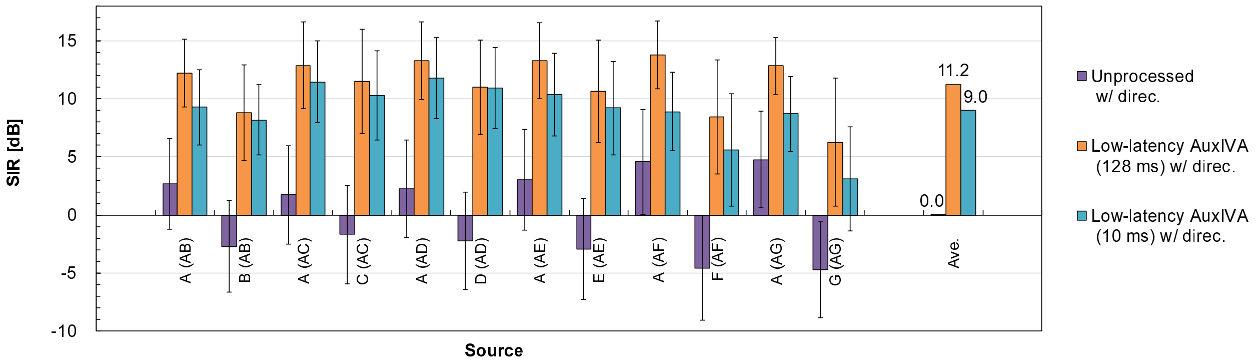


Figure 10: *Separation performance with directional microphones.*

ference in the sensitivity characteristic associated with the directivity. As post-processing for the evaluation, we derive a compensation filter by minimizing the following cost function:

$$J(\boldsymbol{C}(\omega)) = \sum_{\tau} \left| r(\omega,\tau) - \boldsymbol{C}^H(\omega)\boldsymbol{Y}(\omega,\tau) \right|^2, \qquad (4)$$

where $r(\omega,\tau)$ is the STFT representation of the front microphone signal $x_f(n)$, $\boldsymbol{Y}(\omega,\tau) = \begin{bmatrix} y_1(\omega,\tau) \\ y_2(\omega,\tau) \end{bmatrix}$ is a vector comprising of the STFTs of the separated signals $\boldsymbol{y}_k(\omega,\tau)$, and $\boldsymbol{C}(\omega) = \begin{bmatrix} c_1(\omega) \\ c_2(\omega) \end{bmatrix}$ is a vector of the compensation filter $\boldsymbol{c}_k(\omega)$. By differentiating Eq. (4) with respect to $\boldsymbol{C}(\omega)$ and setting the equation to 0, the compensation filter $\boldsymbol{C}(\omega)$ can be calculated as

$$\begin{bmatrix} c_1(\omega) \\ c_2(\omega) \end{bmatrix} = \begin{bmatrix} \sum_{\tau} |y_1(\omega,\tau)|^2 & \sum_{\tau} y_1(\omega,\tau) y_2^*(\omega,\tau) \\ \sum_{\tau} y_1^*(\omega,\tau) y_2(\omega,\tau) & \sum_{\tau} |y_2(\omega,\tau)|^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{\tau} r^*(\omega,\tau) y_1(\omega,\tau) \\ \sum_{\tau} r^*(\omega,\tau) y_2(\omega,\tau) \end{bmatrix}. \qquad (5)$$

Fig. 8 shows an example of the amplitude response of the compensation filter $c_k(\omega)$ for a separated signal with directivity. It was found that the amplitude response was compensated by 6 dB per octave slope.

4.3. Results

Figure 9 shows the separation performance with omnidirectional microphones for the low-latency online AuxIVA algorithm with algorithmic delays of 128 and 10 ms. The bars and the error bars indicated the averaged SIR and the standard deviation on each ten trials, respectively. On the horizontal axis, A(AB) denotes source A in a mixture of source A and source B. The average SIR with the algorithmic delay of 10 ms is 8.0 dB, which is 1.9 dB less than that for the delay of 128 ms. In particular, the SIR tends to degrade toward the front-back direction such as at 30°, 150°, and 180°.

On the other hand, Fig. 10 shows the separation performance with directional microphones. The average SIR with the algorithmic delay of 10 ms is 9.0 dB, which is 2.2 dB less than that for the delay of 128 ms. The resultant SIRs with the directional microphones show better separation performance, which was on average 1.0 dB greater than that with omnidirectional microphones. In particular, the improvement of the SIR increases toward the front-back direction.

5. Conclusion

In this paper, we evaluated the separation performance of low-latency online AuxIVA with directional microphones for binaural hearing aids. When the directivity was processed from an adjacent pair of omnidirectional microphones, the sensitivity of the response at lower frequencies was attenuated by 6 dB per octave compared with the omnidirectional microphones. To compare them fairly, the difference in the sensitivity characteristics was compensated by post-processing. From the evaluation results, the performance of the proposed algorithm with directional microphones in the case of 10 ms latency was 9.0 dB in terms of the SIR, which is only a 2.2 dB performance loss from the case of 128 ms latency. Moreover, the average SIR with directional microphones was about 1.0 dB better than that with omnidirectional microphones. Future work will focus on listening tests to verify the proposed system.

6. Acknowledgments

This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP16H01735.

7. References

[1] Y. Mori, T. Takatani, H. Saruwatari, K. Shikano, T. Hiekata, and T. Morita, "High-presence hearing-aid system using DSP-based real-time blind source separation module," in *Proc. ICASSP*, 2007.

[2] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *Proc. ISCCSP*, 2010.

[3] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, pp. 1260–1277, 2006.

[4] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, pp. 601–608, 2006.

[5] T. Kim, T. Eltoft, and T. W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," in *Proc. ICA*, pp. 165–172, 2006.

[6] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, pp. 70–79, 2007.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, pp. 189–192, 2011.

[8] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IWAENC*, 2012.

[9] N. Ono, "Auxiliary-function based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA ASC*, 2012.

[10] M. A. Stone, B. C. J. Moore, K. Meisenbacher, and R. P. Derleth, "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," *Ear Hear.*, vol. 29, pp. 601–617, 2008.

[11] R. Heurig and J. Chalupper, "Acceptable processing delay in digital hearing aids," *Hear. Rev.*, vol. 17, pp. 28–31, 2010.

[12] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *Proc. EUSIPCO*, pp. 222–226, 2007.

[13] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. CAS I*, vol. 57, no. 7, pp. 1431–1438, 2010.

[14] T. Taniguchi, N. Ono, A. Kawamata, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, pp. 107–111, 2014.

[15] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *J. Am. Acad. Audiol.*, vol. 11, pp. 330–336, 2000.

[16] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proc. ICASSP*, 2017.

[17] J. M. Kates, *Digital Hearing Aids*. Plural Publishing, 2008.

[18] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[19] "Real World Computing Project News Speech Corpus," http://research.nii.ac.jp/src/en/RWCP-SP99.html. [Accessed: 11 Sep. 2016].

[20] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

On the Predictability of the Intelligibility of Speech to Hearing Impaired Listeners

Mark Huckvale, Gaston Hilkhuysen

Speech, Hearing & Phonetic Sciences, University College London, U.K.

m.huckvale@ucl.ac.uk, g.hilkhuysen@ucl.ac.uk

Abstract

What information do we need to know about listeners to predict their performance on a speech intelligibility task and how well can we predict intelligibility anyway? This paper performs a meta-analysis on two speech intelligibility studies of hearing-impaired listeners in which we evaluate different approaches to building a predictive model of intelligibility. The model has two components: a cochlear loss term based on a number of psychoacoustic measures of hearing, and a supra-cochlear loss term to explain residual performance variation. These models are trained using a method of cross-validation to determine how well they might perform on new listeners and new tasks. We found that cochlear loss could only explain 40% of the variability in performance across hearing-impaired listeners, while the supra-cochlear loss can account for a further 20-40% depending on the task. The combined cochlear and supra-cochlear loss terms allow good estimates of intelligibility scores in the data, with speech reception thresholds on a novel listening task being predictable to within 1dB on average.

Index Terms: hearing impairment, speech intelligibility, psychoacoustics, metrics.

1. Introduction

The availability of increasingly powerful computational resources in miniaturized form allows for more advanced signal processing algorithms to be applied within hearing aids. These techniques hold promise for improved speech intelligibility for hearing impaired (HI) users in everyday noisy and reverberant environments.

However, the development of some advanced signal processing algorithm is not enough. It is also necessary to ensure that processing adapts to the requirements of the listener and to the requirements of the listening situation [1]. For a given impaired listener, the advanced aid needs to choose between the relative benefits of equalization, compression, noise reduction, dereverberation, beam-forming or speech enhancement for every listening situation. The challenge is not just to find good signal processing approaches, but to understand how they benefit the listener to ensure they are optimized for the listener and the listening environment.

In previous work we have investigated the utility of speech intelligibility metrics for predicting the impact of speech signal processing on intelligibility for normally-hearing (NH) listeners [2]. We evaluated predictions from intrusive signal metrics of intelligibility against the actual performance of listeners. We showed that metrics like STOI [3] and NCM+ [4] gave fair predictions of the likely speech intelligibility to a listener from analysis of the differences between the clean signal and the processed noisy/reverberant signal. Typically intelligibility could be predicted within 2dB SNR [2].

For these metrics to be useful for finding the best signal processing approaches in hearing aids, they need to be developed in two directions: firstly they need to be made non-intrusive, that is capable of working from the noisy signal alone, and secondly they need to take into account the impact of hearing impairment. Well-established means for converting intrusive metrics to non-intrusive use statistical learning methods applied to large databases of speech materials rated by the intrusive metric [5]. How best to modify these metrics to make predictions for HI listeners, however, is still an active area of research.

A typical approach to take hearing loss into account within intelligibility metrics is to incorporate information about the listener into the front-end signal processing: for example a front-end filterbank might be modified to accommodate degradations in frequency sensitivity (auditory thresholds), frequency selectivity (auditory filter bandwidths) and dynamic range (recruitment) [6]. While this approach seems sensible, it relies on the assumption that the difference between NH and HI listeners is well predicted by characteristics of their hearing loss. In turns out that this is not the whole story. When a group of HI listeners are assessed (as we show later in this paper) there is considerable residual variation in performance compared to NH listeners even after taking their hearing loss into account. There is an echo of the Anna Karenina principle: normally hearing listeners are all alike; every hearing impaired listener is hearing impaired in their own way.

A number of explanations are proposed for why HI listeners are more variable than NH listeners. It could be to do with correlations between hearing loss and cognitive decline [7], or that imperfect auditory representations require more cognitive effort to process which tests the ability of the listener to recruit additional processing capacity or working memory [8]. Or it could be that some of the phonological fine tuning used by NH listeners to discriminate phonemes is degraded to different degrees in different listeners.

HI listener variation that is not predictable from characteristics of their hearing loss is a problem for speech intelligibility metrics, since manipulation of the front-end of the metric may not be enough. For the design of metrics to predict the benefit of speech enhancement to HI listeners this is a real problem since the accuracy of a metric based only on hearing loss could be worse than the likely differences between processing approaches (e.g. approaches may only vary by 1dB in effective SNR but estimated intelligibility for a listener might vary by 2dB).

In this paper we study the variability of speech intelligibility performance by HI listeners in terms of two loss functions. The first relates to those aspects related to psychoacoustic measurements of their hearing loss. We call this their cochlear loss. The second relates to everything else, we call this their supra-cochlear loss. The paper then has three goals:

a) What proportion of the variability of HI listener performance on speech intelligibility tasks is predictable from their cochlear loss?
b) What proportion of the variability of an HI listener on a new intelligibility task is predictable from an estimate of their supra-cochlear loss obtained from another intelligibility task?
c) To what degree is supra-cochlear loss independent of the nature of the listening task?

Our approach is a meta-analysis of two existing data sets in which both psychoacoustic measures and speech intelligibility scores are available for a group of HI listeners. We assume that the psychoacoustic measures reflect cochlear processing. We build the best predictive models of performance from these psychoacoustics and interpret the remaining performance variation on some task as supra-cochlear loss. We then explore how estimates of listeners' supra-cochlear losses vary with intelligibility task.

The structure of the paper is as follows: in section 2 we shortly describe the contents of the data sets in terms of the speech intelligibility scores and the psychoacoustic descriptors available in each. We refer to the original papers for details. In section 3 we describe the modelling approach and the performance measures used. In section 4 we present the results of the meta-analysis and in section 5 discuss their implications.

2. Data sets

2.1. Bethesda Data Set

The Bethesda data set was collected by Summers et al [9] at the Walter Reed National Military Medical Center, Bethesda, MD. The listeners on the test comprised 10 normally-hearing and 18 hearing-impaired subjects. It includes the following psychophysical measurements (code in brackets).

- Pure-tone hearing thresholds at 250, 500, 1000, 1500, 2000, 3000, 4000, 6000 & 8000Hz. (H)
- Degree of peripheral amplitude compression at 500, 1000, 2000 & 4000Hz. (C)
- Auditory filter bandwidths at 500, 1000, 2000 & 4000Hz measured at both 70 and 80 dB SPL. (B)
- Frequency modulation detection thresholds measured at 500, 1000, 2000 & 4000Hz. (F)

The intelligibility of speech-in-noise to each listener was measured using IEEE sentences with both speech-shaped noise and amplitude modulated speech-shaped noise at signal-to-noise ratios (SNRs) of -6, -3, 0 and +3 dB. The speech, presented at 92 dB SPL for all listeners, was not equalized to match auditory thresholds.

For subsequent analysis the speech test scores for each listener were converted to Speech Reception Thresholds (SRT). The % scores were first converted to log-odds ratios and then linear regression was used to find the SNR value for the listener which gave a log-odds of 0 (i.e. 50%). In addition all frequency measurements were converted to log Hertz before modelling.

Since our analysis is focused on the HI listeners, the data points of the NH listeners were combined into one average listener.

2.2. Salamanca Data Set

The Salamanca data set was collected by Johannesen et al [10] at the Universidad de Salamanca, Spain. It consists of test scores on 68 hearing-impaired listeners. The following measurements were made of each listener's hearing ability (code in brackets):

- Pure-tone hearing thresholds at 500, 1000, 2000, 4000 & 6000Hz (H)
- Estimate of cochlear mechanical gain loss (also referred to as outer-hair cell loss, OHC) expressed in decibels (dB). (O)
- Basilar-membrane compression exponent (BMCE). It was defined as the slope (in dB/dB) of an inferred cochlear input/output curve over its compressive segment. (C)
- Frequency modulation detection thresholds (FMDTs), defined as the minimum detectable excursion in frequency for a pure tone carrier at 1500Hz. (F)

Speech intelligibility was assessed for speech-shaped noise (SSN) and a time-reversed two-talker masker (R2TM) using HINT sentences. Performance was recorded in terms of SRT score. Speech materials were presented with linear, frequency-specific amplification to compensate for listeners' audiometric losses.

3. Method

Our goal is to model the effects of cochlear and supra-cochlear deficits on speech intelligibility performance as measured in terms of speech reception threshold. To build a model of cochlear loss we perform a regression on the psychoacoustic measurements of each listener to predict their SRT score for each listening task. Since we do not know the form of that function we use support-vector regression (SVR) [11] that makes no assumptions about the form of the function other than listeners with similar psychoacoustics are likely to have similar scores. SVR determines a subset of the data set that can be used as examples (support vectors) against which a new listener can be compared to best predict their score. The final score is then just the linear combination of support vector scores weighted by their distance to the new vector. Since any given listener may be chosen to be one of the support vectors, we must model the data set using cross-validation, where each listener is left out in turn and a predicted score is made from a model trained from the remaining listeners. To build the cochlear loss model, the psychoacoustic measures were divided into four sets as coded in section 2, and each feature set was tested in isolation and in combination with all other feature sets. Features are normalised before modelling. A grid-search is used to find the best SVR hyper-parameters, and final predictions are computed from 10 modelling runs.

Once we have obtained a predicted score for each listener we can compare the prediction against the actual score and determine two performance measures: R^2, the proportion of variance in scores explained by the prediction, and mean absolute error (MAE) of prediction, which answers the question how far away on average is the prediction from the correct answer.

The difference between actual and predicted scores for a listener was used as an estimate of their supra-cochlear loss.

To explore the size and variability of the supra-cochlea loss term, we can calculate this for each one of the listening tasks in the data sets, and evaluate it on the other. We compare actual scores and the prediction from the estimated cochlear loss on each task together with the estimated supra-cochlear loss from the other task in terms of R^2 and MAE.

Finally we can calculate how much the estimate of the supra-cochlear loss varies across the two listening tasks to explore the extent to which the supra-cochlear loss is dependent upon the nature of the task.

4. Results

4.1. Prediction of Cochlear Loss

Table 1 shows the MAE of prediction of the SRTs for speech-shaped noise and modulated speech-shaped noise for hearing impaired listeners in the Bethesda data set for each combination of psychoacoustic features. Table 2 shows the MAE of prediction of the SRTs for speech-shaped noise masker and for a reversed two-talker masker for hearing impaired listeners in the Salamanca data set.

Table 1. SRT Prediction from Psychoacoustics for Bethesda data in MAE (dB)

Group	Features	SSN	Modulated SSN
Baseline	None	1.982	3.215
Single	H	1.462	2.121
Single	C	1.851	2.871
Single	B	1.848	2.957
Single	F	1.539	2.524
Double	H+C	1.493	2.129
Double	H+B	1.477	***2.027***
Double	H+F	***1.434***	2.192
Double	C+B	1.903	3.037
Double	C+F	1.664	2.718
Double	B+F	1.717	2.507
Triple	H+C+B	1.476	2.135
Triple	H+C+F	1.505	2.251
Triple	H+B+F	1.486	2.085
Triple	C+B+F	1.736	2.628
All	H+C+B+F	1.557	2.174

Tables 1 and 2 show that incorporation of psychoacoustic features into the model can improve the prediction of speech intelligibility scores over a baseline prediction based on the mean of the other listeners. For the Bethesda data set, the MAE reduces from 1.982 to 1.434dB for SSN, and from 3.215 to 2.027 for Modulated SSN. The reduction on the Salamanca data set is much smaller, from 1.137 to 1.018dB for SSN, and from 1.244 to 1.006dB for reversed two-talker masker.

The best feature set combinations were different for the different data sets and tasks; these are indicated in bold in the tables. The SRT predictions for the best performing models on the Bethesda data set are plotted in Figure 1. The SRT predictions for the best-performing models on the Salamanca data set are plotted in Figure 2. The proportion of variance explained by the best performing models is shown in the plots..

Table 2. SRT Prediction from Psychoacoustics for Salamanca data in MAE (dB)

Group	Features	SSN	R2TM
Baseline	None	1.137	1.244
Single	H	1.115	1.083
Single	O	1.129	1.155
Single	F	1.101	1.036
Single	C	1.022	1.219
Double	H+O	1.119	1.111
Double	H+F	1.061	1.010
Double	H+C	1.045	1.046
Double	O+F	1.085	1.060
Double	O+C	1.073	1.118
Double	F+C	***1.018***	1.090
Triple	H+O+F	1.092	1.034
Triple	H+O+C	1.049	1.035
Triple	H+F+C	1.040	1.009
Triple	O+F+C	1.060	1.031
All	H+O+F+C	1.041	***1.006***

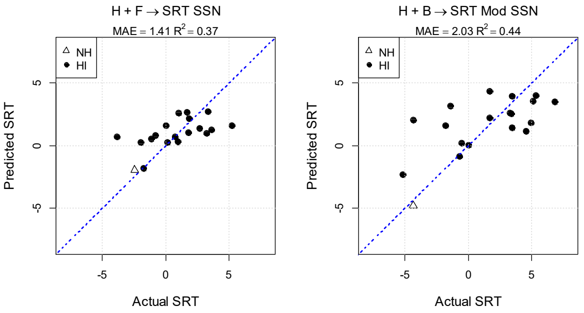


Figure 1. Prediction of SRT from best psychoacoustic features for Bethesda data set (left = SS noise, right = modulated SS noise)

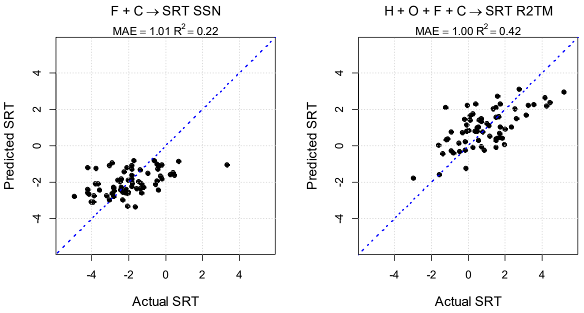


Figure 2. Prediction of SRT from best psychoacoustic features for Salamanca data set (left = SS Noise, right = reversed 2-talker masker)

4.2. Prediction of Supra-Cochlear Loss

The supra-cochlear loss term for each listener for each task is then calculated as the difference between the actual SRT and the SRT predicted from the best feature set for the task. Figures 3 and 4 show the predicted SRT after inclusion of the supra-cochlear loss term. In each case the loss term is computed for the other task. In both data sets and for both tasks, the prediction error is reduced by the inclusion of the supra-cochlear loss, with the MAE reducing to about 1dB for the Bethesda data set and 0.8dB for the Salamanca data set.

Figure 5 shows the correlation between the supra-cochlear loss terms across the two tasks for each of the two data sets. The graphs suggest that the supra-cochlear loss varies by around 1dB on average across the pair of tasks.

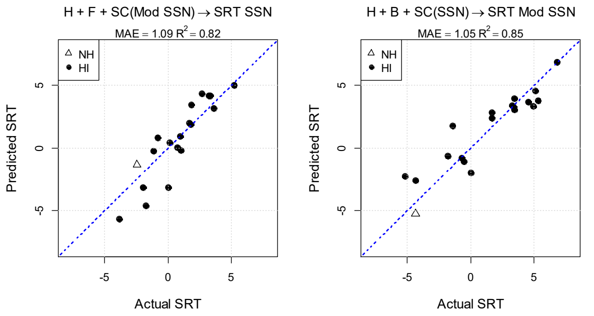


Figure 3. SRT prediction after supra-cochlear loss included in Bethesda data set. Left: SSN score after MSN calibration, right: Mod SSN score after SSN calibration.

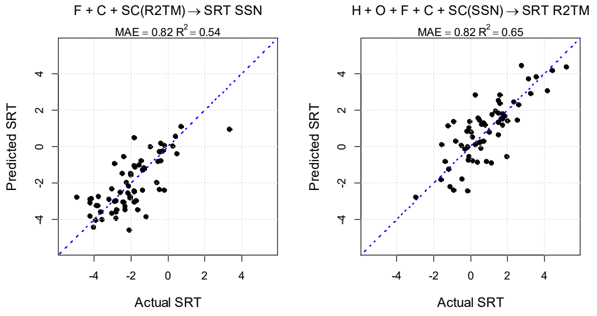


Figure 4 . SRT Prediction after supra-cochlear loss included in Salamance data set. Left: SSN after R2TM calibration, right R2TM after SSN calibration.

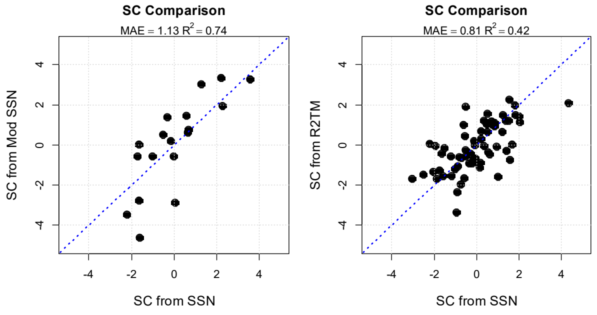


Figure 5. Comparison of supra-cochlear loss across listening tasks. Left Bethesda data set, right Salamanca data set.

5. Discussion

This paper has shown how SRT predictions for HI listeners may be significantly improved using an SVR model of cochlear loss based on the available psychoacoustic measures. Prediction accuracy was better in the Salamanca data set probably because the intelligibility scores were collected with equalization for listener thresholds and so were less variable to begin with. This equalization also explains the different importance given to features in the model, with thresholds being very important features for the Bethesda data set, while for the Salamanca data set, the choice of features made little difference in terms of MAE.

Cochlear loss alone only explained at best 40% of the variability in test scores across HI listeners. Inclusion of a supra-cochlear loss term (calculated from the other task) into the model explains a further 40% of the variation for the Bethesda data set, and a further 20% for the Salamanca data set. The difference is explained by Figure 5, which shows that the two tasks in the Bethesda data set are more similar than those in the Salamanca data set.

Taking both loss terms together we have shown that we can predict second test score performance from first test score performance to within 1dB MAE. This seems within the likely prediction error of a speech signal intelligibility metric.

The residual variability in prediction might come from different sources: (i) experimental error in the collection of the psychoacoustic measures or the intelligibility scores; (ii) the effects of cochlear loss on the task other than that explained by the particular set of psychoacoustic measurements available, or (iii) the task dependency of supra-cochlear loss caused by interactions between the task and cognitive deficits. This interaction might also have arisen if variation in cognition had impact on the collection of psychoacoustic measurements themselves.

In the future, better modelling might arise from: (i) a wider range of psychoacoustic measures – although the evidence presented here suggests that such measures are highly correlated with one another; (ii) a wider range of intelligibility tests per listener to unpack the reasons why supra-cochlear loss is dependent on characteristics of the task; (iii) repeated testing of listeners to obtain estimates of measurement error.

Overall the analysis presented here seems promising for the development of speech signal intelligibility metrics for hearing impaired listeners provided these include a supra-cochlear calibration term for each listener. This might be estimated by incorporating a standardised speech intelligibility test alongside standard psychoacoustic tests in their clinical assessment. The study also makes clear that further work is required to understand the causes of variability in the intelligibility of speech to the hearing impaired.

6. Acknowledgements

Thanks to Van Summers and Enrique Lopez-Poveda for supplying us with the Bethesda and Salamanca data sets.

This work is supported by U.K. Engineering and Physical Sciences Research Council award: “Environment-aware Listener-Optimized Binaural Enhancement of Speech (E-LOBES)” to Imperial College London & University College London.

7. References

[1] B. Kollmeier & J. Kiessling "Functionality of hearing aids: state-of-the-art and future model-based solutions", International Journal of Audiology, DOI: 10.1080/14992027.2016.1256504

[2] G. Hilkhuysen, N. Gaubitch, M. Brookes, M. Huckvale, "Effects of noise suppression on intelligibility II: An attempt to validate physical metrics", J.Acoust.Soc.Am., 135 (2014) 439-50.

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

[4] G. Hilkhuysen, N. Gaubitch, M. Brookes, M. Huckvale, "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios", J.Acoust.Soc.Am. 131 (2012) 531-539.

[5] D. Sharma, G. Hilkhuysen, N. Gaubitch, P. Naylor, M. Brookes, M. Huckvale, "Data driven method for non-intrusive speech intelligibility", 18th European Signal Processing Conference (EUSIPCO-2010) Aalborg, Denmark, August 23-27, 2010.

[6] J. Kates, K. Arehart, "The hearing aid speech perception index (HASPI)", Speech Communication 65, 75-93. doi: 10.1016/j.specom.2014.06.002

[7] C. Füllgrabe, B. Moore, M. Stone, "Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition", Front Aging Neurosci. 2015;6:347. doi: 10.3389/fnagi.2014.00347.

[8] P. Souza, K. Arehart, T. Neher, "Working memory and hearing aid processing: Literature findings, future directions, and clinical applications", Front. Psychol. 2015;6:1894. doi: 10.3389/fpsyg.2015.01894.

[9] V. Summers, M. Makashay, S. Theodoroff, M. Leek, "Suprathreshold auditory processing and speech perception in noise: hearing-impaired and normal-hearing listeners", J.Am. Acad. Audiol 24 (2013) 274-292.

[10] P. Johannesen, P. Perez-Gonzalez, S. Kalluri, J. Blanco, E. Lopez-Poveda, "The Influence of Cochlear Mechanical Dysfunction, Temporal Processing Deficits, and Age on the Intelligibility of Audible Speech in Noise for Hearing-Impaired Listeners", Trends in Hearing 2016, Vol. 20: 1–14.

[11] A. Smola, B. Scholkopf, "A tutorial on support vector regression", Statistics and Computing 14: 199–222, 2004.

CCi-MOBILE platform for cochlear implant and hearing-aid research

Hussnain Ali, Ammula Sandeep, Juliana Saba, and John H.L. Hansen

Center for Robust Speech System - Cochlear Implant Laboratory,
The University of Texas at Dallas, USA
{hussnain.ali, sxa160230, jns109020, john.hansen}@university.edu

Extended Abstract

Improvements in sound processing technology have played a critical role in the advancement of cochlear implant (CI) and hearing-aid (HA) technology. Since the inception of CIs and HAs, investigators have relied on research tools and interfaces to conduct perceptual studies. Research interfaces commonly provided by the manufacturers either have limited functionalities or are not suitable for conducting a broad range of experiments. Portability, wearability, and ease of programmability limits existing research interfaces to benchtop/laboratory use only. Real-world, long-term subject evaluations are needed to assess true potential of novel sound processing strategies.

Our center (CRSS-CILab) over past ten years has been involved in the development of portable research platforms/tools for speech and hearing research [1]. Our latest effort is to leverage computing capabilities of emerging smartphones and tablets for sound processing needs. The functional structure of the platform is shown in Fig. 1. The digital acoustic signal is sampled from Behind-the-Ear (BTE) units and transmitted to the computing platform via a USB-serial port of a custom-developed interface board. The computing platform processes the acoustic signal through a sound coding strategy and generates a set of stimulation data. This data is sent back to the interface board where it is simultaneously delivered to the RF transmission coils (electric stimulation) and hearing aid transducers (acoustic stimulation). In case of electric stimulation, the data is first encoded (using the transmission protocols of the CI device) in the FPGA, before streaming to the implant (see Fig. 1).

The platform can be used for both unilateral and “time-synchronized” bilateral stimulation. Time-synchronized bilateral stimulation means that biphasic pulses and acoustic signal on both left and right ears arrive at the exact same time. In addition, the platform can be configured to provide both electric and acoustic stimulation (EAS) concurrently. Acoustic stimulation can be delivered to ipsilateral as well as contralateral ears, thereby giving 4 channels of time-synchronized stimulation simultaneously in two modes. From operational stand-point, the platform can be used in both real-time and bench-top modes. The real-time mode works similar to a clinical body-worn processor to conduct experiments in free field using the BTE microphone. The bench-top mode, on the other hand, can stream pre-processed stimuli (e.g., audio files) from a desktop PC in laboratory environment. The bench-top mode can also be used to conduct psychoacoustics or psychophysics experiments.

One of the unique and powerful capabilities of the platform is the ability to use it as a real-time speech processor in MATLAB environment. Researchers have access to real-time microphone signals, implement custom algorithms, and stream stimuli for subjective evaluation in real-time from MATLAB. Furthermore, by using smartphones/tablets, there is an additional flexibility to develop and run custom applications (Apps) that are tuned to specific experiments. The touch-screen capability and graphical controls on the smartphone provide an interactive user-interface for modifying processing parameters on the go and enable user input in real-time.

The CCi-MOBILE platform was evaluated acutely with eight post-lingually deafened adult CI users. The assessment of speech recognition was accomplished with AzBio and IEEE sentences presented at different SNR levels as well as with CNC words/phonemes. Study participants were tested in free-field, both with their clinical processor and CCi-MOBILE. The results from acute evaluation indicate that on all measures of test material, CCi-MOBILE platform (μ=59.86±16.02) was not statistically different from each individual’s clinical processors (μ=56.38±17.96). These results indicate that performance with the CCi-MOBILE is comparable to the clinical processor, and that it holds potential for conducting reliable speech assessments in future studies.

The CCi-MOBILE is one-of-a-kind research platform, and is orders of magnitude more flexible and computationally powerful than existing commercially available processors. It will aid in bridging scientific research with commercial applications. The research platform is intended to be an open-source contribution to the cochlear implant and hearing-aid field and will be distributed to the research community on a non-profit model.

References

[1] H. Ali, A. Lobo and P. Loizou, "Design and Evaluation of a PDA-based Research Platform for Cochlear Implants," IEEE Transactions on Biomedical Engineering , vol. 60, no. 11, pp. 3060-3073, 2013.

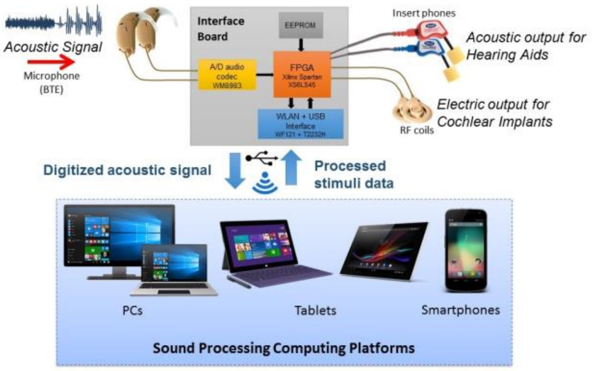

Acoustic Signal
Microphone (BTE)
Interface Board
EEPROM
A/D audio codec WM8983
FPGA Xilinx Spartan XS6LS45
WLAN + USB Interface WF121 + T2232H
Insert phones
Acoustic output for Hearing Aids
Electric output for Cochlear Implants
RF coils
Digitized acoustic signal
Processed stimuli data
PCs
Tablets
Smartphones
Sound Processing Computing Platforms

Fig. 1. High-level description of the CCi-MOBILE research platform.

Towards a measure of the differences in cochlear implant stimulation strategies

Rafael Chiea[1], *Bernardo Murta*[1], *Gustavo Mourao*[1], *Stephan Paul*[1], *Julio Cordioli*[1], *Hussnain Ali*[2], *John Hansen*[2]

[1]Federal University of Santa Catarina, Brazil
[2]University of Texas at Dallas, USA

rafael.chiea@posgrad.ufsc.br, bernardo.murta@lva.ufsc.br,
[2]hussnain.ali, john.hansen@utdallas.edu

Index Terms: Cochlear implants, sound coding, objective metrics

1. Extended abstract

Development and evaluation of new algorithms which aim to improve speech intelligibility and sound perception quality for cochlear implant (CI) users are an important research field. In general, validation of signal processing strategies in CIs is based on subjective evaluations [1]. Human testing is generally time consuming and is subject to availability of sufficient number of CI users to participate in research. In addition, it requires specialized hardware to interface with implanted electronics as well as clinical implementation of speech processing strategies. Due to limited availability of hardware and software, not all researchers have access to the required tools to assess validity and effectiveness of their research ideas.

Validation of signal processing implementations of research ideas does not necessarily requires human speech assessments. Mathematical formulations can be used to characterize and compute normalized metrics which can enable signal processing engineers to compare their encoding schemes with standard implementations. Previous research in this domain, e.g. by Yousefian and Loizou [2], aimed at developing metrics that use the envelope of the processed and clean speech to estimate speech intelligibility. These metrics can be used for comparing strategies to a limited extend, as they are more suitable to evaluate robustness of noise reduction algorithms. Currently, there are not any standard metrics available to compare the output of different sound-processing strategies/implementations.

The goal of this study was to make a first step on the development of an objective tool to compare implementations of cochlear implant strategies. Such comparison may allow the validation of new implementations of existing strategies. The comparison between strategies is performed by the analysis of their electrodograms for a fixed set of input audios, and user parameters (such as threshold and comfort levels, pulse rate and shape). Two N-by-M matrices $\mathbf{A}$ and $\mathbf{A}^{\text{ref}}$ represent, respectively, the electrodogram of the strategy under test and the one of the reference strategy. Here rows are related to the signal in each one of the N electrodes and columns are related to the M time frames of stimulation.

First, the euclidean distance, d_i, between the i^{th} row of the two matrices is calculated, resulting in a N-by-1 array $\mathbf{d}$, according to equation (1). $a_{i,j}$ is the element in row i and column j of the matrix $\mathbf{A}$.

$$\mathbf{d}_i = \sqrt{\sum_{j=1}^{M} \left(a_{i,j} - a_{i,j}^{\text{ref}}\right)^2} \tag{1}$$

The sum of all elements of $\mathbf{d}$ leads to error associated with the audio file f, $ED_f = \sum_{i=1}^{N} \mathbf{d}_{f,i}$ and ED is the distribution of errors associated with all the files as the metric to compare two strategies.

In the present work, we have considered Advance Combination Encode (ACE) strategy as an example to demonstrate the effectiveness of the proposed approach. Three implementations of ACE strategy are compared: clinical implementation as outlined in Cochlear Corp.'s Nucleus Matlab Toolbox, an open-source version developed at the University of Texas at Dallas [3] and custom coded version developed at the Federal University of Santa Catarina [4], based on published data. The above metric was computed for several acoustic inputs of varying complexity and user parameters. These consisted of tones with varying intensities, calibrated chirp signals, consonants, vowels, and speech sentences. The output from these stimuli were used to generate reference electrodogram matrices with the *Nucleus Matlab Toolbox* (NMT).Student's t-test was used to compare error distributions of the two implementations, ED_{UTD} and ED_{UFSC}, to test for equivalence.

It is noted here that different implementations may result in distinct processing delays that may generate time misalignment issues between the pulses on electrodograms. In order to address this, a pre-processing for time alignment is performed. Also, slight variations in channel gains may result in variations in (timing and amplitude of) pulses across channels. The approach outlined in this paper could potentially allow the validation of new implementations of existing strategies, before performing any human testing. Future work will focus on quantifying the error in terms of its perceptual significance (intelligibility and quality), by means of subjective studies in order to understand how this metric relates to human speech perception. This may indicate possible adjustments on the metric, in order to also allow the comparison between different strategies.

2. References

[1] J. Wouters, H. J. McDermott, and T. Francart, "Sound coding in cochlear implants: From electric pulses to hearing," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 67–80, 2015.

[2] N. Yousefian and P. C. Loizou, "Predicting the speech reception threshold of cochlear implant listeners using an envelope-correlation based measure," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3399–3405, 2012.

[3] H. Ali, A. P. Lobo, and P. C. Loizou, "Design and evaluation of a personal digital assistant-based research platform for cochlear implants," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 11, pp. 3060–3073, 2013.

[4] K. Werner, R. Chiea, J. Cordioli, and S. Paul, "Analysis of cochlear implant vocoder simulation including the current spread effect in the presence of background noise." *In Proc. DAGA, Aachen, Germany*, pp. 36–39, 2016.

Noise Suppression Algorithm Based on Loudness Management for Preserving Speech Components

Nobuhiko Hiruma[1], *Hidetoshi Nakashima*[2], *Yoh-ichi Fujisaka*[1]

[1]Rion Co., Ltd., Japan
[2]National Institute of Technology, Kumamoto College, Japan

n-hiruma@rion.co.jp, nakashi@kumamoto-nct.ac.jp, y-fujisaka@rion.co.jp

1. Introduction

A hearing aid amplifies the input sound from a microphone in conformity with the hearing threshold of a user; however, it simultaneously enhances undesired environmental sounds such as the noise of an air conditioner or a car. According to MarkeTrak V [1], such kinds of noise are often perceived by the hearing aid user with a strong feeling of discomfort, and thereby, the user perceives the continuous use of the hearing aid to be inconvenient.

Numerous algorithms have been proposed and implemented for hearing aids as well as cell phones for the purposes of speech enhancement, reduction of discomfort due to noise [2, 3], and alleviation of listening effort [4]. However, it is widely known today that speech intelligibility is not significantly improved by using conventional algorithms, despite the improvement in sound quality [5]. Considering that a hearing aid is an effective communication device for the hearing-impaired, the most important issue for users in their daily lives is to improve the speech recognition rate during communication. Noise reduction algorithms based on spectral subtraction estimate the spectrum of the noise component and then subtract it from the observed signal spectrum; even hence, if the speech spectrum is masked in the noise, it would result in a decrease in the output signal or level of loudness. The degradation of output level for certain frequency components might affect the speech quality depending on the hearing level of the patients, because the output sound level will be lower than the patients threshold of hearing, though this degradation has a positive effect in terms of reducing the annoyance of patients.

In this study, a noise reduction algorithm based on loudness management, which eliminates unnecessary spectral subtraction, is proposed. The concept of our algorithm is that noise components are preserved as much as possible by keeping loudness level of speech signal for positive effect of speech recognition.

2. Proposed System

Figure 1 represents a block diagram of our proposed system. Here, the observed signal $x(t)$ is defined with speech signal $s(t)$ and noise $n(t)$ as

$$x(t) = s(t) + n(t) \tag{1}$$

where t denotes the time index. The above Eq.1 can be written in the frequency domain as

$$X(\omega) = S(\omega) + N(\omega), \tag{2}$$

where ω denotes the angular frequency. The estimated speech signal spectrum is calculated using estimated noise power [6] $|N'(t,\omega)|$, as

$$|S'(t,\omega)|^2 = |X(t,\omega)|^2 - |N'(t,\omega)|^2. \tag{3}$$

For the calculation of partial loudness based on ISO 532B, the observed and estimated speech signal spectra, $|X(t,\omega)|^2$ and $|S'(t,\omega)|^2$, are transformed into one-third octave band energy, $E_x(t,m)$ and $E_{s'}(t,m)$, respectively. Here, m is one-third octave band index.

The partial loudness of the observed signal, $L_x(t,k)$, estimated speech signal, $L_{s'}(t,k)$, and output signal, $L_y(t,k)$, are calculated based on DIN45631 [7] using $E_x(t,m)$ and $E_{s'}(t,m)$. Here, k denotes bark band index and the unit of the partial loudness is sone. In this algorithm, noise suppression gain is determined to make $L_y(t,k)$ correspond to $L_{s'}(t,k)$. Hence, the difference between $L_x(t,k)$ and $L_{s'}(t,k)$ is obtained as

$$L_d(t,k) = L_x(t,k) - L_{s'}(t,k), \tag{4}$$

where $L_d(t,k)$ means suppression gain in sone for the correspondence between $L_y(t,k)$ and $L_{s'}(t,k)$. The suppression gain $G(t,w)$ is obtained by using $L_d(t,k)$ and look-up table for conversion from sone to dB, and then it is transformed into time domain to get FIR filter coefficient $w(t)$ by inverse fast Fourier transform (IFFT). The output signal $y(t)$ is obtained by convolution of $x(t)$ and $w(t)$.

3. Simulation

To confirm the feasibility of the proposed system, a numerical simulation was conducted using MATLAB. The temporal variation of the partial loudness level of the third bark is shown in Figure 2. In this figure, the blue, black, and red lines represent partial loudness of the observed signal $L_x(t,k)$, true speech $L_s(t,k)$, and estimated speech $L_{s'}(t,k)$. Figure 3 shows the result of instantaneous partial loudness level in each bark. In this figure, the green line denotes the partial loudness level of output signal $L_y(t,k)$. It appears that the black and red lines are almost similar. These results indicate that the partial loudness level of estimated speech is almost the same as that of true speech.

4. Evaluation

Evaluation of the output signal in the proposed system is performed using the short-time objective intelligibility (STOI) [8]. For the sake of comparison, the output of conventional noise reduction is also evaluated. In this evaluation, pink noise is added and its level varies from -10 dB to 10 dB of signal-to-noise ratio (SNR) in steps of 2 dB. Figure 4 shows the result of this simulation. The red and blue lines represent the score of output signal of the proposed and conventional methods, respectively.

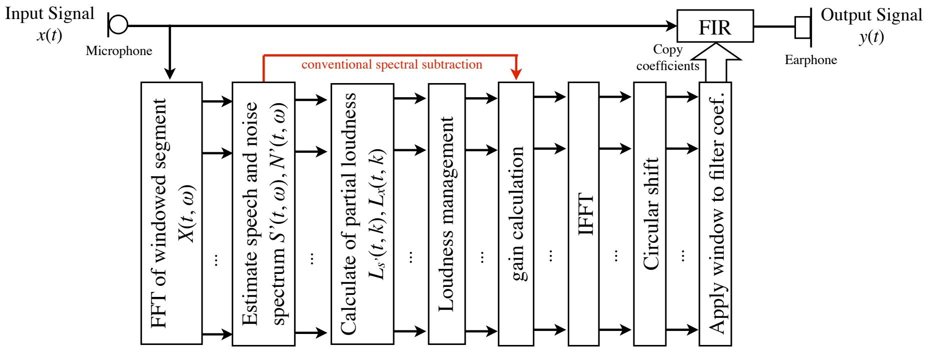

Input Signal
$x(t)$
Microphone
conventional spectral subtraction
FIR
Copy coefficients
Output Signal
$y(t)$
Earphone
FFT of windowed segment $X(t,\omega)$
Estimate speech and noise spectrum $S'(t,\omega), N'(t,\omega)$
Calculate of partial loudness $L_{s'}(t,k), L_x(t,k)$
Loudness management
gain calculation
IFFT
Circular shift
Apply window to filter coef.

Figure 1: *Block diagram of proposed system.*

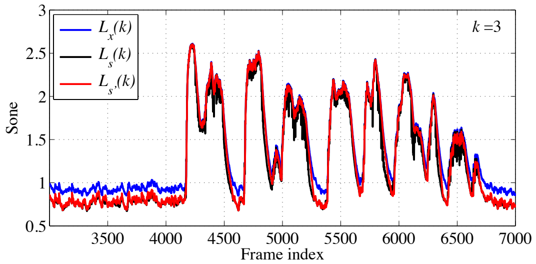


Figure 2: *Temporal variation of the partial loudness levels.*

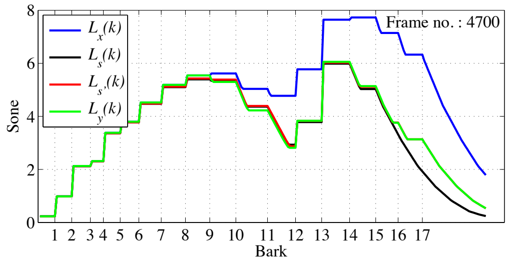


Figure 3: *Instantaneous partial loudness levels.*

The green and black lines represent the score of speech component of the output signal by applying the gain function to these two methods, which is obtained by the convolution of speech signal with the gain function in the proposed and conventional methods.

For all the investigated SNR conditions, the trend of STOI scores for the proposed method is almost the same as that for the conventional method, although the proposed method included extra noise, which is expected to include weak speech spectral components. On the other hand, the scores of the speech part with the gain function of the proposed method are superior to that of the conventional method for low SNR conditions. This result reveals that the proposed method has gain control based on partial loudness, and improves the speech signal quality effectively, while including the extra noise. We think that there is some possibility about positive effect of speech recognition by using the extra noise, which has the same intonation as the estimated speech [9].

5. Conclusions

In this paper, a noise suppression algorithm based on loudness management is proposed. The results of our evaluation using STOI indicate that the score of the clean-speech output by applying the gain function in the proposed method is improved compared with that of the conventional method, while the score of noisy-speech by the proposed method is almost the same as that of the conventional method. These results imply that speech components are preserved; however, a subjective evaluation has not been performed.

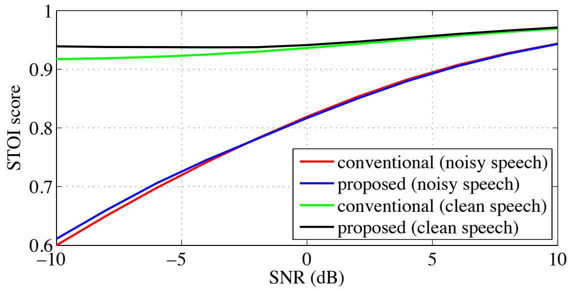


Figure 4: *SNR vs. STOI.*

6. References

[1] S. Kochkin, “MarkeTrack V: Why my hearing aids are in the drawer: The consumers’ perspective,” *The hearing journal*, vol.53, pp. 34-41, 2000.

[2] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

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

[4] A. Sarampalis, S. Kalluri,B. Edwars and E. Hafler, “Objective measures of listening effort: effects of background noise and noise reduction,” *J. Speech Lang Hear Res*. vol. 52, no. 5, pp. 1230-1240, 2009.

[5] 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.

[6] S. Rangachari, and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *IEEE Trans. on Acoust., Speech, and Signal Processing*. vol. 48, no. 2, pp. 220-231, 2006.

[7] E. Zwicker, H. Fastl ,U. Widmann, K. Kurakata, S. Kuwano, and S. Noamba, “Program for calculating loudness according to DIN 45631 (ISO532B),” *Journal of the Acoustical Society of Japan*. vol. 12, no. 1, pp. 39-42, 1991.

[8] 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.

[9] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wyhonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*. vol. 270, no. 5234, pp. 303-304, 1995.

Do models hear the noise? Predicting the outcome of the German matrix sentence test for subjects with normal and impaired hearing

David Hülsmeier[1], *Mareike Buhl*[1], *Nina Wardenga*[2], *Anna Warzybok*[1], *and Marc René Schädler*[1]

[1]Medizinische Physik and Cluster of Excellence Hearing4all, CvO Universität Oldenburg, 26111 Oldenburg, Germany
[2]Department of Otolaryngology and Cluster of Excellence Hearing4all, Hannover Medical School, Hannover, Germany

{david.huelsmeier, mareike.buhl, a.warzybok, marc.r.schaedler}@uni-oldenburg.de
wardenga.nina@mh-hannover.de

Abstract

Impaired hearing negatively affects the speech reception of an increasing proportion of aging societies world-wide. Recently, the German matrix sentence test was used with a fixed noise level to quantify the effect of impaired hearing on speech reception thresholds in 315 ears with differently pronounced hearing loss. Two domains with different linear dependences of the outcome on the pure-tone average hearing loss were identified. One domain where listening in noise dominates the results and one where the individual capacity for listening in "quiet" mainly dominates. The aim of this work was to test to which extent this behavior can be predicted by two different speech intelligibility models based on the individual audiograms. Therefore, the framework for auditory discrimination experiments (FADE) and the speech intelligibility index (SII) were individualized with the 315 audiograms, and the predicted outcomes were compared to the empirical data. Both models were found to predict the characteristic change in the slope, where FADE under-estimated and the SII over-estimated the linear dependence in the listening-in-noise domain. Over-all, FADE predictions were found to be more accurate than SII-based predictions with root-mean-square prediction errors (RMSE) of 5.6 dB and 6.8 dB, respectively. A group-wise analysis revealed that for special cases (e.g., steep hearing loss) with typically high pure-tone averages FADE provides much more accurate predictions (RMSE=6.7 dB) than the SII (RMSE=22.6 dB), while for low pure tone averages the root-mean-square prediction error is typically one to two dB lower with the SII. The conclusion of this paper is that the effect of impaired hearing on speech perception can be only partly explained with the absolute hearing threshold, which is primarily visible in the listening-in-noise domain. Probably a supra-threshold component that is moderately correlated with the absolute hearing threshold of hearing loss is responsible for up to 50% of the slope in the listening-in-noise domain.

Index Terms: speech intelligibility prediction, matrix sentence test, modeling approaches

Towards Multi-modal Hearing Aid Design and Evaluation in Realistic Audio-Visual Settings: Challenges and Opportunities

Amir Hussain[1], *Jon Barker*[2], *Ricard Marxer*[2], *Ahsan Adeel*[1], *William Whitmer*[3], *Roger Watt*[1] *and Peter Derleth*[4]

[1]University of Stirling, UK, [2]University of Sheffield, UK,
[3]Unviersity of Nottingham, UK, [4]Sonova AG, Switzerland

[1]ahu@cs.stir.ac.uk [2]j.p.barker@sheffield.ac.uk

Abstract

A limited number of research developments in the field of speech enhancement have been implemented into commercially available hearing-aids. However, even sophisticated aids remain ineffective in environments where there is overwhelming noise present. Human performance in such situations is known to be dependent upon input from both the aural and visual senses that are then combined by sophisticated multi-level integration strategies. In this paper, we consider the opportunities and challenges presented by hearing-aid development in an audio-visual (AV) speech context. First, we posit the case for new multimodal AV algorithms that enhance speech quality and intelligibility with the aid of video input and low-latency combination of audio and visual speech information. Second, we consider the challenges that the AV setting presents to hearing aid evaluation. We argue that to meaningfully reflect everyday usage, hearing aid evaluation needs to be performed in an audio-visual setting regardless of whether hearing aids are directly using visual information themselves. We consider the need for new AV speech in noise listening tests, and for research into techniques for predicting objective AV speech quality and intelligibility. Finally, an AV speech enhancement evaluation challenge is proposed as a starting point for stakeholder discussion.

Index Terms: audio-visual speech, speech enhancement, speech intelligibility assessment

1. Introduction

The multimodal nature of speech is well established. Speech is produced by the vibration of the vocal cords being filtered according to the configuration of articulatory organs. Due to the visibility of some of these articulators (i.e., lips, teeth and tongue), there is an inherent and perceptible relationship between audio and visual speech properties. Pioneering work [1, 2, 3] demonstrated that listeners exploit this relationship, unconsciously lip reading to improve the intelligibility of speech in noise [4]. Further, looking at a speaker makes speech more detectable in noise [5], i.e., as if audio cues are being visually enhanced [6].

Embracing the multimodal nature of speech presents both opportunities and challenges for hearing assitive technology: on the one hand there are opportunities for the design of new multimodal algoirthms; on the other hand multimodality challenges the current standards for hearing aid evaluation, which generally consider the perception of the audio signal in insolation.

This paper will first consider the potential benefits of designing fully audio-visual hearing devices. In particular, we consider the design of a new breed of device that employs both microphones and video sensors. Such a device has the potential to extract information from the pattern of the speaker's face and lip movements and to employ this information as an additional input to speech enhancement algorithms. In Section 2 we discuss the AV-COGHEAR project that is aiming to build and test prototypes of this technology.

In Section 3 we turn attention to the challenge of hearing device evaluation. Our main concern in this regard is that standard evaluation strategies, which use an audio-only setting, may not be predictive of a device's performance when used in real multimodal conditions. This is just as much true for devices that use audio-only input as it is for audio-visual devices. We consider the requirements of a fully multimodal evaluation and conclude in Section 4 by making a proposal for an open multimodal speech enhancement challenge that we hope will stimulate fresh research in this area.

2. Audio-Visual speech enhancement

2.1. Background

Despite decades of research, there are are few speech enhancement algorithms that can reliably increase the intelligibility of speech corrupted by complex noises typical of everyday listening conditions. For example, spectral subtraction can be very effective for reducing the perception of noise in stationary conditions, but the apparently 'cleaner' processed speech turns out to be no easier to understand. If multiple microphones are available then beamforming algorithms can lead to genuine speech intelligibility improvements but even these techniques are hard to employ in an unpredictable noise environment. Consequently hearing aid algorithms achieve most of their benefit simply by amplifying the signal into the audible range, and offer little advantage for speech listening when speech is present in high levels of background noise.

There is reason to believe that, in contrast to audio-ony algorithms, audio-visual speech enhancement approaches may be able to offer consistent intelligibility gains – especially for hearing impaired listeners. To understand why visual features may be beneficial it is important to understanding why noise renders speech less intelligible in the first place. The commonly understood view is that the noise sources reduce speech intelligibility by energetically masking the target source. Visual signals can then restore intelligibility by delivering phonetic information that has been obliterated in the masked regions. However, this is only part of the picture. Intelligibility is also governed by *informational* masking (IM), i.e., the degree to which the auditory system is able to, i) segregate spectro-temporal (ST) regions that are speech dominated from those that are background-dominated, and ii) focus attention on the target regions.

IM is amplified by even mild hearing impairment, leading to large speech intelligibility losses in social situations where speech is present in the noise background. It has been suggested that this is partly due to the loss of precision with which 'grouping cues' are encoded - i.e., signal properties such as periodicity and apparent location that allow for a signal to be sequentially organised, [7]. Schwartz et al. [8] have shown that visual cues can be exploited at a pre-phonetic stage, reducing IM. Essentially, visual cues can supplement auditory grouping cues, providing a signal that directs attention to the ST regions dominated by the target source.

We believe that listeners may benefit from an AV hearing device that is able to mimic the IM releasing function of visual cues. For example, the device would use the visual information to direct the audio signal processing to amplify speech signal components and attenuate noise components.

2.2. The AV-COGHEAR project

The ongoing UK Engineering and Physical Sciences Research Council (EPSRC) funded AV-COGHEAR project, collaboratively led by Stirling and Sheffield Universities, is a first attempt at developing a cognitively-inspired, adaptive and context-aware approach for combining audio and visual cues (e.g., from lip movement) to deliver speech intelligibility enhancement [9].

The project's overarching goal is the development of next-generation multi-modal hearing aids and listening devices that have the potential to be a disruptive technology redefining user expectations. Beyond hearing aid devices we foresee impact in a number of areas including: cochlear implant signal processing, speech recognition systems, auditory systems engineering in general, and clinical, computational, cognitive and auditory neuroscience. A preliminary deep-learning-driven, multi-modal speech enhancement framework pioneered at Stirling [10] is currently being significantly extended to incorporate innovative perceptually-inspired models of auditory and AV scene analysis developed at Sheffield [11]. Further, novel computational models and theories of human vision developed at Stirling are being deployed to enable real-time tracking of facial features. Contextual multimodality selection mechanisms are being explored, and collaborations with SONOVA and MRC IHR, will facilitate envisaged delivery of a clinically-tested software prototype.

In the literature, much progress has been made to develop enhanced speech processing algorithms capable of improving speech quality. In contrast, little work has been conducted to design algorithms that can improve speech intelligibility. In this project, our hypothesis is that it is possible to combine visual and acoustic input to produce a multimodal hearing device that is able to significantly boost speech intelligibility in the everyday listening environments, in which traditional audio-only hearing devices prove ineffective.

To test this hypothesis, we are collaboratively working to develop and clinically validate a next-generation cognitively-inspired, AV hearing-device software prototype, capable of real-time implementation, which will autonomously adapt to the nature and quality of its visual and acoustic environmental inputs. In this context, we have currently developed two contrasting approaches to speech enhancement developed respectively at Stirling and Sheffield: (1) A deep lip-reading driven Weiner filtering approach, shown in Figure 1 and (2) an audio-visual analysis-resynthesis approach, depicted in Figure 2 [12]. The preliminary objective and subjective evaluation has revealed the potential and reliability of the proposed AV technology as compared to the state-of-the-art audio only speech processing techniques.

3. Speech enhancement evaluation in realistic AV settings

3.1. Background

The development of hearing devices that utilise both audio and visual information, highlights the growing need for hearing devices to be evaluated in realistic multimodal settings. In the literature, there exist several standards for evaluating hearing aid algorithms in audio only settings, ranging from the Connected Speech Test (CST) (CST; [13]), the Speech Intelligibility Index (SSI) [14], to Kates' extension to the SSI [15]. However, there are no established standards for evaluating hearing-aid algorithms in audio-visual settings. Note, an audio-visual extension of the CST [16] was proposed shortly after the audio-only test but has not been widely adopted.

Evaluating a hearing device in an audio-only setting may produce a misleading view of the speech intelligibility benefits it will provide. Except in a few naturally audio-only situations, (e.g., telephone conversations), hearing aid users who are struggling to understand speech in noise will be closely attending the speaker's lips. These listeners will therefore experience a visual benefit which will improve their aided-performance. Note, this is true regardless of whether the hearing aid is using the visual signal itself. The size of this visual benefit needs to be accounted for.

Visual-benefit would not be a problem for hearing aid evaluation if the size of the benefit was independent of the hearing aid algorithm. If this was true then the ranking of algorithms would remain the same and a best algorithm could still be chosen. However, this is unlikely to be the case. For example, consider an algorithm that emphasises aspects of the acoustic signal that are redundantly encoded in the visual signal. This algorithm might provide large benefits in an audio-only evaluation but then be shown to proffer little benefit in a setting where the user sees the lip movements directly.

Despite the clear necessity for hearing aid speech performance to be evaluated from an AV perspective, there has been very little work in this direction. Recently, Wu and Bentlier [17] examined how visual cues impact directional benefit and preference for the directional microphone hearing-aid. The authors administered two speech recognition in noise tests to assess directional benefit: (1) the AV version of the Connected Speech Test (CST; [13, 16]) and (2) the Hearing in Noise Test [18] to investigate the impact of visual cues on the directional benefit. It was reported that visual cues significantly improved speech recognition performance to its ceiling level and reduced the directional benefit and preference for directional processing.

3.2. Challenges for audio-visual hearing aid evaluation

In this section we outline the main challenges facing audio-visual speech intelligibility testing.

3.2.1. Audio-visual HA performance predictors

In an ideal world, hearing aid algorithms could be evaluated cheaply using algorithms that would predict the intelligibility and/or quality of the processed speech. The processed audio speech signal and its video counterpart could be fed into an objective test that would accurately predict intelligibility of the signal. Unfortunately, this is an unrealistic proposition which

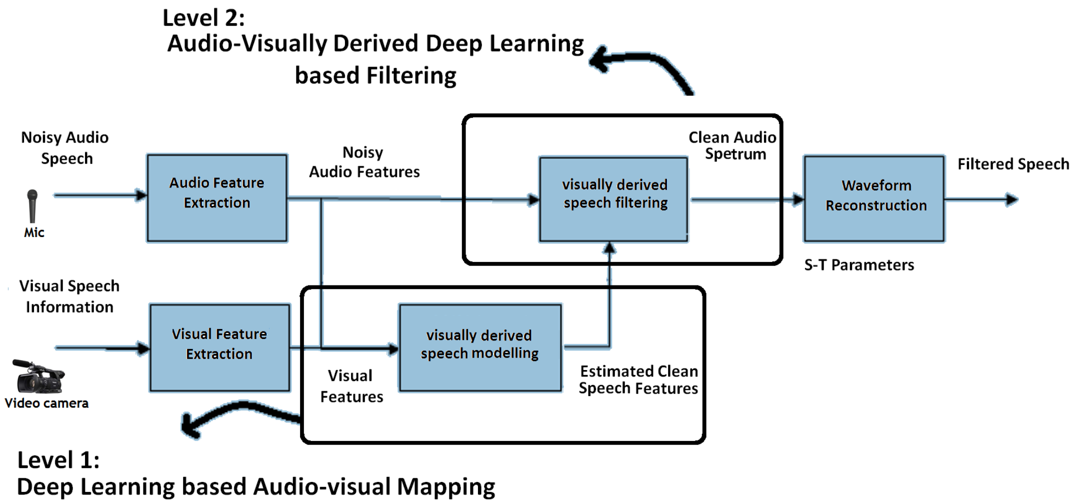

Level 2:
Audio-Visually Derived Deep Learning based Filtering
Noisy Audio Speech
Mic
Audio Feature Extraction
Noisy Audio Features
visually derived speech filtering
Clean Audio Spetrum
Waveform Reconstruction
Filtered Speech
S-T Parameters
Visual Speech Information
Video camera
Visual Feature Extraction
Visual Features
visually derived speech modelling
Estimated Clean Speech Features
Level 1:
Deep Learning based Audio-visual Mapping

Figure 1: *Deep lip-reading driven Weiner filtering (Stirling)*

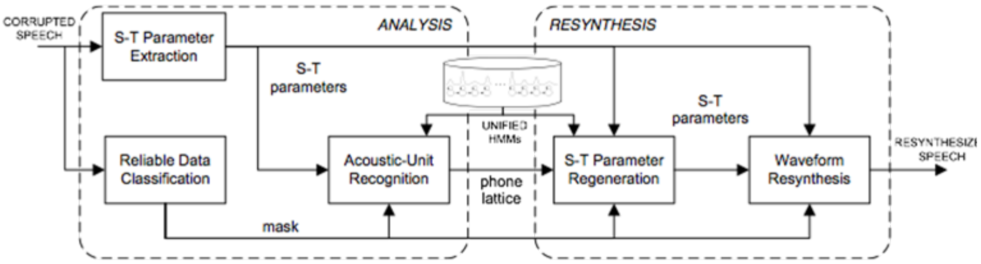


Figure 2: *Audio-visual analysis-resynthesis framework (Sheffield)*

remains a challenge even for audio-only evaluation.

There have been many proposed metrics for objective speech quality and intelligibility prediction. Algorithms are categorised as intrusive or non-intrusive depending on whether they require a clean speech reference signal or not, respectively. We can assume that for hearing-aid development a reference signal can be available and therefore intrusive algorithms can be applied. These include the normalized covariance metric (NCM) [19] and short-time object intelligibility (STOI) [20] which predict intelligibility and perceptual evaluation of speech quality (PESQ) [21] which predicts speech quality. Although different in detail, these algorithms all operate by making a weighted comparison between an auditorily-inspired representation of the reference and corrupted signal.

More recently developed predictors have been especially designed for hearing aid (HA) processing. These include the HA speech quality index (HASQI) [22], the HA speech intelligibility index (HASPI) [23], and an extension of the perception-model-based quality prediction method (PEMO-Q) [24] adapted for hearing impairment (PEMO-Q-HI) [25]. These predictors again compare a reference and a processed signal in an auditory model space, however, their auditory models can be tuned to mimic the effects of a listener's hearing impairment, (e.g., raised thresholds, filter broadening, etc).

A particular problem with these approaches is that their performance can be sensitive to the type of processing performed by the aid (see [26] for a review). For example, non-linear frequency compression (NFC) – a recent development in hearing aids which warps the signal spectrum to fit the listener's usable frequency range – can generate big apparent differences between the reference and processed signal. Unless the metric is designed to expect NFC and compensate for this frequency warping it will predict the NFC processed signal to have low quality/intelligibility. The fundamental problem here is that the metrics are necessarily built on shallow models of speech perception. The resulting need to fit the prediction models to hearing aid algorithms is surely problematic if they are expected to effectively evaluate novel and unanticipated approaches to hearing aid signal processing.

When considering audio-visual intelligibility the situation is worse. There are no adequate models of how acoustic and visual speech information are combined during speech perception. There are currently no models that can predict effects such as informational masking release in non-stationary masker conditions. Without this understanding there is no basis on which to start building models of AV speech intelligibility.

3.2.2. *Difficulties with speech-in-noise listening tests*

There are a large number of standard speech-in-noise listening tests that can be used to directly measure the intelligibility benefit of a hearing device. They include adaptive and fixed SNR tests. The former include the Hearing in Noise Test

(HINT) [27], QuickSIN [28], Words in Noise (WIN) [29, 30] and Bamford-Kowal-Bench SIN (BKB-SIN) [31, 32]. These tests automatically adapt the SNR to the threshold at which communication breaks down (i.e., at which a fixed percentage of words are incorrectly recognised). They are quick and easy to administer but have the drawback of not being able to provide information about performance at other SNRs above and below this threshold. Fixed SNR tests such as CST [13] and Speech Perception in Noise Test (SPIN) [33] measure the percentage of words correctly recognised at a presentation SNR. However, they are susceptible to ceiling effects, i.e., no benefit can be measured once all words are recognised correctly. Choosing an appropriate SNR can be difficult.

Tests vary with respect to the type of speech material - some using isolated words and others complete sentences. Sentences are regarded as more appropriate materials for intelligibility measurement as they reflect real speech and produce steeper psychometric functions that more accurately estimate threshold SNRs. Sentences need to be carefully designed to be phonetically balanced and to have low predictability. So-called 'matrix tests' achieve this by using randomly chosen words from the closed-set in each word position in a sentence: e.g., a sentence might be composed as: <name>, <verb>, <number>, <adjective>, <noun> with 10 choices for each slot leading to 100,000 possible sentences. Such tests have been designed for many different languages, e.g. German [34, 35], Spanish [36], English [37], etc. There has been recent success in predicting sentence test performance in a wide range of noise conditions using statistical techniques adapted from the speech recognition community [38].

Speech-in-noise tests have primarily been designed for use in clinical settings for fitting a device to a given user. Comparing two devices for a given user is a well-posed problem. However, if a conclusion is required about which device is better in a more general sense, then obvious problems emerge. The question is impossible to answer independently of some characterisation of the user, i.e., the precise nature and degree of their hearing deficit. The test would then require a pool of listeners matching this characterisation that is large enough to average out remaining individual differences. This is particularly problematic given that audiograms – the standard characterisation of hearing deficit – can be poor predictors of speech in noise performance (and even poorer predictors of audio-visual speech recognition ability).

The difficulties experienced with audio-only testing are compounded in audio-visual settings, specifically: the visual cues may make it more likely to encounter ceiling effects; the test-retest scores are likely to be more variable given the increased cognitive complexity of the task; there is a starker contrast between controlled, well-framed, clearly articulated visual input and video characteristic of everyday conversational settings; and selecting homogeneous listener pools is more challenging as there are large and unpredictable individual differences in visual speech benefit amongst listeners.

3.2.3. The need for a realistic AV corpus

Reliable evaluation of future AV speech filtering technology will require subjective intelligibilty assessment. This raises the question of what type of speech material to use. Although there exist a number of small well-controlled audio-visual speech corpora, such as BANCA [39], AVICAR [40], VidTIMIT [41], and Grid [42], there is a need for evaluation of multi-modal speech enhancement systems using realistic audiovisual speech data. Audio-visual datasets are required in which speakers are speaking more naturally than in many existing corpora, including conversational speech and imperfect visual data. This is represented by the speaker moving their head, obscuring their face, and also different levels of background noise to take account of the Lombard effect (where speakers naturally adjust their speech to take account of different levels of background noise). To our knowledge, there is no corpus available that contains a sufficient range of AV speech data, or variety of A and V noise (i.e., acoustic noise, speaker movement and occlusion, etc.).

4. Conclusion – Towards an open AV evaluation framework

Future multimodal hearing devices will demand new approaches to evaluation. There will be a need to reconsider how hearing aid algorithms are evaluated during development so as to incorporate visual input, subject to real-time, low-latency constraints. There is also need to reconsider how devices are prescribed and fitted to patients. It will no longer be appropriate to use audio-only speech-in-noise tests. For the new devices the key question will be how the AV processing of the device interacts with the AV processing of the user when presented with realistic AV input.

New standards will only emerge from discussion across the sector involving manufacturers, health care professionals, audiologists and patients. We hope the 2017 CHAT Workshop will provide a starting point for this activity. To stimulate progress we plan to organise an open competition for AV speech enhancement evaluation, which could be run as part of an international INTERSPEECH Workshop in 2018. The competition could run along similar lines to the ISA Blizzard Challenge where the cost of participation pays for evaluation of the algorithms.

We conclude by presenting a tentative proposal for hearing algorithm evaluation with the aim of seeking feedback from the community.

The evaluation campaign will have two phases:

Phase 1 - development of enhancement algorithms: The proposal would be to use the existing AV Grid corpus [42] as the source material. This has a design similar to a matrix test and has already had extensive use in audio-only speech intelligibility studies. We would then mix this corpus with everyday complex noise backgrounds such as café and street noises recorded in the CHiME-3 corpus [43]. Participants would be invited to apply their algorithms (audio-only or audio-visual). We would then measure subjective intelligibility and subjective quality using a large bank of paid listening subjects who would be presented with the processed audio alongside the video. The evaluators could be NH or a homogenous group of HI listeners.

Phase 2 - development of new objective measures: Phase 1 will generate a large amount of listener data, i.e., showing how listeners have responded to variously-processed noisy AV speech samples. This data can then be used to test models that predict AV intelligibility and speech quality. We could immediately evaluate existing audio-based predictors which would be expected to underestimate AV performance. The challenge would then be to develop new AV predictors that extend these models. We would release a subset of the data from Phase 1 for model development and retain a hidden set for model evaluation.

5. Acknowledgements

This research is funded by the EPSRC project AV-COGHEAR, EP/M026981/1.

6. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[2] N. P. Erber, "Auditory-visual perception of speech," *Journal of Speech and Hearing Disorders*, vol. 40, no. 4, pp. 481–492, 1975.

[3] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, vol. 36, no. 4-5, pp. 314–331, 1979.

[4] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisualspeech perception," *Attention, Perception, & Psychophysics*, vol. 60, no. 6, pp. 926–940, 1998.

[5] J. Kim and C. Davis, "Investigating the audio–visual speech detection advantage," *Speech Communication*, vol. 44, no. 1, pp. 19–30, 2004.

[6] K. W. Grant and P.-F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1197–1208, 2000.

[7] D. Ruggles, H. Bharadwaj, and B. G. Shinn-Cunningham, "Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication," *Proceedings of the National Academy of Sciences*, vol. 108, no. 37, pp. 15 516–15 521, 2011.

[8] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.

[9] A. Abel, R. Marxer, J. Barker, R. Watt, W. Whitmer, P. Derleth, and A. Hussain, "A data driven approach to audiovisual speech mapping," in *Advances in Brain Inspired Cognitive Systems. BICS 2016*, C. Liu, A. Hussain, B. Luo, K. Tan, Y. Zeng, and Z. Zhang, Eds. Spring, 2015, vol. 10023, pp. 331–342.

[10] A. Abel and A. Hussain, "Novel two-stage audiovisual speech filtering in noisy environments," *Cognitive Computation*, vol. 6, no. 2, pp. 200–17, 2014.

[11] J. Barker and X. Shao, "Energetic and informational masking effects in an audio-visual speech recognition system." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 3, pp. 446–458, Mar. 2009.

[12] J. Carmona, B. J., A. Gomez, and N. Ma, "Speech spectral envelope enhancement by HMM-based analysis/resynthesis," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 563–66, 2013.

[13] R. Cox, G. Alexander, and C. Gilmore, "Development of the Connected Speech Test," *Ear and Hearing*, vol. 8, no. 5, pp. 119S–126S, 1987.

[14] ANSI S3.5-1997, "American National Standard methods for the calculation of the Speech Intelligibility Index," American National Standards Institute, New York, Tech. Rep., 1997.

[15] J. Kates, "Coherence and the Speech Intelligibility Index," *The Journal of the Acoustical Society of America*, vol. 115, no. 2604, pp. 1085–99, 2004.

[16] R. Cox, G. Alexander, C. Gilmore, and P. K.M., "The Connected Speech Test version 3: Audiovisual administration," *Ear and Hearing*, vol. 10, no. 1, pp. 29–332, 1989.

[17] Y. Wu and R. Bentlier, "Impact of visual cues on directional benefit and preference: Part I–laboratory tests." *Ear and Hearing*, vol. 31, no. 1, pp. 22–34, 2010.

[18] M. Nilsson, S. Soli, and J. Sulivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–99, 1994.

[19] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

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

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[22] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[23] ——, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[24] R. Huber and B. Kollmeier, "PEMO-QA new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[25] R. Huber, V. Parsa, and S. Scollie, "Predicting the perceived sound quality of frequency-compressed speech," *PloS one*, vol. 9, no. 11, p. e110260, 2014.

[26] 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.

[27] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[28] M. C. Killion, P. A. Niquette, G. I. Gudmundsen, L. J. Revit, and S. Banerjee, "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2395–2405, 2004.

[29] R. H. Wilson, "Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance," *Journal of the American Academy of Audiology*, vol. 14, no. 9, pp. 453–470, 2003.

[30] R. H. Wilson and C. A. Burks, "Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: A clinic protocol," *Journal of rehabilitation research and development*, vol. 42, no. 6, p. 839, 2005.

[31] P. Niquette, J. Arcaroli, L. Revit, A. Parkinson, S. Staller, M. Skinner, and M. Killion, "Development of the BKB-SIN test," in *Proc. annual meeting of the American Auditory Society*, 2003.

[32] J. Bench, A. Kowal, and J. Bamform, "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," *British Journal of Audiology*, vol. 13, pp. 108–112, 1987.

[33] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[34] K. Wagener, T. Brand, V. Kühnel, , and B. Kollmeier, "Entwicklung und evaluation eines satztests fürdie Deutsche sprache I: Design des Oldenburger Satztests (Development and evaluation of a sentence test for the German language I: Design of the Oldenburg Sentence Test)," *Z. Fuer Audiologie, Audiological Acoust*, vol. 38, no. 2, pp. 4–15, 1999.

[35] ——, "Entwicklung und evaluation eines satztests fürdie Deutsche sprache II: Optimierung des Oldenburger Satztests (development and evaluation of a sentence test for the German language II: Optimization of the Oldenburg Sentence Test)," *Z. Fuer Audiologie, Audiological Acoust*, vol. 38, no. 2, pp. 44–56, 1999.

[36] S. Hochmuth, T. Brand, M. Zokol, F. Zenker Castro, N. Wardenga, and B. Kollmeier, “A Spanish matrix sentence test for assessing speech reception thresholds in noise.” *Int. J. Audiol.*, vol. 51, no. 7, pp. 536–544, 2012.

[37] M. Zokoll, A. Warzybok, R. Carroll, B. Kreisman, P. Allen, K. Wagener, and B. Kollmeier, “Design, optimization, and evaluation of an American English matrix sentence test in noise.”

[38] B. Kollmeier, M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand, “Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by plomp with a quantitative processing model,” *Trends in Hearing*, vol. 20, pp. 1–17, 2016.

[39] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée *et al.*, “The BANCA database and evaluation protocol,” in *International conference on Audio-and video-based biometric person authentication*. Springer, 2003, pp. 625–638.

[40] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. S. Huang, “AVICAR: audio-visual speech corpus in a car environment.” in *INTERSPEECH*, 2004, pp. 2489–2492.

[41] C. Sanderson, “The VidTIMIT database,” IDIAP, Tech. Rep., 2002.

[42] 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.

[43] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third chimespeech separation and recognition challenge: Dataset, task and baselines,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

Evaluation of the benefit of neural network based speech separation algorithms with hearing impaired listeners

Gaurav Naithani[1] , *Tom Barker*[1], *Giambattista Parascandalo*[1†]
Lars Bramsløw[2], *Niels Henrik Pontoppidan*[2], *and Tuomas Virtanen*[1]

[1]Tampere University of Technology, Finland
[2]Eriksholm Research Centre, Oticon A/S, Denmark

[1]{gaurav.naithani, thomas.barker, giambattista.parascandolo, tuomas.virtanen}@tut.fi
[2]{labw, npon}@eriksholm.com

Abstract

Source separation is a useful technology for improving the benefit from hearing aids. However, most of the existing approaches to evaluating source separation rely on computational methods, and do not consider the effect of the algorithm on the end user. We seek to address this mismatch by quantifying the benefit of two state-of-the-art deep neural network (DNN) based source separation techniques, in terms of actual speech intelligibility benefits evaluated via subjective listening tests with 15 hearing impaired (HI) listeners, as well as more established computational metrics by which most source separation algorithms are currently compared. We present here our proposed source separation approach which is a novel application of the 'Convolutional Recurrent Neural Network' (CRNN) deep learning architecture, and compare it with feedforward deep neural network (FDNN) approach. We evaluate these approaches on two talker mixtures from Danish hearing in noise test (HINT) database. We are particularly interested in speech separation in this work as the hearing-impaired listeners have problems understanding speech in the presence of one or more competing voices.

Index Terms: source separation, deep neural networks, low latency, hearing aids

1. Introduction

Source separation is an important technology for improving hearing aid performance, and recently, large advances in this domain have been achieved using a range of techniques using 'deep neural networks (DNN)' – whereby mapping of an input to a target output is realised through learning complicated non-linear relationships which are captured within the network parameters. These approaches achieve state-of-the-art performance even at very low latency, which is critical for hearing aids [1]. It has been postulated (e.g., in [2]) that delays larger than 10 ms are objectionable to hearing impaired (HI) listeners. The algorithmic delay of the DNN based approach used in our work is 8 ms. This low-latency performance is therefore one of the critical design features when considering source separation for hearing aids.

Alongside low-latency performance, another primary goal of a hearing-aid algorithm is to improve speech intelligibility, yet most of the current evaluation methods do not address this need with respect to hearing-impaired listeners. Typically, source separation algorithms, and the literature which reports them, focuses primarily on the performance of the algorithms in terms of separated source energy (e.g., source to distortion ratio (SDR) [3]), predicted perceptual quality (PEASS [4]), or predicted intelligibility. The existing predicted intelligibility metrics such as short term objective intelligibility (STOI [5]) and extended short term intelligibility (ESTOI [6]) are based on models of normal hearing and tested on normal hearing listeners, so they may not be accurate predictors of algorithm performance for use in hearing aids.

Overall, current trends for developing and evaluating source separation as a general technology do not adequately consider the needs of its use specifically for hearing aids. We therefore seek to address this in both development of low-latency source separation and evaluation strategy and present our findings to date here.

2. DNN for source separation

We use the time-frequency masking paradigm of source separation whereby a DNN is used to predict time-frequency mask corresponding to the target speaker. The input features are short-time Fourier transform (STFT) coefficients and output is soft ratio mask defined as the ratio of magnitude spectrum of the target speaker and sum of magnitude spectra of constituent sources in the acoustic mixture (e.g., in [7, 8]). The predicted time-frequency mask is multiplied with mixture spectrum to yield the target speaker spectrum.

We investigate convolutional recurrent neural network (CRNN) for source separation, originally proposed in [7]. The motivation of using this architecture is to combine the feature extraction property of convolutional layers from the input, i.e., time-frequency representation of the acoustic mixture in our case, and the ability of recurrent layers (with long short term memory (LSTM) units [9]) to model long term temporal dependencies. We compare this architecture to a feedforward DNN architecture similar to the one used in [10]. Table 1 shows the hyperparameters used for the two architectures. Note that in case of FDNN, frames spanning previous temporal context of 32 ms is fed to the input for estimation of the current frame, as was done in [10]. For more details on hyperparameter selection for the two architectures, please refer [7]. Output neurons for both topologies use sigmoid activations while hidden units for CRNN are rectified linear units and FDNN are sigmoid units. Max pooling is used after each convolutional layer in CRNN but only along frequency axis. Dropout regularization of 0.4 is used. For training DNNs Keras library [11] is used.

3. Evaluation

The dataset used for training and evaluation of neural networks is an extended version of the Danish hearing in noise test (HINT) dataset developed by [12]. The extended version consists of three male and three female speakers, each of them

[*] The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources used in experiments reported in this paper.

[†] The author is currently with Max Planck Institute for Intelligent Systems.

Table 1: *Hyperparameters used for the FDNN and CRNN. The pooling scheme represents max pooling operation along time and frequency axes.*

FDNN			CRNN						
hidden layers	hidden neurons	previous context	conv. layers	recurr. layers	recurr. neurons	conv. filters	pooling scheme	sequence length	conv. kernel size
4	1024	32 ms	3	1	256	256	*1 by 2*	512 ms	3 × 3

recorded speaking 13 lists consisting of 5 word natural sentences [13]. We use four lists for training and one list for validation. The remaining eight lists are used for testing. The test mixtures are prepared by summing the signals corresponding to the two talkers. The evaluation of the methods is based upon: 1) Computational metrics of separation, i.e., source to distortion ratio (SDR), and extended short term objective intelligibility (ESTOI), the latter being better suited to our task as interferer in our case (i.e., for two talker mixtures) is non stationary; and 2) Word recognition tests with hearing impaired listeners.

For subjective listening tests, a target-masker (TM) set up is used where one of the constituent speaker serves as the target signal. A cue is provided before the playback to indicate which of the speaker sentence the listener must reproduce. The listening test scores are percentage of correct word scores reproduced by the listener, transformed according to [14] to remove floor and ceiling effects. The study involves 15 hearing-impaired listeners with moderate to severe sloping hearing losses. In addition to the two DNN test conditions, we have two more test conditions: one where unprocessed mixture is presented (referred as *Sum*) and the other where the ground truth source is presented (referred as *Separate*). A comparison between these four test conditions is made.

4. Results and conclusions

Table 2 reports SDR and ESTOI values corresponding to FDNN and CRNN, for three speaker pairs: M1 F1, M1 M2, and F1 F2. CRNNs here showed a slightly better average ESTOI scores than FDNN. The subjective listening test, as depicted in Figure 1, showed a significant benefit of 35 % points with the DNN methods in comparison to the *Sum* condition. The difference between the two DNN modes was not found statistically significant albeit a slightly higher mean accuracy was observed for CRNN as compared to FDNN. It is interesting to observe that ESTOI metric showed similar pattern but the difference in performance between the two DNN architecture is not large enough to infer if the ESTOI metric is a good predictor of intelligibility performance for HI listeners.

The obtained results in this study show that DNN based algorithms have significant potential for improving speech intelligibility for HI listeners in tasks where a speech signal of interest is to be attended to in the presence of a masker speech signal. A more exhaustive description of listening test results will be reported in [15].

Table 2: *Performance metrics for the two DNN architectures.*

Speaker pair	FDNN		CRNN	
	SDR	ESTOI	SDR	ESTOI
M1 F1	7.42	0.77	7.44	0.79
M1 M2	5.96	0.76	6.06	0.78
F1 F2	5.40	0.71	5.56	0.72

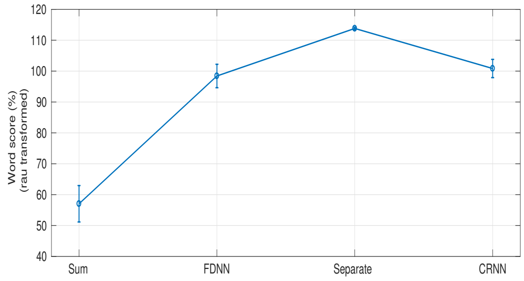


Figure 1: *Word recognition rates for the two DNN architectures for TM task. The vertical bars denote 0.95 confidence intervals.*

5. References

[1] L. Bramsløw, "Preferred signal path delay and high-pass cut-off in open fittings," *International journal of audiology*, vol. 49, no. 9, pp. 634–644, 2010.

[2] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[3] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[6] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[7] G. Naithani, T. Barker, G. Parascandolo, L. Bramsløw, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, (in press).

[8] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 1562–1566.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] G. Naithani, G. Parascandolo, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Low-latency sound source separation using deep neural networks," in *IEEE Global Conference on Signal and Information Processing*, 2016.

[11] F. Chollet, "Keras." GitHub, 2016, available at *https://github.com/fchollet/keras*.

[12] J. B. Nielsen and T. Dau, "The Danish hearing in noise test." *International journal of audiology*, vol. 50, no. 3, pp. 202–8, 2011.

[13] L. Bramsløw, M. Vatti, R. K. Hietkamp, and N. H. Pontoppidan, "A new competing voices test paradigm to test spatial effects and algorithms in hearing aids," in *International Hearing Aid Research Conference (IHCON)*, 2016, p. 1.

[14] G. Studebaker, "A "rationalized" arcsine transform," *Journal of Speech and Hearing Research*, vol. 28, no. September, pp. 455–462, 1985.

[15] L. Bramsløw, G. Naithani, T. Barker, A. Hafez, N. H. Pontoppidan, and T. Virtanen, "Hearing impaired listeners benefit from deep neural network source separation in competing voice scenarios," manuscript in preparation for the Journal of The Acoustical Society of America.

Blind Reverberation Power Estimation Using Exponential Averaging with Attack and Release Time Constants for Hearing Aids

Kotoyo Nozaki[1*], *Yasuhiro Oikawa*[1**], *Yusuke Ikeda*[2], *Yoh-ichi Fujisaka*[3], *Masahiro Sunohara*[3]

[1]Waseda University, Japan
[2]Tokyo Denki University, Japan
[3]RION Co., Ltd., Japan

*kotokoto3726@asagi.waseda.jp, **yoikawa@waseda.jp

Abstract

Dereverberation processing is necessary for hearing-aid systems with a limitation of computational cost because reverberation degrades speech intelligibility in some reverberation environments. The spectral subtraction (SS) method is a simple and well-known technique as not only a noise-reduction method but also a dereverberation method. In dereverberation methods that are based on SS for hearing aids, it is desirable to estimate the reverberation power with blind processing. The SS-based blind-estimation method was proposed by Sunohara et al. using exponential averaging with attack and release time constants for single-channel speech signals.

In this paper, we evaluate the estimation accuracy of the reverberation components of the proposed method. The estimation error, which is the difference between the true and estimated reverberation power was used for evaluation, that was compared the results obtained with the method proposed by Lebart et al., which is a non-blind SS-based dereverberation method. From the results, the reverberation power was more correctly estimated, especially in the case of long reverberation, and the estimation error of the proposed method was about a half of the well-known non-blind method by Lebart.

Index Terms: reverberant speech, blind estimation, hearing aids, exponential averaging, spectral subtraction

1. Introduction

Speech recognition is difficult owing to the reverberation. In particular, hearing-impaired persons have difficulties when listening to a speech in reverberation environments. For example, their recognition rates of speech decrease as the reverberation time increases [1]. For severely hearing-impaired listeners, it becomes hard to follow conversations in noisy or reverberant conditions even when speaking with only one person, as Moore et al. reported [2].

According to Folkeard et al.[3], listening may be improved performing dereverberation processing in hearing aids for hearing-impaired persons. Many researchers have studied dereverberation processing for a long time, and many methods have been proposed. Neely et al. proposed a method using inverse filtering of the room impulse response [4], and Gannot et al. employed a signal subspace approach [5]. Although these methods are effective for dereverberation, it is hard for them to be implemented into hearing aids because of the limited of computational resource available. A method that is considered to be relatively simple is the spectral subtraction (SS) method [6], which is effective not only for noise reduction but also for dereverberation [7]. Löllmann proposed a dereverberation method based on the SS method for hearing-aid systems[8].

The accurate estimation of reverberant components is important for the SS method. There are two kinds of estimation methods—namely, non-blind methods and blind methods. For non-blind estimation methods [9, 10], information about the sound field, such as the reverberation time, is required beforehand. On the other hand, ordinary blind estimation methods do not require previous information about the sound field [11, 12, 13]. However, most of them require multiple microphones to estimate the reverberant components; hence, it is difficult to implement them into hearing aids. Recently, we proposed a dereverberation system with the blind estimation method using only single-channel speech signals based on exponential averaging with attack and release time constants [14, 15].

In this paper, we evaluate our proposed dereverberation system with blind estimation of reverberation power was evaluated. True reverberation power was calculated from the impulse responses without direct sound to obtain an estimation error, which is the difference between true and estimated reverberation powers. The estimation error of the proposed estimation method was compared with that of a typical non-blind method proposed by Lebart [9].

2. Dereverberation Method

2.1. Spectral Subtraction

The SS method was proposed by Boll in 1979 as a noise-reduction method [6]. It is also known to be valid for dereverberation. Then, Kinoshita et al. evaluated the validity of the SS method for the reduction of latter reverberant components caused by evaluation experiments using spectrogram and automatic speech recognition (ASR) [7]. The above-mentioned method is relatively simple compared with other methods, such as inverse filtering or the signal subspace approach [4, 5]. The procedure of the SS method is as follows:

Step:1 Perform a Fourier transformation on the input signal.

Step:2 Estimate the noise/reverberation power of the input signal.

Step:3 Subtract the estimated noise/reverberation power from the power of the input signal.
(This subtracted power is treated as the gain.)

Step:4 Multiply the gain by the input power.

In the second step in the procedure, it is necessary to accurately estimate the reverberation power for dereverberation. In the dereverberation method proposed by Lebart, the reverberation power is estimated using an impulse response model [9], and it is a well-known non-blind estimation method.

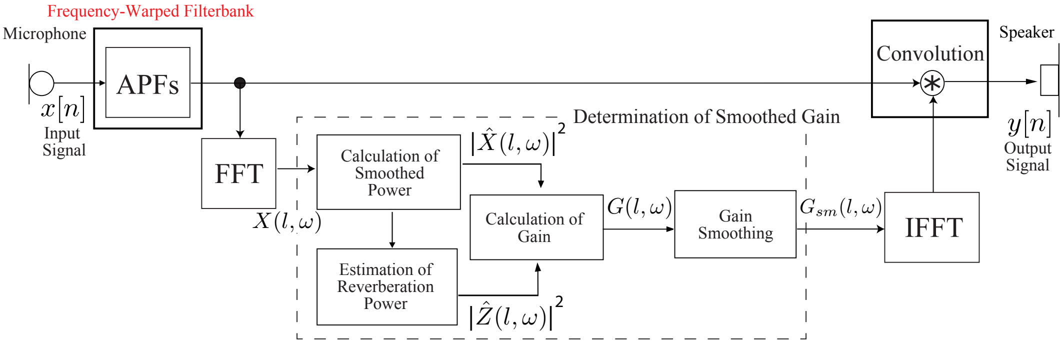


Figure 1: *System overview. The input signal flows through all-pass filters (APFs) with the warping parameter to analyze the input signal. The gain function $G(l,\omega)$ is determined by the estimated reverberation power.*

2.2. Exponential Averaging with Attack and Release Time Constants

Figure 1 shows the signal flow of our dereverberation system, which includes our proposed estimation method for reverberation power. An input signal from the microphone is analyzed using a frequency-warped filterbank (FWF) [16, 17], and transformed into the frequency domain using FFT. Sunohara et al. applied the weighted overlap-add (WOLA) filterbank[18] in a part of analysis/synthesis[14]. After the analysis, the smoothed signal power is calculated, and then the reverberation power is also estimated from the smoothed signal power. The gain function, which is derived from the above two powers, was smoothed in order to avoid musical noise issues in the output signal. The smoothed gain function was then transformed into the time domain and convolved with the warped signal that was passed through in each all-pass filter (APF).

Exponential averaging is one of the smoothing methods that were proposed by Roberts [19], and it is frequently used as a low-pass filter (LPF) on the field of signal processing. The weighting coefficient changes depending on the relationship between the estimated reverberation power and the input signal power. This method is given as follows.

Input signal $x(t)$ is generated by convolving the source signal $s(t)$ with the room impulse response $h(t)$, which includes the reverberation component,

$$x(t) = s(t) * h(t), \tag{1}$$

where $*$ represents the convolution operation. The frequency-analyzed input signal obtained by FWF is represented as $X(l,\omega)$, where l and ω denote the time and frequency indexes, respectively. The smoothed power $|\hat{X}(l,\omega)|^2$ is given as

$$|\hat{X}(l,\omega)|^2 = \beta|X(l,\omega)|^2 + (1-\beta)|\hat{X}(l-1,\omega)|^2, \tag{2}$$

where the parameter $\beta\,(0 < \beta \leq 1)$ represents the weighting value in exponential averaging. $|\hat{X}(l,\omega)|^2$ includes the reverberation power $|\hat{Z}(l,\omega)|^2$. Sunohara et al. estimated the reverberation power [14],

$$|\hat{Z}(l,\omega)|^2 = \gamma(l)|\hat{X}(l,\omega)|^2 + \{1-\gamma(l)\}|\hat{Z}(l-1,\omega)|^2, \tag{3}$$

where $\gamma(l)$ is a parameter that determines the time constant of exponential averaging

$$\gamma(l) = \begin{cases} \gamma_{\text{at}}, & (|\hat{X}(l,\omega)|^2 > |\hat{Z}(l-1,\omega)|^2) \\ \gamma_{\text{re}}, & (\text{otherwise}) \end{cases}. \tag{4}$$

Finally, the gain function $G(l,\omega)$ is obtained from the smoothed power of the input signal $|\hat{X}(l,\omega)|^2$ and reverberation signal $|\hat{Z}(l,\omega)|^2$,

$$G(l,\omega) = \begin{cases} \sqrt{\frac{|\hat{X}(l,\omega)|^2 - |\hat{Z}(l,\omega)|^2}{|\hat{X}(l,\omega)|^2}} & (|\hat{X}(l,\omega)|^2 > |\hat{Z}(l-1,\omega)|^2) \\ 0 & (\text{otherwise}) \end{cases}. \tag{5}$$

Then, the gain function $G(l,\omega)$ is smoothed to $G_{sm}(l,\omega)$ using the exponential averaging, as shown in Eq.(2). $G_{sm}(l,\omega)$ is transformed by IFFT and convolved with the signal analyzed by APFs as shown in Fig.1.

2.3. Well-known non-blind estimation method

Lebart used the power spectral density (PSD) to estimate the reverberation power from the reverberation time, which is a well-known non-blind estimation method [9].

$$|\tilde{Z}(l,\omega)|^2 = e^{-2\Delta l_0}|\hat{X}(l-l_0,\omega)|^2, \tag{6}$$

where l_0 is the typical duration for which it can be assumed that the signal is stationary. In this paper, let l_0 be 50 ms and

$$\Delta = \frac{3\ln 10}{T_r}, \tag{7}$$

where T_r is the reverberation time.

3. Experiment

In this section, a numerical experiment is conducted in order to evaluate the accuracy of proposed method by comparing the estimated reverberation power with true one as the estimation error.

3.1. Method

The reverberation speeches for the experiment were created by convolving anechoic speeches with an impulse response. The reverberation was modeled based on Polack's model [20]. In Polack's model, it is assumed that the room impulse response $h(t)$ is generated using an unsteady stochastic process

$$h(t) = \begin{cases} 0, & t < 0 \\ b(t)e^{-\Delta t}, & t \geq 0 \end{cases}, \tag{8}$$

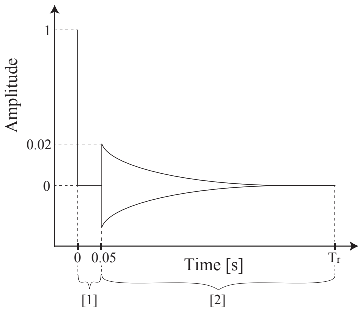


Figure 2: *Pattern diagram of the room impulse response based on the model proposed by Polack [20]. The first part $h_d(t)$ represents a direct sound and the second part $h_r(t)$ represents a reverberation component. The border between the direct sound and the reverberation component was 50 ms.*

where $b(t)$ is a steady Gaussian noise with an average of 0, and Δ is shown in Eq.(7).

Impulse responses were created based on this model. A direct sound is assigned at the beginning of the impulse response, and the reverberation decay starts from 50 ms after the direct sound. Let $h_d(t)$ be the first part before 50 ms and $h_r(t)$ be the second part after 50 ms. The impulse response with direct sound and the reverberation component is defined by

$$\hat{h}(t) = \begin{cases} 0, & t < 0 \\ h_d(t), & 0 \leq t < 50 \text{ ms} \\ h_r(t), & 50 \text{ ms} \geq t \end{cases}, \tag{9}$$

$$h_d(t) = \begin{cases} 1, & t = 0 \\ 0, & 0 \geq t \end{cases}, \tag{10}$$

$$h_r(t) = b(t)e^{-\Delta t}, \tag{11}$$

where $b(t)$ is the same as shown in Eq. (8). Figure 2 shows a pattern diagram of the impulse response.

For the evaluation, we used the difference between the estimated reverberation power and the true reverberation power, which is the estimation error. First, a true reverberation speech without direct sound was generated by convolving an anechoic speech with $h_r(t)$,

$$z_{true}(t) = s(t) * h_r(t), \tag{12}$$

which is $z_{true}(t)$ a true reverberation speech. It was transformed to $|Z_{true}(l,\omega)|$ in the frequency domain by FFT, and $|Z_{true}(l,\omega)|$ was smoothed according to Eq.(2),

$$|\hat{Z}_{true}(l,\omega)|^2 = \beta|Z_{true}(l,\omega)|^2 + (1-\beta)|\hat{Z}_{true}(l-1,\omega)|^2. \tag{13}$$

Table 1: *Experimental condition.*

Parameter	**Value**
γ_{at} (time constant)	3.125×10^{-5} (2.0 s)
γ_{re} (time constant)	6.250×10^{-4} (1.0×10^{-1} s)
Gain estimation parameter: attack (time constant)	1.250×10^{-2} (5.0×10^{-3} s)
Gain estimation parameter: release (time constant)	9.375×10^{-4} (6.7×10^{-2} s)
Smoothing parameter: β	1.250×10^{-3}
Lower bound of gain reduction: η	0
Sampling frequency [kHz]	16
Reverberation time [s]	1.0, 1.5, 2.0
FFT length [samples] (time)	32 (2 ms)

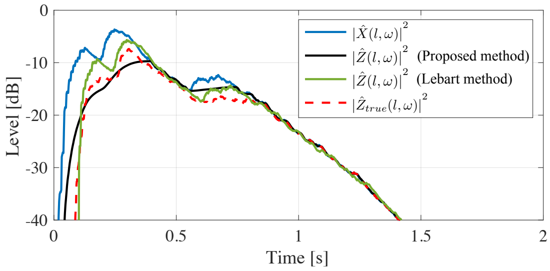


Figure 3: *Level of estimated powers. This figure shows the level of powers of the input signal (blue line), the reverberation signals estimated by the proposed method (black line) and by the non-blind method proposed by Lebart (green line), and the true reverberation signal (red broken line) at the 8th frequency band in FWF, whose central frequency was 1100 Hz. The reverberation time T_r was 1.5 s.*

Second, the sum of the estimated reverberation power $|\hat{Z}(l,\omega)|^2$ and the true reverberation power $|\hat{Z}_{true}(l,\omega)|^2$ was calculated for each method, after which the sum of the difference between the estimated and true reverberation powers $S_{diff}(\omega)$ was calculated,

$$S_{est}(\omega) = \sum_l |\hat{Z}(l,\omega)|^2, \tag{14}$$

$$S_{true}(\omega) = \sum_l |\hat{Z}_{true}(l,\omega)|^2, \tag{15}$$

$$S_{diff}(\omega) = \sum_l ||\hat{Z}(l,\omega)|^2 - |\hat{Z}_{true}(l,\omega)|^2|. \tag{16}$$

In this analysis, the summation was calculated for 3.0 s from the beginning. Finally, the ratio of the sum of the difference power $S_{diff}(\omega)$ and the sum of true reverberation power $S_{true}(\omega)$ was defined as the error ratio in the reverberation power estimation $Z_{ER}(\omega)$,

$$Z_{ER}(\omega) = \frac{S_{diff}}{S_{true}}. \tag{17}$$

We compared the values of Z_{ER} of the proposed method and the well-known Lebart method at each reverberation time. The speech sources were selected from the familiarity-controlled word-lists (FW03) [21], which contains words spoken by 2 male and 2 female Japanese speakers. In this experiment, 100 words were used for each speaker, and the total number of words was 400. After calculating $Z_{ER}(\omega)$ for each sound, histograms were approximated with the kernel distribution using MATLAB, and the expected value $E(\omega)$ was then calculated for each frequency band. The number of bins on the histograms was 100, and the parameters are defined in Table 1.

3.2. Results

Figure 3 represents one of the estimated powers when the reverberation time T_r was 1.5 s. This figure shows each power of the input signal, reverberation signals estimated by the proposed method and Lebart method, and the true reverberation signal. The power of the reverberation signal estimated by the Lebart method is larger than that of the reverberation signal estimated by our proposed method, especially at the beginning part. It is also larger than the power of the true reverberation signal.

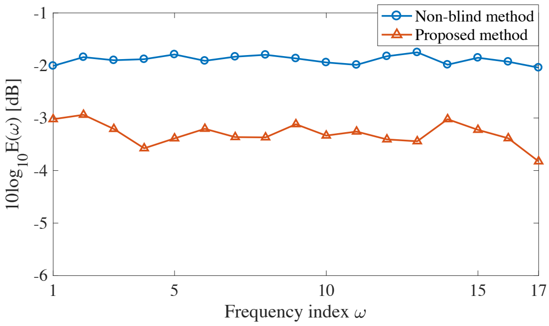


Figure 4: *Experimental results obtained when the reverberation time T_r was 1.0 s. The vertical axis represents $E(\omega)$ of the logarithmic scale and the horizontal axis represents the frequency index ω in FWF.*

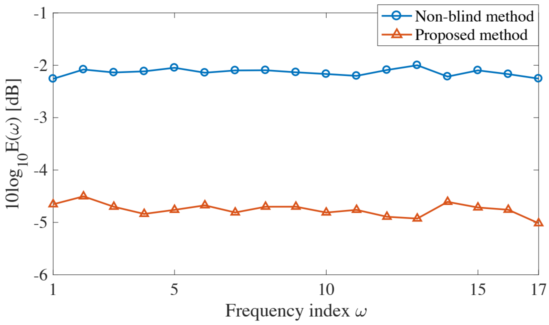


Figure 5: *Experimental results obtained when the reverberation time T_r was 1.5 s. The vertical axis represents $E(\omega)$ of the logarithmic scale and the horizontal axis represents the frequency index ω in FWF.*

Figure 4, 5 and 6 show the error ratio $E(\omega)$ of the logarithmic scale at each reverberation time, 1.0 s, 1.5 s, and 2.0 s, respectively. In each figure, the blue line shows $E(\omega)$ in the non-blind method, and the orange line shows $E(\omega)$ in the proposed method. For the proposed method, $E(\omega)$ exceeded that of the non-blind method proposed by Lebart at each reverberation time and in each frequency band. In the case of long reverberation condition, the value of $E(\omega)$ for the proposed method was around 3 dB lower than that of the non-blind method proposed by Lebart. This means that in the proposed method, the value of $E(\omega)$ was about one half of Lebart's method. It was also shown that by using our proposed method, the reverberation power is more correctly estimated than the other method. Because the value of $E(\omega)$ in the proposed method decreases as the reverberation time increases, the proposed method may be more effective as the reverberation time increases.

4. Conclusions

We evaluate the implementation of our newly proposed blind reverberation power-estimation method which was proposed for the purpose of implementing that into hearing-aid devices. The estimated reverberation power of the proposed method was compared with that of well-known non-blind methods. We performed experiments for three kinds of reverberation times. The results obtained suggested that the proposed method is more effective than the well-known non-blind method, regardless of the reverberation time. In the future, the speech signal processed by this proposed method will be evaluated using objective/subjective experiments. Furthermore, reverberation power-estimation methods for impulse responses including early reflection sounds will be studied.

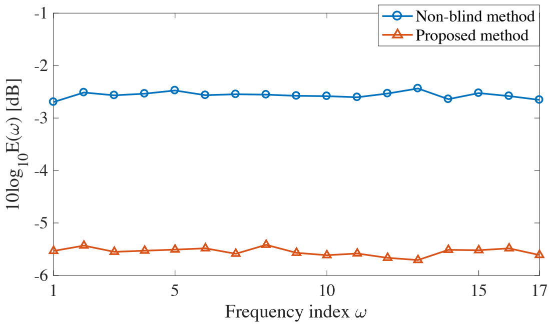


Figure 6: *Experimental results obtained when the reverberation time T_r was 2.0 s. The vertical axis represents $E(\omega)$ of the logarithmic scale and the horizontal axis represents the frequency index ω in FWF.*

5. References

[1] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech, Lang., and Hear. Res.*, vol. 17, no. 4, pp. 724–739, 1974. [Online]. Available: + http://dx.doi.org/10.1044/jshr.1704.724

[2] B. C. J. Moore, *Cochlear hearing loss.* John Wiley and Sons, 2007.

[3] L. V. S. S. Folkeard, P, "Using a de-reverberation program to improve speech intelligibility and reduce perceived listening effort," *Hearing Review*, vol. 24, no. 4, pp. 32–33, 2017.

[4] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, 1979.

[5] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1074–1090, Jan 2003. [Online]. Available: http://dx.doi.org/10.1155/S1110865703305049

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[7] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition," in *INTERSPEECH*, 2005.

[8] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, p. 437807, 2009.

[9] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

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

[11] K. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 5, pp. 1579–1591, 2007.

[12] L. Wang, K. Odani, and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, p. 12, 2012.

[13] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[14] M. Sunohara, M. Nakaichi, and Y. Kondou, "Simple dereverberation method for hearing aid users (in Japanese)," *Meet. Acoust. Soc. Jpn.*, pp. 853–854, 2015.

[15] K. Nozaki, Y. Ikeda, Y. Oikawa, Y. Fujisaka, and M. Sunohara, "Low latency dereverberation with considering speech naturalness for hearing aids," *Joint Meet. Acoust. Soc. Am. Acoust. Soc. Jpn.*, 2016.

[16] A. Harma, M. Karjalainen, L. Savioja, V. Valimaki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *J. Audio. Eng. Soc.*, vol. 48, no. 11, pp. 1011–1031, 2000.

[17] J. M. Kates and K. H. Arehart, "Multichannel dynamic-range compression using digital frequency warping," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 3003–3014, Jan 2005. [Online]. Available: http://dx.doi.org/10.1155/ASP.2005.3003

[18] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*, ser. Prentice-Hall Signal Processing Series: Advanced monographs. Prentice-Hall, 1983. [Online]. Available: https://books.google.co.jp/books?id=X_NSAAAAMAAJ

[19] S. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.

[20] J. D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, 1988. [Online]. Available: http://www.theses.fr/1988LEMA1011

[21] S. Amano, K. Kondo, Y. Suzuki, and S. Sakamoto, "Speech data set for word intelligibility test based on word familiarity (FW03)," *NII Speech Resources Consortium*, 2006.

Predicting the benefit of binaural noise reduction algorithms with FADE

Marc Rene Schädler, David Hülsmeier, Mareike Buhl, Birger Kollmeier and Anna Warzybok

Universität Oldenburg and Cluster of Excellence Hearing4all

Abstract

The simulation framework for auditory discrimination experiments (FADE) was adopted and validated to predict the individual aided speech-in-noise recognition performance of listeners with normal and impaired hearing with and without a given noise suppression strategy. FADE uses a simple automatic speech recognizer (ASR) to estimate the lowest achievable speech reception thresholds (SRTs) from simulated speech recognition experiments in an objective way, independent from any empirical reference data. Empirical data from the literature was used to evaluate the model in terms of predicted SRTs and benefits in SRT with the German matrix sentence recognition test when using eight single and multi-channel binaural hearing aid pre-processing algorithms. To allow individual predictions of SRTs in binaural conditions, the model was extended by implementing a simple "better ear" approach and individualized by taking into account the audiograms. In a realistic binaural cafeteria condition, FADE explained about 90% of the variance of the empirical SRTs for a group of normal-hearing listeners and predicted the corresponding benefits with a root-mean-square prediction error of 0.6 dB. This high prediction accuracy highlights the potential of the current approach for the objective assessment of benefits in SRT without any a-priori knowledge about the empirical data. The predictions for the group of hearing-impaired listeners explained 75% of the variance of the empirical SRTs, while the individual predictions explained less than 25%. This indicates that additional individual factors, such as, e.g., a supra-threshold processing deficiency, should be taken into account for improving the accuracy of individual predictions with impaired hearing. A competing talker condition clearly showed one limitation of current ASR technology, as the empirical performance with SRTs lower less than -20 dB could not be predicted.

Index Terms: speech recognition; aided hearing; objective evaluation; model

An iOS-based speech audiometry for self-assessment of hearing status

Hongying Yang, Tao Song, Mingjie Song, Xihong Wu, Jing Chen

Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of
Machine Perception (Ministry of Education), Peking University, Beijing, China

yanghy@pku.edu.cn, chenj@cis.pku.edu.cn

Abstract

Speech audiometry is one of the methods for evaluating hearing status, and recently it’s tended to be implemented on mobile phones with the development of internet and smart devices. Although several applications include the function of speech audiometry, the effectiveness of this function was rarely reported.

In this work, we developed an Apple iOS-based application (Hearing Assistant) with functions of pure-tone audiometry and speech audiometry for people speaking Mandarin Chinese. The speech material was designed to reflect listeners’ audibility at specified frequencies. The reliability of the speech audiometry was analyzed and discussed according to the users’ data collected through this application.

The results showed a significant but week correlation between pure-tone thresholds and scores of speech audiometry, indicating the speech audiometry can be used to discriminate impaired and normal hearing but it’s not accurate enough to evaluate hearing thresholds quantitatively at each test frequency.

Index Terms: speech intelligibility, hearing impairment, mobile devices, iOS-based application

1. Introduction

Recently, there is increasing interest in developing a suitable method for self-assessment of hearing situation and conducting hearing compensation on smartphones. Some studies have concentrated on implementing pure tone audiometry (PTA) on smartphones, which lead to some apps (uHear, EarTrumpet, etc). However, PTA has critical requirement on test environment and stimuli calibration. Factors influencing PTA test on smartphones have been studied in previous works, including the test program [1], the test accuracy [2], the time cost [3], the reference sound level [4], the user types [5, 6], and the usage of specific extra-devices [7].

Some applications are available for speech audiometry with different test materials and procedures, for instance, uHear uses Acceptable Noise Level Speech Recognition Test, Siemens Hearing Test uses nonsense word recognition test in noise. All of these applications can show speech intelligibility and suggest brief rating of hearing impairment, but unlike PTA test, they can’t supply any prescription for frequency-dependent hearing compensation. The present work aimed to develop an iOS-based speech audiometry for predicting the degree of hearing impairment at the specified frequencies.

The method of supra-threshold word identification was adopted in the speech audiometry, as this test is relatively robust for the variation of sound level. It has been reported that the speech perception performance remains constant for both normal-hearing and hearing-impaired people when sound level is in a reasonable range [8], and the speech recognition score decreases when stimuli level exceeds comfortable level [9]. A paradigm of identifying the spoken word in a confused word pair was used in the audiometry, in which the confused word pair was shown as text on the mobile screen. To select appropriated words as speech materials and produce confused word pairs, acoustic features of Mandarin initials and finals were studied.

The distinctive features of confused Mandarin initials and finals has been systematically studied in [10]. Based on this work, we produced monosyllabic word pairs to assess frequency-dependent hearing impairment by acoustic simulation with normal hearing listeners in a previous study [11]. The monosyllabic words were in consonant-vowel (CV) structure. For each pair, the syllables shared the same consonant or vowels, and kept the other phoneme different but easy to confuse, e.g. “/ba4/ vs./pa4/” or “/tao2/ vs. /tou2/” (the number represents the tonal pattern, and they were the same within a pair). Hearing impairment was simulated by processing stimuli with a serial of band-stop filters at different cut-off frequencies. Normal-hearing listeners took part in the experiment and their task was to select the correct syllable within each pair after one syllable was displayed. The results showed a significant correlation between frequency ranges of the band-stop filters (where hearing impairment occurred) and the frequency range of the distinctive feature for each pair, indicating the feasibility to predict the frequency-dependent hearing impairment by using speech audiometry with confused words pairs. However, these results were based on hearing-loss simulation in laboratory, and the frequency range was mainly below 2000 Hz.

In the present study, test material was determined as disyllabic words, as they’re more intelligible than mono-syllabic for Mandarin Chinese. The distinctive feature of the word pairs was determined as the frequency and intensity of formants, since speech formants are the most important acoustic feature for identifying a phoneme. Additionally, the general syllable structure of Chinese words, consonant-vowel, makes this acoustic feature quite critical for word identification. The first work for analyzing the strength of speech formant along frequencies among phonemes was finished by Fant and his colleagues [12] based on Swedish language. Whereafter, the conception of “Speech Banana” was developed, which showed the mean frequency of the first three formants and its mean intensity of a phoneme by a scatterplot in audiogram, and it has been available for English and Mandarin Chinese [13]. Generally, Speech Banana for Mandarin Chinese showed that the intensities of consonants are lower than vowels and the frequency of consonants (the frequency with local maximum energy) are higher than the frequency of vowels (the mean frequency of first three formant). It’s also reported that the perception of consonants is harder than the perception of vowels, and it is easily affected by background noise or hearing impairment [14]. These studies suggest that the word

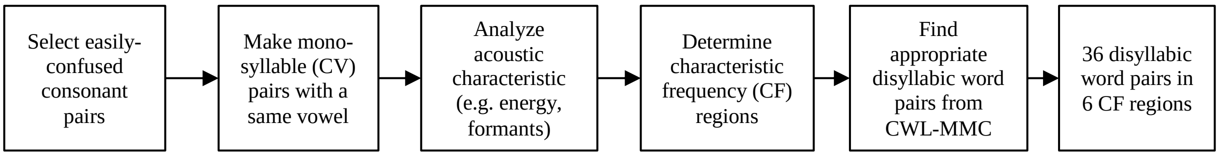

Select easily-confused consonant pairs
Make mono-syllable (CV) pairs with a same vowel
Analyze acoustic characteristic (e.g. energy, formants)
Determine characteristic frequency (CF) regions
Find appropriate disyllabic word pairs from CWL-MMC
36 disyllabic word pairs in 6 CF regions

Figure 1: *The illustration shows the procedure for word pairs stimuli selection.*

identification within a confused pair would be dominant by the consonant identification, rather than by the vowel.

Therefore, for producing the speech materials, we firstly selected mono-syllable pairs that have CV structure and the same vowel. For each pair, consonants which are easily confused were selected, due to their energy distribution showing big difference for one specific frequency (or a narrow frequency range). In other words, the two confused consonants were adjacent along frequency but separate along hearing level in Speech Banana. These specific frequency or frequency region was defined as the characteristic frequency (CF) of the CV syllable pair. And then disyllabic word pairs consist of these CV syllable pairs were produced based on the following rules: 1) for each pair of disyllable words, one pair of confused CV syllable was assigned to the words, and the other syllable of the words were kept the same, e.g. “Ji2Zhong1 vs. Qi2Zhong1” (“concentrate vs. among” in English); 2) all words were common words in everyday life; 3) these words were assumed to be spondaic as two syllables of spondee were equally important and easy to identify. Word-pairs selected in this way could ensure that the only distinctive feature was concentrated at one certain CF. And then the wrong identification of the confused words can reflect the elevation of hearing threshold at the CF. Figure 1 shows the diagram of the whole procedure for the word pairs selection.

An iOS-based application (Hearing Assistant) was developed and released in China Apple Store, including functions of PTA and speech audiometry test. The effectiveness of the PTA test of this application has been testified and reported in [15]. To evaluate the validity of the speech audiometry test, the results of the speech audiometry were compared with that of PTA test among three user groups: normal-hearing (NH) group, hearing-impaired (HI) group and the all-user (AU) group, according to the data collected from 235 application users.

2. Method

2.1. Material and stimuli

Based on the speech Banana of Mandarin Chinese, the consonants were divided into six groups according to their locations along frequency axis [13], corresponding to six CF regions, around 250, 500, 1k, 2k, 4k, and 6k Hz. These CFs were determined to be matched with the test frequencies of PTA. Then, thirty-six disyllabic word pairs were selected from Common Word List of Modern Mandarin Chinese (CWL-MMC) [16], to produce 6 word pairs for each of the six CFs. These words are often common-used in daily conversation and their occurrence-frequencies were also balanced to ensure they could be equally identified when they were both audible. For each disyllabic word pair, their only difference is the consonant of the confused syllable, so the CF of syllables can also represent the CF the word pairs. Please notice, although there is one syllable the same between the two words within a pair, the corresponding characters of the two words can be different, because a syllable in Chinese Mandarin may have multiple meanings, which are dependent on the corresponding Chinese characters. These words were selected from the CWL-MMC by programming based on the rules above, and the output were checked manually.

The 72 words (36 pairs) were spoken by a young female native Chinese speaker, and recorded digitally onto a computer disk at the sampling rate of 44.1 kHz with 24-bit quantization. The amplitude of each stimulus was normalized in terms of root-mean-square (rms) pressure.

2.2. Implementation of iOS-based speech test

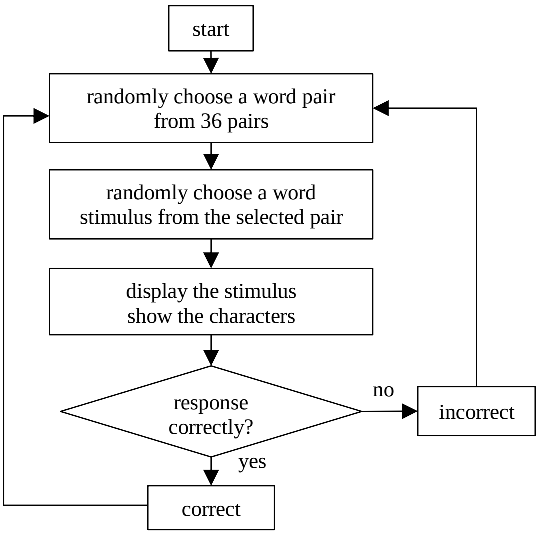

start
randomly choose a word pair from 36 pairs
randomly choose a word stimulus from the selected pair
display the stimulus show the characters
response correctly?
no
incorrect
yes
correct

Figure 2: *The illustration shows the procedure of speech audiometry test in the application.*

The procedure of iOS-based speech test was as follows. At first, the listener pressed a “start” button to initial the procedure. Then the App began to play a random-chosen word stimulus through the earphone, and at the same time, the Chinese character of the word pair were shown on the screen. The listener was instructed to select the characters of the word he/she heard on the screen by pushing a button. The correct/incorrect response was accounted. Afterwards, another word pair was tested till the 36 pairs were all finished. Figure 2 shows diagram of this procedure. A progressing bar indicated the percent that had been finished. Usually, the whole test could be finished in ten minutes. After the test, the speech recognition scores were mapped to hearing level at each CF in the audiogram, where a higher score corresponded to a lower hearing level and a lower score correspond to a higher hearing level. A suggestion of the hearing situation was also showed with text.

Before this procedure, all users had to finish the PTA test, otherwise the function of speech audiometry could not be operated. To ensure the speech stimuli were displayed at a comfortable level, the sound level was adjusted at 30 dB higher than the mean hearing threshold for each user.

2.3. Implementation of iOS-based PTA

The procedure of the iOS-based PTA test was as described in the previous work [15]. Listeners were instructed to push a "begin" button to initiate the test, and push the button "I heard" if a pure tone was heard. The ascending method introduced in the standard of ISO 8253-1 [17] was implemented.

There was a preliminary calibration before the App was released on the App Store, and the consistency across different Apple devices has been testified by previous studies [15, 18]. These works confirmed that a relatively reliable PTA results could be collected from application users. Users can download the App from App Store and use the earphones to finish the tests with instruction, and there is no need for self-calibration.

2.4. Data collection and the users

The data used for analysis were from the 300 users who registered at the first five months. To ensure the reliability of the results, the data was prescreened by deleting some invalid data, which were collected from listeners who did not finish the whole test or did not go through the test as instructed, according to the following rules:

1) If the results of a PTA test were 90 dB HL for all frequencies, which means the button was not pressed at all and the PTA test procedure was not finished correctly, the data of this test was thought as invalid and deleted.
2) If the results of a speech test are 0 for all frequency range, which means the button was not pressed at all and the speech test procedure was not finished correctly, the data of this test was thought as invalid and deleted.

According to the criterions above, the data of 65 users were deleted. The data of 235 users were remained and divided into two groups by calculating the mean hearing threshold averaged across 0.5, 1, 2, 4 kHz [19]. Please notice that some invalid data might be remained even following the rules above, since these tests were unsupervised and it was difficult to ensure the user finish the task in a right way. The listeners whose mean hearing threshold lower than or equal to 25 dB HL fell into the normal-hearing (NH) group, and the listeners whose hearing threshold exceeded 25 dB HL fell into hearing-impaired (HI) group. Finally, there were 71 and 164 users for the NH and HL groups, respectively, and the valid data for the all 235 users was named as AU (all users).

3. Results

3.1. Data analysis

The mean speech intelligibility, that is the mean score of speech audiometry averaged across 6 CF regions, and the mean hearing threshold of PTA averaged across 0.5, 1, 2, 4 kHz for each user are shown in Figure 3. Every circle/cross represents one user's data, and circles represent the users of NH group; crosses represent the users of HI group. The distribution patterns of data are different between NH and HI groups. The data of NH group distribute mainly around 1 for speech intelligibility except some outliers, but the data distribute more widely along speech intelligibility for HI group.

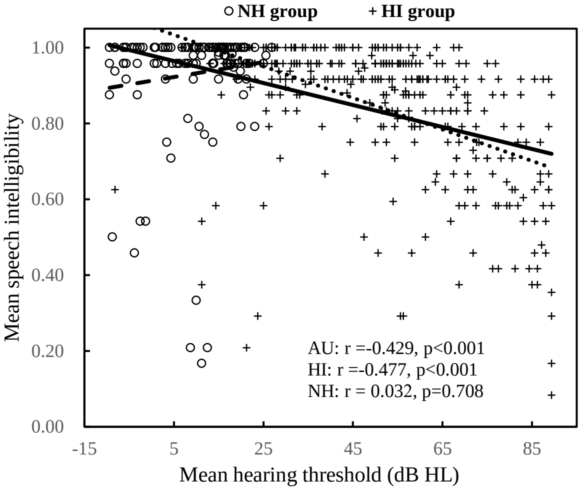


Figure 3: *The scatterplot represents mean speech intelligibility and mean hearing threshold for all users. The solid line is fitted for AU group, the dot line for HI and the dashed line for NH group.*

A correlation analysis between the mean speech intelligibility and mean hearing thresholds was conducted for the AU, NH and HI groups, respectively. The results revealed that the correlation was significant for AU group (r=-0.429, p<0.001) and HI group (r=-0.477, p<0.001). The negative r values indicated that the lower mean speech intelligibility was relevant to the higher mean hearing threshold. These results are consistent with previous studies that there is correlation between the phoneme recognition or word recognition performance and PTA, especially for hearing impaired listeners [20]. However, the correlation is not significant for NH group (r=0.032, p=0.708). Because the speech audiometry is specifically designed for HI listeners and the task is too easy for the normal-hearing, it is reasonable that their performance of speech audiometry is mainly around 1 and not significantly correlated to the hearing threshold.

When the users' performance was analyzed for each CF, the hearing thresholds at 3 kHz and 8 kHz of PTA were excluded, as the CFs for speech audiometry could only correspond to the other 6 frequencies. The correlation analysis between the speech intelligibility and the hearing thresholds for each of the six CFs was conducted, separately, and the r values for three user groups were listed in Table 1. The "whole" means that the analysis was conducted with the data for the all 6 CFs. The values which were marked with one asterisk represent the correlation is significant at the 0.05 level, and the values with two asterisks represent the significance at 0.01 level. The results showed that there was significant correlation between speech intelligibility and hearing threshold at each frequency for both HI and AU group, but not for NH group. This results were consistent with the results based on the mean speech intelligibility and mean hearing threshold. The speech intelligibility and the hearing threshold at each CF for HI group are shown in Figure 4. Every data point represents one HI user's data at one CF, and different symbols represent different frequency regions. The distribution patterns of each CF are similar. These results confirmed that the correlation between speech recognition scores and hearing thresholds was similarly in each CF region. For the significant correlations, the r values were all around -0.3, suggesting that the correlation was relatively weak at all CFs and it might be not accurate enough to evaluate audibility quantitatively at each test frequency.

Table 1: *The correlation of speech intelligibility and hearing thresholds in different CF regions for AU, NH and HI groups.*

Frequency (Hz)	**R**		
	AU	NH	HI
250	-.335**	0.092	-.329**
500	-.337**	0.178*	-.365**
1000	-.340**	-0.119	-.341**
2000	-.343**	0.013	-.385**
4000	-.282**	0.029	-.275**
6000	-.317**	0.020	-.313**
whole	-.413**	0.046	-.316**

**.Correlation is significant at the 0.01 level(2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

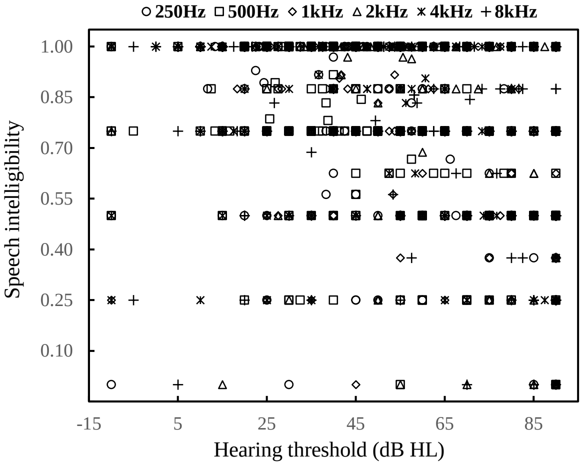


Figure 4: *The scatterplot represents speech intelligibility and hearing thresholds at each CFs.*

In summary, the significant correlation between speech intelligibility and hearing thresholds in HI group indicated that the speech audiometry with the word identification test in our App was valid in part and can help to assess individual hearing abilities on smart phones. However, although the correlation was significant, it is mild, and further work is required to improve the accuracy of the speech audiometry.

4. Discussion

The significant correlation values between speech intelligibility and hearing thresholds are all lower than 0.5, indicating the use of speech audiometry to evaluate the loss of hearing threshold could not be accurate as PTA. It could be explicated from three aspects.

Firstly, there is difficulty for user data's prescreening. Present screening criterion could only exclude the results from the tests that are not response correctly entirely, while the results from users who might response partly incorrectly could not be monitored. Besides, results from the tests that were conducted in noisy environment might also become the outliers for the whole group.

Secondly, when stimuli are disyllabic words and displayed without noise, the task of speech identification is still easy for those listeners with mild or moderate hearing impairment. For example, the proportion of the users who get speech intelligibility bigger than 0.6 is more than 90% of all the listeners. Hence the differences for individuals whose hearing impairment levels were mild or moderate were not reflected in the current results, and resulted into a relatively small correlation. In future work it is necessary to design a task whose difficulty could be adaptively adjusted according to listeners' proceeding performance.

Thirdly, the sound level for displaying speech stimuli was adjusted based on the user's individual hearing threshold. This manipulation was to ensure the listener can hear sound at a comfortable level individually, and to avoid the floor effect of speech intelligibility due to the very weak sound for HI users. However, this manipulation also brought in a sort of compensation for hearing impairment [21]. As a consequence, the performance of HI users was underestimated, and the correlation between the speech intelligibility and the hearing threshold was reduced. Moreover, for some subjects with mean hearing thresholds exceeding 40 or 50 dB HL, the stimuli level might induce a new problem for the speech perception performance, as it was reported that speech recognition scores could be reduced significantly with increasing sound level for some hearing-impaired subjects, due to loudness recruitment [22, 23]. It was also reported that the uncomfortable sound level could be high for some HI listeners [24].This individual difference between HI listeners was not taken into account in the present work.

The word pairs used in present work were selected based on Speech Banana of Mandarin Chinese, in which the location of a consonant along frequency was determined by the frequency at which the maximum energy is in the spectrogram. However, the energy of some consonants spread in a wide frequency range, and some other acoustic cue, e.g. duration, could be important for the word identification. This issue need more detailed and accurate analysis in the future work, to upgrade the present speech audiometry.

Listeners' performance for speech identification depends on several factors, including hearing threshold, frequency resolution and time resolution of auditory processing, and even high-level cognitive processing. It is important to discriminate the effects of these factors when speech audiometry is used to evaluate hearing impairment. Future work is required to use more suitable speech stimuli and more efficient test procedure for applying speech audiometry on smartphones.

5. Conclusions

A new iOS-based speech audiometry was produced, and the hearing impairment at 6 frequency regions (250-6000 Hz) could be evaluated by the identification of 36 confused disyllabic-word-pairs for people speaking Mandarin Chinese. The validity of this speech audiometry was analyzed by comparing with PTA based on 235 users. The results showed that the speech intelligibility has significant correlation with hearing threshold for hearing-impaired listeners but no remarkable correlation for normal-hearing listeners for each frequency region.

6. Acknowledgements

The work was supported by the National Natural Science Foundation of China (No. 61473008 and No. 11590773), and a Newton alumni funding by the Royal Society, UK.

7. References

[1] D. J. Van Tasell and P. Folkeard, "Reliability and accuracy of a method of adjustment for self-measurement of auditory thresholds," *Otology & neurotology*, vol. 34, no. 1, pp. 9, 2013.

[2] R. Al-Abri, M. Al-Balushi, A. Kolethekkat, D. Bhargava, A. Al-Alwi, H. Al-Bahlani, and M. Al-Garadi, "The Accuracy of IOS Device-based uHear as a Screening Tool for Hearing Loss: A Preliminary Study From the Middle East," *Oman Medical Journal*, vol. 31, no. 2, pp. 142, 2016.

[3] R. Levent, S.O. Tarik, E. Hulya, O. Ustun, and Y. M. Deniz, "Smartphone Based Audiometric Test for Confirming the Level of Hearing: Is It Useable in Underserved Areas?," *Journal of International Advanced Otology*, vol. 12, no. 1, pp. 61-66, 2016.

[4] M. Masalski, L. Kipiński, T. Grysiński, and T. Kręcicki, "Hearing Tests on Mobile Devices: Evaluation of the Reference Sound Level by Means of Biological Calibration," *Journal of Medical Internet Research*, vol. 18, no. 5, pp. e130, 2016.

[5] J. C. Yeung, S. Heley, Y. Beauregard, S. Champagne, and M. A. Bromwich, "Self-administered hearing loss screening using an interactive, tablet play audiometer with ear bud headphones," *International Journal of Pediatric Otorhinolaryngology*, vol. 79, no. 8, pp. 1248, 2015.

[6] S. Abughanem, O. Handzel, L. Ness, M. Benartziblima, K. Faitghelbendorf, and M. Himmelfarb, "Smartphone-based audiometric test for screening hearing loss in the elderly," *European Archives of Oto-Rhino-Laryngology*, vol. 273, no. 2, pp. 333-339, 2016.

[7] M. Debevc, I. Kožuh, and H. Meier, "A Usability Requirements Analysis for Wireless Interaction and Connectivity for Elderly Hearing Aid Users," in *Human Factors in Computing and Informatics: First International Conference, SouthCHI 2013,*: Maribor, Slovenia, 2013 pp. 260-271.

[8] G. Walker and D. Byrne, "Reliability of speech intelligibility estimation for measuring speech reception thresholds in quiet and in noise," *Australian Journal of Audiology*, vol. 7, no.1, pp. 23-31, 1985.

[9] G.A. Studebaker, R.L. Sherbecoe, D.M. Mcdaniel, and C.A. Gwaltney, "Monosyllabic word recognition at higher-than-normal speech and noise levels," *Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2431-2444, 1999.

[10] I. Mcloughlin, "Vowel intelligibility in Chinese," *IEEE Transactions on Audio Speech & Language Processing*, vol. 18, no. 1, pp. 117-125, 2010.

[11] J. Chen, H. Y. Yang, and X. H. Wu, "Evaluation of frequency-dependent hearing loss by words identification," *Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1995-1995, 2016.

[12] G. Fant, *Acoustic analysis and synthesis of speech with applications to Swedish*. LM Ericcson, 1959.

[13] Y.C. Hung and Y.C. Ma. "Development of a Chinese Speech Banana," in *10th Asia Pacific Symposium on Cochlear Implants and Related Sciences*. *Beijing, China.*, 2015.

[14] P. M. Zurek and L. A. Delhorne, "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," *Journal of the Acoustical Society of America*, vol. 82, no. 5, pp. 1548, 1987.

[15] Y. Xing, Z. Fu, X.H. Wu, and J. Chen, "Evaluation of Apple iOS-based automated audiometry," in *22nd International Congress on Acoustics*: Buenos Aires. 2016.

[16] Y.M. Li and T.K. Wang, *Lexicon of Common Words in Contemporary Chinese*. The Commercial Press, 2008.

[17] International Organization for Standardization, "ISO 8253-1. Acoustics-Audiometric test methods-Part 1: Basic pure tone air and bone conduction threshold audiometry," Geneva. 1989.

[18] A.V. Hasselt, "Clinical evaluation of a computerized self-administered hearing test," *International Journal of Audiology*, vol. 51, no. 8, pp. 606-610, 2012.

[19] J. G. Clark, "Uses and abuses of hearing loss classification," *Asha*, vol. 23, no. 7, pp. 493-500, 1981.

[20] D. R. Moore, M. Edmondson-Jones, P. Dawes, H. Fortnum, A. Mccormack, R. H. Pierzycki, and K. J. Munro, "Relation between speech-in-noise threshold, hearing loss and cognition from 40-69 years of age," *PLoS ONE*, vol. 9, no. 9, pp. e107720-e107720, 2013.

[21] B. C. J. Moore, J. Marriage, J. Alcántara, and B.R. Glasberg, "Comparison of two adaptive procedures for fitting a multi-channel compression hearing aid," *International Journal of Audiology*, vol. 44, no. 6, pp. 345-357, 2005.

[22] D.D. Dirks, C.A. Kamm, J.R. Dubno, and T.M. Velde, "Speech recognition performance at loudness discomfort level," *Scandinavian Audiology*, vol. 10, no. 4, pp. 239-46, 1981.

[23] J. Jerger and S. Jerger, "Diagnostic significance of PB word functions," *Archives of Otolaryngology*, vol. 93, no. 6, pp. 573-580, 1971.

[24] H. Fletcher, "The Perception of Speech and Its Relation to Telephony," *Science*, vol. 108, no. 2816, pp. 682, 1950.

Important role of temporal cues in speaker identification for simulated cochlear implants

Zhi Zhu[1], *Ryota Miyauchi*[1], *Yukiko Araki*[2], *Masashi Unoki*[1]

[1]Japan Advanced Institute of Science and Technology, Japan
[2]Kanazawa University, Japan
[1]{zhuzhi,ryota,unoki}@jaist.ac.jp, [2]yukikoa@staff.kanazawa-u.ac.jp

Abstract

Speaker identification is still challenging issue for cochlear implant (CI) users due to the poor spectral cue provided by the CI device. To optimize CI systems for the users, it is important to understand the role of temporal modulation cues in speaker identification, as the CI device provides temporal modulation cues as primary cues. This study investigates the relative contributions of spectral and temporal cue on speaker identification by using noise-vocoded speech (NVS) as a CI simulation. In the experiment, speaker identification was conducted in normal-hearing listeners as a function of the number of channels (4, 8, and 16) and upper limitation of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) in NVS. The number of channels and upper limitation of envelope frequency present the spectral and temporal resolution of NVS separately. Results showed that the performance of speaker identification was not affected by spectral resolution significantly, at least in the limited set of stimuli in the present study. In addition, the results also showed that the performance was more sensitive to temporal resolution. It is suggested that temporal modulation cues contribute to speaker identification and have the potential to improve speaker identification if enhanced.

Index Terms: temporal modulation cue, speaker identification, noise-vocoded speech, cochlear implant

1. Introduction

The temporal envelope of speech has been proved to be an important cue in perceiving linguistic information included in the speech. Shannon et al. showed that the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for listeners to the recognize of linguistic information [1]. The modulation frequency bands from 4 to 16 Hz have been shown to be important regions in speech recognition [2]. Also, cochlear-implant (CI) users can achieved good performance in speech recognition, as the CI device can provide sufficient temporal cues. However, human speech includes not only linguistic information, but also nonlinguistic information such as speaker individuality. CI users cannot accurately identify speakers as the CI device provides poor spectral cues [3].

To optimize CI systems for their users, the role of temporal modulation cues in speaker identification must be understood. It is necessary to know which aspects of the temporal modulation cues have the potential to improve speaker identification if enhanced. Luo and Fu successfully enhanced the tone recognition on the NVS scheme by manipulating the amplitude envelope to more closely resemble the F0 contour [4]. Their results showed the possibility of enhancing the recognition of nonlinguistic information by modifying the temporal envelope.

Traditional research about speaker identification by humans has focused on spectral cues based on speech production. The formant frequencies have been found to carry not only information about vowels but also information regarding speaker individuality [5]. Kitamura et al. indicated that speaker individuality exists mainly in the frequency bands higher than 2212 Hz of the speech spectral envelope [6]. The fundamental frequency contours are also shown to be important cues in speaker identification [7]. Generally, the speaker individualities related to fundamental frequency and spectral envelope can be thought of as results of the individual differences of vocal organs. Unfortunately, current CI devices cannot encode the spectral and fundamental frequency information of speech sufficiently for speaker identification.

As CI listeners using the temporal envelope of speech as a primary cue, Vongphoe and Zeng evaluated whether temporal cues are sufficient to support both speech recognition and speaker identification [3]. Their results showed a disassociation between speech and speaker recognition using primarily temporal cues: CI users performed well at vowel recognition but poorly at speaker recognition. On the other hand, Gonzalez and Oliver investigated speaker identification as a function of the number of channels in both noise and sin-wave vocoded speech as CI simulations [8]. The performance of speaker identification was shown to be poorer with fewer number of channels of noise-vocoded speech (NVS). However, Krull et al. showed that training resulted in improved identification of speakers in CI simulations [9]. Moreover, child CI users succeeded in differentiating their mothers' utterances from those of other people [10]. CI users's differentiation of speakers was facilitated by long-term familiarity. It is suggested that the temporal modulation information has possibility to be an effective cue for CI users to distinguish speakers.

In a previous study, the relative contributions of spectral and temporal cues in vocal emotion recognition for NVS is clarified by varying the the number of channels and upper limitation of envelope frequency systematically [11]. As the result, the temporal resolution of NVS affected the vocal emotion recognition significantly. Moreover, we examined word and speaker recognition using NVS while systematically varying the upper limit of the modulation frequency [12]. The results suggested that the temporal resolution of NVS should contribute to the speaker recognition. However, the role of temporal cues in speaker identification is still unknown.

This paper aims to clarify the role of temporal cues in speaker identification with NVS as a CI simulation. In the experiment, speaker identification was conducted by normal-hearing listeners as a function of the number of channels (4, 8, and 16) and upper limitation of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) in NVS. The number of channels and upper limitation of envelope frequency present the spectral and temporal resolutions of NVS separately. The experimental

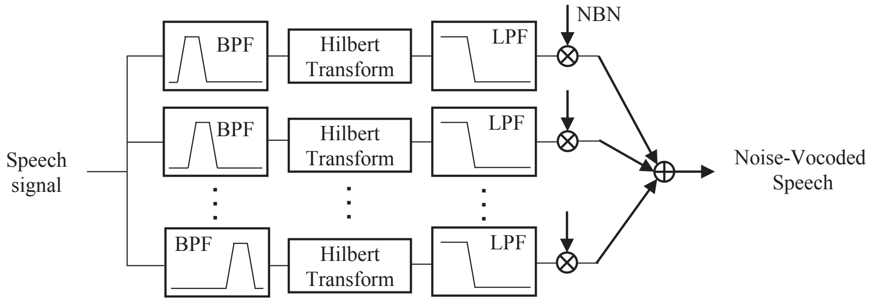


Figure 1: *Signal processing method for noise-vocoded speech. BPF: bandpass filter; LPF: low-pass filter; NBN: narrow-band noise*

paradigm used in this study can clarify the important modulation frequency band for speaker identification. The potential to improve speaker identification by enhance the temporal modulation cues is then discussed.

2. Speech data and signal processing

2.1. Speech data

The speech data used in this study were selected from the ATR Japanese speech database set C and recorded at a 20 kHz sampling frequency. Each sentence was uttered for about 4 to 5 seconds.

In this study, the XAB method was used in the speaker identification experiment. In the XAB method, one trial consists of three different speech signals (X, A, and B). The speakers of A and B are different, and the speaker of X is also the speaker of either A or B. Participants are asked to select which speaker, A or B, is more similar to the speaker of X. It is assumed that the similarity of a speaker pair will affect the results of experiment. The speaker pair with high similarity may be difficult to be distinguish, even when the spectral and temporal cues were preserved. On the contrary, the speaker pair with low similarity may be still easy to be distinguish, even if the cues related to speaker identification were reduced. This kind of bias is not desirable.

Kitamura et al. measured the perceptual similarity of speaker individualities of 20 female and 20 male Japanese speakers in ATR speech database set C [13]. Two same sentences with different speakers were presented to normal-hearing listeners, and the listeners were asked to select the similarity of these two speakers from 1 to 5. The perceptual similarity of speakers is considerable to generate some undesirable bias in the XAB test. Therefore, in order to remove the impact of similarity, the speaker pairs of speech data used in this study have perceptual similarity closest to the average value of perceptual similarity (female: 1.87, and male: 1.99) measured by Kitamura et al. [13]. The 5 female and 5 male speaker pairs used in this study and their perceptual similarities are shown in Table 1. All 20 speakers are different and the speakers of each pair have the same gender. 6 sentences of each speaker were used to generate the NVS stimuli.

2.2. Signal processing

Figure 1 schematically illustrates a schematic diagram of the signal processing to generate NVS. First, to reduce the effect of the average intensity, the active speech levels of all speech signals were normalized to -26 dBov by using the P.56 speech voltmeter [14]. Speech signal was first divided into several frequency bands with a band-pass filterbank. The bandwidth and boundary frequencies of the band-pass filters (6th-order Butterworth Infinite Impulse Response (IIR) filter) were defined using $\mathrm{ERB_N}$ (Equivalent Rectangular Bandwidth) and $\mathrm{ERB_N}$-number scale [15]. The $\mathrm{ERB_N}$-number scale is comparable to a scale of distance along the basilar membrane so that the frequency resolution of the auditory system can be faithfully replicated by dividing frequency bands in accordance with the $\mathrm{ERB_N}$-number. The relationship between $\mathrm{ERB_N}$-number and acoustic frequency is defined as follows:

Table 1: *Speaker pairs selected from ATR database and their average similarity index measured by Kitamura et al. [13]. Left and right halves show female and male speaker pairs, respectively.*

Speaker pair		Similarity	Speaker pair		Similarity
F407	F306	1.87	M509	M318	1.99
F611	F418	1.86	M603	M409	1.98
F606	F605	1.875	M508	M113	2.00
F720	F213	1.88	M519	M211	2.01
F709	F614	1.83	M520	M517	1.97

$$\mathrm{ERB_N} - \mathrm{number} = 21.4\log_{10}\left(\frac{4.37f}{1000}+1\right) \quad (1)$$

where f is acoustic frequency in Hz. The boundary frequencies of the band-pass filters were defined from 3 to 35 $\mathrm{ERB_N}$-number with bandwidth as 2, 4, or 8 $\mathrm{ERB_N}$. Therefore, the numbers of channels of the band-pass filterbank were 16, 8, or 4. The number of channels presents the frequency resolution of NVS: higher frequency resolution is obtained with more number of channels.

Then, the temporal envelope of the output signal from each band-pass filter was extracted by using a Hilbert transformation and performing a low-pass filter (2nd-order Butterworth IIR filter). The cut-off frequency of the low-pass filter determined the upper limit of envelope frequency that presents the temporal resolution of NVS. To investigate the role of temporal envelope cues for speaker identification, the conditions of the cut-off frequencies of the low-pass filter were 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz. Moreover, there was an additional "0" Hz condition where only the direct current component of the Hilbert envelope was extracted.

Finally, the temporal envelope in each channel served to amplitude modulation with the band-limited noise which was generated by band-pass filtering white noise at the same boundary frequency. All amplitude-modulated band-limited noises were summed to generate the NVS stimulus. The NVS was widely used as a CI simulation, as the spectral cues of speech were reduced.

3. Experimental procedure

Nine native Japanese speakers (two female and seven male) participated in this experiment. All participants had normal hearing (hearing losses of the participants were below 12 dB in the frequency range from 125 to 8000 Hz).

This experiment was carried out by using the XAB method. One trial consisted of three different speech signals (X, A, and B). The contents of stimuli X, A, and B were as follows:

- X: Noise-vocoded speech
- A: Noise-vocoded speech with the same speaker as X
- B: Noise-vocoded speech with a different speaker from X.

Participants were asked to compare the speakers of A and B with the speaker of X to select which speaker was more similar to the speaker of first speech X. Both stimulus with XAB and XBA orders were presented to counterbalance any effects due to the order of presentation. All the speaker pairs of A and B are shown in the Table 1.

A total of 3 different number of channels (4, 8, and 16) and 9 upper limits of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) created 18 NVS conditions. The original speech was also presented as a control condition. The participants were allowed to listen to each stimulus only once. Before the experiment, 10 stimuli were presented to the participants to familiarize participants with the CI simulation and the experimental environment. The stimuli used in the experiment were different from that used in the practice. The number of stimuli was 560 and all stimuli were presented totally randomized.

The experiment was conducted while the participants were in a sound-proof room. The sound pressure level of background noise was lower than 25.8 dB. The stimuli were simultaneously presented to both ears of a participant through a PC, audio interface (RME, Fireface UCX), and a set of headphones (SENNHEISER HDA 200). The sound pressure levels were calibrated to be the same among participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

4. Results

Figure 2 shows the average value of speaker recognition rates, and the error bars indicate ±1 standard error of the mean. Under the original speech condition, the recognition rate was close to 95 %. Participants performed nearly perfectly in speaker identification with the original speech. The results of NVS stimuli showed that the performance of speaker identification improved as the upper limit of envelope frequency increased. The results for 4-band NVS were lower than 8 or 16-band NVS at some upper limits of envelope frequency. However, the performance was not obviously affected by the number of channels.

A repeated-measures analysis of variance (ANOVA) was conducted on the results with the number of channels and upper limit of envelope frequency as the factors. It is confirmed that there was a significant main effect of the upper limit of envelope frequency ($F(8, 64) = 23.8631, p < 0.01$). However, there was no significant main effect of the number of bands ($F(2, 16) = 3.3230, p = 0.29$) and there was also no significant interaction between the two factors ($F(16, 128) = 1.1608, p = 0.16$). These results showed that the performance of speaker identification was significantly affected by the temporal resolution, which suggest that temporal modulation cues contribute to speaker identification. The performance was less sensitive to the spectral resolution, however, at least in the limited set of stimuli in the present study.

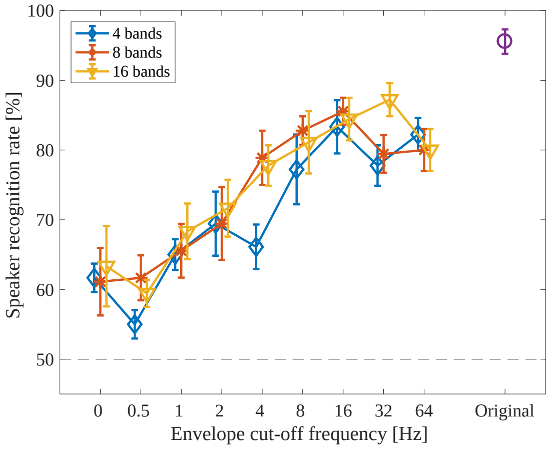


Figure 2: *Speaker recognition rates in all 27 NVS conditions and original speech condition. Error bars indicate* ±1 *standard error of the mean.*

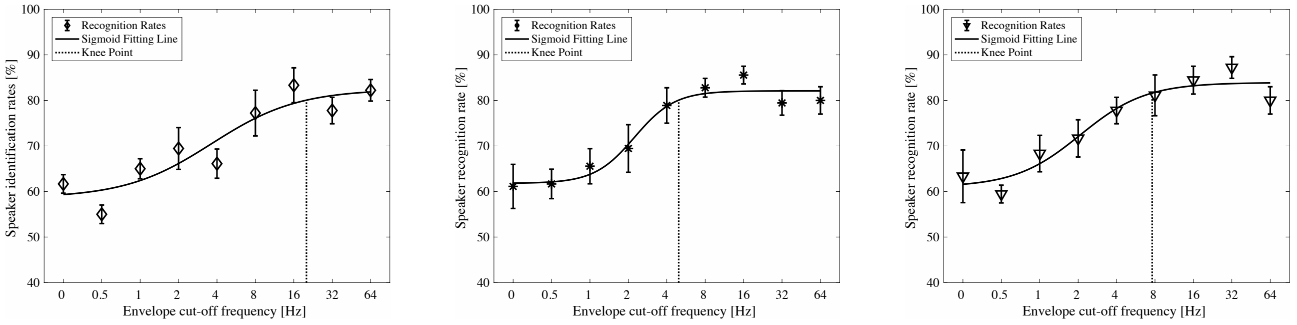
5. Discussion

5.1. Effect of spectral resolution

The speaker identification rates of 4-band NVS are lower in some conditions of the upper limit of envelope frequency. However, the number of channels did not affect the performance of speaker identification significantly. These results were different from the results of previous studies in which the performance was improved as the number of channels increased [3][8]. One difference between the present study and previous studies is that the upper limit of envelope frequencies in this study was lower. In previous studies, the cut-off frequencies of the low-pass filter were 500 Hz [3] and 160 or 400 Hz [8]. The modulation frequency bands between about 50 and 500 Hz are related to the periodicity information about fundamental frequency [16], which is not included in the stimuli used in the present study. One possible explanation may be that the temporal cues related to the periodicity information in higher modulation frequency bands are more sensitive to the number of channels. The main target of this study is to clarify the role of temporal cues in lower modulation frequency bands that include the information about variations of intensity, duration, attack, decay, and segmental cues of speech.

5.2. Effect of upper limit of envelope frequency

This study is intended to clarify the role of temporal modulation cues in speaker identification. Specifically, the important modulation frequency bands for speaker identification are investigated. To identify the important modulation frequency bands, a



(a) 4 bands

(b) 8 bands

(c) 16 bands

Figure 3: *Speaker recognition rates in each condition of number of channels and their sigmoid fitting lines.*

sigmoid function was used to fit the data of the experiment. The sigmoid function was mathematically defined as follows:

$$y = \frac{a}{1 + e^{b(x-c)}} + d \qquad (2)$$

where x is the upper limit of envelope frequency and y is the percent-correct scores. The parameters a, b, c, and d were calculated on the basis of the method of least squares. Moreover, the upper limit of envelope frequency at which 90% of the performance plateau was defined as a knee point. The results of fitting lines and knee points of each condition of the number of channels are shown in Fig. 3. The coefficients of determinations R^2 of the fitting results in 4, 8, and 16-band NVS were 0.86, 0.95, and 0.93.

The knee point of 4-band NVS was about 20.09 Hz which was higher than those of 8-band NVS (4.96 Hz) and 16-band NVS (7.60 Hz) . As the spectral cues provided by 4-band NVS was poor, participants may primarily use the temporal modulation cues to recognize the speaker rather than spectral cues. However, it still should be mentioned that there was no significant interaction between the number of channels and the upper limit of envelope frequency. More Xu and Pfingst measured both consonant and vowel recognition as a function of the number of channels (1 to 16) and upper limit of envelope frequency (1 to 512) [17]. The knee points of vowel recognition in different numbers of channels conditions are all below about 4 Hz. The knee points of consonant recognition are from 4 to 16 Hz, which are closer to the knee points for speaker identification in this study. Tachibana et al. conducted a experiment of NVS sentence recognition with various of upper limits of envelope frequency [18]. They found that an increase in the upper limit of envelope frequency from 4 to 8 Hz improved the correct response rate more that increasing the upper limit of envelope frequency from 8 to 16 Hz. Both studies showed that the duration and segmental cues included in such modulation frequency band below about 16 Hz are important in the perception of linguistic information. In this study, these duration and segmental cues of the temporal envelope are also suggested to be used in speaker identification. These segmental cues related to the rhythm, tempo, and the speaking style of the speaker which should be different with different speakers.

The results of this paper have shown that the temporal modulation cues contribute to speaker identification and that the temporal modulation information below about 20 Hz seems to be important. In the future, the modulation spectral features [19] related to speaker individuality and the effect of modifying such modulation spectral on speaker identification will be investigated. In a previous study, we confirmed that the vocal emotion of NVS can be converted by modifying the modulation spectrogram of temporal envelope [20]. Whether the speaker individuality information of NVS can be converted by modifying the modulation spectrogram should also be discussed further.

6. Summary

This study aimed to clarify the role of temporal cues in speaker identification with noise-vocoded speech (NVS) as a cochlear implant (CI) simulation. Speaker identification was conducted by normal-hearing listeners as a function of the number of channels (4, 8, and 16) and the upper limitation of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) in NVS. The result showed that speaker identification rates improved significantly as the upper limit of envelope frequency increased. However, the performance was not obviously affected by the number of channels. The modulation frequency bands below about 20 Hz were shown to be important in speaker identification with 4-band NVS. In conclusion, it is suggested that temporal modulation cues contribute to speaker identification and have the potential to improve speaker identification if enhanced. It is important to understand not only which parts but also exactly what kinds of features of temporal envelope have possibility to be important cues for speaker identification.

7. Acknowledgements

This work was supported by a Grant in Aid for Scientific Research (A) (No. 25240026), Innovative Areas (No. 16H01669) from MEXT, Japan, and the Mitsubishi Research Foundation. This work was also supported by JSPS KAKENHI Grant Number JP 17J08312.

8. References

[1] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech Recognition with Primarily Temporal Cues,” *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[2] R. Drullman, J. Festen and R. Plomp, “Effect of reducing slow temporal modulations on speech reception,” *Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, .

[3] M. Vongphoe, and F. G. Zeng, “Speaker recognition with temporal cues in acoustic and electric hearing,” *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1055–1061, 2005.

[4] X. Luo, and Q. Fu, “Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants,” *Journal of the Acoustical Society of America*, vol. 116, pp. 3659–3667, 2004.

[5] R. E. Remez, J. M. Fellowes, and P. E. Rubin, “Talker identification based on phonetic information,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, pp. 651–666, 1997.

[6] T. Kitamura, and M. Akagi, “Speaker individualities in speech spectral envelopes,” *Journal of Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 283–289, 1995.

[7] M. Akagi, and T. Ienaga, “Speaker individuality in fundamental frequency contours and its control,” *Journal of Acoustical Society of Japan (E)*, vol. 18, no. 2, pp. 73–80, 1997.

[8] J. Gonzalez, and J. Oliver, “Gender and speaker identification as a function of the number of channels in spectrally reduced speech,” *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 461–470, 2005.

[9] V. Krull, X. Luo and K. Kirk, “Talker-identification training using simulations of binaurally combined electric and acoustic hearing: generalization to speech and emotion recognition,” *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3069–3078, 2012.

[10] T. Vongpaisal, S. E. Trehub, E. G. Schellenberg, P. Lieshout, and B. C. Papsin, “Children with cochlear implants recognize their mother’s voice,” *Ear and Hearing*, vol. 31, no. 4, pp. 555–566, 2010.

[11] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, “The role of spectral and temporal cues for vocal emotion recognition by cochlear implant simulations,” *Acoustics ’17*, 2017. (in press)

[12] Z. Zhu, Y. Nishino, R. Miyauchi, and M. Unoki, “Study on linguistic information and speaker individuality contained in temporal envelope of speech,” *Acoustical Science & Technology*, vol. 37, no. 5, pp. 258–261, 2017.

[13] T. Kitamura, T. Nakama, H. Ohmura, and H. Kawamoto, “Measurement of perceptual speaker similarity for sentence speech in ATR speech database,” *Journal of Acoustical Society of Japan*, vol. 71, no. 10, pp. 516–525, 2015.

[14] Intl. Telecom. Union, “Objective measurement of active speech level,” *ITU-T,* P.56, Switzerland, 1993.

[15] B. C. J. Moore, *An introduction to the psychology of hearing*, 6th Edition, London, Elsevier, pp. 74–80, 2013.

[16] S. Rosen, “Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects,” *Philosophical Transactions: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[17] L. Xu, and P. Pfingst, “Spectral and temporal cues for speech recognition: Implications for auditory prostheses,” *Hearing Research*, vol. 242, pp. 132–140, 2008.

[18] R. Tachibana, Y. Sasaki, and H. Riquimaroux, “Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech,” *Acoustical Science and Technology*, vol. 34, no. 4, pp. 263–270, 2013.

[19] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, “Modulation spectral features for predicting vocal emotion recognition,” *INTERSPEECH2016*, pp. 262–266, 2016.

[20] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, “Feasibility of Vocal Emotion Conversion on Modulation Spectrogram for Simulated Cochlear Implants,” *EUSIPCO2017*, 2017. (in press)

Better hearing in noise with binaural prostheses inspired by the contralateral medial olivocochlear reflex

Enrique A. Lopez-Poveda

University of Salamanca, Salamanca, Spain

Abstract

In natural hearing, cochlear mechanical compression is dynamically adjusted via the medial olivocochlear efferent reflex (MOCR). These adjustments probably help understanding speech in noisy environments and are not available to the users of cochlear implants (CIs). I will present a bilateral CI sound processing strategy that reinstates the effects of the contralateral MOCR to CI users using frequency-specific, contralaterally controlled dynamic compression. The new strategy significantly facilitates understanding speech in competition with noise in bilateral and unilateral listening conditions, and enhances spatial release from masking. The strategy may be usefully applied in hearing prostheses.

Further details may found in Lopez-Poveda et al. (2016) *Ear & Hearing* 37(3):e138–e148, and Lopez-Poveda et al. (2017) *Hearing Research* 348:134–137. Work funded by MINECO (ref. BFU2015-65376-P), FEDER, and MED-EL GmbH.

Perceptual contribution of fundamental frequency contour and its implication to assistive hearing devices for Chinese-speaking hearing-impaired users

Fei Chen

Southern University of Science and Technology, China

Abstract

For Chinese speech perception, tonal information mainly carried by fundamental frequency (F0) contour plays an important role to lexical tone identification. Recent work showed that for Mandarin sentence recognition at ideal listening conditions (e.g., in quiet), the distortion of F0 contour can be compensated by other contextual cues, and hence F0 contour may be treated as relatively redundant cue for Mandarin sentence intelligibility in quiet. However, at adverse listening conditions (e.g., in noise), tone contour is important for the recognition of Mandarin sentences. The present work assessed the contribution of F0 contour to Mandarin speech perception simulating two scenarios of 1) high-frequency hearing loss and 2) understanding frequency-compressed speech.

In Experiment 1, the wideband Mandarin speech was first processed to have a flat F0 contour, and then low-passed filtered, which simulated the high-frequency hearing loss. In Experiment 2, the F0-flattened Mandarin sentences were processed by a non-linear frequency compression strategy, which compressed the spectral information up to 700 Hz to 500 Hz (Chen and Chan, JASA, 2016). The processed stimuli were presented to normal-hearing Mandarin-speaking listeners to recognize, and the intelligibility scores were compared with those of counterpart conditions with normal F0 contour.

Results showed that flattening F0 contour significantly reduced the intelligibility of low-pass filtered and frequency-compressed Mandarin speech, compared with conditions with normal F0 contour. This indicates the importance of preserving F0 contour for Mandarin speech perception in conditions of high-frequency hearing loss and designing assistive hearing techniques (e.g., frequency compression) for Chinese-speaking hearing-impaired users.

Index Terms: Chinese speech perception; fundamental frequency contour; assistive hearing devices; frequency compression

Towards Next-Generation Lip-Reading Driven Hearing-Aids: A preliminary Prototype Demo

Ahsan Adeel, Mandar Gogate, Amir Hussain

Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, UK

E-mail: {aad, mgo, ahu}@cs.stir.ac.uk

Abstract

Speech enhancement aims to enhance the perceived speech quality and intelligibility in the presence of noise. Classical speech enhancement methods are mainly based on audio only processing which often perform poorly in adverse conditions, where overwhelming noise is present. This paper presents an interactive prototype demo, as part of a disruptive cognitively-inspired multimodal hearing-aid being researched and developed at Stirling, as part of an EPSRC funded project (COG-AVHEAR). The proposed technology contextually utilizes and integrates multimodal cues such as lip-reading, facial expressions, gestures, and noisy audio, to further enhance the quality and intelligibility of the noise-filtered speech signal. However, the preliminary work presented in this paper has used only lip-reading and noisy audio. Lip-reading driven deep learning algorithms are exploited to learn noisy audio-visual to clean audio mappings, leading to enhanced Weiner filtering for more effective noise cancellation. The term context-aware signifies the device's learning and adaptable capabilities, which could be exploited in a wide-range of real-world applications, ranging from hearing-aids, listening devices, cochlear implants and telecommunications, to need for ear defenders in extreme noisy environments. Hearing-impaired users could experience more intelligible speech by contextually learning and switching between audio and visual cues. The preliminary interactive Demo employs randomly selected, real noisy speech videos from YouTube to qualitatively benchmark the performance of the proposed contextual audio-visual approach against a state-of-the-art deep learning based audio-only speech enhancement method.

Index Terms: Speech Enhancement, Cognitively-Inspired, Multimodal, Lip-Reading, Sentiment Features, Deep Learning

1. Introduction

The extensive speech enhancement requirement in wide-range of real-world applications and the advent of advanced signal processing methods have opened new ways to explore and develop more efficient and advanced speech processing technologies. Over the past few decades, several speech enhancement methods have been proposed, ranging from the state-of-the-art statistical, analytical, and classical optimization approaches, to advanced deep learning based methods. The classic speech enhancement methods are mainly based on audio only processing [1][2][3][4]. Recently, researchers have also proposed deep learning based advanced speech recognition [5] and enhancement [6] methods. However, most of the speech enhancement methods are based on single channel (audio only) processing, which often perform poorly in adverse conditions [7]. In this research, we aim to leverage the audio-visual (AV) nature of speech which is inherently multimodal and capable of improving intelligibility in noise [8][9][10] [11][12][13] and have modelled lip-reading as a regression problem for speech enhancement. Specifically, we envision developing a multimodal hearing device that significantly improves both speech quality and intelligibility in everyday and extreme listening environments.

The inherent multimodal nature of the speech is well established in the literature and it is well understood that how speech is produced by the vibration of vocal chords with respect to the articulatory organs configuration. The correlation between the articulatory organs (visible properties) and speech has been shown in several ways in literature using biological, psychological, and mathematical experiments [8][11][10][14]. Therefore, the clear visibility of some of the articulatory organs such as lips, teeth, and tongue could be effectively utilized to extract the clean speech out of the noisy audio signal. In addition, the visual features such as facial expressions and body language also play a vital role in speech perception. The major advantage of using visual cues for generating clean audio feature is their natural noise immunity (i.e. visual speech representation always remains unaffected by the acoustic noise) [15].

In the literature, most of the proposed lip-reading approaches have modelled lip-reading as a classification problem for speech recognition. However, limited work has been conducted to model lip-reading as a regression problem for speech enhancement. In this research, we envision cognitively-inspired, context-aware multimodal speech processing technology based on lip-reading regression model. The technology is aimed at helping users in noisy environments, by contextually learning and switching between audio and visual cues. The initial aim of this research is to develop a proof of concept prototype of the audio-visual hearing-aid technology. An early prototype demo has been developed for online evaluation and feedback. The prototype demo is benchmarked against the state-of-the-art audio-only approach (reported in IEEE Spectrum Magazine 2017) that applied cutting-edge machine learning based on deep neural networks [16]. The preliminary objective and subjective testing has revealed the potential and reliability of the proposed technology as compared to the state-of-the-art audio only speech enhancement techniques.

The rest of the paper is organized as follows: Section II presents an overview of the proposed context-aware multimodal speech processing technology, including multimodal feature extraction, audiovisual mapping, and noisy audio filtering methods. Section III presents the qualitative speech enhancement testing of the proposed technology. Finally, Section IV concludes this work.

2. Proposed Next-Generation Context-Aware Multimodal Technology

The proposed novel cognitively-inspired, multimodal approach aims to contextually exploit and integrates multimodal cues, such as lip-reading and audio features. The term context-aware signifies the technology's contextual learning and adaptable capabilities, which can be employed in next-generation multi-modal applications, including assistive technology such as hearing-aids, cochlear implants, and listening devices. The disruptive technology is capable of contextually enhancing speech intelligibility in extreme noisy environments, so can also be useful for users in situations where ear defenders are worn, such as emergency and disaster response and battlefield environments. In applications such as teleconferencing, video signals could be used to filter and enhance acoustic signals arriving at the receiver-end. People with visual impairment who are unable to see visual cues can also benefit from the proposed technology, particularly in emergency situations. Preliminary simulation results, including an interactive online prototype, demonstrate the potential of the proposed multimodal speech enhancement technology for enabling transformative applications in extreme environments. An abstracted processing of the proposed technology is depicted in Fig. 1, where the multimodal (audio-visual) system has integrated the aforementioned cues for speech processing. The proposed (audio-visual) system extracts the available multimodal features contextually to estimate the clean audio features and then exploits them for real-time speech enhancement. More technical details are comprehensively presented in [17]

2.1. Dataset

For preliminary analysis, the widely used Grid [18] and CHiME corpuses [19] are used to extract lip-reading and noisy audio features. The visual features are extracted using Grid Corpus, whereas CHiME2 is used for extracting audio features. The work could easily be extended to include other visual features such as gestures, facial expressions, body language etc., which is a part of future work and will be presented in upcoming publications. From both the corpuses, an audiovisual (AV) dataset is built by preprocessing the utterances and extracting the audio and visual features. The preprocessing includes sentence alignment and incorporation of prior visual frames. The sentence alignment is used to remove the silence time from the video to restrict the model from learning redundant or insignificant information. The sentence alignment process enforced the model to learn the correlation between the spoken word and corresponding visual representation, rather than over learning the silence. Secondly, the prior visual frames are used to incorporate the temporal information, which ultimately helped the learning engine to better correlate the visual features to corresponding speech features. The audio and visual features extraction procedure is shown in Fig. 2. The audio feature extraction procedure includes sampling, segmentation, Hamming windowing, Fourier transformation, and FB audio features calculation. The visual feature (lip-reading) extraction procedure includes frames extraction, viola-jones lip detector, object tracker, lip cropping, 2D-DCT/convoluted features. Once the dataset is built, it could then be fed into a deep learning model such as LSTM to learn the correlation between audio and visual features.

2.2. Multimodal Features Extraction

2.2.1. Audio Features

The audio features are extracted using the widely used log-filterbank (FB) vectors and Mel-frequency cepstral coefficients (MFCC). The input audio signal is sampled at 50kHz and segmented into N 16ms frames with 800 samples per frame and 62.5% increment rate. Afterwards, a Hamming window and Fourier transformation are applied to produce 2048-bin power spectrum. Finally, a 23-dimensional log-FB is applied, followed by the logarithmic compression to produce 23-D log-FB signal. For MFCC calculation, DCT of the log-auditory-spectrum is obtained.

2.2.2. Visual Features

The visual features include only lip movements in this preliminary work. The lip movement features are extracted using 2D-DCT based standard and widely used visual feature extraction method. Firstly, the video files are processed to extract a sequence of individual frames. Secondly, the Viola-Jones lip detector [20] is used to identify the Region-of-Interest (ROI) in terms of a bounding box. Finally, the object tracker [21] is used to track the lip regions across the sequence of frames. The visual extraction procedure produced a set of corner points for each frame, where the lip regions are then extracted by cropping the raw image for desired visual features, followed by 2D-DCT calculation. More details are comprehensively presented in [22].

2.3. Audiovisual Mapping and Clean Audio Features Estimation

For successful implementation of the proposed technology, one of the essential steps include the estimation of clean audio power spectrum (i.e. audiovisual speech mapping). The audiovisual speech mapping aims to approximate the audio features given only visual information. In the proposed approach, multimodal features (i.e. lip movements) are mapped to the clean audio features using long-short-term memory network. Once the deep learning models are successfully trained and validated, the model predicts the audio log-FB vectors given only the noisy audio and visual features, which are then exploited by the proposed noisy speech filtering framework for speech enhancement. More detail on AV mapping is comprehensively presented in [22], where multilayer perceptron (MLP) was used for AV mapping.

2.4. Enhanced visually derived Wiener filtering

In signal processing, Wiener Filter is the state-of-the-art filter that helps to produce an estimate of a clean audio signal by linear time-invariant (LTI) filtering of an observed noisy audio signal. In the proposed approach, an enhanced visually derived Wiener filtering (EVWF) for speech enhancement has been used. The EVWF effectively exploited the estimated low dimensional clean audio features (through lip-reading) to estimate the high dimensional clean audio power spectrum. Specifically, the EVWF transformed the estimated low dimensional clean audio features into high dimensional clean audio power spectrum using inverse FB transformation. Afterwards, the Wiener filter is calculated (using the estimated audio features) and applied to the magnitude spectrum of the noisy input audio signal, followed by the inverse fast Fourier transform (IFFT), overlap, and combining processes to produce the enhanced magnitude spec-

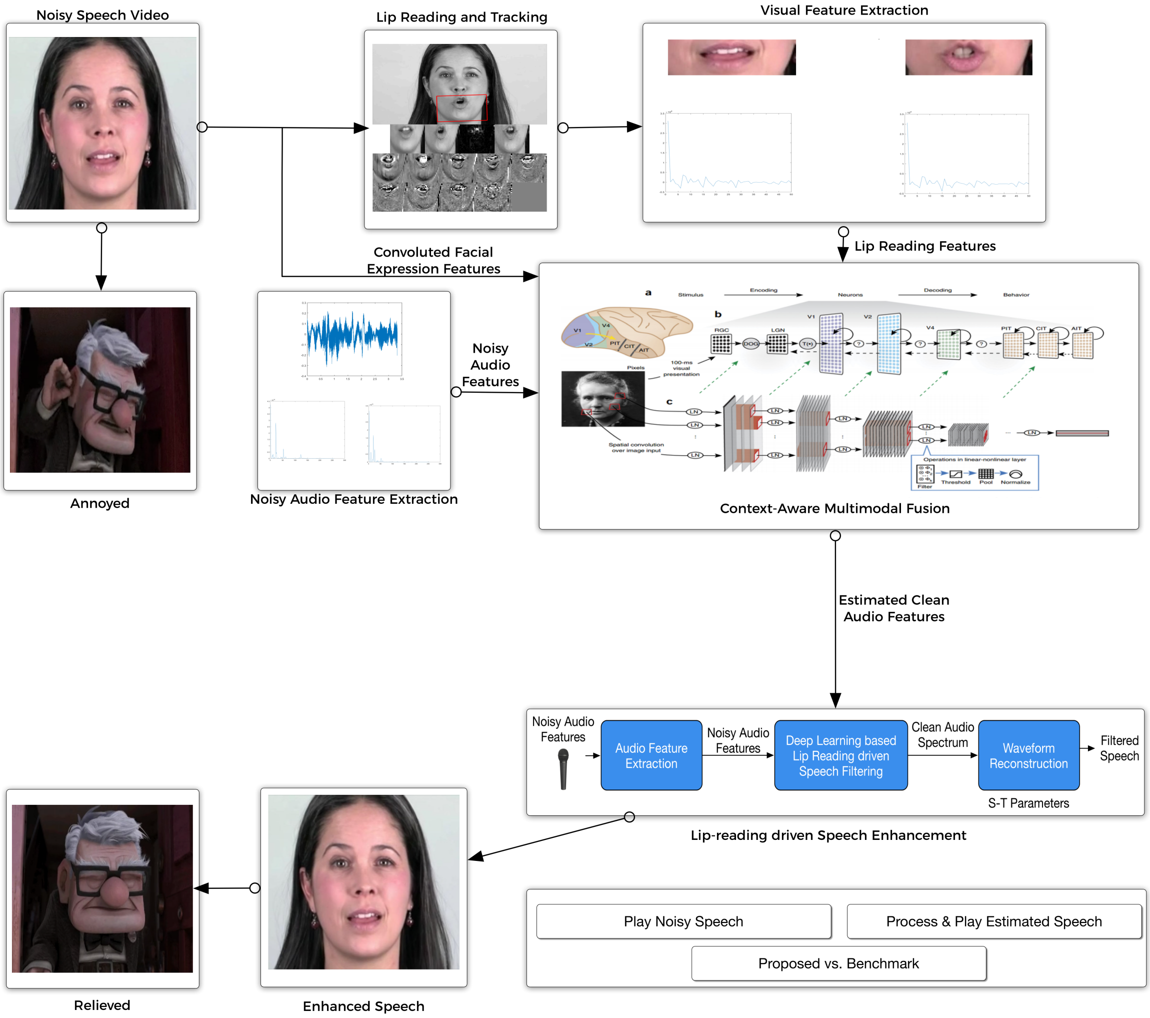


Figure 1: *Prototype Demo: Lip-reading driven deep learning approach for speech enhancement: The device is capable of helping users in noisy environments to experience intelligible clean speech, by contextually learning and switching between audio and visual cues. Audio cues include noisy speech only, whereas visual cues include lip movements.*

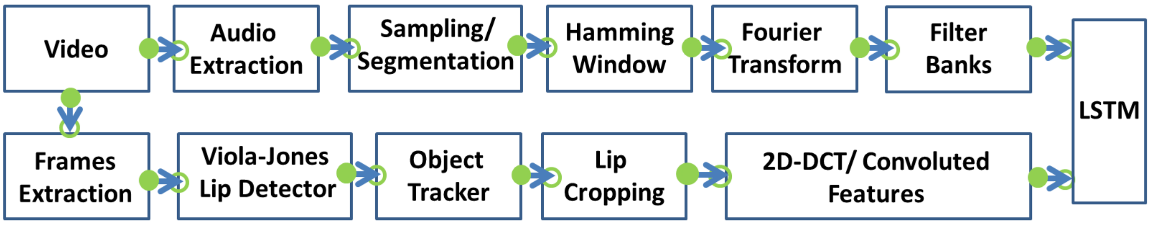


Figure 2: *Audio-Visual Dataset Generation Procedure*

trum.

3. Qualitative Speech Enhancement Testing

The speech enhancement quality of the proposed technology is tested by feeding the randomly selected, real noisy speech videos from YouTube, into the audio-visual speech enhancement system. In the online Demo, the user can listen to the noisy audio, by clicking the *PLAY NOISY SPEECH* button. The visually-enhanced/estimated speech can be perceived by clicking the *PROCESS & PLAY ESTIMATED SPEECH* button. Users can also get a sneak pre-view of behind the scene speech enhancement processing, by clicking individual processing components, such as lip-reading, visual feature extraction, and noisy audio features extraction. The preliminary interactive Demo and the processed multimodal speech demonstrates the potential of context-aware audio-visual hearing-aids, based on big data and deep learning technology.

The second part of the Demo invites listeners to qualitatively compare the new audio-visual approach with a latest audio-only, deep learning benchmark system, recently reported in the IEEE Spectrum Magazine, 2017 [16]. The authors in [16] proposed a DNN-based supervised speech segregation system. The samples of speech before and after enhancement are available at ($http : //cogbid.cs.stir.ac.uk/cogavhearingdemo$) for both DNN-based supervised speech segregation and our proposed multimodal approach. The demo has also demonstrated that how the hearing-aid users perceive the speech. In the sample speech utterances, the proposed multimodal approach has shown better and consistent speech enhancement as compared to the DNN-based supervised speech segregation system. In addition, the proposed multimodal system has recovered both the pitch and clean speech as compared to the DNN-based supervised speech segregation approach, preserving the naturalness of the speech among male/female/infant voices.

4. Conclusion and Future Work

In this paper, an interactive prototype demo is presented as part of a disruptive cognitively-inspired multimodal speech processing technology. In the online demo, the preliminary speech enhancement results and comparisons with the state-of-the-art deep learning based audio-only method have demonstrated the potential and reliability of the proposed speech processing technology. We believe that the disruptive technology is capable of contextually enhancing speech intelligibility in everyday life and even in extreme noisy environments such as emergency and disaster response, and battlefield environments. In addition, the technology could also be utilized in applications such as teleconferencing, where video signals could be used to filter and enhance acoustic signals arriving at the receiver-end. In future, we intend to investigate the performance of the proposed speech processing technology in more realistic real-world scenarios.

5. Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1 (CogAVHearinghttp://cogavhearing.cs.stir.ac.uk). In accordance with EPSRC policy, all experimental data used in the project simulations is available at http://hdl.handle.net/11667/81. The authors would also like to gratefully acknowledge Dr Andrew Abel at Xi'an Jiaotong-Liverpool University for building and providing the audio-visual dataset. Lastly, we gratefully acknowledge the support of NVIDIA Corporation for donating the Titan X Pascal GPU for this research.

6. References

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

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

[3] W. K. Pratt, "Generalized wiener filtering computation techniques," *IEEE Transactions on Computers*, vol. 100, no. 7, pp. 636–641, 1972.

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

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[7] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.

[8] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[9] N. P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 423–425, 1969.

[10] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.

[11] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, vol. 36, no. 4-5, pp. 314–331, 1979.

[12] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *British journal of audiology*, vol. 21, no. 2, pp. 131–141, 1987.

[13] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[14] M. L. Patterson and J. F. Werker, "Two-month-old infants match phonetic information in lips and voice," *Developmental Science*, vol. 6, no. 2, pp. 191–196, 2003.

[15] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[16] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[17] A. Adeel, M. Gogate, B. Whitmer, and R. Watt, "A novel lip-reading driven deep learning approach for speech enhancement," *Emerging Topics in Computational Intelligence, IEEE Transactions on*, 2017.

[18] 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.

[19] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The chime corpus: A resource and a challenge for computational hearing in multisource environments." in *Interspeech*. Citeseer, 2010, pp. 1918–1921.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[21] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[22] A. Abel, R. Marxer, J. Barker, R. Watt, B. Whitmer, P. Derleth, and A. Hussain, "A data driven approach to audiovisual speech mapping," in *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings 8*. Springer, 2016, pp. 331–342.

Home Environmental Sound Alert System for Deaf and Hard-of-Hearing Users

Hye-Seung Cho, Hyoung-Gook Kim

Kwangwoon University, Seoul, Rep. of Korea

{hye_seung401; hkim}@kw.ac.kr

Abstract

Sound signals provide a great deal of information about their sound sources. However, deaf and hard-of-hearing people are not able to access this important information. Therefore, an assistive technology is required that automatically recognizes sound information and converts it into usable information for hearing-impaired people. In this paper, a home environmental sound alert system for deaf and hard-of-hearing users is presented. The system detects the sound generated in the home environment, converts the sound into text, and provides this text to the user. The core component of the environmental sound alert system is an accurate sound event detection mechanism. For precise sound event detection, we proposed an improvement method including signal estimation, channel selection, and a bidirectional gated recurrent neural network.

Index Terms: environmental sound alert, sound event detection, wireless sensor network, gated recurrent neural network

1. Introduction

Sensory abilities such as vision and hearing are very important in human life. In particular, certain information in our society usually depends on communication via sound. This social practice renders important information inaccessible to many deaf and hard-of-hearing people [1]. Therefore, to provide hearing-impaired people with information about sound, different assistive technologies have been developed. For instance, the light system that flash the light when somebody rings the doorbell is one of the representative assistant infrastructures. However, such approaches are only targeted to specific events and they are ineffective when multiple sound events are generated at the same time.

In this paper, we propose a home environmental sound alert for deaf and hard-of-hearing people based on sound event detection (SED). The proposed system is composed of a wireless sensor network (WSN) and the user's smart device.

The environmental sound alert system needs to detect and recognize sound events accurately in various situations in real life. Therefore, we also proposed a method to improve the performance of the SED of the proposed system. The proposed method is comprised of signal estimation, channel selection, and a bidirectional gated recurrent neural network (GRNN) [2]. However, even if a high-performance SED is applied, detection errors can occur due to various external factors. In order to minimize the risk caused by these errors, the proposed system provides both the event detection results and their probabilities.

The outline of this paper is as follows. In Section 2, the proposed system and the detail of the SED method are explained. Experimental results are presented in Section 3 and conclusions are given in Section 4.

2. Proposed System

Fig. 1 schematically illustrates the proposed environmental sound alert system based on SED.

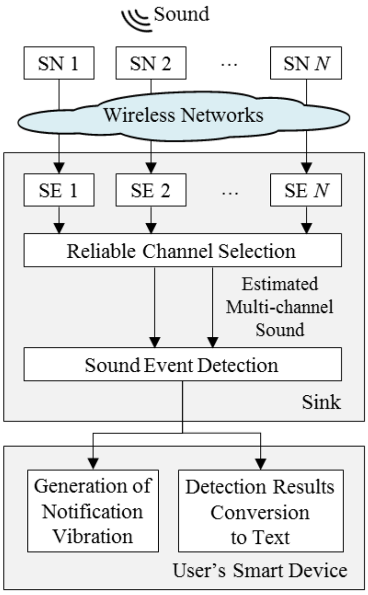

Sound
SN 1
SN 2
···
SN *N*
Wireless Networks
SE 1
SE 2
···
SE *N*
Reliable Channel Selection
Estimated Multi-channel Sound
Sound Event Detection
Sink
Generation of Notification Vibration
Detection Results Conversion to Text
User's Smart Device

Figure 1: *Schematic illustration of the proposed system.*

The wireless sensor nodes (SNs) can simultaneously capture sounds generated in a room. Each SN is equipped with a single microphone. The microphone at each SN receives mixed or polyphonic sounds. The recorded mixed sounds are encoded and transmitted via sound packets to the sink through networks with wireless links that are associated with packet loss. When each microphone packet arrives at the sink via the wireless networks, it is decoded into a signal frame. Lost packets are recovered by packet loss concealment in signal estimation (SE).

Then, a set of microphones of which the signals are the most highly correlated with each other are chosen among the multi-channel microphones to increase computational efficiency and achieve better performance. In this paper, we used a signal-based channel selection (CS) method using a multi-channel cross-correlation coefficient (MCCC) [3]. The basic concept of this approach is to treat the channel that is uncorrelated with the other channels as being unreliable and to select only a subset of microphones with the most correlated signals.

After CS, the signals of the selected two-channels are then used for environmental SED. Motivated by the human auditory system (using two ears), we extracted a noise-reduced spectrogram and time difference of arrival (TDOA) [4] from the two-channel audio information. The features were used as high-resolution spectral inputs to train the bidirectional GRNN (BGRNN). The BGRNN is one of the most recent neural networks, and demonstrates good performance in sequence modeling. It provides a fast and stable convergence rate compared to the long short-term memory recurrent neural networks (LSTM-RNN).

Detected sound event labels and probabilities are sent to the user device. This information is sorted in the order of highest to lowest accuracy and converted to text. The vibration for notification is also generated. The user is then notified with a vibration and text notification letting them know which sound event has occurred.

3. Experimental Results

In this section, the performance of the proposed SED method of the proposed system is evaluated using real life audio.

The living area used in our experiments was a 30 m^2 apartment. The rooms were equipped with sound sensors. Real sounds were recorded with sound sensors from various everyday environments. The sound corpus contained 10 sound classes. The polyphony percentages of the test set among the annotated frames were as follows: 83.5%, 10.6%, 4.3%, and 1.6% at polyphonic levels of 1, 2, 3, and 4, respectively. These samples for each class were distributed randomly (60% in training set, 20% in validation set, and 20% in test set). The performance of the proposed method was compared to that of different classifiers in combination with different features as follows, where NR, ST, and 2 denote the noise reduction, spectrogram, and two-channels, respectively:

Baseline (BL): The baseline system used MFCC coefficients (20 static, 20 delta, and 20 acceleration) extracted from one-channel audio. Before feature extraction, SE and CS were performed, although NR was not applied. A Gaussian mixture model (GMM) was used for SED.

Proposed Method (PM): The PM was composed of SE, CS, NR2, ST2, TDOA, and BGRNN. The input layer of the BGRNN comprised 40 units and three hidden layers with 200 GRU units. 3-layer BGRNNs were initialized with orthogonal weights and rectifier activation functions. The network was trained by binary cross-entropy as loss function.

Method 1 (M1): M1 was composed of SE, CS, NR, ST, TDOA, and BGRNN. One-channel features per frame were applied to the BGRNN classifier.

Method 2 (M2): M2 was composed of CS, ST, TDOA, and BGRNN. One-channel spectrogram features without SE or NR per frame were extracted and applied with 1 TDOA to the BGRNN classifier.

Method 3 (M3): M3 was composed of SE, CS, NR2, ST2, TDOA, and GRNN. Instead of using the BGRNN classifier of M1, a GRNN was used as the classifier with two-channel features.

Method 4 (M4): M4 was composed of SE, CS, NR2, ST2, TDOA, and LSTM-RNN. The LSTM-RNN was used as a classifier with two-channel features. The input layer of the LSTM-RNN comprised 40 units and two hidden layers with 200 LSTM units. The network was trained by binary cross-entropy as loss function.

For the evaluation metrics of system performance for SED, we used error rate (ER) and F-scores calculated in one second segments [5]. Experimental results show the baseline system (BL) has an event average error rate (ER) of 1.1 and F-score of 64.5%. The PM significantly outperforms the baseline system in terms of ER and F. These results confirm that SE, CS, and NR significantly contributed to the detection of overlapping sound events. In addition, BGRNN achieves better classification results than GMM, GRNN, and LSTM-RNN, which were trained on the same audio features. Spatial and noise-reduced spectrogram features from the multi-channel audio show considerable improvements over those when only mono-channel audio was used with the same classifier.

Table 1: *Performance Comparison for different combinations of classifiers and feature.*

Methods	ER	F(%)
BL	1.1	64.5
PM	0.63	86.7
M1	0.71	83.2
M2	0.98	74.9
M3	0.65	86.3
M4	0.67	85.9

4. Conclusions

In this paper, we proposed a home environmental sound alert system for deaf and hard-of-hearing users. The proposed system provides the user with information of all of the detected sound events and their probabilities. We also presented an improved sound event detection scheme for a reliable and effective alert system. Experimental results demonstrate the potential effective use of the proposed alert system in practical situations. Future work will focus on extending the BGRNN to improve detection accuracy for various sound event detection based systems.

5. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2015R1D1A1A01059804).

6. References

[1] D. Bragg, N. Huynh, and R. E. Ladner, “A personalizable mobile sound detector app design for deaf and hard-of-Hearing users,” *Proceedings of the ACM SIGACCESS, October 24–26, Reno, NV, USA*, 2016, pp. 3–13.

[2] M. Zoehrer and F. Pernkopf, “Gated recurrent networks applied to acoustic scene classification and acoustic event detection,” *Detection and Classification of Acoustic Scenes and Events 2016*, 2016.

[3] M. Wölfel, “Channel selection by class separability measures for automatic transcriptions on distant microphones,” *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 582–585.

[4] D. Pavlidi et al., “Real-time multiple sound source localization and counting using a circular microphone array,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[5] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.

ASR-based Measures for Microscopic Speech Intelligibility Prediction

Mahdie Karbasi[*], *Dorothea Kolossa*[*+]

[*]Cognitive Signal Processing Group, Ruhr-Universität Bochum, 44801 Bochum, Germany
[+]Kavli Institute for Theoretical Physics, UC Santa Barbara, USA
{mahdie.karbasi, dorothea.kolossa}@rub.de

Abstract

Automatic and accurate prediction of human speech perception performance is of great benefit for developing speech processing algorithms. Automatic speech recognizers (ASR) can be designed with the goal of mimicking human performance in speech recognition, hence, they can also be employed for predicting the intelligibility of speech. This paper presents two new objective measures for predicting speech intelligibility at the word level. The normalized likelihood difference (NLD) and the time alignment difference (TAD) are the proposed measures, extracted utilizing the hidden Markov models (HMMs) trained for an ASR system. Experimental results show that the proposed measures can accurately predict the normal-hearing listeners' performance in a keyword recognition task.

Index Terms: speech intelligibility prediction, automatic speech recognition, microscopic approach, objective measure

1. Introduction

The number of applications of devices working with speech signals is growing every day. For instance, many researchers are developing speech processing algorithms for hearing aids, which are widely needed in our aging societies. For these developments, it has always been a requirement to assess the intelligibility or quality of the signal at hand before or after processing. Partially automating this task rather than purely relying on listening tests is beneficial considering the time and cost required in human intelligibility assessment.

In the last decades, many objective measures have been published, which aim to predict the speech intelligibility from a macroscopic point of view. Well-known objective measures like the speech intelligibility index (SII) [1], speech transmission index (STI) [2], short time objective intelligibility (STOI) [3], and mutual-information-based models [4] compare the degraded speech with a reference in long segments, e.g. over an entire sentence, and predict only the average number of speech units, like words, heard correctly. The speech-based envelope power spectrum model (sEPSM) [5] is another example of macroscopic measures, which uses an auditory model to analyze the speech signal and computes the signal-to-noise ratios in modulation frequency bands as a measure of intelligibility. This model was later extended to mr-sEPSM [6] and sEPSM-corr [7] in order to account for non-linear degradations as well.

Macroscopic measures typically require longer input signals in order to obtain a sufficient accuracy in intelligibility prediction. In contrast to such methods, microscopic approaches process smaller segments of speech and attempt to predict the individual listener's response to a speech signal on a word-by-word or phoneme-by-phoneme basis. As an example, the microscopic method proposed in [8] uses an auditory model to extract features from speech signals and the dynamic time warping algorithm to compare the features extracted from a degraded signal to its clean counterpart for predicting the intelligibility of single words.

In another microscopic framework [9], Kollmeier et al. have considered the outputs of an ASR system as predictors of speech perception in both normal-hearing and hearing-impaired listeners. In this method, in contrast to the previously mentioned intelligibility prediction methods, it is not required to have access to the clean signal as a reference for predicting the speech intelligibility. Also, this method can benefit from the language knowledge implemented as a grammar in ASR systems. In [10], it has been shown that in listening tests, humans are taking advantage of their prior knowledge about the characteristics of speech units such as phonemes. Therefore, the authors have suggested to take the phonetic information into account in the design of instrumental quality or intelligibility measures. Otherwise, comparing the processed speech only to a signal-based reference can lead to unreasonably low quality estimates in scenarios like artificial speech bandwidth extension. A non-intrusive prediction of intelligibility has been introduced in [11, 12] that uses either the oracle transcriptions or the ASR-recognized transcriptions of the speech signal and synthesizes the clean features, required inside an intrusive intelligibility prediction method.

Microscopic methods promise to be more precise in estimating intelligibility and in diagnosing problems due to specific phoneme confusions. We have previously proposed an approach [13] that uses the logarithm of likelihood ratio of the true and the ASR-recognized word as an objective metric for the intelligibility prediction. In this paper, we introduce two other new heuristic measures for predicting the speech intelligibility from a microscopic point of view. The proposed measures are extracted utilizing an HMM-based ASR system, which will be explained and inspected in detail in the following sections.

2. ASR-based Microscopic Intelligibility Measures

Within the process of recognizing a speech signal, an HMM-based ASR system can compute some intermediate features that are indicative of its confidence. The likelihood of N-best state sequences or N-best word choices are primary examples [14]. Consequently, it can be hypothesized that such features contain information about the intelligibility of speech units as well. Moreover, it can be stated that the less intelligible a speech signal is, the more errors are expected in the ASR output. Hence, the time alignment information, estimated during the recognition process, can be used as another source of information on the intelligibility of speech units. In order to exploit such HMM-based features in the context of speech intelligibility prediction, the normalized likelihood difference (NLD) and the time alignment difference (TAD) are introduced in this paper.

Prior to extracting the NLD and TAD, an HMM-based

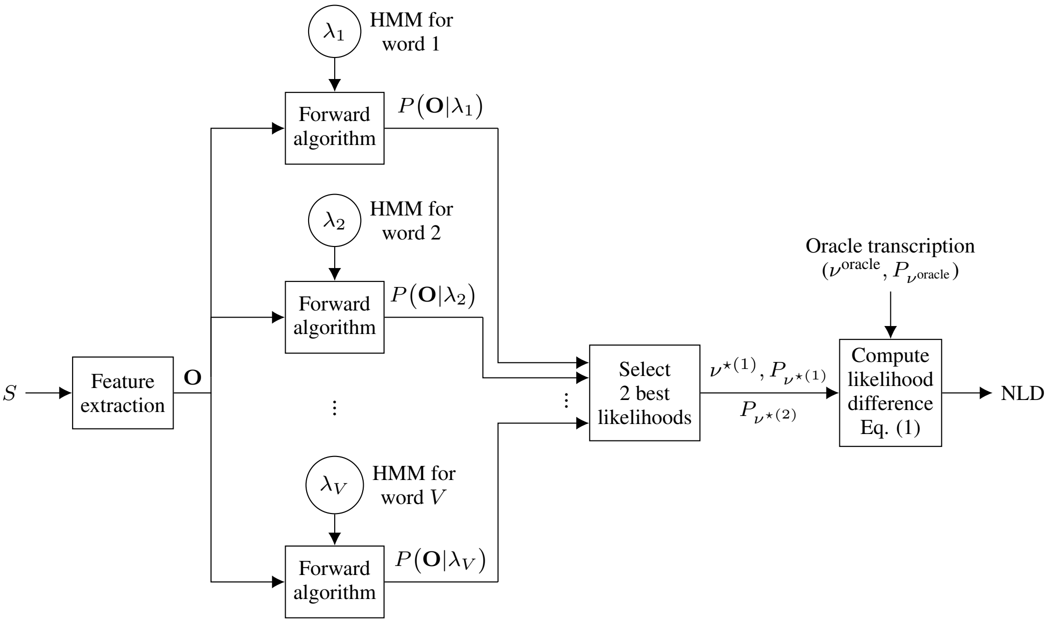

λ_1
HMM for word 1
Forward algorithm
$P(\mathbf{O}|\lambda_1)$
λ_2
HMM for word 2
Forward algorithm
$P(\mathbf{O}|\lambda_2)$
S
Feature extraction
$\mathbf{O}$
λ_V
HMM for word V
Forward algorithm
$P(\mathbf{O}|\lambda_V)$
Select 2 best likelihoods
$\nu^{\star(1)}, P_{\nu^{\star(1)}}$
$P_{\nu^{\star(2)}}$
Oracle transcription
$(\nu^{\text{oracle}}, P_{\nu^{\text{oracle}}})$
Compute likelihood difference Eq. (1)
NLD

Figure 1: *Block diagram of the first proposed speech intelligibility measure, the NLD.*

speech recognition system must be trained. Here, for each word ν in the vocabulary, one HMM λ_ν is built. The parameters of each HMM can be estimated by optimizing the likelihood of the training set observation vectors for the associated word or by discriminative training. Based on these trained models, Figure 1 and Figure 2 illustrate the schematic diagram of extracting the proposed measures NLD and TAD, respectively. The detailed description of these measures is provided below. It is notable that the introduced measures, here, are extracted at the word level, however, it is possible to extend the current framework for predicting the perception of phonemes as well.

2.1. Normalized Likelihood Difference (NLD)

In order to extract the NLD per word, the speech signal is segmented into the constituent words and each segment is fed into the system as an input. The first step in extracting the NLD is to apply a feature extraction method to the input signal S and estimate the observation sequence $\mathbf{O} = \{\mathbf{o}_1\mathbf{o}_2\ldots\mathbf{o}_T\}$. Then, the model likelihoods given the observation sequence are computed for all possible words, $P(\lambda_\nu|\mathbf{O}), 1 \le \nu \le V$. All model likelihoods are sorted to find the first $\nu^{\star(1)}$ and second $\nu^{\star(2)}$ most likely word:

$$\begin{aligned}\nu^{\star(1,2)} &= \underset{1\le\nu\le V}{\arg\max}^{(1,2)}\big[P(\lambda_\nu|\mathbf{O})\big]\\ &= \underset{1\le\nu\le V}{\arg\max}^{(1,2)}\Big[\frac{P(\mathbf{O}|\lambda_\nu)P(\lambda_\nu)}{P(\mathbf{O})}\Big]\end{aligned} \tag{1}$$

Here, $P(\lambda_\nu|\mathbf{O})$ is the likelihood of the word model λ_ν given the observation sequence $\mathbf{O}$ and V is the number of all possible words.

Since the probability of the observation sequence, $P(\mathbf{O})$, is independent of word models and the prior probability of each model, $P(\lambda_\nu)$, is equal here for all possible words, Equation (1) can be reformulated to

$$\nu^{\star(1,2)} = \underset{1\le\nu\le V}{\arg\max}^{(1,2)}\big[P(\mathbf{O}|\lambda_\nu)\big]. \tag{2}$$

As shown in Figure 1 and according to Equation (2), following the feature extraction, the forward algorithm is applied to the observation sequence in order to compute the probability of the input sequence given the model λ_ν for all possible words. Next, all probabilities are sorted and the words with the first and second highest probability are selected. The word with the highest probability $\nu^{\star(1)}$ is compared to the oracle transcription of the signal ν^{oracle}. If they are equal, the NLD is defined to be the difference between the likelihoods of the first $\lambda_{\nu^{\star(1)}}$ and the second $\lambda_{\nu^{\star(2)}}$ best word models given the observation sequence and normalized with the best likelihood. Otherwise, the order of the difference between the two model likelihoods is interchanged:

$$\text{NLD} = \begin{cases} \dfrac{P(\lambda_{\nu^{\star(1)}}|\mathbf{O}) - P(\lambda_{\nu^{\star(2)}}|\mathbf{O})}{P(\lambda_{\nu^{\star(1)}}|\mathbf{O})}, & \text{if } \nu^{\star(1)} = \nu^{\text{oracle}} \\ \dfrac{P(\lambda_{\nu^{\text{oracle}}}|\mathbf{O}) - P(\lambda_{\nu^{\star(1)}}|\mathbf{O})}{P(\lambda_{\nu^{\star(1)}}|\mathbf{O})}, & \text{if } \nu^{\star(1)} \neq \nu^{\text{oracle}} \end{cases} \tag{3}$$

Similar to Equation 2, here, $P(\lambda_\nu|\mathbf{O})$ can be replaced by $P(\mathbf{O}|\lambda_\nu)$ and the NLD is computed in practice using the probability of the observation sequence given the word models.

2.2. Time Alignment Difference (TAD)

As a second metric, we consider the time alignment difference (TAD) between the recognized and oracle transcriptions of the input signal. The computational steps of this measure are shown in Figure 2. Like the NLD, the TAD is estimated at the word level, but the input to this method is the entire sentence. At first, a feature extraction algorithm is applied to the input signal. Then, a decoder is employed to perform a continuous speech recognition on the observation sequence $\mathbf{O}$. Lastly, the recognized time alignment for a specified keyword is compared to its oracle time alignment and their relative difference is computed as the TAD measure

$$\text{TAD} = \frac{|Tb_{Rec} - Tb_{Oracle}| + |Te_{Rec} - Te_{Oracle}|}{L}, \tag{4}$$

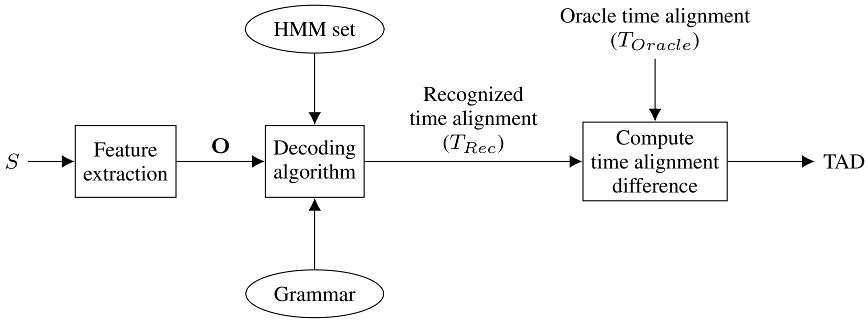


Figure 2: *Block diagram of the second proposed speech intelligibility measure, the TAD.*

where Tb_{Rec} and Tb_{Oracle} are the recognized and the oracle beginning frame index of a single word, respectively. Similarly Te_{Rec} and Te_{Oracle} represent the recognized and the oracle ending frame index of the same word and lastly, L is the word length in frames.

3. Experiments

3.1. Speech Intelligibility Database

The original Grid corpus [15] and its noisy version [16] have been used in the following experiments. The original corpus contains 34000 clean speech signals in total, recordings of 34 English speakers made at the University of Sheffield. Each Grid utterance is a 6-word sentence with a fixed grammar: <Verb (4)- Color (4)- Preposition (4)- Letter (25)- Digit (10)- Adverb (4)>, where the numbers in parentheses represent the number of available choices for each word type.

In addition to the clean Grid database, there is also a noisy version, which has been created by adding speech-shaped noise to the clean signals at 12 different signal-to-noise ratios (SNRs) from -14 dB up to 6 dB in steps of 2 dB, plus 40 dB (labeled as clean). For the noisy database, the results of a listening test conducted on 20 listeners with normal hearing are available. Each participant has listened to 2000 utterances and has been asked to recognize three keywords, the color, letter, and digit [16].

3.2. Experimental Setups

The first step in all experiments is the extraction of features from the speech signals. As features, the first 13 Mel frequency cepstral coefficients (MFCCs) plus their first (Δ) and second order derivatives ($\Delta\Delta$) were used. Hamming-windowed frames with a length of 25 ms and a frame shift of 10 ms were chosen for the MFCC extraction algorithm. The sampling frequency was set to 25000 Hz in all experiments.

For ASR, each word was modeled using a linear left-to-right HMM, resulting in 51 whole-word HMMs plus one silence model. The number of states were chosen as three times the number of phonemes of the modelled word. A 2-mixture diagonal covariance GMM represents the state output distribution of all HMMs.

In order to be able to use the entire data collected in the listening test for evaluating the intelligibility measures, all experiments were carried out with 5-fold cross validation. During each fold, the speech database was divided into the disjoint training (60%), development (20%), and test (20%) sets. To raise the accuracy, noise-dependent models were trained separately at each SNR and development sets were used to assess the accuracy of HMMs during the training.

To evaluate the proposed intelligibility measures, single Gaussian models (GM) were utilized to predict the intelligibility of the Grid keywords. For each test, two GMs were trained; one to represent the distribution of the intelligibility measure for correctly recognized words and another one representing the distribution of the same intelligibility measure but for words misrecognized by human listeners. Hence, in this framework, intelligibility measures were used as input features for GMs. After training, GMs were employed to predict whether an input speech signal can be recognized correctly at the word level. Here, the development set data were used for training the GMs and the test set data were used to evaluate them.

3.3. Evaluation

In the current work, the proposed intelligibility measures are assessed for predicting the normal-hearing listeners' performance. The results, presented below, are averaged over all 20 listeners available in the Grid database. Table 1 contains the accuracy of the proposed intelligibility measures NLD and TAD in predicting human keyword recognition results, averaged over 12 SNRs. Also, the accuracy of the ASR system in the same task is given in this table, which is computed by a direct comparison of ASR and human recognition outcomes. In addition to the mentioned methods, the well-known macroscopic intelligibility measure STOI [3] was used as a baseline to predict the intelligibility of Grid keywords. Please note that the STOI method needs longer units of speech for its computations and normally, it can not be used for computing the intelligibility of a single word. Therefore, we have augmented the length of every speech signal corresponding to a word by repeating the same signal several times to allow for a computation of the STOI per word. This repetition was implemented in the one-third octave band domain. After framing and extracting the one-third octave band representation of the signal in each frame, all frames were repeated several times for both degraded and reference speech signals without inserting any silence gaps. Therefore, no artifacts have been introduced to the signal, which might have disadvantaged the STOI method.

Considering the results in Table 1, it is evident that both proposed measures, NLD and TAD, have higher accuracies, on average, in comparison to the STOI and to the direct use of the ASR output. Moreover, the average accuracy of the TAD exceeds that of the NLD measure. A statistical significance analysis using Fisher's exact test [17] has shown that the TAD is statistically different from all competitors at a significance level of 0.01. Furthermore, the comparison of the NLD and ASR results with that of the STOI using the same test has shown that

Table 1: *Average accuracy of all considered intelligibility measures in predicting the keyword recognition performance of 20 normal-hearing listeners.*

ASR	STOI	NLD	TAD
78.70	77.24	78.76	**80.92**

both methods are statistically different from the STOI, at a significance level of 0.01 as well.

An SNR-based comparison of the above intelligibility prediction methods is provided in Figure 3. One can observe that the TAD has the highest accuracy in most SNRs down to -8 dB. The STOI has a comparable performance to that of the TAD at 2 dB and higher SNRs but its accuracy drops steeply in the middle of the plot. The NLD and ASR have a similar pattern and are less accurate at higher SNRs than the STOI and TAD measures. The NLD is performing better than the ASR in most SNRs except for the very high (greater than 4 dB) and very low (smaller than -12 dB) ones.

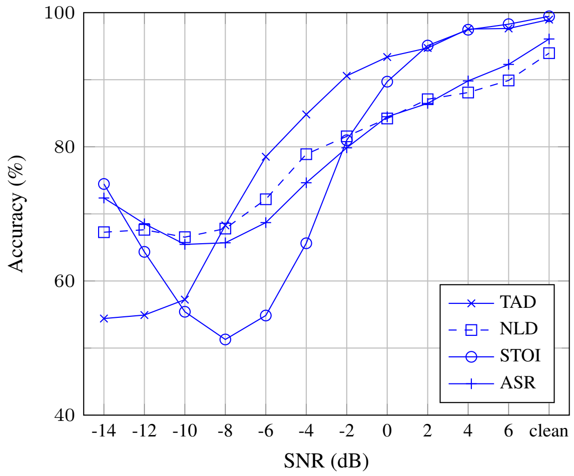


Figure 3: *Accuracy of all considered intelligibility measures per SNR in predicting the the keyword recognition performance of 20 normal-hearing listeners.*

4. Conclusions and Future Work

In this work, we have introduced two new intelligibility measures, NLD and TAD, both derived from a simple ASR system. These measures are proposed for predicting the intelligibility from a microscopic point of view. The NLD is computed based on the likelihood difference of the 2 best word choices and the TAD depends on the time alignment information. It was shown that, on average, both measures outperform the STOI method as well as the direct ASR system output. The TAD achieves a higher accuracy than the NLD. In some SNRs, both measures have lower accuracies in comparison to the baseline methods which needs more analysis. Extracting and appending complementary information to the proposed measures, and employing a discriminatively trained, DNN-based ASR can be considered as possible solutions for elevating the accuracy of our measures. The capacity of these measures for predicting the individual performance of hearing-impaired listeners will be examined in future works. Also, an extension of the NLD based on the likelihoods of n-best word hypotheses should be investigated in future work, with the goal of predicting likely word confusions.

5. Acknowledgements

This research has received funding from the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°[317521]. It was also supported in part by the National Science Foundation under Grant No. NSF PHY-1125915. The authors would like to thank Jon Barker for providing a noisy version of the Grid database with comprehensive listening test results.

6. References

[1] *Methods for the Calculation of the Speech Intelligibility Index*, S3.5-1997, ANSI, New York, NY, USA, 1997.

[2] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

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

[4] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2014.

[5] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the envelope power signal-to-noise ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. 2384–2384, Apr. 2011.

[6] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, 2013.

[7] 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, pp. 2670–2679, 2016.

[8] T. Jürgens, S. Fredelake, R. M. Meyer, B. Kollmeier, and T. Brand, "Challenging the speech intelligibility index: macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners," in *Proc. Interspeech 2010*, Sep. 2010, pp. 2478–2481.

[9] B. Kollmeier, M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand, "Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with fade empowering the attenuation and distortion concept by Plomp with a quantitative processing model," *Trends in Hearing*, vol. 20, pp. 1–17, 2016.

[10] T. Fingscheidt and P. Bauer, "A phonetic reference paradigm for instrumental speech quality assessment of artificial speech bandwidth extension," in *Proc. 4th International Workshop on Perceptual Quality of Systems*, 2013.

[11] M. Karbasi, A. H. AbdelAziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *Proc. ICASSP 2016*, Mar. 2016, pp. 624–628.

[12] M. Karbasi, A. H. Abdelaziz, H. Meutzner, and D. Kolossa, "Blind non-intrusive speech intelligibility prediction using twin-HMMs," in *Proc. Interspeech 2016*, Sep. 2016, pp. 625–629.

[13] M. Karbasi and D. Kolossa, "A microscopic approach to speech intelligibility prediction using auditory models," in *Proc. DAGA 2015*, Mar 2015, pp. 16–19.

[14] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[16] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[17] A. Agresti, "A survey of exact inference for contingency tables," *Statistical science*, pp. 131–153, 1992.

Exploiting deep learning to inform spectral contrast enhancement for hearing-impaired listeners

Ning Ma[1], *Guy J. Brown*[1], *Jon Barker*[1], *Michael Stone*[2]

[1]Department of Computer Science, University of Sheffield, UK
[2]School of Psychological Sciences, University of Manchester, UK
{n.ma, g.j.brown, j.p.baker}@sheffield.ac.uk, michael.stone@manchester.ac.uk

Abstract

Hearing-impaired (HI) listeners have great difficulty in understanding speech when there is noise, music, or people talking in the background. Despite of some advances that have been made in the field, current algorithms adopted in modern hearing aids have had limited success in improving speech intelligibility, as background noise is amplified along with people's voice of interest. Previously many studies [1, 2, 3, 4, 5] have focused on enhancing spectral contrast of sound in order to increase speech intelligibility in noise. These techniques typically employ a digital processing method for altering spectral contrast – the difference in amplitude between spectral peaks and valleys. However, the effects of such processing are usually small (for example, [4] report a relative improvement of 8% in intelligibility for one set of parameters only at a signal-to-masker ratio of -6 dB, but not higher).

One of the reasons that such methods are not very effective is that spectral contrast is enhanced without distinguishing target speech from background noise. Such blind processing does not necessarily improve auditory grouping, a likely process listeners use to organise sound mixtures into auditory streams belonging to individual sources. In some cases grouping cues may be degraded by such processing [4].

Recently, machine learning methods that utilise a deep neural network (DNN) have shown breakthrough performance in speech and language processing. DNNs can learn to identify the spectro-temporal regions in which speech dominates the noise (referred to as a "mask")[6, 7, 8]. The speech-dominating parts are amplified whereas those in which background noise dominates are discarded. Such noise-filtering methods have been shown improved speech intelligibility for hearing-impaired listeners [9, 10].

There are two reasons, however, to believe that filtering out all background noise may not be the optimal solution for improving speech intelligibility for hearing-aid users. First, improved signal-to-noise ratios do not necessarily translate into improved speech intelligibility. Identification of reliable speech regions is challenging in daily listening environments where a large and variable number of sound sources are present, and mislabelling of speech-dominating parts can often degrade speech intelligibility. Second, even if we could perfectly filter out the background noise, some of them are often desired to which listeners might want to switch attention. For example, when talking to someone at a train station, the listener might also want to pay attention to the announcement in the background.

In this study we look beyond traditional spectral contrast enhancement and propose an approach in which deep learning is used to inform spectral contrast enhancement. At first, a DNN based method is adopted to identify the spectro-temporal mask dominated by the target speech. Previous studies have used relatively simple DNN architectures and input features [9, 10]. The recent research conducted at Sheffield has demonstrated that incorporation of pitch-related features in a long short term memory (LSTM) network is capable of learning long-term dependencies which is particular effective when the background noise is not stationary as in daily listening environments [8]. In the subsequent spectral contrast enhancement, the time-frequency components belonging to a same source are processed coherently. This could include not just enhancing the spectral contrast for the target speech, but also reducing the spectral contrast. Such a method is analogous to reducing the depth of field of a lens in photography, thus emphasising the target subject while de-emphasising the background.

The aim of this study is to determine the benefit of the proposed deep learning methods on speech intelligibility for hearing-impaired listeners. The objectives include:

- To measure the benefit of improved deep learning methods for speech masks estimation on speech intelligibility for hearing-impaired listeners;
- To determine the effect of mask-informed spectral contrast enhancement.

We plan to conduct two sets of listening experiments. In the first set of experiments, the estimated probabilistic mask will be used to directly resynthesise enhanced signals, by weighting time-frequency (T-F) bins accordingly before overlap-adding signals over all frequencies. Three different DNNs will be used including the state-of-the-art baseline system proposed in [10] and two proposed systems [8]. In the second set of experiments, the best performing mask estimation method from the first experiment will be selected and used to inform spectral contrast enhancement. In this case spectral contrast enhancement is applied only to the T-F bins that are more likely to be dominated by speech. The other T-F bins are left intact so that the background sound are not completely filtered out. The results will be compared with those by direct resynthesis from the first set of experiments.

We will measure the effect of such processing on speech intelligibility by measuring the percent correct identification of keywords in sentences presented in both speech-shaped noise and a second talker. Two groups of listeners will be invited to take part in this experiment: normal-hearing listeners and hearing-impaired listeners. Two signal-to-masker ratios (SMRs) will be used for each set of tests.

Index Terms: spectral contrast enhancement, deep learning, hearing impairment, speech understanding in noise

1. References

[1] M. A. Stone and B. C. J. Moore, "Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligi- bility and quality," *J. Rehab. Res. Devel.*, vol. 29, pp. 39–56, 1992.

[2] T. Baer, B. C. J. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality and response times," *J. Rehab. Res. Devel.*, vol. 30, pp. 49–72, 1993.

[3] J. M. Alexander, R. L. Jenison, and K. R. Kluender, "Real-time contrast enhancement to improve speech recognition," *PLoS ONE*, vol. 6, no. 9, 2011.

[4] J. Chen, T. Baer, and B. C. J. Moore, "Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2987–98, 2012.

[5] W. Nogueira, T. Rode, and A. Buchner, "Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants," *J. Acoust. Soc. Am.*, vol. 139, no. 2, pp. 728–739, 2016.

[6] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[7] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. ICASSP*, 2014, pp. 3709–3713.

[8] N. Ma, R. Marxer, J. Barker, and G. Brown, "Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 490–495.

[9] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–3038, 2013.

[10] J. Chen, Y. Wang, S. Y. S.E., D. Wang, and E. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, pp. 2604–12, 2016.

Noise reduction using an eyeglass-frame microphone array based on DOA estimation by LASSO

Kenji Ozawa[1], *Shion Yokouchi*[1], *Masanori Moirse*[1]

[1]Faculty of Engineering, University of Yamanashi, Japan
{ozawa, t13cs059, mmorise}@yamanashi.ac.jp

Abstract

In our previous study, we have proposed a practical method for noise reduction using a microphone array. The method initially estimates the direction of arrival (DOA) and waveform of a noise signal, then subtracts the estimated noise from the output of a reference microphone to restore a target signal. However, the method is effective only when there is one noise source. This study expands the method by using the least absolute shrinkage and selection operator (LASSO) algorithm. When there are multiple noise sources, the DOA of the most dominant noise is estimated by the LASSO to reduce the noise effectively. The results from computational simulation experiments show the efficiency of the proposed method when the microphone array is mounted on an eyeglass frame.

Index Terms: microphone array, DOA estimation, LASSO, eyeglass frame

1. Introduction

Microphone arrays are effective for improving speech intelligibility in a noisy environment. As for hearing aids, a simple installation of a microphone array is to mount it on an eyeglass frame. However, classical delay-and-sum (DAS) beamforming provides very small amounts of noise reduction at low frequencies because the array length is comparable to the wavelength of a low frequency component involved in speech. The performance has been progressed by using adaptive array processing [1] and the superdirective array technique [2]. These advanced processing techniques need a high calculation cost, while the cost should be reduced for portable devices such as hearing aids.

We have proposed a practical method to suppress noise using a microphone array [3]. However, the method is effective only when there is one noise source. This study expands the method by using the least absolute shrinkage and selection operator (LASSO) algorithm [4].

2. Proposed system

2.1. System structure

The basic structure of the expanded system is the same as that in [3]. Figure 1 shows the block diagram of the system when there are four microphones. The microphones are located on an eyeglass frame (Fig. 2) where the first one from an edge is defined as the reference microphone (M$_\mathrm{R}$). A sound arriving from perpendicular to the eyeglass frame is regarded as a target signal and sounds arriving from other directions are treated as noise signals. This system first estimates the DOA and waveform of a noise signal observed at the $\mathrm{M_R}$ independently across frequency bins, and then subtracts the estimated waveform of the noise signal from the $\mathrm{M_R}$ output to suppress the noise.

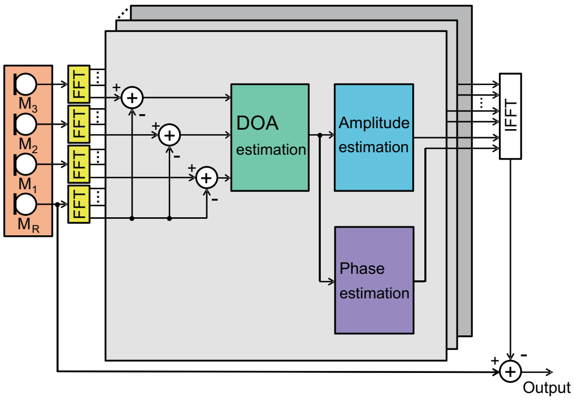


Figure 1: *Block diagram of the proposed system [3].*

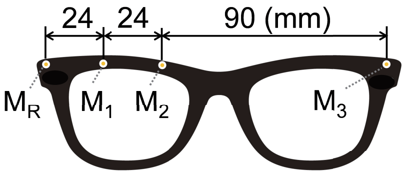


Figure 2: *Microphone array mounted on an eyeglass frame.*

A signal observed at the $\mathrm{M_R}$, $M_\mathrm{R}(\omega)$, is a superimposition of a target signal, $S(\omega)$, and a noise signal, $N(\omega)$.

$$M_\mathrm{R}(\omega) = S(\omega) + N(\omega) \tag{1}$$

If we assume that one noise sources is located far away from the array, the signal observed at the i-th microphone, $M_i(\omega)$, can be described as follows:

$$M_i(\omega) = S(\omega) + N(\omega)e^{-j\omega\tau_i} \tag{2}$$

where τ_i is the time delay from the $\mathrm{M_R}$. By subtracting $M_\mathrm{R}(\omega)$ from $M_i(\omega)$, the residual signal from the i-th microphone, $R_i(\omega)$, is given by

$$\begin{aligned} R_i(\omega) &= M_i(\omega) - M_\mathrm{R}(\omega) \\ &= N(\omega)\left(e^{-j\omega\tau_i} - 1\right) \\ &= 2\left|N(\omega)\right| \sin\left(-\frac{\omega\tau_i}{2}\right) e^{j\left(\mathrm{Arg}[N(\omega)] - \frac{\omega\tau_i}{2} + \frac{\pi}{2}\right)} \end{aligned} \tag{3}$$

where $|N(\omega)|$ and $\mathrm{Arg}[N(\omega)]$ describe the amplitude and phase of the noise, respectively. Thus, the amplitude and phase of $R_i(\omega)$ are given as follows:

$$\begin{aligned} \left|R_i(\omega)\right| &= 2\left|N(\omega)\right| \sin\left(-\frac{\omega\tau_i}{2}\right), \\ \mathrm{Arg}[R_i(\omega)] &= \mathrm{Arg}[N(\omega)] - \frac{\omega\tau_i}{2} + \frac{\pi}{2}. \end{aligned} \tag{4}$$

First, we estimate the DOA of the noise signal, θ_N, by

$$\theta_N = \sin^{-1}\left(\frac{-2c(\mathrm{Arg}[R_{i+1}(\omega)] - \mathrm{Arg}[R_i(\omega)])}{\omega(d_{i+1} - d_i)}\right) \quad (5)$$

where d_i denotes the distance between $\mathrm{M_R}$ and M_i [3].

Next, we estimate $N(\omega)$ based on Eq. (4) using $\tau_i = d_i \sin(\theta_N)/c$ where c is the velocity of sound.

$$\begin{aligned} |N(\omega)| &= \frac{|R_i(\omega)|}{2\sin\left(-\frac{\omega\tau_i}{2}\right)}, \\ \mathrm{Arg}[N(\omega)] &= \mathrm{Arg}[R_i(\omega)] + \frac{\omega\tau_i}{2} - \frac{\pi}{2} \end{aligned} \quad (6)$$

The above derivation indicates that the noise signal at the $\mathrm{M_R}$ is easily estimated. Thus, the calculation cost is low.

Finally, the waveform of the noise is synthesized by superimposing an inverse fast Fourier transform, IFFT[$N(\omega)$], of all frequency bins onto it, which is then subtracted from the output of the $\mathrm{M_R}$ to restore the target signal.

2.2. Expansion of the system

As described above, the DOA of noise, θ_N, is given by Eq. (5). When there are two noise sources at the same time, however, this equation estimates an intermediate direction of the two sources. As a result, the amplitude and phase of the superimposed noise cannot be estimated correctly by Eq. (6) and the performance of noise reduction is degraded.

To overcome this problem, this study expands the method to focus the most dominant noise in a temporal frame in signal processing. Because speech signals are sparse in the time-frequency domain, the dominant signal can be specified for every frequency bin in a temporal frame. Thus, the DOA of a dominant signal can be regarded as a function of frequency, $\theta_N(\omega)$. We decided to adopt the LASSO algorithm to specify the DOA of a dominant noise source. If the experimental conditions such as the number of microphones are enough, this algorithm estimates not only the DOA but also the amplitude and phase of the noise [5]. However, we estimate only the DOA of a dominant source because the present condition of four microphones with a short array seems not enough. Based on the estimated DOA, the amplitude and phase of the dominant noise are calculated by Eq. (6) for every frequency bin independently.

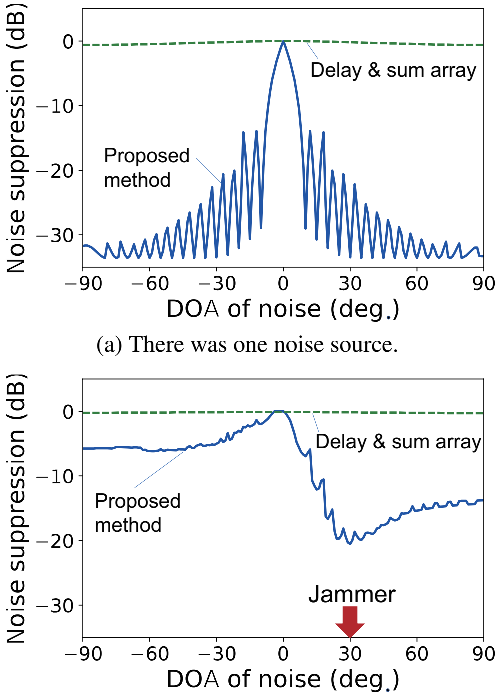
3. Evaluation of the proposed system

We conducted computer simulation experiments to evaluate the performance of the expanded system. The target, noise, and jammer signals were “Thank you very much. (female)”, “Hello, hello. (male)”, and “Welcome to Japan. (female)”, respectively. Off-axis noises were made with phase-shifting digital filters. The sampling frequency was 16 kHz, and the FFTs were conducted for 512-point temporal frames with the hanning window (frame shift: 256 points). The noise DOAs were estimated independently across frequency bins for higher than 1500 Hz, and their median was set for lower frequency bins. The LASSO estimation was implemented using the function of MultiTaskLasso (alpha = 0.5) in a Python machine learning library, scikit-learn. The bases of LASSO consist of the following vectors:

$$\left[0,\ e^{-j\omega\tau_1} - 1,\ e^{-j\omega\tau_2} - 1,\ e^{-j\omega\tau_3} - 1\right]^T$$

where these vectors were prepared for every 5°.

Figure 3 shows the obtained amount of noise suppression that is defined as the decrease in noise power in the $\mathrm{M_R}$ output. When there is one noise source, the performance is much better than the DAS beamforming (Fig. 3(a)). Because the LASSO bases were prepared in 5° steps, the amount of noise reduction is larger if the DOA of noise is near one of the bases. When another noise source (jammer) is fixed to 30°, the performance becomes poorer for the contralateral bearings (Fig. 3(b)). However, it is better than our previous system [3] in which little suppression is observed for the contralateral bearings. Thus, we can conclude that the expansion was succeeded.



(a) There was one noise source.

(b) There were two noise sources where one was fixed at 30°.

Figure 3: *Performance of noise suppression as a function of the DOA of a noise signal.*

Because the system was implemented in Python interpreter language to use a smart library of machine learning, the processing is not realtime operation. The next step of the project will be implementation of the system using a compiler language.

4. Acknowledgements

This work was supported by JSPS KAKENHI Grant (JP16K06384) and the Cooperative Res. Project (H28/A10) of the Res. Inst. of Electr. Commun., Tohoku University.

5. References

[1] J. E. Greenberg and P. M. Zurec, “Evaluation of an adaptive beamforming method for hearing aids,” *J. Acoust. Soc. America*, vol. 91, no. 3, pp. 1663–1676, 1992.

[2] J. M. Kates, “Superdirective arrays for hearing aids,” *J. Acoust. Soc. America*, vol. 94, no. 4, pp. 1930–1933, 1993.

[3] K. Ozawa, K. Amano, M. Morise, G. Shimizu, S. Oode and K. Ono, “Superdirective microphone array based on DOA and waveform estimations of noise,” *Proc. of the 5th IEEE Global Conference on Consumer Electronics (GCCE 2016)*, pp. 119–120, 2016.

[4] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Royal Statistical Soc. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] A. Xenaki and P. Gerstoft, “Compressive beamforming,” *J. Acoust. Soc. America*, vol. 136, no. 1, pp. 260–271, 2014.

Auditory profiling through computational data analysis

Raul H Sanchez, Federica Bianchi, Michal Fereczkowski, Sébastien Santurette and Torsten Dau

Hearing Systems group, DTU, Denmark

Abstract

Nowadays, the pure-tone audiogram is the main tool used to characterize the degree of hearing loss and to fit hearing aids. However, the perceptual consequences of a hearing loss are typically associated not only with a loss of sensitivity, but also with a loss of clarity (distortion loss) that is not captured by the audiogram. Detailed characterization of hearing deficits can be complex and it has to be simplified in order to efficiently investigate the specific compensation needs of individual listeners. The aim of this study is to characterize individual hearing deficits by means of a test battery that allows to capture the diverse aspects of hearing loss, considering not only the loss of sensitivity but also supra-threshold distortions.

It was hypothesized that any listeners hearing can be characterized along two dimensions: distortion type I and distortion type II. While distortion type I can be linked to factors affecting audibility, distortion type II is considered as a non-audibility-related distortion, or clarity loss. To evaluate our hypothesis, the data from two studies was re-analyzed using a data-driven approach. Both studies carried out an extensive battery of psychoacoustic tests on potential hearing-aid users. The new analysis was based on an archetypal analysis and uses unsupervised learning to identify extreme patterns in the data which provide the basis for different auditory profiles. Subsequently, a decision tree was obtained that enables a simple classification of the listeners into one of the profiles.

This novel approach provided evidence for the existence of four different "auditory profiles" in the data. The most significant predictors for the profile identification were related to temporal processing, peripheral compression, and speech perception. The current approach is promising for identifying the most relevant tests for auditory profiling and considering new fitting strategies based on the individuals deficits.

Index Terms: hearing deficits; hearing aids; hearing profile; temporal processing; supra-threshold distortions

Audio-Visual Speech Recognition for a Person with Severe Hearing Loss Using Deep Canonical Correlation Analysis

Yuki Takashima[1], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1], *Kiyohiro Omori*[2]

[1]Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan
[2]Hyogo Institute of Assistive Technology, Kobe, Japan

y.takasima@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

Recently, we proposed an audio-visual speech recognition system based on a neural network for a person with an articulation disorder resulting from severe hearing loss. In the case of a person with this type of articulation disorder, the speech style is quite different from that of people without hearing loss, making a speaker-independent acoustic model for unimpaired persons more or less useless for recognizing it. Our proposed system has shown high performance; however, some problems remain. Although the feature extraction networks are trained using the phone labels as the target class, it is difficult to obtain the correct alignment for their speech. Also, it is necessary to consider a gap between audio and visual feature spaces to treat the different modalities. In this paper, we propose a feature extraction method using deep canonical correlation analysis to tackle these weaknesses. The effectiveness of this approach was confirmed through word-recognition experiments in noisy environments, where our feature extraction method outperformed the conventional methods.

Index Terms: Speech recognition, multimodal, deep canonical correlation analysis, assistive technology

1. Introduction

In recent years, a number of assistive technologies using information processing have been proposed; for example, sign language recognition using image recognition technology [1] and text reading systems from natural scene images [2]. In this study, we focused on communication-assistive technology for a physically unimpaired person to enable him or her to communicate with a person with an articulation disorder resulting from severe hearing loss.

Some people with hearing loss who have received speech training, or who lost their hearing after learning to speak, can communicate using spoken language. However, in the case of automatic speech recognition (ASR), their speech style is so different from that of people without hearing loss that a speaker-independent (audio-visual) ASR model for unimpaired persons is of little use for recognizing such speech as described in Section 5.1. Matsumasa *et al.* [3] researched an ASR system for articulation disorders resulting from cerebral palsy, and reported the same problem. Najnin*et al.* [4] investigated the relationship between a hearing-impaired individual's speech and his hearing loss.

The performance of speech recognition systems generally degrades in a noisy environment. For people with hearing loss, because they do not hear ambient sound, they cannot control the volumes of their voices and their speaking style in a noisy environment, and it is difficult, those who are physically unimpaired, to recognize utterances using only the speech signal. In such cases, we try to read the lips of the speaker to compensate for the reduction in recognition accuracy. For people with hearing problems, lip reading is one communication skill that can help them communicate better. In the field of speech processing, audio-visual speech recognition has been studied for robust speech recognition under noisy environments [5, 6, 7]. In this paper, we investigate an audio-visual speech recognition approach for articulation disorders resulting from severe hearing loss.

Recently, we proposed bottleneck feature extraction [8] from audio and visual features for a hearing-impaired person using convolutive bottleneck networks (CBN), which stack multiple layers of various types (such as a convolution layer, a pooling layer, and a bottleneck layer) [9] forming a deep network. Thanks to the convolution and pooling operations, we can train the convolutional neural network (CNN) robustly to deal with the small local fluctuations of an input feature map. In some tandem approaches using deep learning, an output layer plays a classification role, and output units are used as a feature vector for a recognition system, where phone labels are used as a teaching signal for an output layer. However, in the case of an articulation disorder, the phone label estimated by forced alignment may not be correct. An approach based on CBN [10] uses a bottleneck layer as a feature vector for a recognition system, where the number of units is extremely small compared to the adjacent layers, following the CNN layers. Therefore, the bottleneck layer is a better feature than an output layer, which is strongly influenced by some wrong phone labels because it is expected that the bottleneck layer can aggregate the propagated information and extract fundamental features included in an input map. In this paper, we investigate another approach to tackle this alignment problem—unsupervised learning.

In multi-view learning, deep canonical correlation analysis (DCCA) [11], which is nonlinearly-extended canonical correlation analysis (CCA), has been proposed. DCCA has two deep neural networks and simultaneously learns nonlinear mappings (both networks) of two modalities that are maximally correlated. CCA is a statistical method for dealing with the correlation between sets of two variables, finding linear projection vectors. Unlike CCA, DCCA is a parametric method, and it can learn the complex transformations of two views. DCCA has been applied to several audio classification tasks [12, 13], and improved [14]. The DCCA objective function is optimized in an unsupervised manner over the actual data; therefore, it is not necessary to use some wrong phone labels for training networks.

In most multimodal speech recognition systems, audio and visual features are integrated by just concatenating these features. Because the audio and visual features are intrinsically different, and a gap between audio and visual feature spaces may cause undesirable effects in speech recognition. Applying

DCCA, gaps between two feature spaces are reduced, and we expect to obtain more complementary features for speech recognition. We will show in this paper that our proposed feature can achieve better recognition performance in noisy environments.

The rest of this paper is organized as follows: In Section 2, we review CCA and DCCA. In Section 3, our proposed method is explained. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Preliminaries

In this section, we review CCA and DCCA, where two views represent the audio and visual features.

2.1. Canonical Correlation Analysis

Let $\mathrm{X}_{audio} \in \mathbb{R}^{d_1 \times N}$, $\mathrm{X}_{visual} \in \mathbb{R}^{d_2 \times N}$ denote audio and visual features with N samples where the sample mean of these matrices is normalized to zero, and d_1 and d_2 represent the dimension of the audio and visual features, respectively. In CCA, the correlation coefficient is calculated as follows:

$$\begin{aligned}\rho(\boldsymbol{a}, \boldsymbol{b}) &= \mathrm{corr}(\boldsymbol{a}^\top \mathrm{X}_{audio}, \boldsymbol{b}^\top \mathrm{X}_{visual}) &(1)\\ &= \frac{\boldsymbol{a}^\top \Sigma_{av} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \Sigma_{aa} \boldsymbol{a}} \sqrt{\boldsymbol{b}^\top \Sigma_{vv} \boldsymbol{b}}}, &(2)\end{aligned}$$

where $\boldsymbol{a} \in \mathbb{R}^{d_1}$, $\boldsymbol{b} \in \mathbb{R}^{d_2}$ are the projection vectors, which are parameters of CCA, and $\Sigma_{av} \in \mathbb{R}^{d_1 \times d_2}$, $\Sigma_{aa} \in \mathbb{R}^{d_1 \times d_1}$, $\Sigma_{vv} \in \mathbb{R}^{d_2 \times d_2}$ are the cross-covariance matrices of X_{audio} and X_{visual}, the covariance matrix of X_{audio} and X_{visual}, respectively. Since $\rho(\boldsymbol{a}, \boldsymbol{b})$ is invariant to scalling of $\boldsymbol{a}$ and $\boldsymbol{b}$, we assume that each standard variance of denominator in Eq. (2) has one; that is the projections are constrained to have unit variance,

$$\max_{\boldsymbol{a}, \boldsymbol{b}} \boldsymbol{a}^\top \Sigma_{av} \boldsymbol{b} \ \text{ subject to } \ \boldsymbol{a}^\top \Sigma_{aa} \boldsymbol{a} = \boldsymbol{b}^\top \Sigma_{vv} \boldsymbol{b} = 1 \qquad (3)$$

If we use $L \leq min(d_1, d_2)$ pairs of linear projection vectors, the projection matrices for audio and visual features are formed as $\mathrm{U} \in \mathbb{R}^{d_1 \times L}$ and $\mathrm{V} \in \mathbb{R}^{d_2 \times L}$, respectively. We obtain the following formulation to identify the projection matrices A and B:

$$\begin{aligned}&\text{maximize } \ \mathrm{tr}(\mathrm{A}^\top \Sigma_{av} \mathrm{B}) &(4)\\ &\text{subject to } \ \mathrm{A}^\top \Sigma_{aa} \mathrm{A} = \mathrm{B}^\top \Sigma_{vv} \mathrm{B} = \mathrm{I},\end{aligned}$$

where $\mathrm{tr}(\cdot)$ and I indicate the sum of the elements on the main diagonal and the unit matrix, respectively.

The optimal objective value is the sum of the top k singular values of $\mathrm{T} = \Sigma_{aa}^{-1/2} \Sigma_{av} \Sigma_{vv}^{-1/2}$. The optimal projection matrices are given by $(\mathrm{A}, \mathrm{B}) = (\Sigma_{aa}^{-1/2} \mathrm{U}_k, \Sigma_{vv}^{-1/2} \mathrm{V}_k)$, where $\mathrm{U}_k \in \mathbb{R}^{d_1 \times k}$ and $\mathrm{V}_k \in \mathbb{R}^{d_2 \times k}$ are the first k left- and right-singular vectors of T. Indeed, the covariance matrices Σ_{aa} and Σ_{vv} are estimated from data using regularization so that they are constrained to the nonsingular matrix.

2.2. Deep Canonical Correlation Analysis

DCCA computes the representations of the two views by passing them through multiple stacked layers of nonlinear transformation. Given the audio and visual features (X_{audio},X_{visual}), the outputs of the audio and visual neural networks are written as $f(\mathrm{X}_{audio}; \theta_1) \in \mathbb{R}^{o \times N}$, $f(\mathrm{X}_{visual}; \theta_2) \in \mathbb{R}^{o \times N}$, respectively. θ_1, θ_2 indicate parameters of the audio and visual networks, respectively. DCCA computes the total correlation as follows:

$$corr(\boldsymbol{a}^\top f(\mathrm{X}_{audio}; \theta_1), \boldsymbol{b}^\top f(\mathrm{X}_{visual}; \theta_2)) = \mathrm{tr}(\mathrm{T}^\top \mathrm{T})^{\frac{1}{2}}, \qquad (5)$$

where $\mathrm{T} = \hat{\Sigma}_{aa}^{-1/2} \hat{\Sigma}_{av} \hat{\Sigma}_{vv}^{-1/2}$ as reviewed in section 2.1. $\hat{\Sigma}_{av} = \frac{1}{N-1} \mathrm{X}_{audio} \mathrm{X}_{visual}^\top$ and $\hat{\Sigma}_{aa} = \frac{1}{N-1} \mathrm{X}_{audio} \mathrm{X}_{audio}^\top + r_1 \mathrm{I}$ are the covarince matrices with regularization constant $r_1 > 0$, similarly for $\hat{\Sigma}_{vv}$. The goal of DCCA is to jointly learn parameters $\{\theta_1, \theta_2, \boldsymbol{u}, \boldsymbol{v}\}$ for both views, such that the correlation is as high as possible. The parameters $\{\theta_1, \theta_2\}$ are trained using back-propagation. The gradient of Eq. 5 can be computed as follows:

$$\begin{aligned}&\frac{\partial corr(\boldsymbol{a}^\top f(\mathrm{X}_{audio}; \theta_1), \boldsymbol{b}^\top f(\mathrm{X}_{visual}; \theta_2))}{\partial f(\mathrm{X}_{audio}; \theta_1)}\\ &= \frac{1}{N-1}(2\nabla_{aa} \mathrm{X}_{audio} + \nabla_{av} \mathrm{X}_{visual}) \qquad (6)\end{aligned}$$

where $\nabla_{ab} = \hat{\Sigma}_{aa}^{-1/2} \mathrm{U} \mathrm{V}^\top \hat{\Sigma}_{vv}^{-1/2}$ and $\nabla_{aa} = -\frac{1}{2} \hat{\Sigma}_{aa}^{-1/2} \mathrm{U} \mathrm{D} \mathrm{V}^\top \hat{\Sigma}_{aa}^{-1/2}$, and the derivative with respect to $f(\mathrm{X}_{visual}; \theta_2)$ has a symmetric expression.

General DNN objective functions are written as the expectation (or sum) of error functions (e.g., squared loss) calculated for each training sample. This property naturally suggests stochastic gradient descent (SGD) for optimization, where gradients are estimated for a few training examples (a mini-batch) and iteratively updated parameters. However, in DCCA (Eq. 5), it is necessary to estimate the covariance matrices for the training samples. Andrew et al. [11] used a full-batch algorithm (L-BFGS) for optimization. This is undesirable for applications with large training sets, as each gradient step computed on the entire training set can be very expensive in both memory and time. To mitigate this problem, Wang et al. [12] showed that it works well, even for this type of objective, if larger mini-batches are used. It is considered that a large mini-batch has enough information to estimate covariances. Hence, in this paper, we also configure a larger mini-batch size.

3. Related Works

Deep learning has had recent successes for acoustic modeling [15]. Deep neural networks (DNNs) contain many layers of nonlinear hidden units. The key idea is to use greedy layer-wise training with restricted Boltzmann machines (RBMs) followed by fine-tuning. Ngiam *et al.* [16] proposed multimodal DNNs that learn features over audio and visual modalities. Mroueh *et al.* [17] improved this method and proposed an architecture considering the correlations between modalities. Ninomiya *et al.* [6] investigated integration of bottleneck features using multi-stream hidden Markov models (HMMs) for audio-visual speech recognition.

CNNs also have demonstrated impressive performance on several tasks, such as image analysis [18, 19, 20] and spoken language [21] and music recognition [22]. In our previous work [8], we showed that the features extracted from CNNs lead to effective results for speech recognition thanks to the properties of the local receptive field and the shift invariant. Therefore, in this paper we do not use DNNs, but CNNs, for nonlinear mappings of two modalities.

Recently, multimodal learning has been researched in relation to discovering useful information about the world. If such methodology can be used to develop an accurate system, we would be able to obtain non-verbal information that cannot, at

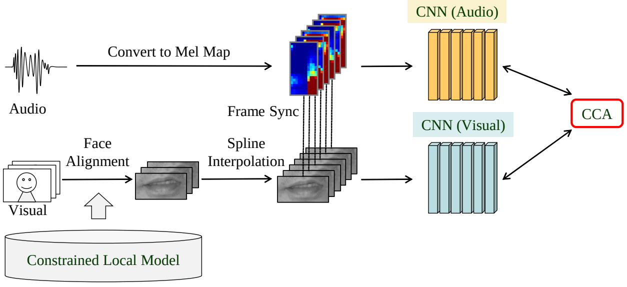


Figure 1: *Deep CCA using CNNs*

this time, be explained expressly and cannot be obtained from discriminative models. Unsupervised learning is an approach to can be used to handle this problem. For multimodal fusion tasks, several approaches have been proposed [23, 24] where modalities are modeled using a generative model that is based on an RBM. For speech recognition tasks, the reproducibility of the input data is not necessary due to the fact that a DCCA approach is concise. In this paper, we employ DCCA with CNNs as a robust feature extractor for the fluctuation of the speech uttered by a person with cerebral palsy.

4. Multimodal Feature Extraction Using DCCA

4.1. Flow of the Proposed Method

Figure 1 shows the flow of our proposed feature extraction. To employ advantages of our previous work [8], we use CNNs for the mappings of DCCA instead of DNNs. Hereafter, $f(\cdot;\theta)$ in section 2.2 indicates a CNN operation where the input is two-dimensional.

First, we prepare the input features for training a CNN from lip images and speech signals uttered by a person with hearing loss. For the audio signals, after calculating short-term mel spectra from the signal, we obtain mel-maps by merging the mel spectra into a 2D feature with several frames, allowing overlaps.

The visual signals of the eyes, mouth, nose, eyebrows, and outline of the face are aligned using the point distribution model (PDM), and its model parameter is estimated by constrained local model (CLM). Then, a lip image is extracted, and the extracted lip image is interpolated to fill the sampling rate gap between visual features with respect to audio features. In this paper, we adopted spline interpolation to the lip images.

The parameters of audio and visual CNNs are jointly learned by back-propagation with SGD where the gradients are calculated by DCCA objective function, starting from random values. Following the training of both CNNs, the input mel-map and lip images are transformed to the output units through each CNN, and projected linearly as follows:

$$\boldsymbol{\alpha}_t = \hat{\Sigma}_{aa}^{-1/2}\mathrm{U}_k f(\mathrm{X}_t;\theta_1) \tag{7}$$

$$\boldsymbol{\beta}_t = \hat{\Sigma}_{vv}^{-1/2}\mathrm{V}_k f(\mathrm{Y}_t;\theta_2), \tag{8}$$

where $(\mathrm{X}_t, \mathrm{Y}_t)$ are two-dimensional input feature for audio and visual at time t, and ($\boldsymbol{\alpha}_t \in \mathbb{R}^k$, $\boldsymbol{\beta}_t \in \mathbb{R}^k$) are the corresponding features, respectively. Then these features are concatenated, and $[\boldsymbol{\alpha}_t^\top\ \boldsymbol{\beta}_t^\top]^\top \in \mathbb{R}^{2k}$ is used as the feature in the training of HMMs for speech recognition.

4.2. Application to Speech Uttered by a Person with Hearing Loss

DCCA has an advantage for speech uttered by a person with hearing loss. In the case of an articulation disorder, the phone label estimated by forced alignment may not be correct. However, several approaches based on DNN use the phone label as the target class to learn parameters. The DCCA accomplishes the training procedure in an unsupervised fashion to find the maximal correlation between two sets of modalities. Therefore, the feature extracted from networks trained by DCCA is not influenced by some wrong phone labels. By using DCCA, audio and visual features are transformed through networks so that output units have a high correlation. In noisy environments, we expect that even if the audio feature is degraded, the transformed feature has adequate robustness because this feature has a high correlation to the visual feature which is noise-invariant.

5. Experiments

5.1. Recognition Results Using a Speaker-independent Acoustic Model

At the beginning, we attempted to recognize utterances using a speaker-independent acoustic model for unimpaired people (This model is included in Julius [25]). The acoustic model consists of a triphone HMM set with 25-dimensional MFCC features (12-order MFCCs, their delta and energy) and 16 mixture components for each state. Each HMM has three states and three self-loops. For a person with hearing loss, a recognition rate of only 3.24% was obtained, but for a physically-unimpaired person, a recognition rate of 88.89% was obtained for the same task. It is clear that the speaking style of a person with hearing loss differs considerably from that of a physically-unimpaired person. Therefore, it is considered that a speaker-dependent acoustic model is necessary for recognizing speech from a person with hearing loss.

5.2. Word Recognition Experiments

5.2.1. Experimental Conditions

Our proposed method was evaluated on word recognition tasks. We recorded utterances of one male person with hearing loss, where the text is the same as the ATR Japanese speech database A-set [26]. We used 2,620 words as training data, and 216 words as test data. The utterance signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 5 msec. For the acoustic-visual model, we used the monophone-HMMs (54 phonemes) with 3 states and 6 mixtures of Gaussians. We compare our audio-visual feature with conventional MFCC+Δ+ΔΔ (36-dimensions) and MFCC+Δ+ΔΔ+ discrete cosine transform (DCT) (66-dimensions). Then, our proposed method and audio-visual features were evaluated in noisy environments. White noise was added to audio signals and their SNR is set to 20dB, 10dB, and 5dB. Audio CNN and HMMs are trained using the clean audio feature.

5.2.2. Architecture of the Networks

We construct deep networks, which consist of a convolution layer, a pooling layer, and fully-connected MLPs. For the input layer of audio CNN, we use a mel-map of subsequent 13-frames with 39-dimensional mel spectrum, and the frame shift is 5 msec. For the input layer of visual CNN, frontal face videos are recorded at 60 fps. Luminance images are extracted from

Table 1: *Filter size, number of feature maps and number of MLPs units for each architecture. The value for C indicates the filter size of the convolution layer that has #1 maps. The convolution layer is associated with the pooling layer. The value of S means the pooling factor. The value for M indicates the number of units for each layer in the MLP part.*

	Input	C	S	#1	M
Audio CNN	39×13	4×2	3×3	13	108, 30, 108
Visual CNN	12×24	5×5	2×2	13	108, 30, 108

the image using CLM and resized to 12×24 pixels. Finally, the images are up-sampled by spline interpolation and input to the CNN.

Table 1 shows parameters used in experiments. We set the bottleneck layer into networks to investigate the performance of bottleneck features. In the training procedure, a learning rate and a momentum are set to be 0.0001 and 0.99, respectively.

5.2.3. *Number of Mini-batch Sizes*

In the preliminary experiment, we compared the effects of changing the number of mini-batches with 50 epochs in a clean environment. Table 2 shows the results when changing the number of mini-batches as 1,200, 1,500, 1,800, 2,100 and 2,400. Through the experiments, we found that the performance improves as the number of mini-batches increased. The reason for the improvement is attributed to being able to estimate the covariance matrix more accurately when using larger mini-batch sizes. In the future experiments, we will use a mini-batch size of 2,100.

Table 2: *Word recognition accuracy for each mini-batch size*

# of mini-batches	1,200	1,500	1,800	2,100	2,400
Recognition accuracy [%]	63.89	65.28	66.20	71.76	71.76

5.2.4. *Results and Discussion*

Figure 2 shows the word recognition accuracies in noisy environments. We compared the audio-visual feature extracted from our proposed method with two conventional features: MFCC+Δ+ΔΔ (MFCC), MFCC+Δ+ΔΔ+DCT (MFCC+DCT). In Figure 2, DCCA and DCCA bottleneck denote the features extracted from the final projection layer and the bottleneck layer (30-dimensions). Comparing DCCA bottleneck with DCCA, the former shows better accuracies. This is because the information that the audio feature has is lost when it is transformed to near the visual space. The DCCA bottleneck feature is better than MFCC+DCT in SNR of 10dB. This might be because the DCCA bottleneck feature obtained more noise-robustness compared with the conventional feature. These results show our proposed method improves performance.

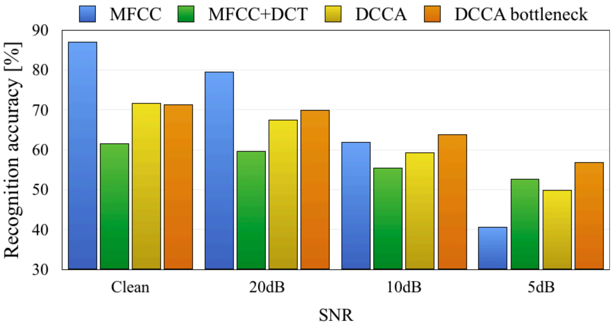


Figure 2: *Word recognition accuracy using HMMs*

Figure 3 shows the word recognition accuracies comparing our proposed method with the previous method [8]. The DCCA framework is the unsupervised learning that is applied to the actual data in order to find the maximal correlation between two sets of modalities without other information. Therefore, the extracted feature might not be able to present the phonological information. Our previous work employed supervised learning using the phone labels. In our experiments, the accuracy of the DCCA bottleneck degraded on average 14% compared to using supervised learning.

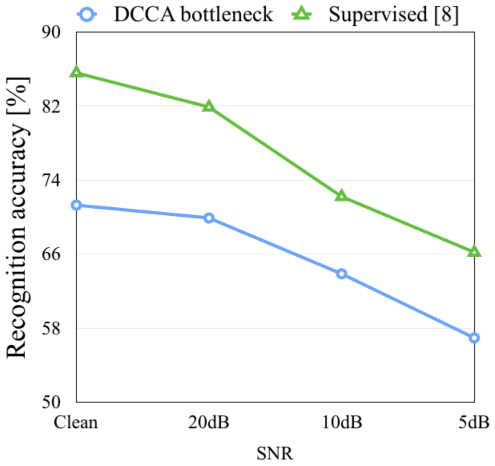


Figure 3: *Word recognition accuracy of unsupervised and supervised training procedure*

6. Conclusions

In this paper, we discussed an audio-visual speech recognition system for a person with an articulation disorder resulting from severe hearing loss based on CNNs. We proposed a feature extraction method using CNNs trained by deep CCA which is optimized in an unsupervised manner. In the DCCA procedure, audio and visual CNNs are trained maximizing the correlation between the units of each output layer. When a noisy input signal is fed to CNNs, although the audio feature is degraded, the visual feature compensates for the degraded audio feature data to increase accuracy. Then, the degradation of accuracy is restrained in high-noisy environments.

In comparison, in experiments between the proposed feature and a conventional unsupervised feature (MFFC+DCT), the proposed feature showed better performances than the conventional one. The improvement was more significant in high-noisy environments. However, the performance of the proposed method was lower than that of the supervised method. This result suggests that using DCCA, the phonological information are not necessarily extracted.

In future work, we will further investigate a better DCCA-based feature extraction which is also highly correlated to the phonological information. Person with an articulation disorder resulting from severe hearing loss need the various applications for communication, for example, voice-to-signal conversion system. Although their speech style is so different from that of people without hearing loss, they can make appropriate lip shapes. Therefore, the voice-to-signal conversion system is able to help the interaction with others. We will also research this system in future work.

7. Acknowledgements

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

8. References

[1] J. Lin, Y. Wu, and T. S. Huang, "Capturing human hand motion in image sequences," in *Workshop on Motion and Video Computing*, 2002, pp. 99–104.

[2] N. Ezaki, M. Bulacu, and L. Schomaker, "Text detection from natural scene images: Towards a system for visually impaired persons," in *ICPR*, 2004, pp. 683–686.

[3] H. Matsumasa, T. Takiguchi, Y. Ariki, I. chao Li, and T. Nakabayashi, "Integration of metamodel and acoustic model for dysarthric speech recognition," *Journal of Multimedia*, pp. 254–261, 2009.

[4] S. Najnin, B. Banerjee, L. L. Mendel, M. H. Kapourchali, J. K. Dutta, S. Lee, C. Patro, and M. Pousson, "Identifying hearing loss from learned speech kernels," in *INTERSPEECH*, 2016, pp. 243–247.

[5] M. Tomlinson, M. Russell, and N. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *ICASSP*, 1996, pp. 821–824.

[6] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," in *INTERSPEECH*, 2015.

[7] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *CoRR*, vol. abs/1611.01599, 2016.

[8] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, and K. Nakazono, "Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss," in *INTERSPEECH*, 2016, pp. 277–281.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665

[10] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *ICSP*, 2014, pp. 505–509.

[11] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[12] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *ICASSP*, 2015, pp. 4590–4594.

[13] N. E.-D. El-Madany, Y. He, and L. Guan, "Multiview learning via deep discriminative canonical correlation analysis," in *ICASSP*, 2016, pp. 2409–2413.

[14] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multi-view representation learning: Objectives and optimization," *CoRR*, vol. abs/1602.01024, 2016.

[15] G. Hinton, D. Li, Y. Dong, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82 – 97, 2012.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696.

[17] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *ICASSP*, 2015, pp. 2130–2134.

[18] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1408–1423, 2004.

[19] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," in *VISAPP (2)*, 2008, pp. 290–294.

[20] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.

[21] G. Montavon, "Deep learning for spoken language identification," in *NIPS Workshop on deep learning for speech recognition and related applications*, 2009.

[22] T. Nakashika, C. Garcia, and T. Takiguchi, "Local-feature-map integration using convolutional neural networks for music genre classification," in *INTERSPEECH*, 2012.

[23] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2949–2980, 2014.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011, pp. 689–696.

[25] "Open-Source Speech Recognition Software Julius," http://julius.sourceforge.jp/.

[26] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

Towards GCC-NMF Speech Enhancement for Hearing Assistive Devices: Reducing Latency with Asymmetric Windows

Sean UN Wood, Jean Rouat

NECOTIS, Department of Electrical and Computer Engineering
Université de Sherbrooke, Québec, Canada
sean.wood@usherbrooke.ca, jean.rouat@usherbrooke.ca

Abstract

We present a modified version of the real-time GCC-NMF stereo speech enhancement algorithm that drastically reduces the inherent system latency by incorporating an asymmetric STFT windowing strategy. Long analysis windows retain the high spectral resolution required by GCC-NMF, while short synthesis windows significantly reduce the overall system latency. We show that GCC-NMF speech enhancement quality is relatively unaffected by this windowing strategy, with the overall objective PEASS score remaining stable for varying system latencies. The asymmetric windowing technique comes at a cost of increased computational load, with shorter synthesis windows requiring a shorter frame advance, thus increasing the number of windows to be processed. We present an analysis of the computational requirements of GCC-NMF to run in real-time on a variety of hardware platforms including the Raspberry Pi and the NVIDIA Jetson TX1. All tested systems are fast enough to achieve latencies at least as low as 24 ms with small NMF dictionaries of 64 atoms, while the fastest NVIDIA K40 GPU system is capable of achieving 6 ms latency with a large dictionary of 1024 atoms.

Index Terms: real-time, latency, speech enhancement, source separation, GCC-NMF, GCC, NMF, GCC-PHAT, CASA

1. Introduction

A wealth of speech enhancement algorithms designed to suppress noise and reverberation have been developed in fields such as speech coding, automatic speech recognition, and source separation. Many such algorithms, however, remain inapplicable in the context of hearing assistive devices due to both inherent algorithmic latency and computational performance on low-power hardware. We address these hurdles here with respect to the real-time GCC-NMF speech enhancement algorithm we introduced recently [1, 2].

Many speech enhancement algorithms including GCC-NMF are built around the short-time Fourier transform (STFT) with which sound is processed in short, overlapping segments of time [3]. A consequence of the traditional STFT is an inherent algorithmic latency where, independent of processing speed, there exists a trade-off between spectral resolution and the delay between the system's input and output. With many algorithms relying on high spectral resolution, latencies greater than 64 ms are common. In the context of assistive listening devices, however, such high latencies are perceived as objectionable echoes as a superposition of both the aided and unaided sounds are heard by the listener [4]. Depending on the type and severity of hearing loss, delays below 15 to 32 ms are likely required to be tolerable [5, 6], with delays less than 10 ms being a reasonable objective in the general case [7, 8].

In this work, we integrate the asymmetric STFT windowing approach proposed by Mauler and Martin [9] into the GCC-NMF speech enhancement system, simultaneously providing high spectral resolution and latencies well below 10 ms, depending on available computational power. An alternative approach to low delay speech enhancement was developed by Löllmann and Vary using low delay filter banks [10, 11]. We begin with a review of the real-time GCC-NMF speech enhancement algorithm in Section 2, followed by a description of the asymmetric STFT windowing method in Section 3. We then demonstrate the robustness of GCC-NMF speech enhancement quality to latency reduction with asymmetric windowing, as well as an analysis of the computational requirements of GCC-NMF on a variety of hardware platforms in Section 4, followed by the conclusion in Section 5.

2. Real-time GCC-NMF

The GCC-NMF stereo speech enhancement algorithm combines the non-negative matrix factorization (NMF) unsupervised dictionary learning algorithm [12] with the generalized cross-correlation (GCC) spatial localization method [13]. GCC-NMF is flexible in terms of microphone separation, where separations ranging from 5 cm to 1 m have been tested previously [1]. NMF provides a parts-based representation of the input mixture signal in terms of *dictionary atoms* in the magnitude frequency domain, while GCC provides an estimate of the time delay of arrival (TDOA) of each dictionary atom, at each point in time. The NMF dictionary atoms estimated to originate from the direction of interest are recombined and used to construct a Wiener-like filter, as is typical for NMF-based speech enhancement [14]. The resulting filter is applied to the mixture signal to yield the system output. For offline speech enhancement, the NMF dictionary may be learned directly from the mixture signal, while in the online case, it is pre-learned from isolated speech and noise signals using a different dataset than used at test time, generalizing to new speakers, acoustic and noise conditions, and recording setups [2].

Online GCC-NMF speech enhancement is performed on a frame-by-frame basis given the pre-learned NMF dictionary W_{fd} (with f indexing frequency and d indexing the dictionary atoms), the complex-valued left and right Fourier-transformed frames V_{lf} and V_{rf}, a set of possible TDOAs indexed by τ, and the target direction τ_s estimated using an accumulated GCC-PHAT localization process [2]. First, the GCC-NMF angular spectrum $G_{d\tau}^{\mathrm{NMF}}$, is constructed for each dictionary atom,

$$G_{d\tau}^{\mathrm{NMF}} = \sum_f W_{fd} \,\mathrm{Re}\left(\frac{V_{lf} V_{rf}^*}{|V_{lf}|\,|V_{rf}|} e^{j2\pi f\tau} \right) \tag{1}$$

where for a given atom d, $G_{d\tau}^{\mathrm{NMF}}$ is a function of τ that will be high for values of τ near the estimated direction of arrival. A binary atom mask M_d is then constructed given the estimated target direction τ_s,

$$M_d = \begin{cases} 1 & \text{if } \left|\tau_s - \text{argmax}_\tau G^{\text{NMF}}_{d\tau}\right| < \epsilon/2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

such that only atoms whose GCC-NMF angular spectrum reaches its peak within a window of size ϵ of the target TDOA are accepted. Finally, a Wiener-like filter is constructed as the ratio of the sum of unmasked atoms $\sum_d W_{fd}M_d$ to the sum of all atoms $\sum_d W_{fd}$, and is used to filter input signals resulting in the enhanced speech signal $\widehat{X}_{cf}$, where c indexes the left and right channels,

$$\widehat{X}_{cf} = \frac{\sum_d W_{fd}M_d}{\sum_d W_{fd}} V_{cf} \tag{3}$$

With the online GCC-NMF speech enhancement algorithm now defined, we proceed to show how the underlying STFT imposes an lower limit on its real-time latency, and how the asymmetric STFT windowing method mentioned previously may be used to drastically reduce this latency.

3. Asymmetric STFT Windowing

3.1. STFT and Latency

The STFT processes sound in short, overlapping segments of time called *frames*. Each frame is multiplied by an *analysis window* prior to computing its Fourier transform. Resynthesis is achieved by taking the inverse Fourier transform of the transformed frame, multiplying the resulting samples by a *synthesis window*, and combining neighbouring frames via the overlap-add (OLA) method. Perfect reconstruction can be achieved if the transform has the constant overlap-add (COLA) property, i.e. if the overlapped sum of the product of the analysis and synthesis windows is constant over time [15]. A commonly used window for analysis and synthesis is the square root of the periodic Hann window, where the periodic Hann function is defined for frame size N as,

$$\text{H}_N[n] = \begin{cases} \frac{1}{2}\left(1 - \cos\left(2\pi\frac{n}{N}\right)\right) & 0 \le n < N \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The above process of overlapped signal windowing with OLA resynthesis induces a latency L_{OLA} equal to the window size N. In order to run in real-time, all processing including the Fourier transform and its inverse, should occur within a single frame advance R, resulting in a total system latency of $N + R$ [15]. We previously presented the real-time GCC-NMF separation system on input signals sampled at 16 kHz, with a window size of 1024 samples with varying frame advance, resulting in a total latency of 64 ms plus 8 ms with a frame advance of 128 samples, for example. As described in Section 1, latencies this large are unsuitable for real-world use in hearing assistive devices.

A first approach to reduce the GCC-NMF system latency is to simply reduce the window size N. This comes at the expense of decreasing the spectral resolution, however, and as we will show in Section 4.2, GCC-NMF speech enhancement quality decreases significantly for small window sizes with this approach. We therefore present another approach to latency reduction based on an asymmetric STFT windowing method that combines long analysis windows with short synthesis windows.

3.2. Asymmetric STFT windowing

Departing from the tradition of symmetric analysis and synthesis windows that have the same duration, asymmetric windowing allows us to simultaneously achieve high spectral resolution and low latency by combining long analysis windows with relatively short synthesis windows. The asymmetric windows we use in this work have been adapted from the more general case proposed by Mauler and Martin [9], though other asymmetric windowing approaches can be found in the literature [16, 17, 18, 19].

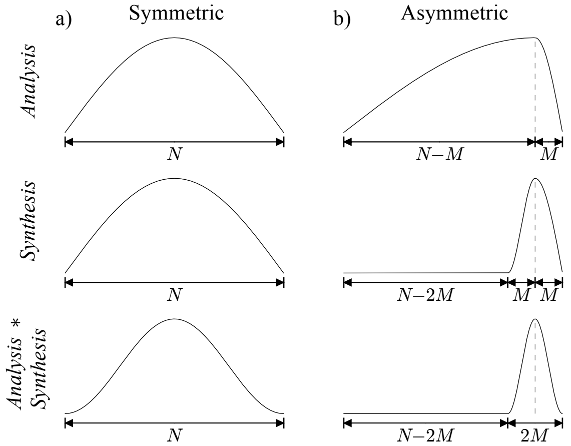


Figure 1: *Comparison of the symmetric and asymmetric STFT window functions for frame size N. a) Traditional symmetric square root Hann analysis and synthesis window functions and their product Hann window, all having duration N. b) Asymmetric window functions, where the analysis window has duration N and is weighted towards the right, while the synthesis window has duration $2M{<}N$, and shares its right edge with the underlying frame. The resulting product of the analysis and synthesis windows is a Hann window of size $2M$ that also shares its right edge with the underlying frame.*

For a given frame size N, the asymmetric analysis and synthesis windows are designed such that their point-wise product is a Hann window of size $2M < N$. This Hann window shares its right edge with the underlying frame, and can be made to be much shorter than the frame itself by choosing $2M \ll N$, as depicted in Figure 1. The analysis window h_A and the synthesis window h_S are defined mathematically as,

$$h_A[n] = \begin{cases} \sqrt{\text{H}_{2(N-M)}}[n] & 0 \le n < N{-}M \\ \sqrt{\text{H}_{2M}}[n{-}\,(N{-}2M)] & N{-}M \le n < N \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$h_S[n] = \begin{cases} \frac{\sqrt{\text{H}_{2M}}[n-(N-2M)]}{\sqrt{\text{H}_{2(N-M)}}[n]} & N{-}2M \le n < N{-}M \\ \sqrt{\text{H}_{2M}}[n{-}(N{-}2M)] & N{-}M \le n < N \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

These window functions are constructed in two parts with respect to the center of the analysis-synthesis product Hann window, i.e. $N{-}M$. To the right of $N{-}M$, both analysis and synthesis windows consist of the right half of a square root Hann window of size $2M$. To the left, the analysis window consists of the left half of a Hann window of size $N{-}M$, while the synthesis window is defined as the ratio of the analysis window and the product Hann window, limited to the range $N-2M \leq n < N-M$.

In Figure 2, we compare the traditional STFT windowing method using square root Hann windows with the asymmetric STFT windowing presented above. The analysis window size is N in both cases, and the asymmetric synthesis window size is set to $N/4$, i.e. with $M=N/8$. In both cases, the window overlap is 50% of the synthesis window, such that perfect reconstruction (PR) is achieved. We note that retaining the relative synthesis window overlap while decreasing the synthesis window size results in a significant increase in the number of windows required for the overlap-add windowing process, thus increasing the computational load. We also note that this approach increases the start-up latency of the STFT, though this may mitigated by simply pre-padding the signal with zeros.

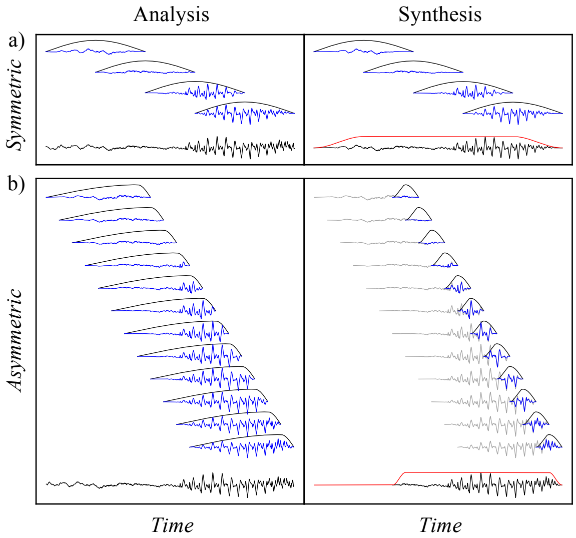


Figure 2: *Comparison of analysis and synthesis processes for a) symmetric and b) asymmetric windowing functions, given a synthesis window overlap of* 50%. *For the analysis stage (left), the input signal and time-shifted analysis window functions are shown in black, with the resulting windowed frames shown in blue. For synthesis stage (right), the time-shifted synthesis window functions shown in black, the reconstructed frames are shown in gray, and the resulting windowed frames used for overlap-add resynthesis are overlaid in blue. The overlap-add result is then shown below in black, with the normalized overlapped sum of the anslysis-synthesis window products in red.*

4. Experiments

In this section, we first compare the effect of latency reduction using the symmetric and asymmetric windowing methods on the learned NMF dictionary atoms, followed by the effect on GCC-NMF speech enhancement quality. We then study the empirical processing time requirements of GCC-NMF for a variety of hardware platforms, to determine the conditions under which the proposed low-latency system may be run in real-time on currently available hardware.

We reuse here the data and evaluation metrics as presented in the development of the real-time GCC-NMF speech enhancement algorithm [2]. Unsupervised training data consists of a small subset of the speech and noise signals from the CHiME challenge [20], taken as 4096 randomly chosen frames divided equally between speech and noise signals from a single microphone. Evaluation data consists of the two-channel mixtures of speech and real-world noise from the SiSEC speech enhancement challenge [21], where the microphones are separated by 8.6 cm, though as mentioned previously, GCC-NMF has been tested for microphone separations ranging from 5 cm to 1 m [1]. Both datasets are sampled at 16 kHz. Speech enhancement quality is quantified with the Perceptual Evaluation methods for Audio Source Separation (PEASS) toolkit [22], designed to better correlate with subjective assessments than the traditional SNR-based metrics. PEASS metrics consist of four scores quantifying the overall enhancement quality, target fidelity, interference suppression, and lack of perceptual artifacts, where higher scores are better for all measures. Future work will include a wider range of evaluations metrics including measures of speech intelligibility, STOI [23] and ESTOI [24]. The default NMF dictionary size for the experiments that follow is 1024, while the default STFT synthesis window overlap is 75%.

4.1. Effect on NMF dictionary atoms

As described in Section 3.1, the inherent latency of the real-time GCC-NMF speech enhancement algorithm using traditional symmetric STFT windowing may be reduced by simply reducing the STFT frame size N. An undesired consequence of this approach, however, is a reduction spectral resolution, as decreasing the STFT frame size results in increasingly wideband spectrograms. In Figure 3a), we depict example NMF dictionary atoms learned for varying symmetric STFT window size, noting that as the window size is decreased, dictionary atoms become increasing wideband, and the spectral details captured with longer duration windows are lost.

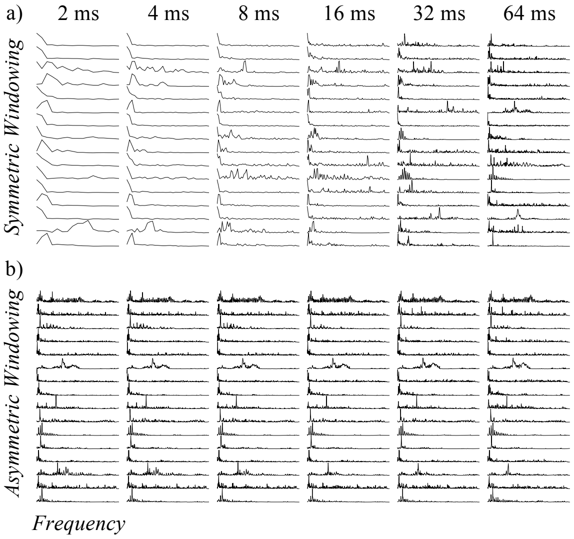


Figure 3: *Example NMF dictionary atoms learned for varying STFT synthesis window size for a) symmetric windowing and b) asymmetric windowing. For each window size, a subset of 16 randomly chosen dictionary atoms are shown from a total of 1024. For symmetric windowing, the analysis window length decreases with the synthesis window, while for asymmetric windowing, the analysis window size remains fixed at 64 ms, with only its shape changing as a function of synthesis window size.*

Contrary to the traditional windowing approach, asymmetric windowing allows us to retain the long-duration analysis windows while decreasing the synthesis window size. As the synthesis window size $2M$ is reduced, the analysis window size remains fixed at the frame size N with its shape increasingly weighted towards the future, as we showed in Figure 1b). In Figure 3b), we present example NMF atoms learned using the asymmetric window approach for varying synthesis window size, where the learned NMF atoms are shown to retain spectral detail, regardless of synthesis window size. As identical training data and random seed is used in all cases, the resulting atoms remain very similar across synthesis window sizes, with only subtle differences in the learned dictionary atoms resulting from the different analysis window shapes.

4.2. Effect on speech enhancement quality

In Figure 4a), we present the PEASS scores on the SiSEC speech enhancement dataset as a function of STFT window size for the symmetric windowing case. We first note that the overall enhancement performance decreases with decreasing window size, with a significant drop in performance for window sizes less than 8 ms. This is likely due to a decreased separability of speech and noise sources with the wideband NMF atoms shown above, resulting in decreased quality of the resulting GCC-NMF speech enhancement. We also note a drastic trade-off between interference suppression and lack of artifacts, where smaller window sizes result in increased interference suppression at the cost of significant artifacts.

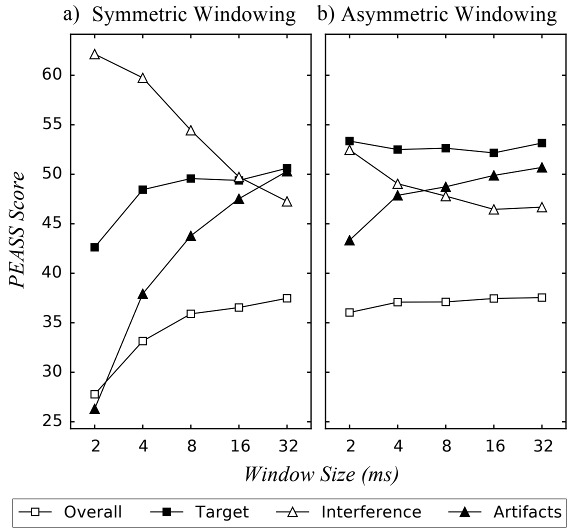


Figure 4: *Effect of STFT synthesis window size on GCC-NMF speech enhancement performance for a) symmetric windowing and b) asymmetric windowing with a fixed analysis window of 64 ms. The PEASS scores correspond to objective measures of overall enhancement quality, target fidelity, interference suppression, and lack of artifacts, where higher scores are better in all cases.*

In Figure 4b), we present the effect of latency on PEASS scores for the SiSEC dataset for the asymmetric windowing approach. The analysis window here is kept fixed at 1024 samples at 16 kHz (64 ms), while the synthesis window size is varied from 512 to 32 samples (32 to 2 ms), with an overlap of 75% of the synthesis window used in each case. We note that the overall PEASS score remains relatively constant for varying synthesis window size, with only a slight reduction for synthesis windows as short as 2 ms. We also note the same trade-off between interference suppression and lack of artifacts as with symmetric windowing, though it is much more tempered for the asymmetric windowing approach. Finally, we note that the target fidelity is consistently higher for the asymmetric windowing case, and remains relatively constant for varying synthesis window size. These results demonstrate that the proposed asymmetric windowing approach is a viable solution to reduce the latency of real-time GCC-NMF to values well below the threshold required for hearing devices while maintaining the enhancement quality of the higher latency symmetric windowing approach.

4.3. Latency and GCC-NMF processing time

We now proceed to study the computational requirements of the GCC-NMF speech enhancement algorithm with asymmetric windowing to determine the conditions under which it may be executed in real-time. As we saw in Section 3.2, the inherent latency of the asymmetric STFT process is equal to the duration of the synthesis window plus the frame advance. For speech enhancement to be performed in real-time, the system must then process a single frame within the time of a single frame advance. This processing time includes the windowing processes, the forward FFT, the GCC-NMF speech enhancement processing itself, the inverse FFT, and the OLA summation.

In Figure 5a), we present the average measured processing time of the online GCC-NMF enhancement algorithm for a *single frame* as a function of the NMF dictionary size, for a variety of hardware platforms. We note that the processing time increases approximately linearly with dictionary size, with the slope varying between hardware platforms. On all systems presented, processing times less than 8 ms are possible, provided a small enough dictionary is used, where enhancement performance decreases smoothly with decreasing dictionary size as we have shown previously [2].

In Figure 5b), we depict the relationship between system latency and available processing time for a single frame, as a function of synthesis window size and overlap. Decreasing either the synthesis window size or the frame advance decreases the system latency at a cost of decreased available processing time. We may combine this information with Figure 5a) to determine, for a given hardware system and dictionary size, the available synthesis window size and overlap values (and resulting latencies), in order for the system to run in real-time. All systems prove fast enough for a synthesis window size of 16 ms with 50% overlap and a dictionary size of 64, resulting in a latency of 24 ms. All systems except the Raspberry Pi may achieve 12 ms latency for small to moderate dictionary sizes, with a window size of 8 ms and 50% overlap. The fastest system (Tesla K40 GPU) can achieve 6 ms latency for dictionaries at least as large as 1024 atoms. These results demonstrate that it is possible to achieve latencies suitable for hearing assistive devices with real-time speech enhancement with GCC-NMF using the asymmetric windowing technique. While these results are promising, the hardware platforms tested remain significantly more powerful than those found in currently available hearing aids. Future work will therefore involve additional implementation optimizations in order to run the system with larger dictionaries on even lower-power devices.

5. Conclusion

We have presented an approach to reducing latency in the real-time GCC-NMF speech enhancement system by incorporating

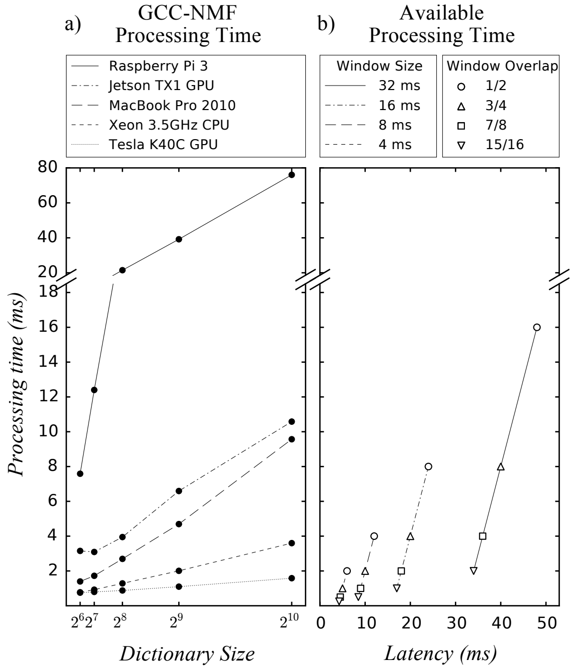


Figure 5: *Real-time GCC-NMF computational requirements with the asymmetric STFT windowing technique, with a) Effect of dictionary size on GCC-NMF mean empirical processing time for a single frame on various hardware platforms given an analysis window size of 64 ms, and b) available processing time for a single frame, given the asymmetric STFT windowing approach, presented for varying synthesis window size and overlap, with the resulting latency as the horizontal axis.*

an asymmetric STFT windowing technique. This asymmetric windowing method provides long duration analysis windows as with the traditional symmetric window approach, maintaining the high spectral resolution required by GCC-NMF, but uses short synthesis windows in order to drastically reduce system latency. We have shown that while speech enhancement performance suffers for the traditional symmetric windowing method when decreasing system latency using shorter windows, the asymmetric windowing approach results in relatively constant performance across a wide range of synthesis window sizes as short as 2 ms given an analysis window size of 64 ms. The computational requirements of the online GCC-NMF algorithm were presented for a variety of hardware platforms including the Raspberry Pi and NVIDIA Jetson TX1, and it was shown that the system may run in real-time with latencies below 24 ms on all platforms, provided the NMF dictionary size is adapted to the computational capabilities of the hardware. Moderately powerful hardware may achieve 12 ms latency with moderate dictionary sizes, while for the most powerful hardware we tested, a Tesla K40 GPU, latencies as low as 6 ms are possible with a large NMF dictionary of 1024 atoms. Source code for this work will be made available at `https://www.github.com/seanwood/gcc-nmf`.

6. Acknowledgements

The authors would like to thank the NSERC discovery grant, and FQRNT (CHIST-ERA, IGLU) for funding our research, and NVIDIA for contributing the Tesla K40 GPU. We would also like to thank William E Audette for inspiring discussions that lead directly to this work, and the anonymous reviewers for their thoughtful feedback.

7. References

[1] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, “Blind speech separation and enhancement with GCC-NMF,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.

[2] S. U. N. Wood and J. Rouat, “Real-time speech enhancement with GCC-NMF,” in *Interspeech 2017*, 2017.

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

[4] M. A. Stone and B. C. Moore, “Tolerable hearing aid delays. I. estimation of limits imposed by the auditory path alone using simulated hearing losses,” *Ear and Hearing*, vol. 20, no. 3, p. 182, 1999.

[5] M. A. Stone and B. C. J. Moore, “Tolerable hearing-aid delays: IV. effects on subjective disturbance during speech production by hearing-impaired subjects,” *Ear and Hearing*, vol. 26, no. 2, pp. 225–235, 2005.

[6] R. Herbig and J. Chalupper, “Acceptable processing delay in digital hearing aids,” *Hearing Review*, vol. 17, no. 1, pp. 28–31, 2010.

[7] J. Agnew and J. M. Thornton, “Just noticeable and objectionable group delays in digital hearing aids,” *JOURNAL-AMERICAN ACADEMY OF AUDIOLOGY*, vol. 11, no. 6, pp. 330–336, 2000.

[8] H. Dillon, G. Keidser, A. O’brien, and H. Silberstein, “Sound quality comparisons of advanced hearing aids,” *The hearing journal*, vol. 56, no. 4, pp. 30–32, 2003.

[9] D. Mauler and R. Martin, “A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement,” in *Signal Processing Conference, 2007 15th European*, 2007, pp. 222–226.

[10] H. W. Löllmann and P. Vary, “Low delay filter-banks for speech and audio processing,” *Speech and audio processing in adverse environments*, pp. 13–61, 2008.

[11] ——, “Low delay noise reduction and dereverberation for hearing aids,” *EURASIP Journal on advances in signal processing*, vol. 2009, no. 1, p. 437807, 2009.

[12] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[13] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

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

[15] J. O. Smith III, *Spectral audio signal processing*. W3K publishing, 2011.

[16] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, “MPEG-4 low delay audio coding based on the AAC codec,” in *Audio Engineering Society Convention 106*. Audio Engineering Society, 1999.

[17] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Geiger, V. Ruoppila, P. Ekstrand, and G. Bernhard, “MPEG-4 enhanced low delay AAC-a new standard for high quality communication,” in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[18] L. Su and H.-t. Wu, “Minimum-latency time-frequency analysis using asymmetric window functions,” *arXiv preprint arXiv:1606.09047*, 2016.

[19] K. T. Andersen and M. Moonen, “Adaptive time-frequency analysis for noise reduction in an audio filter bank with low delay,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 784–795, 2016.

[20] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Computer Speech & Language*, 2016.

[21] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, “The 2016 signal separation evaluation campaign,” in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 323–332.

[22] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046–2057, 2011.

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

[24] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

The Acoustically Transparent Hearing Device: Towards Integration of Individualized Sound Equalization, Electro-Acoustic Modeling and Feedback Cancellation

Florian Denk[1,4], *Steffen Vogl*[3], *Henning Schepker*[2,4],
Birger Kollmeier[1,4], *Matthias Blau*[3,4], *Simon Doclo*[2,4]

[1] Medizinische Physik, University of Oldenburg, Germany
[2] Signal Processing Group, University of Oldenburg, Germany
[3] Institut für Hörtechnik und Audiologie, Jade Hochschule Oldenburg, Germany
[4] Cluster of Excellence *Hearing4all*
{florian.denk, simon.doclo} @uni-oldenburg.de

Abstract

Assistive hearing devices often suffer from a low acceptance by the end user due to poor sound quality. Recently, a novel acoustically transparent hearing device was developed that aims at increasing the acceptance and benefit, also for (near-to) normal-hearing people, by providing better sound quality. The hearing device integrates three microphones and two receivers and can be calibrated in-situ in an attempt to conserve the open-ear sound transmission characteristics of an individual person.
To further improve the quality of acoustic transparency and extend the functionality of the hearing device, we outline the integration of further models and algorithms. Electro-acoustic models of the device can improve adjustment to transparency by providing a better estimate of the pressure at the eardrum with an in-ear microphone. In addition, the multi-microphone device layout allows the development of custom feedback cancellation algorithms by means of a beamformer in order to robustly steer a spatial null towards the hearing device receiver.

1. Introduction

Despite a great improvement in hearing technology in the past decades, the acceptance of assistive hearing devices is still limited, partially due to poor sound quality [1, 2, 3]. This is particularly true for potential first-time users with a mild-to-moderate hearing loss or even (near-to) normal hearing. While they would benefit from features like speech enhancement or amplification in acoustically challenging situations, they are usually not willing to accept a general degradation of the sound quality. Therefore, an important challenge is to develop a device that is acoustically transparent, i.e., that allows hearing comparable to that of the open ear while being capable of providing a desired sound enhancement at the eardrum. These principles can be applied not only to hearing aids, but also to consumer products, e.g., hearables [4, 5].

We recently developed a prototype of an acoustically transparent hearing device that can be individually calibrated aiming to preserve the open-ear sound transmission characteristics of the particular user, even if the ear canal is partially occluded [6]. The used sound equalization approach exploits the microphone positions of a novel vented multi-microphone earpiece, including an in-ear microphone for monitoring the pressure in the ear canal. Acoustical transparency on the perceptual level was verified in a subjective listening experiment [6], and convincing sound quality with the device was observed for normal hearing subjects [7]. Nevertheless, the need for improving transparency in a physical sense was revealed in a recent technical evaluation [8]. Furthermore, other processing stages might interact with the desired goal of acoustic transparency.

After presenting the hardware of the device in Section 2, in this paper we first review the sound equalization approach to achieve acoustic transparency in Section 3, and then present two approaches that aim at improving and completing its functionality towards a full acoustically transparent hearing device. To improve the acoustic transparency feature, a promising approach is to include electro-acoustic models of the device [9]. These models provide an accurate estimate of the sound pressure at the eardrum with an in-ear microphone, which is key to precise sound equalization in a non-occluding fit. Principles and first results comparing the estimated and measured pressure at the eardrum are outlined in Section 4. In addition, the multi-microphone hardware layout facilitates feedback cancellation using a beamformer with a spatial null steered towards the receiver [10, 11, 12] in addition to state-of-the-art adaptive feedback cancellation methods [13]. The principle is briefly introduced and potential interactions of the null-steering approach with the aim of providing acoustic transparency are evaluated in Section 5. Challenges resulting from integrating all approaches are discussed in Section 6.

2. Hardware

The custom in-the-ear type earpiece with relatively open acoustic properties is depicted in Figure 1. A schematic drawing is shown in Figure 2, together with the filter stages in sound equalization (see Section 3) and feedback cancellation (see Section 5), as well as references to the electro-acoustic model (see Section 4).

All electronic components are removably fitted into an individual silicone earmould that fills the concha bottom. In total, the device contains 3 microphones and 2 receivers. Two microphones (Type *Knowles GA-38*) and two balanced armature receivers are located in an acrylic tube referred to as the core, which is inserted into a bore through to the ear canal. The first microphone is located at the inner face of the core and points towards the eardrum ("in-ear microphone" with output voltage y_1 and pressure p_1) and serves to monitor the sound pressure in the ear canal. The second microphone is located at the outer face of the core and points outwards ("entrance microphone", with output voltage y_2 and pressure p_2). The third microphone

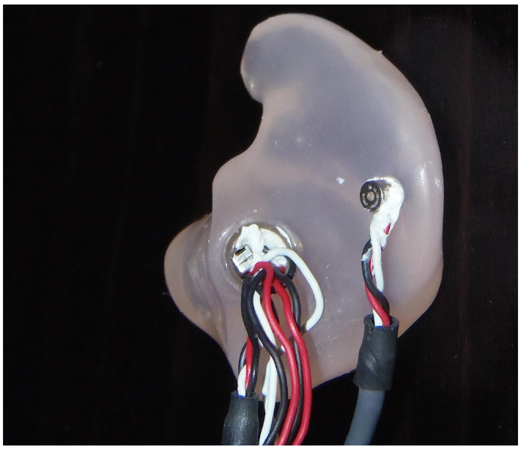


Figure 1: *Assembled earpiece, from [6]. 3 microphones and 2 receivers are fitted into an individual silicone earmould.*

(”concha microphone”, Type *Knowles FG-23329*, with output voltage y_3 and pressure p_3) is placed in the back of the concha by flush insertion into a hole. The two independent receivers are positioned next to the microphones at both ends of the core, but both pointing towards the eardrum. The inner one is a tweeter (*Knowles WBFK-30019*, with input voltage u_1) and the outer one a woofer (*Knowles FK-26768*, with input voltage u_2). Note that although included in the electro-acoustic model, the woofer is not currently used in operation, i.e., it is not considered in sound equalization and feedback cancellation, which is indicated by a dashed line in Figure 2. The hearing device is connected to a PC for real-time signal processing via a soundcard and a custom supply and amplifier box.

The residual space in the core between the microphones and receivers forms a vent, to increase wearing comfort by ventilation and reduction of the occlusion effect [14, 15]. This also implies that sounds below 1 kHz reach the eardrum without considerable attenuation, and the frequency response of the receivers is restricted to above ca. 800 Hz [6].

3. Achieving Acoustic Transparency by Individualized Sound Equalization

3.1. Principles

Acoustic transparency is achieved, when the superposition of direct sound leaking through the core and the electro-acoustically reproduced sound at the eardrum is physically or perceptually equal to the pressure that would be present with an open ear. Achieving acoustic transparency can be separated into two problems: First, the pressure at the eardrum with an open ear has to be estimated based on the available microphone signals to compute the so-called target pressure. Second, the device has to be adjusted such that the target pressure is generated at the eardrum of the individual subject when the device is in the ear, i.e., sound equalization is performed.

In [6], the target pressure was defined as the pressure at the concha microphone, multiplied with an appropriate frequency-dependent gain function. This strategy is justified by observations from spatial audio technology showing that the relative transfer function between a recording point near the (blocked) ear canal entrance and the eardrum of an open ear is not direction-dependent [16]. Thus, the concha microphone approximately contains the direction-dependent portion of the transfer function to the eardrum, and the optimal gain function is the relative transfer function between the concha microphone location and the eardrum in the individual ear. In [6], a flat gain function was used, with the extension that the direct sound leaking through the individual core is considered.

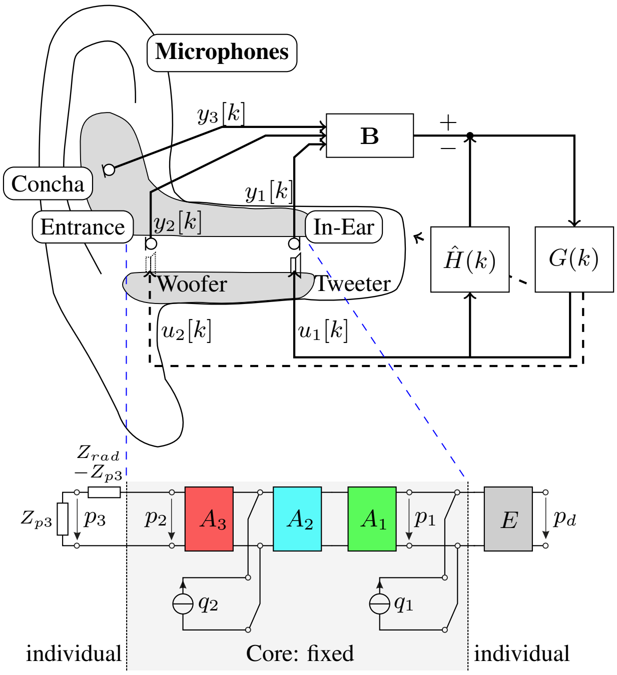


Figure 2: *Top part: Schematic drawing of the hearing device with filter stages for sound equalization and feedback cancellation. Lower part: Corresponding elements and circuit of the electro-acoustic model.*

To achieve sound equalization to the target pressure, the filter G of the hearing device is adjusted in a calibration routine conducted in-situ, i.e., when the device is inserted into the ear. The concha microphone is used to pick up external sound. Assuming that the pressure at the eardrum and the in-ear microphone are similar, the pressure at the eardrum generated by the external sound source and the active device is estimated using the in-ear microphone. Based on the observed deviation from the target pressure, the filter G is adapted until convergence is achieved.

3.2. Current limits and possible extensions

In psychoacoustic experiments with normal-hearing subjects, satisfactory results in terms of acoustic transparency on the perceptual level were observed [6, 7]. However, physical evaluations still reveal some deficits with the current sound equalization approach. Figure 3 shows measurements of the Real-Ear Insertion Gain (REIG) of the transparent hearing device prototype as presented in [6]. The REIG is the difference between the sound pressure at the eardrum measured when the device is inserted and with an open ear. Acoustic transparency on a physical level is achieved if the REIG is 0 dB across all frequencies. The measurements were conducted in a free-field environment in both ears of 12 subjects, and include 3 incident directions in the horizontal plane (azimuth $\theta = 0°, 90°, -135°$).

The measured REIGs deviate from 0 dB, particularly for

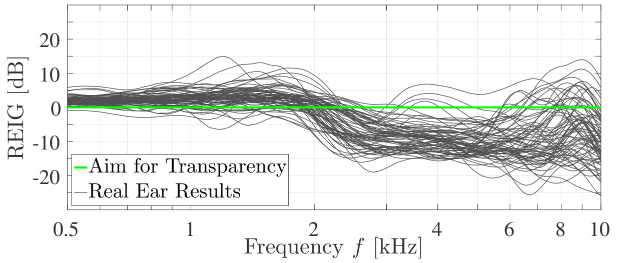


Figure 3: *Real-Ear Insertion Gain (REIG) after 1/6 octave smoothing measured with the hearing device prototype as presented in [6]. The data includes measurements in both ears of 12 subjects, with 3 incident directions in the horizontal plane.*

frequencies above 2 kHz. The error is notably different between subjects and incidence directions, and the variation increases with frequency. This result shows that there is room for improvement in acoustic transparency, which may be tackled with various approaches.

Most of the observed error in the REIG can be explained by two factors: errors in the estimation of the target pressure, and inaccuracies in the sound equalization due to incorrect estimation of the pressure at the eardrum. To estimate the pressure at the eardrum, in [6] the pressure at the in-ear microphone was used, which in most cases introduces an individual estimation error of up to ±20 dB, which is highly variable across frequencies [8]. Thus, the sound equalization error could be reduced if a better estimate of the pressure at the eardrum were available. Electro-acoustic modeling approaches can be used for this purpose, which are treated in Section 4.

In addition, the occurrence of acoustic feedback due to the acoustic coupling between the receiver and the concha microphone has been neglected so far. While this is possible when only the concha microphone is utilized for sound pickup and the applicable gain is limited, appropriate feedback management is a prerequisite when larger amplification than for acoustic transparency is required, or when both external microphones are used for sound pickup, e.g., when implementing a directional microphone. Feedback cancellation techniques tailored to the custom hardware layout are reviewed in Section 5, where possible interactions with acoustic transparency are examined.

4. Electro-Acoustic Model

In our previous work [9], we proposed an electro-acoustic model, which serves to better understand the underlying physical principles of sound transmission in the hearing device, and to estimate quantities at locations where they cannot be directly measured, e.g., the sound pressure at the eardrum. The current focus is to predict the sound pressure at the eardrum p_d in $vivo$, based on measurements using the microphones of the hearing device only.

The model is made up of lumped elements and two-port networks, as depicted in Figure 2. The middle part is the core, which can be regarded as fixed over individual subjects. On the other hand, both terminations, i.e. the external sound field and the residual ear canal, are individual to every ear. The complete model cannot be determined in one step, but is built up in a series of measurements and calculations that are described in the following.

4.1. Model of the Core

First, the model of the core is obtained. It consists of:

- two microphones, characterized by their sensitivity measured prior to assembling the core, each converting its output voltage signal y_m to the corresponding pressure p_m.
- two receivers, which are modeled as ideal volume velocity sources, delivering the flux q_n. This source parameter was also measured prior to assembling the core, according to the technique described by [17].
- the vent, represented by three acoustic transmission lines modeled as two-port networks $A_{1,2,3}$ according to [18]. The three parameters of each transmission line (length, radius and a loss factor) need to be fitted by referring to acoustic measurements. The microphones and receivers are coupled into the vent at locations depicted in Figure 2.

To fit the free parameters of the transmission lines, the assembled core was coupled to a training setup with known termination impedances, and all four transfer functions between the two receivers and microphones 1 and 2 were measured. The medial termination was an IEC711 coupler, while at the lateral end the core was mounted in a baffle. The optimal parameters of the two-port networks were found by minimizing the differences between the measured and modeled transfer functions. Good agreement and computational effectiveness could be achieved with the Nelder-Mead-Simplex [19] algorithm, where the parameter values were constrained to realistic boundaries.

4.2. Model of the Individual Ear

In a second step, a model of the individual ear is estimated. It contains both terminations of the core, as shown in Figure 2. The external sound field (outer termination of the core) is characterized by the radiation impedance Z_{rad}, which can be further split into the transfer impedance Z_{p3} between the outer core end and the concha microphone, and a remaining impedance $Z_{rad} - Z_{p3}$. Z_{rad} is approximated by the physical model of a piston in baffle. The model of the individual ear canal E (medial termination of the core) is individualized based on measurements in the ear of a subject. It is composed of four cascaded acoustic transmission lines with four radii and one total length as parameters, and two parallel load impedances Z_l and $Z_{l,residual}$ located medially (across them p_d is produced). Z_l is a purely resistive frequency-independent load to represent losses, $Z_{l,residual}$ is complex valued and frequency dependent.

Assuming the core model and the outer termination are known and using the transfer function measurements from any of the two receivers to the in-ear microphone, the acoustic impedance Z_{ec} at the point of p_1 (i.e., the in-ear microphone) in the direction towards the ear canal can be calculated. Then, the parameters of the individual ear canal model E are fitted by minimizing both level and phase differences between measured and modeled impedances Z_{ec}, summed across the frequencies from approximately 1 to 15 kHz. Again, the Nelder-Mead-Simplex algorithm was applied with realistic boundaries. Since it was observed that results depended on initial values, 500 random initial values were used and the lowest cost result taken.

Several studies (e.g. [20, 21]) have shown that Z_{ec} - or the reflectance derived from it - can be used to estimate an ear canal model E and ultimately predict the sound pressure at the eardrum p_d.

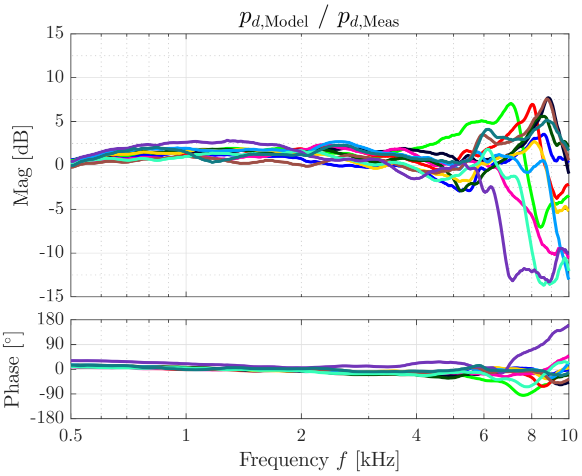


Figure 4: *Deviation between the pressure at the eardrum predicted by the model $p_{d,\text{Model}}$ and the one measured with a probe tube microphone $p_{d,\text{Meas}}$, for twelve subjects, with the woofer as sound source.*

4.3. Evaluation

The individual ear canal model E and the predicted pressure at the eardrum p_d were evaluated by means of probe tube measurements in 12 subjects.

The differences between the model predictions and the measurements of p_d created by the woofer are shown in Figure 4. Below 6 kHz, the agreement in both magnitude and phase is very good. However, for higher frequencies, the differences increase. It should be noted that in this frequency range the probe tube measurements are more likely to be corrupted by errors, as the tube had to be in place together with the earmold which made visual inspection of its position impossible. Furthermore, around 8 kHz the core has a low source impedance, i.e., the impedance at the point where p_1 is measured towards the lateral direction is low compared to typical ear canal impedances Z_{ec}. This reduced measurement accuracy may additionally lead to deviations between the estimated and measured sound pressure.

5. Feedback Cancellation

Acoustic feedback occurs when a signal is picked up by a microphone, amplified, played back by a receiver and picked up again by the microphone, creating a closed-loop system. In hearing devices, adaptive feedback cancellation (AFC) is typically used to reduce the detrimental effect of acoustic feedback, which is most often perceived as howling or whistling. In AFC, an adaptive filter is used to estimate the acoustic feedback path between the hearing device receiver and the microphone, theoretically allowing for perfect feedback cancellation [22]. However, due to the closed-loop electro-acoustic system, the estimate of the acoustic feedback path is generally biased [23, 24]. Several algorithms have been proposed with the aim of reducing this bias, where the so-called prediction-error-method [24] seems most promising. While an AFC algorithm can be applied for any hardware layout, the considered multi-microphone setup (cf. Figure 2) additionally allows for the use of multi-microphone feedback cancellation approaches. This includes a fixed null-steering beamformer that exploits the spatial diversity of the microphones to steer a spatial null towards the position of the hearing device receiver. Note that only the inner receiver of the device is considered here.

Several optimization approaches for calculating the null-steering beamformer coefficients have been proposed, including a robust least-squares design [10, 11] and a robust min-max design [12] aiming at directly maximizing the maximum stable gain of the hearing device, i.e., the gain before the closed-loop system becomes unstable. Furthermore, the benefit of combining a fixed null-steering beamformer and an AFC algorithm based on the prediction-error-method to cancel residual feedback has recently been shown [13]. However, in none of the presented null-steering beamformer optimization approaches [10, 11, 12], the preservation of the pickup microphone directional response that is required for achieving acoustic transparency has been taken into account. This implies that the null-steering beamformer may alter spectral directional cues and bias spatial perception, e.g., sound localization. Therefore, after briefly introducing the optimization procedure, in the following we analyze the directional response of the fixed null-steering beamformer.

We assume time-invariance of the acoustic feedback paths $H_m(k) = H_m$, $m = 1, \ldots, M$ between the receiver and the mth microphone. Assuming the availability of I measurements of the acoustic feedback paths (e.g., obtained by prior measurement), the coefficients of the null-steering beamformer $\mathbf{B}$ are obtained by minimizing the following least-squares cost-function [11]

$$J_{LS}(\mathbf{b}) = \sum_{i=1}^{I} \|(\mathbf{H}^{(i)})^T \mathbf{b}\|_2^2, \tag{1}$$

where $\mathbf{b}$ is the ML_B-dimensional vector of the beamformer coefficients and $\mathbf{H}^{(i)}$ is the $ML_B \times (L_B + L_H - 1)$-dimensional matrix of concatenated convolution matrices of the acoustic feedback paths from the ith measurement, $i = 1, \ldots, I$, with L_B the number of beamformer coefficients for each microphone and L_H the length of the acoustic feedback path. To prevent the trivial solution of $\mathbf{b} = \mathbf{0}$, the beamformer coefficients in a reference microphone m_0 are constrained to correspond to a delay of L_d samples, i.e.,

$$\mathbf{b}_{m_0} = [\, \underbrace{0 \;\; \ldots \;\; 0}_{L_d} \;\; 1 \;\; 0 \;\; \ldots \;\; 0 \,]^T. \tag{2}$$

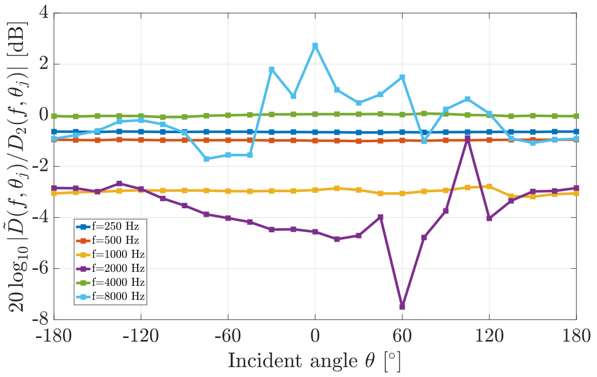


Figure 5: *Directional response of the beamformer output relative to the directional response of the entrance microphone ($m = 2$) as a function of the azimuth θ.*

To obtain the beamformer coefficients, we first measured the acoustic feedback paths of the hearing device in the left ear of a dummy head with adjustable ear canals [25], both in free-field and with a hand very close to the ear, using a sampling rate of 32 kHz. The beamformer coefficients were then computed by minimizing (1) subject to the constraint in (2) for $M = 3$ microphones (in-ear, entrance and concha microphone), $L_B = 32$, $m_0 = 2$ (entrance microphone), $L_d = 16$ and $I = 2$. The resulting added stable gain, i.e., the increase in gain margin compared to using only the entrance microphone ($m = 2$), was 18.3 dB and 22.6 dB for the free-field condition and the hand condition, respectively.

To compute the directional response of the null-steering beamformer for an incoming signal, the acoustic transfer functions to the microphones $D(\theta_j)$, $j = 1, \ldots, J$ were measured for $J = 24$ equidistantly spaced angles θ_j surrounding the dummy head at a distance of approximately 2.5 m in the horizontal plane. Figure 5 shows the directional response of the beamformer $\tilde{D}(\theta_j) = \mathbf{B}^T\mathbf{D}(\theta_j)$ for multiple frequencies relative to the directivity $D_2(\theta_j)$ of the entrance microphone. Ideally, the relative directional response would be equal to 0 dB for all frequencies and incidence angles. However, the response is different from 0 dB for most of the considered frequencies and directions. Nevertheless, for most frequencies and incident angles the null-steering beamformer alters the directivity only by approximately ±4 dB.

6. Discussion and Summary

The principles of an acoustically transparent hearing device presented in [6] and physical evaluation results have been reviewed in Section 3, and possible extensions towards improving and extending its functionality have been presented in Sections 4 and 5. While the good performance of electro-acoustic modeling and customized feedback cancellation for the hearing device has been demonstrated for these approaches individually, a next challenge is the integration of the two approaches with the transparency feature of the hearing device in real-time operation.

Unbiased estimation of the pressure at the eardrum can improve acoustic transparency by improving sound equalization to a target pressure at the eardrum. The electro-acoustic model presented in Section 4 is able to predict the pressure at the eardrum that is generated by the hearing device receiver accurately up to approximately 6-7 kHz in magnitude and phase. However, when estimating the sound pressure at the eardrum in normal operation, the superposition with the direct sound leaking through the vent needs to be considered. Since the model is in principle also able to predict the pressure generated by the receiver at the in-ear microphone, this predicted pressure can be subtracted from the observed pressure at the in-ear microphone to obtain an estimate of the direct sound only. The pressure at the eardrum generated by the direct sound alone can then also be predicted. It should be noted that for integrating the electro-acoustic model with the acoustically transparent hearing device, it is sufficient to extract all relevant transfer functions from the model after calibration measurements.

Although the null-steering beamformer presented and evaluated in Section 5 yields impressive results in terms of feedback cancellation, it was also noted that it introduces a direction-dependent bias compared to the reference microphone. Spectral directional cues contained in the reference microphone signal are thus altered, which may introduce perceptual errors regarding spatial hearing or other undesired artifacts. However, the deviations are in the range of about ±4 dB, and their perceptual relevance is not yet clear. Another issue is the delay of L_d samples introduced by the beamformer, which should be considered when designing the equalization filter G. In principle, this can be achieved by performing the in-situ calibration [6] with the beamformer output as hearing device input signal.

The electro-acoustic models of the individual ear, as well as the null-steering beamformer require knowledge of the transfer functions between the hearing device receivers and microphones. They can be measured in-situ in the individual ear using only the device as part of calibration measurements, which are also necessary to achieve transparency [6]. Gathering these data is therefore no practical obstacle to integrating the electro-acoustic model and the null-steering beamformer into a future version of the prototype.

In conclusion, there seem to be no principal problems hindering the integration of electro-acoustic models and customized feedback cancellation methods into our prototype hearing device. Both are promising approaches to improving the sound equalization to achieve acoustic transparency in our prototype hearing device, as well as increasing its functionality in more realistic application scenarios with higher gain settings and more than one pickup microphone in each side. Future work will hence focus on the implementation of the presented approaches to construct an improved version of our acoustically transparent hearing device.

7. Acknowledgement

This work was supported by the Research Unit FOR 1732 *Individualized Hearing Acoustics* and the the Cluster of Excellence *Hearing4all*, both funded by the German Research Council DFG.

8. References

[1] M. Kinkel, “Hearing aid (ha) demographics: What did change during the last two decades?” in *International Hearing Aid Conference IHCON*, 2016.

[2] M. C. Killion, “Myths about hearing aid benefit and satisfaction,” *Hearing Review*, vol. 11, pp. 32–70, 2004.

[3] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, “Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[4] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, “Augmented reality audio for mobile and wearable appliances,” *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[5] P. F. Hoffmann, F. Christensen, and D. Hammershøi, “Insert Earphone Calibration for Hear-Through Options,” in *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones*, 2013.

[6] F. Denk, M. Hiipakka, B. Kollmeier, and S. M. A. Ernst, “An individualised acoustically transparent earpiece for hearing devices,” *International Journal of Audiology*, vol. Early Online, pp. 1–9, 2017.

[7] F. Denk, B. Kollmeier, and S. M. Ernst, “High-fidelity hearing instruments: Evaluating listening quality of a new prototype using a method for evaluating modified listening (MEML),” in *AES Conference on Headphone Technology*, 2016, pp. 206–213.

[8] F. Denk, L. Haverkamp, B. Kollmeier, and S. M. Ernst, “Evaluation of a high-fidelity hearing device prototype using real-ear measurement,” in *DGA Deutsche Gesellschaft für Audiologie*, 2017.

[9] S. Vogl, T. Sankowsky-Rothe, and M. Blau, "Electro-acoustic modeling of an earpiece with integrated microphones and loudspeakers. (in german, *Elektroakustische Modellierung eines Ohrpassstücks mit integrierten Mikrofonen und Lautsprechern*)," in *Fortschritte der Akustik - DAGA*, 2016, pp. 1163–1166.

[10] H. Schepker, L. T. Tran, S. Nordholm, and S. Doclo, "Acoustic feedback cancellation for a multi-microphone earpiece based on a null-steering beamformer," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China*, Sep. 2016.

[11] H. Schepker, L. T. T. Tran, S. E. Nordholm, and S. Doclo, "A robust null-steering beamformer for acoustic feedback cancellation for a multi-microphone earpiece," in *Proc. 12th ITG Conference on Speech Communication, Paderborn, Germany*, Oct. 2016, pp. 165–169.

[12] H. Schepker, L. T. Tran, S. Nordholm, and S. Doclo, "Null-steering beamformer for acoustic feedback cancellation in a multi-microphone earpiece optimizing the maximum stable gain," in *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA*, Mar. 2017, pp. 341–345.

[13] H. Schepker, L. T. T. Tran, S. Nordholm, and S. Doclo, "Combining null-steering and adaptive filtering for acoustic feedback cancellation in a multi-microphone earpiece," in *Proc. European Signal Processing Conference (EUSIPCO), Kos Island, Greece*, Aug. 2017.

[14] A. Winkler, M. Latzel, and I. Holube, "Open versus closed hearing-aid fittings: A literature review of both fitting approaches," *Trends in Hearing*, vol. 20, p. 2331216516631741, 2016.

[15] M. Blau, T. Sankowsky, A. Stirnemann, H. Oberdanner, and N. Schmitt, "Acoustics of open fittings," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3011, 2008.

[16] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[17] M. Blau, T. Sankowsky, P. Roeske, H. Mojallal, M. Teschner, and C. Thiele, "Prediction of the sound pressure at the ear drum in occluded human cadaver ears," *Acta Acustica United with Acustica*, vol. 96, pp. 554–566, 2010.

[18] D. H. Keefe, "Acoustical wave propagation in cylindrical ducts: Transmission line parameter approximations for isothermal and nonisothermal boundary conditions," *Journal of the Acoustical Society of America*, vol. 75, pp. 58–62, 1984.

[19] J. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, pp. 308–313, 1965.

[20] H. Hudde, A. Engel, and A. Lodwig, "Methods for estimating the sound pressure at the eardrum," *Journal of the Acoustical Society of America*, vol. 106, pp. 1977–1992, 1999.

[21] T. Sankowsky-Rothe, M. Blau, S. Köhler, and A. Stirnemann, "Individual equalization of hearing aids with integrated ear canal microphones," *Acta Acustica United with Acustica*, vol. 101, pp. 552–566, 2015.

[22] T. van Waterschoot and M. Moonen, "Fifty Years of Acoustic Feedback Control: State of the Art and Future Challenges," *Proceedings of the IEEE*, vol. 99, no. 2, pp. 288–327, Feb. 2011.

[23] M. Siqueira and A. Alwan, "Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 443–453, Jul. 2000.

[24] A. Spriet, I. Proudler, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3749–3763, 2005.

[25] M. Hiipakka, M. Tikander, and M. Karjalainen, "Modeling the External Ear Acoustics for Insert Headphone Usage," *Journal of the Audio Engineering Society*, vol. 58, no. 4, pp. 269–281, 2010.

Individuality-Preserving Speech Synthesis System for Hearing Loss Using Deep Neural Networks

Tsuyoshi Kitamura[1], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1], *Kiyohiro Omori*[2]

[1]Graduate School of System Informatics, Kobe University, Japan
[2]Hyogo Institute of Assistive Technology, Kobe, Japan

kitamura@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

Statistic parametric speech synthesis (SPSS) systems [1] are being widely researched in the field of speech processing. We present in this paper a speech synthesis method for people with hearing loss. Because of their disorders, their prosody is often unstable and their speech rate, pitch, and spectrum differ from those of physically unimpaired persons, which causes their speech to be less intelligible and, consequently, makes communication with physically unimpaired persons difficult. In order to deal with these problems, this paper describes an approach that makes use of a novel combination of deep neural networks (DNN)-based text-to-speech synthesis using the DNNs of a physically unimpaired person and a person with hearing loss, while preserving the individuality of a person with hearing loss. Through experimental evaluations, we have confirmed that the proposed method successfully synthesizes an intelligible speech signal from a hard-to-understand signal while preserving the target speaker's individuality.

Index Terms: hearing disorders, speech synthesis system, deep neural networks, assistive technologies

1. Introduction

In this paper, we focus on, as one assistive technology for a person with hearing loss, a speech synthesis system that assists persons in their speech communication. Their speech style may be different from those of persons without hearing loss and the utterances may be less intelligible due to hearing loss. It sometimes makes verbal communication with other people difficult.

A DNN-based speech synthesis system [2], [3], [4] is a text-to-speech (TTS) system that can generate signals from input text data. A TTS system may be useful for persons with hearing loss because the synthesized speech signal may become more intelligible by adjusting the utterance duration, pitch, and spectrum.

In this paper, we propose a DNN-based speech synthesis method for a person with hearing loss. To generate an intelligible synthesized speech signal while preserving the speaker's individuality, we use speech data from both a person with hearing loss and a physically unimpaired person. Because the speech rate of a person with hearing loss may be unstable, the duration model of a person with hearing loss is modified using the DNNs of a physically unimpaired person to stabilize the speech rate. In addition, the F_0 patterns of a person with hearing loss are often unstable. To solve this problem, in the synthesis step, the F_0 features predicted from the networks of a physically unimpaired person are used as the input of the networks of a person with hearing loss after being converted to the average F_0 of the hearing loss person using a linear transformation.

As for the spectral problem associated with a person with hearing loss, the consonant parts of utterances are sometimes unclear or unstable. To resolve the consonant problem, we generate the spectrum for some consonants from the acoustic model of a physically unimpaired person and the vowel spectrum from the acoustic model of a person with hearing loss in order to preserve the speaker's individuality.

The rest of this paper is organized as follows; In Section 2, an introduction to related work about assistive technology is presented. In Section 3, a speech synthesis system using deep neural networks is presented. Section 4 presents the proposed speech synthesis system for a person with a hearing disorder. In Section 5, in order to confirm the effectiveness of our method, the experimental data are evaluated. Finally, the conclusions are drawn in Section 6.

2. Related Works

To assist people with articulation disorders, a number of assistive technologies using information processing have been proposed. As one of the techniques used for statistic parametric speech synthesis, the Hidden Markov model (HMM)-based TTS approach [5], has been studied for a long time and a number of assistive technologies using a HMM-based TTS system have been proposed; for example, Veaux used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting form Amyotrophic Lateral Sclerosis (ALS) [6]. They have proposed a reconstruction method for degenerative speech disorders using an HMM sound synthesis system. In this method, the subject's utterances were used to adapt an average voice model pre-trained on many speakers. Creer also adapted the average voice model of multiple speakers to severe dysarthria data [7], and Khan used such an adaptation method on a laryngectomy patient's data [8]. The authors of this paper also proposed a HMM-based TTS system for people with articulation disorders [9].

Recently, deep learning has had success in speech synthesis in regard to naturalness and sound quality compared with hidden Markov models [1]. Deep neural networks contain many layers of nonlinear hidden units and represent a mapping function from linguistic features to acoustic features. In the field of speech processing technology, speech recognition (lip reading using deep learning) has also had success [10].

Recently, to develop sound quality and naturalness, the architectures of the DNN have been improved; for example, using long-short-term-memory to take the continuity of speech into account [11], and using i-vectors to adapt the average voice model of multiple speakers [12]. In the adaptation task, a small amount of speech data is required to create synthesized speech because it is difficult for a person with an articulation disorder to say many sentences.

In this paper, we employ trajectory training [3] to train DNN. Trajectory training is regarded as a successful method and has been widely used in recent years for various tasks.

3. DNN-based Speech Synthesis

Fig. 1 shows the overview of the basic approach to text-to-speech (TTS) synthesis based on deep neural networks. The figure shows the synthesis parts of a DNN-based TTS system. In the training part of networks, inputs, the linguistic features extracted from an input text-by-text analysis, are mapped to the output acoustic features (spectral, F0, and aperiodicity) using back propagation. In the synthesis part, the linguistic features are mapped to the output acoustic features using forward propagation. In the parameter generation part, the output acoustic parameters including static, delta, and acceleration parameters are generated into smooth parameter trajectories using the speech parameter generation algorithm [13]. In the waveform synthesis part, a vocoder, such as STRAIGHT [14] or WORLD [15], is used to synthesize speech signals from the smooth parameter trajectories. State duration densities are modeled by the method used in HMM-based speech synthesis systems [16] to control rhythm and tempo, where state durations of each phoneme HMM are modeled by a multi-dimensional Gaussian distribution.

DNN-based speech synthesis comprises the training part and the synthesis part. Acoustic features consist of D-dimensional static features $\mathbf{c}_t = [c_t(1), c_t(2), ..., c_t(D)]$ and corresponding dynamic features $\Delta\mathbf{c}_t$ and $\Delta^2\mathbf{c}_t$, written as

$$\mathbf{o}_t = [\mathbf{c}_t^{\mathrm{T}}, \Delta\mathbf{c}_t^{\mathrm{T}}, \Delta^2\mathbf{c}_t^{\mathrm{T}}]^{\mathrm{T}} \tag{1}$$

Dynamic features are computed from the sequence of static features. The sequence of acoustic features $\mathbf{o} = [\mathbf{o}_1^{\mathrm{T}}, \mathbf{o}_2^{\mathrm{T}}, ..., \mathbf{o}_T^{\mathrm{T}}]$ can be caluculated from the sequence of static features $\mathbf{c} = [\mathbf{c}_1^{\mathrm{T}}, \mathbf{c}_2^{\mathrm{T}}, ..., \mathbf{c}_T^{\mathrm{T}}]$ by

$$\mathbf{o} = \mathbf{W}\mathbf{c} \tag{2}$$

where T is the number of frames included in the sequence and $\mathbf{W}$ is a matrix used to extend static features $\mathbf{c}$ to acoustic features $\mathbf{o}$ [13].

In the training part, the input text is analyzed and transformed into labels, which contain linguistic contexts. The networks learn the complex mapping function from linguistic features $\mathbf{x}_t$ to acoustic features $\mathbf{o}_t$, where the frame-level mean square errors between the predicted acoustic features $\hat{\mathbf{o}}_t$ and the observed acoustic features $\mathbf{o}_t$ are minimized using the back-propagation algorithm.

In the synthesis part, output features include static, delta, and acceleration features. To generate the smooth parameter trajectories, the maximum likelihood parameter generation (MLPG) algorithm [17] is used to take the dynamic features as constraints. The smooth parameter trajectory $\hat{\mathbf{c}}$ is given by

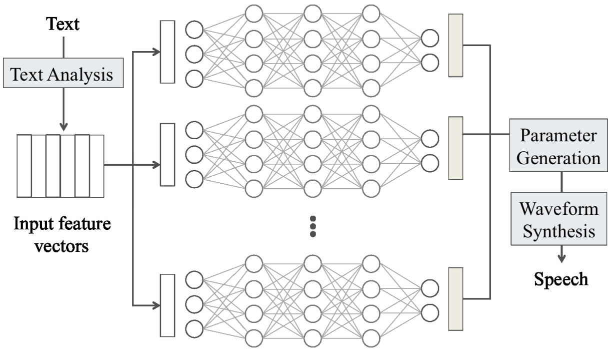


Figure 1: *A flow of speech synthesis using deep neural networks.*

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} P(\mathbf{o}|\boldsymbol{\lambda}) = \arg\max_{\mathbf{c}} \mathcal{N}(\mathbf{W}\mathbf{c}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \bar{\mathbf{c}} \tag{3}$$

where $\boldsymbol{\lambda}$ is the model parameter and $\mathcal{N}(|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean vectors μ and covariance matrix $\boldsymbol{\Sigma}$. The smooth parameter trajectories calculated by the MLPG algorithm can be written by (4).

$$\hat{\mathbf{c}} = (\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{4}$$

In the synthesis part, $\boldsymbol{\mu}$ is the frame obtained by performing a forward propagation and $\boldsymbol{\Sigma}$ is computed from the training data. We can reconstruct the speech waveform from the smooth parameter trajectory $\hat{\mathbf{c}}$ by using a vocoder.

3.1. Trajectory training

To take the interaction between the static and dynamic features into account, the trajectory training minimizes the utterance-level trajectory error than the frame-level error [3]. This training criterion is called the minimum generation error (MGE). The Euclidean distance between the predicted trajectory $\hat{\mathbf{c}}$ (calculated by MLPG) and the observed static trajectory is called the trajectory error. The objective function is written as

$$L = (\hat{\mathbf{c}} - \mathbf{c})^{\mathrm{T}}(\hat{\mathbf{c}} - \mathbf{c}) = (\mathbf{R}\hat{\mathbf{o}} - \mathbf{c})^{\mathrm{T}}(\mathbf{R}\hat{\mathbf{o}} - \mathbf{c}) \tag{5}$$

where

$$\hat{\mathbf{R}} = (\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \tag{6}$$

Mean-variance normalization is performed to $\hat{\mathbf{c}}$ and $\mathbf{c}$ before calculated the trajectory error. The mean and variance values are calculated from the training data in advance. The parameters of DNN are trained by back-propagation using gradient, as is the case with conventional frame-level training.

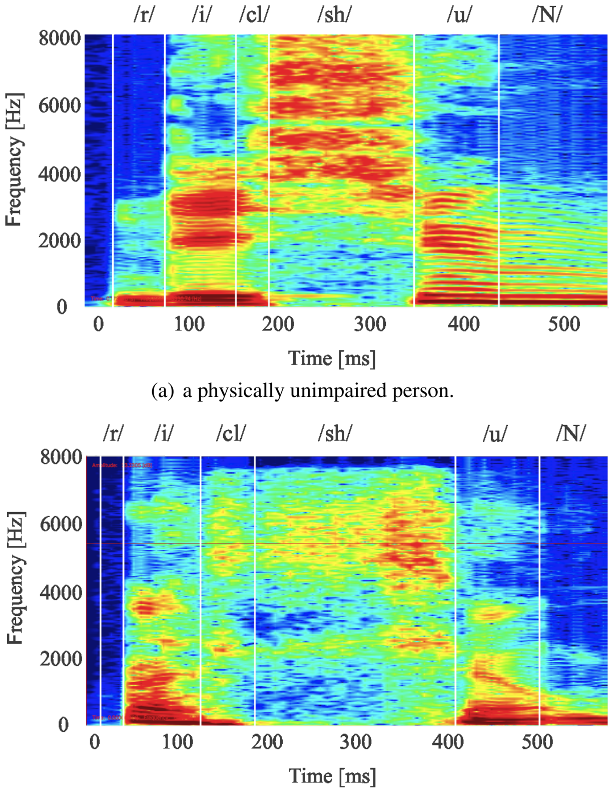
4. DNN-based Speech Synthesis for a Person with Hearing Loss

In our method, the voice of two people, a person with hearing loss and a physically unimpaired person, are used to generate a more intelligible synthesized speech signal that preserves the individuality of the person with hearing loss. Fig. 2 shows the original spectrograms for the word "/r/ /i/ /cl/ /sh/ /u/ /N/""/r/ /i/ /cl/ /sh/ /u/ /N/" of a physically unimpaired person and a person with hearing loss.

As shown in Fig. 2, the high-frequency spectral power of a person with hearing loss is weaker compared to that of a physically unimpaired person. In addition, the duration of a person with hearing loss is unstable that some phones (ex: "/cl" and "/sh/") are too long compared to other phones although the speech length is almost the same as that of a physically unimpaired person. This may be one of the reasons behind the unintelligibility. Therefore, in our method, a more intelligible synthesized speech signal that preserves the speaker's individuality is generated by using the features of both a person with hearing loss and a physically unimpaired person.

4.1. $\mathrm{F_0}$ model modification

Fig. 3 shows the overview of the approach to $\mathrm{F_0}$ modification. As the $\mathrm{F_0}$ patterns of a person with hearing loss are often unstable, we modify it using the $\mathrm{F_0}$ features of a physically unimpaired person.



(a) a physically unimpaired person.

(b) a person with hearing loss.

Figure 2: *Sample spectrograms of /r i cl sh u N/.*

In the training part, first, DNNs of a physically unimpaired person and a person with hearing loss are trained independent of each other. For the DNNs of a physically unimpaired person, the input is the linguistic features and the output is the spectral, aperiodicity, and F_0 features of a physically unimpaired person. For the DNNs of a person with hearing loss, the input is the linguistic features and the F_0 features (static, delta, and acceleration) and the output is spectral features and aperiodicity features of a person with hearing loss.

As shown in Fig. 3, in the synthesis part, first, linguistic features are mapped to the spectral, aperiodicity, and F_0 features using the DNNs of a physically unimpaired person. The output F_0 features of a physically unimpaired person are converted to those of a person with hearing loss by using the linear transformation in Eq. (7) and then, they are used as the input for networks of a person with hearing loss.

$$\hat{w}_t = \frac{\sigma_x}{\sigma_w}(w_t - \mu_w^{(F_0)}) + \mu_x^{(F_0)} \tag{7}$$

where w_t represents a log-scaled F_0 of a physically unimpaired person at the frame t, $\mu_w^{(F_0)}$ and σ_t represent the mean and standard deviation of w_t, respectively. $\mu_x^{(F_0)}$ and σ_x represent the mean and standard deviation of log-scaled F_0 of a person with hearing loss, respectively.

4.2. Duration model modification

The speech rhythm and tempo of a person with hearing loss differ from those of physically unimpaired persons, and this causes their speech to be less intelligible. To solve this problem, the

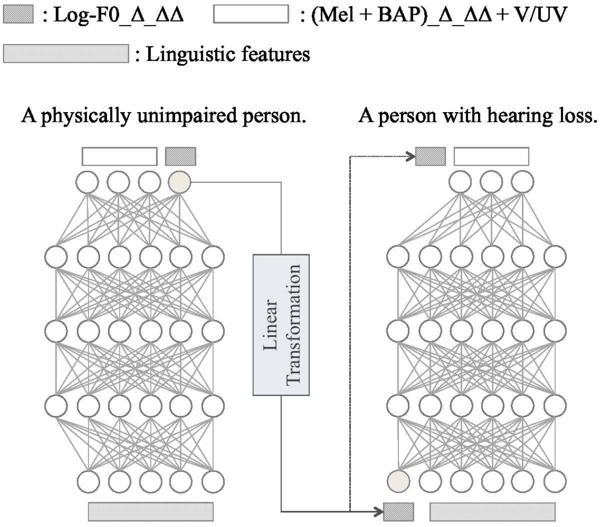

▨ : Log-F0_Δ_ΔΔ
□ : (Mel + BAP)_Δ_ΔΔ + V/UV
: Linguistic features
A physically unimpaired person.
A person with hearing loss.
Linear Transformation

Figure 3: *A flow of the F_0 modification method.*

speech rhythm and tempo of a physically unimpaired person are used in the synthesis part. However, as the average speech rate contains rich speaker individuality, the average speech rate of the synthesized speech signal is fit to that of a person with hearing loss. To implement these ideas, the duration model is modified as follows:

$$y_i = t_i - \mu_w^{(Dur)} + \mu_x^{(Dur)} \tag{8}$$

$$\mu_w^{(Dur)} = \frac{\sum_{i=1}^{I} \mu_{ti}}{I} \tag{9}$$

$$\mu_x^{(Dur)} = \frac{\sum_{i=1}^{I} \mu_{xi}}{I} \tag{10}$$

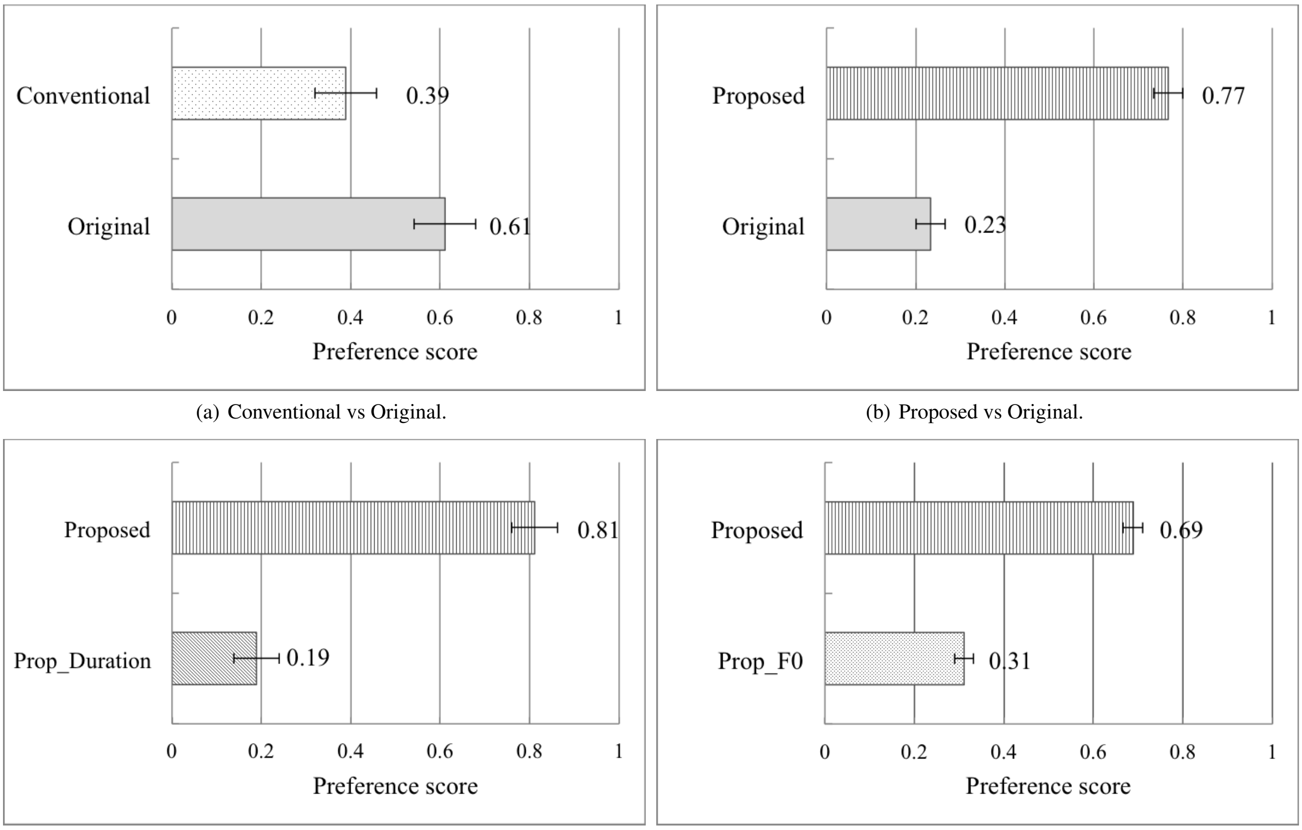
In Eq. (8), t_i represents the value of the i-th node in the duration model of a physically unimpaired person. In Eqs. (9) and (10), I represents the total number of nodes in the model, u_{ti} represents the mean the value of the i-th node in the model of a physically unimpaired person, and u_{xi} represents the mean the value of the i-th node in the model of a person with hearing loss.

5. Experiments

5.1. Experimental conditions

We prepared the training data for two men. One is a physically unimpaired person, and the other is a person with hearing loss. We used 503 sentences from the ATR Japanese database B-set for a physically unimpaired person and we recorded 503 sentences uttered by a person with hearing loss. 450 and 53 utterances were used for training and development, respectively. In addition, we recorded 10 sentences uttered by a person with hearing loss for testing. The speech signal was sampled at 16kHz and the frame shift was 5 msec. Acoustic and prosodic features were extracted using WORLD [15]. As spectral parameters, the 0-th through the 49-th mel-cepstrum coefficients [18], and their dynamic and acceleration coefficients were used. As excitation parameters, log-F_0 and 25 band-filtered aperiodicity [19] were used, along with their dynamic and acceleration coefficients.

In order to confirm the effectiveness of our method, four systems were compared.



(a) Conventional vs Original.

(b) Proposed vs Original.

(c) Proposed vs Prop_Dur.

(d) Proposed vs Prop_F_0

Figure 4: *Preference scores for the listening intelligibility based on subjective evaluations.*

- **Conventional**: DNN-based speech synthesis system using trajectory training
- **Prop_Dur**: **Conventional** + "Duration model of a person with hearing loss was modified in Eq. (8)"
- **Prop_F_0**: **Conventional** + "F_0 modification based on section 4.1"
- **Proposed**: **Prop_Dur + Prop_F_0**

In **Conventional**, input features consisted of 395 features, which comprised 386 binary and 9 numeric features. Binary features were derived from categorical linguistic contexts including quinphone identities, accent type, position of phone, mora, word, and so on. Numeric features include frame position information. Output features consisted of 50 mel-cepstrum coefficients, log-F_0, and 25 band-filtered aperiodicity, their deltas and accelerations coefficients, and a voiced/unvoiced value (3 + (50 + 25 + 1) + 1 = 229). Input features were normalized to the range 0.0-1.0 based on min-max and output features were normalized to zero mean and unit variance. To reduce the training cost, silence frames were removed from the training data of a person with hearing loss. The architecture of the networks was 4-hidden-layers, with each hidden layer containing 700 units. The sigmoid activation function was used for hidden layers, and the linear activation function was used for the output layer. In order to complement some consonant parts of a person with hearing loss, the consonants /s, sh, k, t, ts, z, ch/ were replaced by those of a physically unimpaired person.

In order to evaluate the models, we evaluated the listening intelligibility and the speaker similarity by listening to voices recorded and synthesized under the five conditions (original speech, **Conventional**, **Prop_Dur**, **Prop_F_0**, **Proposed**). A total of 9 Japanese speakers took part in the listening test using headphones. For speaker similarity, a DMOS (Degradation Mean Opinion Score) test was performed. In the DMOS test [20], the original speech signal was used as the reference signal and the option score was set to a 5-point scale (5: degradation is inaudible, 4: degradation is audible but not annoying, 3: degradation is slightly annoying, 2: degradation is annoying, 1: degradation is very annoying). For the listening intelligibility, a paired comparison test was carried out, where each subject listened to pairs of speech signals converted by two methods, and then selected which sample was more intelligible.

5.2. Experimental results

Fig. 4 shows the preference score on the listening intelligibility, where the error bar shows a 95% confidence score. As shown in Fig. 4, our proposed method obtained a higher score than the original recorded speech signal, **Prop_F_0** and **Prop_Dur**. The synthesized speech of **Conventional** is less intelligible than the original recorded speech, but the synthesized speech of **Proposed** is more intelligible than the original speech signal and **Conventional**. Also, as shown in Fig. 4 (c) and (d), the modification of both the F_0 and duration model will result in synthesizing more intelligible speech signals.

Fig. 5 shows the results of the DMOS testing on speaker similarity, where the error bars show a 95% confidence score. As shown in Fig. 5, the synthesized voice from **Conventional** is

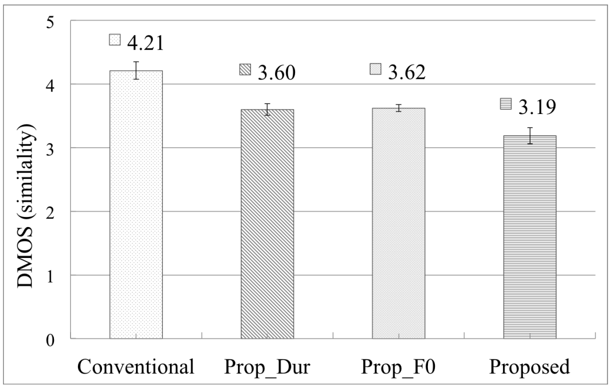


Figure 5: *Speaker similarity to the hearing loss person based on subjective evaluations.*

the most similar to the original voice of the person with hearing loss. Also, it was confirmed that the speaker individuality of a person with hearing loss was lost when using features of a physically unimpaired person. The DMOS score of the proposed method was 3.19 (4: degradation is audible but not annoying, 3: degradation is slightly annoying) and this means speaker individuality is slightly annoying but preserved.

Therefore, from Figs. 4 and 5, it is confirmed that our proposed method generates synthesized signals that are intelligible and include the individuality of a person with hearing loss.

6. Conclusions

In this paper we have proposed a text-to-speech synthesis method using deep neural networks for a person with hearing loss. In our method, to generate more intelligible synthesized sounds while preserving the individuality of a person with hearing loss, a novel combination approach of deep neural networks was employed. The F0 features of a person with hearing loss were modified using those of a physically unimpaired person. The duration model of a physically unimpaired person was used to modify the model of a person with hearing loss. In order to complement some consonant parts of a person with hearing loss, the consonant parts were replaced by those of a physically unimpaired person. The experimental results showed that our method was highly effective in improving the listening intelligibility of speech spoken by a person with hearing loss. In future research, we will complement the vowel parts of the spectral parameters in the training part.

7. Acknowledgements

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

8. References

[1] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *IEEE ICASSP*, 2013, pp. 7962–7966.

[3] Z. Wu and S. King, “Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[4] T. Toda and S. Young, “Trajectory training considering global variance for HMM-based speech synthesis,” in *IEEE ICASSP*, 2009, pp. 4025–4028.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Sixth European Conference on Speech Communication and Technology*, 1999.

[6] C. Veaux, J. Yamagishi, and S. King, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders,” in *Interspeech*, 2012, pp. 967–970.

[7] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, “Building personalised synthetic voices for individuals with severe speech impairment,” *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[8] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, “Reconstructing the voice of an individual following laryngectomy,” *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[9] R. Ueda, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis,” in *ACM ICMI*, 2015, pp. 343–346.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, and T. Sainath, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[11] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, “TTS synthesis with bidirectional lstm based recurrent neural networks,” in *Interspeech*, 2014, pp. 1964–1968.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors.” in *IEEE ASRU*, 2013, pp. 55–59.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *IEEE ICASSP*, vol. 3, 2000, pp. 1315–1318.

[14] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[15] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling for hmm-based speech synthesis,” in *ICSLP*, vol. 98, 1998, pp. 29–32.

[17] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “The effect of neural networks in statistical parametric speech synthesis,” in *IEEE ICASSP*, 2015, pp. 4455–4459.

[18] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *IEEE ICASSP*, vol. 1, 1992, pp. 137–140.

[19] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” in *MAVEBA*, 2001, pp. 59–64.

[20] “Methods for subjective determination of transmission quality.” p. 800.

Comparison of RTF Estimation Methods between a Head-Mounted Binaural Hearing Device and an External Microphone

Nico Gößling, Daniel Marquardt, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics and
Cluster of Excellence Hearing4All, Oldenburg, Germany

{nico.goessling,daniel.marquardt,simon.doclo}@uni-oldenburg.de

Abstract

Besides noise reduction, an important objective of a binaural speech enhancement algorithm is the preservation of the binaural cues of both the desired speech source as well as the undesired noise in order to preserve the spatial impression of the acoustic scene for the listener. Recently, it has been shown for the binaural MVDR beamformer with partial noise estimation (MVDR-N) that by combining head-mounted hearing devices with an external microphone it is possible to improve the noise reduction performance while achieving the same binaural cue preservation. While the relative positions of the head-mounted microphones can be assumed to be stationary this assumption does not hold for the external microphone, which can change its relative position due to head movements or direct movement of the listener or the external microphone. In this paper, we compare the influence of different methods for estimating the relative transfer functions of the desired speech source between the head-mounted microphones and the external microphone on the noise reduction and binaural cue preservation performance of the binaural MVDR-N beamformer.

Index Terms: binaural cues, noise reduction, external microphone, interaural coherence, relative transfer functions

1. Introduction

Noise reduction algorithms for head-mounted hearing devices (e.g., hearing aids) are crucial to improve speech quality and intelligibility in background noise. Binaural devices, consisting of one or more microphones on each side of the head of the listener, are able to exploit not only spectral but also spatial information on both sides of head [1–3]. Besides noise reduction, preserving the binaural cues of all present sound sources is an important task of a binaural noise reduction algorithm in order to ensure that the listener's spatial impression is not distorted by the algorithm.

For a single desired speech source, the binaural multichannel Wiener filter (MWF) [2, 4] has been shown to preserve the binaural cues of the desired speech source. However, it typically distorts the binaural cues of the noise, such that the residual noise is perceived as coming from the same direction as the desired speech source which is obviously undesired. As an extension, the binaural MWF with partial noise estimation (MWF-N) has been proposed [2, 4, 5], which aims at preserving the speech component and a scaled version of the noise component in the reference microphones of the left and the right hearing device. It has been shown that the mixing parameter in the binaural MWF-N allows to trade off noise reduction and binaural cue preservation performance of the noise component [4].

In this paper we consider the binaural minimum variance distortionless response (MVDR) beamformer with partial noise estimation (MVDR-N) [2, 4–6], which can be considered as a special case of the binaural MWF-N only performing spatial processing. Recently, the use of one or more external microphones (eMics) in combination with head-mounted hearing devices (HHDs) have been explored [7–13]. It has been shown that using an eMic can increase both noise reduction and binaural cue preservation performance, depending on the position of the eMic [10, 12].

To implement the binaural MVDR beamformer, an estimate of the relative transfer functions (RTFs) of the desired speech source between all microphones and the reference microphones on both HHDs are required. Instead of using *reverberant* RTFs, one can also use *anechoic* RTFs. When an estimate of the direction-of-arrival (DOA) of the desired speech source is available these anechoic RTFs can be easily constructed for the head-mounted microphones, e.g., based on measurements or head models. However, even when the DOA of the desired speech source (relative to head) is known, this can not be used to compute the (anechoic or reverberant) RTF between the reference microphones and the eMic, since the position of the eMic is not known. Hence, the (anechoic or reverberant) RTF needs to be estimated from the microphone signals.

In this paper, we investigate the influence of three different RTF estimation methods [14–17] on the noise reduction and binaural cue preservation performance of the binaural MVDR-N beamformer for a scenario with one desired speech source surrounded by diffuse multi-talker noise in a reverberant environment. As will be seen, the so-called covariance whitening [14, 15, 17] outperforms the others in terms of noise reduction and binaural cue preservation performance.

2. Configuration and notation

2.1. Signal model

Consider the multiple-input binaural-output (MIBO) system depicted in Fig. 1, consisting of a HHD with M_L microphones on the left side of the head, a HHD with M_R microphones on the right side of the head and an additional eMic, located somewhere else in the room at an unknown position. The m-th microphone signal in the left HHD $Y_{L,m}(\omega)$ can be written in the frequency-domain as

$$Y_{L,m}(\omega) = X_{L,m}(\omega) + N_{L,m}(\omega), \quad m = 1, \ldots, M_L, \quad (1)$$

with $X_{L,m}(\omega)$ the speech component and $N_{L,m}(\omega)$ the noise component. The m-th microphone signal in the right HHD $Y_{R,m}(\omega)$ can be written similarly. The eMic signal $Y_e(\omega)$ can

This work was supported in part by the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG), and by the joint Lower Saxony-Israeli Project ATHENA.

be written as

$$Y_e(\omega) = X_e(\omega) + N_e(\omega), \tag{2}$$

with $X_e(\omega)$ the speech component and $N_e(\omega)$ the noise component in the eMic signal. For conciseness, we will omit the frequency variable ω in the remainder of the paper whenever possible. All microphone signals can be stacked in an M-dimensional vector, with $M = M_L + M_R + 1$, as

$$\mathbf{y} = [Y_{L,1} \ \dots \ Y_{L,M_L} \ Y_{R,1} \ \dots \ Y_{R,M_R} \ Y_e]^T, \tag{3}$$

which can be written as

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \tag{4}$$

where the vectors $\mathbf{x}$ and $\mathbf{n}$ are defined similarly as (3).

For a single desired speech source the speech vector $\mathbf{x}$ is given by

$$\mathbf{x} = \mathbf{a}S, \tag{5}$$

where the vector $\mathbf{a}$ contains the acoustic transfer functions (ATFs) between the desired speech source and all microphones and S is the (dry) speech signal. Please note that in the time-domain the vector $\mathbf{a}$ corresponds to the room impulse responses (RIRs) between the desired speech source and all microphones and hence includes reverberation.

Without loss of generality, we define the first microphone of both HHDs as the reference microphones. For ease of notation, the reference microphone signals $Y_{L,1}$ and $Y_{R,1}$ are further denoted as Y_L and Y_R and can be written as

$$Y_L = \mathbf{e}_L^T \mathbf{y}, \quad Y_R = \mathbf{e}_R^T \mathbf{y}, \tag{6}$$

where $\mathbf{e}_L$ and $\mathbf{e}_R$ denote M-dimensional zero vectors with $\mathbf{e}_L(1) = 1$ and $\mathbf{e}_R(M_L + 1) = 1$. Similarly, the eMic signal can be written as $Y_e = \mathbf{e}_e^T \mathbf{y}$, with $\mathbf{e}_e = [0 \dots 1]^T$. Using (6), the reference microphone signals can be written as

$$Y_L = \underbrace{A_L S}_{X_L} + N_L, \quad Y_R = \underbrace{A_R S}_{X_R} + N_R, \tag{7}$$

where $A_L = \mathbf{e}_L^T \mathbf{a}$ and $A_R = \mathbf{e}_R^T \mathbf{a}$ denote the ATFs between the reference microphones and the desired speech source. The anechoic ATFs (not including reverberation) are denoted as $\bar{A}_L$ and $\bar{A}_R$. The RTF vectors for the left and the right HHD, relating the ATF vector $\mathbf{a}$ to the reference microphones [15, 16], are defined as

$$\mathbf{h}_L = \frac{\mathbf{a}}{A_L}, \quad \mathbf{h}_R = \frac{\mathbf{a}}{A_R}. \tag{8}$$

The speech and noise correlation matrices are given by

$$\mathbf{R}_x = \mathcal{E}\left\{\mathbf{x}\mathbf{x}^H\right\} = \Phi_s \mathbf{a}\mathbf{a}^H, \tag{9}$$

$$\mathbf{R}_n = \mathcal{E}\left\{\mathbf{n}\mathbf{n}^H\right\}, \tag{10}$$

with $\mathcal{E}\{\cdot\}$ the expectation operator, ^{H} the conjugate transpose and $\Phi_s = \mathcal{E}\left\{|S|^2\right\}$ the power spectral density (PSD) of the speech signal. The noise correlation matrix is assumed to be full rank and hence invertible. By assuming statistical independence between $\mathbf{x}$ and $\mathbf{n}$, the correlation matrix of the microphone signals can be written as

$$\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_n. \tag{11}$$

The (binaural) output signals of the left and the right HHD are obtained by filtering *all* microphone signals, including the external microphone signal, with the complex-valued filter vectors $\mathbf{w}_L$ and $\mathbf{w}_R$, respectively, i.e.,

$$Z_L = \mathbf{w}_L^H \mathbf{y}, \quad Z_R = \mathbf{w}_R^H \mathbf{y}. \tag{12}$$

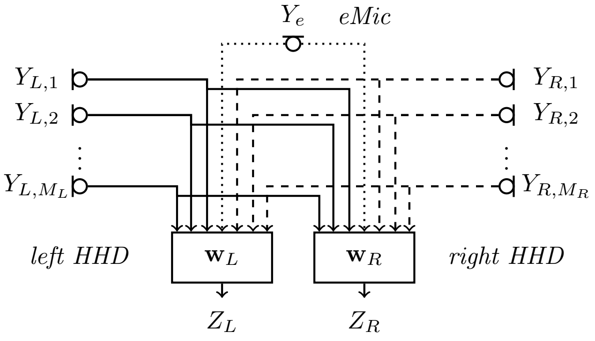


Figure 1: *MIBO system consisting of two head-mounted hearing devices and an external microphone*

The speech component in the output signals is given by $Z_{x,L}$ and $Z_{x,R}$.

2.2. Binaural cues

In addition to monaural cues, binaural cues are used by the listener to localize sound sources and to get a sense of the surrounding sound field [18, 19]. For coherent (directional) sound sources the most descriptive binaural cues are the interaural level difference (ILD) and the interaural time difference (ITD). The interaural coherence (IC) is important for source localization in multi-source and reverberant environments since it determines the reliability of the ILD and ITD cues [19, 20].

The input interaural transfer function (ITF) of the speech component is defined as

$$\text{ITF}_x^{\text{in}} = \frac{\mathcal{E}\left\{X_L X_R^*\right\}}{\mathcal{E}\left\{|X_R|^2\right\}} = \frac{\mathbf{e}_L^T \mathbf{R}_x \mathbf{e}_R}{\mathbf{e}_R^T \mathbf{R}_x \mathbf{e}_R}. \tag{13}$$

The output ITF of the speech component is similarly defined as

$$\text{ITF}_x^{\text{out}} = \frac{\mathcal{E}\left\{Z_{x,L} Z_{x,R}^*\right\}}{\mathcal{E}\left\{|Z_{x,R}|^2\right\}} = \frac{\mathbf{w}_L^H \mathbf{R}_x \mathbf{w}_R}{\mathbf{w}_R^H \mathbf{R}_x \mathbf{w}_R}. \tag{14}$$

The input ILD of the speech component is defined as the power ratio of the speech component in the left and the right HHD [4], i.e.,

$$\text{ILD}_x^{\text{in}} = \frac{\mathcal{E}\left\{|X_L|^2\right\}}{\mathcal{E}\left\{|X_R|^2\right\}} = \frac{\mathbf{e}_L^T \mathbf{R}_x \mathbf{e}_L}{\mathbf{e}_R^T \mathbf{R}_x \mathbf{e}_R}. \tag{15}$$

The output ILD of the speech component is similarly defined as

$$\text{ILD}_x^{\text{out}} = \frac{\mathcal{E}\left\{|Z_{x,L}|^2\right\}}{\mathcal{E}\left\{|Z_{x,R}|^2\right\}} = \frac{\mathbf{w}_L^H \mathbf{R}_x \mathbf{w}_L}{\mathbf{w}_R^H \mathbf{R}_x \mathbf{w}_R}. \tag{16}$$

The ITD can be calculated from the ITF as [4]

$$\text{ITD} = \frac{\angle \text{ITF}}{\omega}, \tag{17}$$

with $\angle$ denoting the phase. The input noise IC is defined as

$$\text{IC}_n^{\text{in}} = \frac{\mathbf{e}_L^T \mathbf{R}_n \mathbf{e}_R}{\sqrt{(\mathbf{e}_L^T \mathbf{R}_n \mathbf{e}_L)(\mathbf{e}_R^T \mathbf{R}_n \mathbf{e}_R)}}. \tag{18}$$

The output noise IC is similarly defined as

$$\text{IC}_n^{\text{out}} = \frac{\mathbf{w}_L^H \mathbf{R}_n \mathbf{w}_R}{\sqrt{(\mathbf{w}_L^H \mathbf{R}_n \mathbf{w}_L)(\mathbf{w}_R^H \mathbf{R}_n \mathbf{w}_R)}}. \tag{19}$$

The (real-valued) magnitude-squared coherence (MSC) is defined as $\text{MSC} = |\text{IC}|^2$.

3. Binaural noise reduction

In this section we introduce a binaural noise reduction approach that uses all microphones to spatially filter the microphone inputs. The binaural MVDR-N beamformer [2, 4–6] minimizes the output noise PSD while preserving the speech component in the reference microphone signals (and hence the binaural cues of the speech component) and a scaled version of the noise component in the reference microphone signals. The constraint optimization problem for the left filter can be formulated as

$$\min_{\mathbf{w}_L} \mathcal{E}\left\{|\mathbf{w}_L^H\mathbf{n} - \eta N_L|^2\right\} \quad \text{s.t.} \quad \mathbf{w}_L^H\mathbf{a} = A_L. \tag{20}$$

The solution for the left filter is given by [6, 21]

$$\mathbf{w}_{\eta,L} = (1-\eta)\overbrace{\frac{\mathbf{R}_n^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{R}_n^{-1}\mathbf{a}}A_L^*}^{\mathbf{w}_{0,L}} + \eta\mathbf{e}_L, \tag{21}$$

$$= (1-\eta)\frac{\mathbf{R}_n^{-1}\mathbf{h}_L}{\mathbf{h}_L^H\mathbf{R}_n^{-1}\mathbf{h}_L} + \eta\mathbf{e}_L, \tag{22}$$

with $0 \leq \eta \leq 1$ a real-valued mixing parameter. The solution for the right filter is similar to (22) by substituting R for L. Using (22) in (12), the output of the binaural MVDR-N beamformer can be interpreted as a mixture between the binaural MVDR beamformer output (scaled with $1 - \eta$) and the (noisy) reference microphone signal (scaled with η).

For $\eta = 0$ the binaural MVDR-N beamformer is equal to the binaural MVDR beamformer $\mathbf{w}_{0,L}$ [2, 3, 22] and hence preserves the ILD and ITD cues of the desired speech source [4]. However, it has been shown in [6] that for the binaural MVDR beamformer the output noise MSC is equal to 1 and hence the surrounding noise field is perceived as coming from the same direction as the desired speech source. For $\eta = 1$ the binaural MVDR-N beamformer output is equal to the reference microphone signals in (6) and hence preserves the binaural cues of both the desired speech source and the noise component, although no noise reduction is achieved. Hence, the binaural MVDR-N beamformer trades off noise reduction against binaural cue preservation of the noise component using the mixing parameter η.

Since accurately estimating the ATF vector $\mathbf{a}$ is known to be difficult [23], several methods for estimating the RTF vectors $\mathbf{h}_L$ and $\mathbf{h}_R$ have been proposed [14–17] and hence the usage of (22) is preferred. If all microphone positions are known and a reliable DOA estimation is available, one can also use measured [24] or simulated [25] *anechoic* RTF vectors. While this is a reasonable (and robust) approach when only using the head-mounted microphones, the exact position of the eMic is usually not known. Hence, at least for the eMic, other methods, e.g., estimated RTFs between the reference microphones and the eMic need to be considered.

Due to robustness, we use anechoic RTFs for the head-mounted microphones (assuming the DOA θ to be known) and estimated RTFs only for the eMic, i.e.,

$$\boxed{\tilde{\mathbf{h}}_L = \begin{bmatrix}\bar{\mathbf{h}}_L(\theta)\\ H_{e,L}\end{bmatrix}, \quad \tilde{\mathbf{h}}_R = \begin{bmatrix}\bar{\mathbf{h}}_R(\theta)\\ H_{e,R}\end{bmatrix}} \tag{23}$$

where $\bar{\mathbf{h}}_L(\theta)$ and $\bar{\mathbf{h}}_R(\theta)$ denote the M_L- and M_R-dimensional anechoic (measured or simulated) RTF vectors which depend on the DOA θ for the left and the right HHD, respectively, and $H_{e,L}$ and $H_{e,R}$ denote the estimated (anechoic or reverberant) RTFs between the HHD reference microphones and the eMic. The construction of the RTF vectors is schematically depicted in Fig. 2.

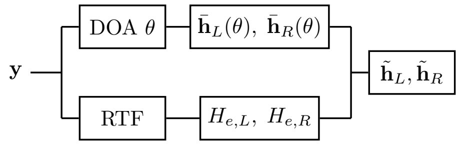

y
DOA θ
$\bar{\mathbf{h}}_L(\theta),\ \bar{\mathbf{h}}_R(\theta)$
RTF
$H_{e,L},\ H_{e,R}$
$\tilde{\mathbf{h}}_L, \tilde{\mathbf{h}}_R$

Figure 2: *Proposed construction of the RTF vectors*

By using anechoic RTFs for both the head-mounted microphones and the eMic the RTF vectors are connected by a simple factor, i.e., $\tilde{\mathbf{h}}_L = \tilde{\mathbf{h}}_R\frac{\bar{A}_L}{\bar{A}_R}$ and hence are parallel. This leads to the aforementioned mapping of the noise component to the position of the desired speech source and hence the output noise MSC being equal to 1. By mixing anechoic and reverberant RTFs, i.e., estimating reverberant RTFs for the eMic, the RTF vectors are not parallel, which leads to partial cue preservation of the noise component even when the mixing parameter η is set to 0 as will be seen in the experimental results in Section 5.

4. RTF estimation methods

In this section we describe three different methods to estimate the RTFs $H_{e,L}$ and $H_{e,R}$ between the head-mounted reference microphones and the eMic which are then used in (23). Although only the estimators for $H_{e,L}$ are discussed, the estimators for $H_{e,R}$ can again simply be obtained by substituting R for L. Using the speech correlation matrix in (9), the RTF between the left reference microphone and the eMic is given by

$$H_{e,L} = \frac{\mathbf{e}_e^T\mathbf{R}_x\mathbf{e}_L}{\mathbf{e}_L^T\mathbf{R}_x\mathbf{e}_L} = \frac{A_e}{A_L}. \tag{24}$$

4.1. Biased approach

Assuming a reasonable large SNR, the speech correlation matrix in (24) can simply be replaced by the (noisy) correlation matrix of the microphone signals $\mathbf{R}_y$ in (11), leading to the biased estimator

$$\boxed{H_{e,L}^{\mathrm{b}} = \frac{\mathbf{e}_e^T\mathbf{R}_y\mathbf{e}_L}{\mathbf{e}_L^T\mathbf{R}_y\mathbf{e}_L} = \frac{\mathcal{E}\left\{Y_eY_L^*\right\}}{\mathcal{E}\left\{|Y_L|^2\right\}}} \tag{25}$$

Generally, by using the biased estimator in (25) to estimate $H_{e,L}^{\mathrm{b}}$ and $H_{e,R}^{\mathrm{b}}$, the RTF vectors in (23) are not parallel.

4.2. MVDR pre-processed RTF estimation

An alternative approach to estimate the RTFs was proposed in [13], where it was proposed to pre-process the head-mounted microphones using an MVDR beamformer. The binaural MVDR beamformer only using the HHDs can be written in terms of the anechoic RTFs vectors $\bar{\mathbf{h}}_L(\theta)$ and $\bar{\mathbf{h}}_R(\theta)$ as

$$\mathbf{w}_{\mathrm{H},L} = \begin{bmatrix}\frac{\mathbf{R}_{n,\mathrm{H}}^{-1}\bar{\mathbf{h}}_L(\theta)}{\bar{\mathbf{h}}_L^H(\theta)\mathbf{R}_{n,\mathrm{H}}^{-1}\bar{\mathbf{h}}_L(\theta)} & 0\end{bmatrix}^T, \tag{26}$$

where $\mathbf{R}_{n,\mathrm{H}}$ is the $(M-1) \times (M-1)$-dimensional noise correlation matrix only using the head-mounted microphones. The MVDR pre-processed (biased) RTF estimate is then given by

$$\boxed{H_{e,L}^{\mathrm{pp}} = \frac{\mathcal{E}\left\{Y_e\mathbf{y}^H\mathbf{w}_{\mathrm{H},L}\right\}}{\mathcal{E}\left\{\mathbf{w}_{\mathrm{H},L}^H\mathbf{y}\mathbf{y}^H\mathbf{w}_{\mathrm{H},L}\right\}} = \frac{\mathbf{e}_e^T\mathbf{R}_y\mathbf{w}_{\mathrm{H},L}}{\mathbf{w}_{\mathrm{H},L}^H\mathbf{R}_y\mathbf{w}_{\mathrm{H},L}}} \tag{27}$$

By substituting $\mathbf{w}_{\mathrm{H},R}$ in (27) it can easily be shown that $H_{e,L}^{\mathrm{pp}} = H_{e,R}^{\mathrm{pp}}\frac{\bar{A}_R}{\bar{A}_L}$ and hence, by using the pre-processed estimator in

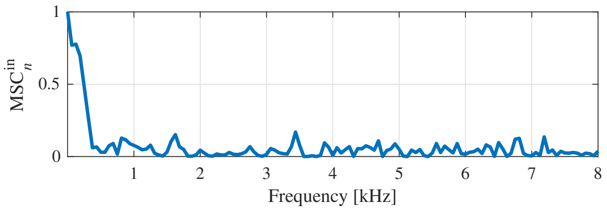


Figure 3: *Input noise MSC generated by four loudspeakers*

(27) the RTF vectors in (23) are parallel, what leads to the mapping of the residual noise to the position of the desired speech source.

4.3. Covariance whitening

Covariance whitening is a well-known approach to estimate RTFs [14, 15, 17]. The noise correlation matrix can be factorized into a lower triangular matrix $\mathbf{L}$ and its conjugate transpose $\mathbf{L}^H$ using the Cholesky decomposition, i.e., [15, 17]

$$\mathbf{R}_n = \mathbf{L}\mathbf{L}^H, \quad \mathbf{R}_n^{-1} = \mathbf{L}^{-H}\mathbf{L}^{-1}. \tag{28}$$

Using (28), the pre-whitened correlation matrix of the microphone signals is given by

$$\mathbf{R}_y^{\mathrm{w}} = \mathbf{L}^{-1}\mathbf{R}_y\mathbf{L}^{-H}. \tag{29}$$

The eigenvalue decomposition (EVD) of this pre-whitened matrix is given by

$$\mathbf{R}_y^{\mathrm{w}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^H, \tag{30}$$

with $\mathbf{V}$ an $M \times M$-dimensional matrix containing the eigenvectors and $\boldsymbol{\Lambda}$ an $M \times M$-dimensional diagonal matrix containing the eigenvalues. Using the eigenvector $\mathbf{v}_{\max}$ that corresponds to the largest eigenvalue, the RTF can be estimated as [15, 17]

$$\boxed{H_{e,L}^{\mathrm{cw}} = \frac{\mathbf{e}_e^T\mathbf{L}\mathbf{v}_{\max}}{\mathbf{e}_L^T\mathbf{L}\mathbf{v}_{\max}}} \tag{31}$$

Compared to the MVDR pre-processed approach in Section 4.2, the covariance whitening approach aims at estimating the *reverberant* RTFs and hence, the RTF vectors in (23) are not parallel.

5. Experimental results

5.1. Setup

All signals were recorded in a laboratory with variable acoustics ($7\ \mathrm{m} \times 6\ \mathrm{m} \times 2.7\ \mathrm{m}$) where the reverberation time was set to about $350\ \mathrm{ms}$. We used two behind-the-ear (BTE) hearing aid dummies each having two microphones with an inter-microphone distance of about $7.6\ \mathrm{mm}$, and an external microphone, i.e., $M = 5$ microphones in total. The hearing aids were placed on the ears of a head-and-torso simulator (HATS) that was placed in the middle of the room. The desired speech source was played back by a loudspeaker placed at about $2\ \mathrm{m}$ distance to the middle of the head at an angle of about $35°$, i.e., on the right side of the HATS. The background multi-talker noise was realized by four loudspeakers in the corners of the room that were facing the corners and playing back uncorrelated multi-talker noise. Fig. 3 shows the measured input noise MSC using the first microphone of each hearing aid as reference microphone. The speech and noise signals were recorded separately such that we were able to mix them at different input SNRs afterwards. The external microphone was placed at $0.5\ \mathrm{m}$ distance to the desired speech source parallel to the viewing direction of the HATS.

$\mathrm{iSNR}_R^{\mathrm{in}}$ [dB]	-10	-5	0	5
$\mathrm{iSNR}_L^{\mathrm{in}}$ [dB]	-14.5	-9.5	-4.5	0.5
$\mathrm{iSNR}_e^{\mathrm{in}}$ [dB]	-2.5	2.5	7.5	12.5

Table 1: *Input intelligibility-weighted SNRs*

For the anechoic RTF vectors $\bar{\mathbf{h}}_L(\theta)$ and $\bar{\mathbf{h}}_R(\theta)$ used in (23) we used the database presented in [24] who used similar hearing aid dummies in an anechoic room. We assumed a DOA of $35°$ and chose the respective measurements from the database. The processing was done at a sampling rate of $16\ \mathrm{kHz}$ using an STFT-based weighted overlap-add framework with a frame length of $16\ \mathrm{ms}$ (256 samples) and a frame shift of 50%. The input signals consisted of $2\ \mathrm{s}$ noise-only followed by $18\ \mathrm{s}$ of speech-plus-noise. The noise correlation matrix $\hat{\mathbf{R}}_n$ was estimated during the noise-only part, whereas the microphone signal correlation matrix $\hat{\mathbf{R}}_y$ was estimated during the speech-plus-noise part. The RTFs between the reference microphones and the external microphone were estimated using $\hat{\mathbf{R}}_n$ and $\hat{\mathbf{R}}_y$, cf. Section 4. The obtained filters in (22) were applied to the complete signal

We evaluated four different filters, namely

- the binaural MVDR beamformer in (26) only using the head-mounted microphones ($\mathbf{w}_{\mathrm{H}}$)
- the binaural MVDR-N beamformer in (22) using either the biased RTF estimate in (25) ($\mathbf{w}_\eta^{\mathrm{b}}$), the pre-processed RTF estimate in (27) ($\mathbf{w}_\eta^{\mathrm{pp}}$) or the covariance whitening RTF estimate in (31) ($\mathbf{w}_\eta^{\mathrm{cw}}$)

As objective performance measures we used the intelligibility-weighted SNR (iSNR) [26] improvement for the left and the right hearing aid relative to the reference microphone signals, the MSC error comparing the input noise MSC (cf. Fig. 3) with the output noise MSC, and the ILD and ITD errors comparing the input speech ILD and ITD with the output speech ILD and ITD. All measures have been averaged over all frequencies. We set up two experiments where we changed either the input iSNR or the mixing paramter η.

5.2. Experiment 1

In the first experiment we varied the input iSNR in the right reference microphone ($\mathrm{iSNR}_R^{\mathrm{in}}$) from $-10\ \mathrm{dB}$ to $5\ \mathrm{dB}$ in steps of $5\ \mathrm{dB}$. This led to the input iSNRs for the left reference microphone ($\mathrm{iSNR}_L^{\mathrm{in}}$) and the eMic ($\mathrm{iSNR}_e^{\mathrm{in}}$) as shown in Table 1. The mixing parameter was set to $\eta = 0$ such that the filter in (22) is equal to the binaural MVDR beamformer $\mathbf{w}_{0,L}$.

The results are depicted in Fig. 4. As can be observed, the performance of the filter $\mathbf{w}_{\mathrm{H}}$ does not depend on the input iSNR, whereas for the filters that exploit an RTF estimate between the reference microphones and the eMic the input iSNR influences the performance. The binaural MVDR beamformer using the covariance whitening RTF estimate $\mathbf{w}_\eta^{\mathrm{cw}}$ clearly outperforms all other filters in all objective measures.

It can be observed especially for the right iSNR improvement that the covariance whitening RTF estimate is less affected by a low input iSNR. For all values of the right input iSNR the covariance whitening RTF estimate leads to the highest output iSNR for both the left and the right side. Further, the filter using the pre-processed RTF estimate $\mathbf{w}_\eta^{\mathrm{pp}}$ always leads to a higher output iSNR than the filter using the biased RTF estimate $\mathrm{w}_\eta^{\mathrm{b}}$. The filter $\mathbf{w}_{\mathrm{H}}$ always leads to the lowest output iSNR.

For the MSC error of the noise component the filters using

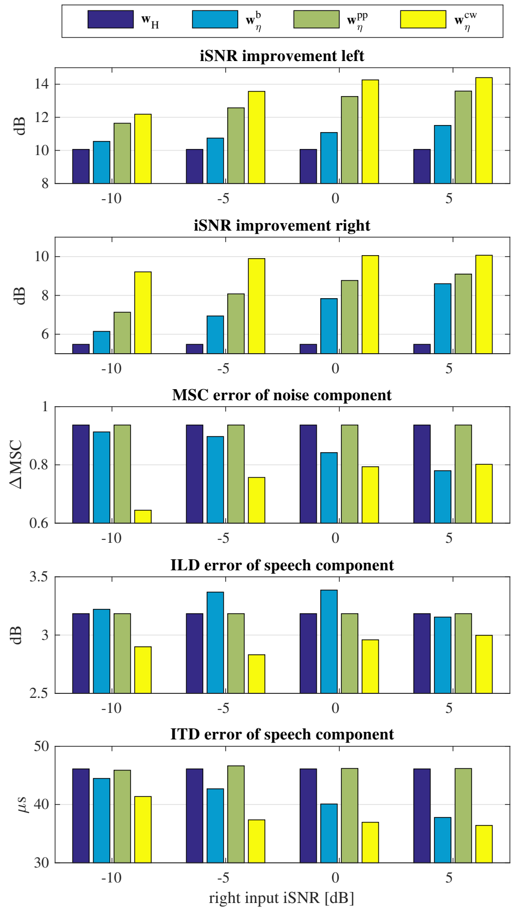


Figure 4: *Results of the first experiment where the input* iSNR *in the right reference microphone has been changed and the mixing paramter has been set to* $\eta = 0$

parallel RTF vectors ($\mathbf{w}_\mathrm{H}$ and $\mathbf{w}_\eta^\mathrm{pp}$) lead to a constant value, whereas the filters using non-parallel RTF vectors ($\mathbf{w}_\eta^\mathrm{b}$ and $\mathbf{w}_\eta^\mathrm{cw}$) lead to smaller errors. The MSC error of the noise component decreases with increasing right input iSNR for the filter using the biased estimate $\mathbf{w}_\eta^\mathrm{b}$ and increases for the filter using the covariance whitening estimate $\mathbf{w}_\eta^\mathrm{cw}$. While $\mathbf{w}_\eta^\mathrm{cw}$ outperforms $\mathbf{w}_\eta^\mathrm{b}$ for low right input iSNRs, the biased approach leads to the smallest MSC error for the highest right input iSNR and hence outperforms all other filters in this condition.

The ILD error of the speech component does not vary much with changes of the right input iSNR, but $\mathbf{w}_\eta^\mathrm{cw}$ outperforms all other filters in all conditions.

The ITD error of the speech component is constant over all conditions for the filters using the parallel RTF vectors ($\mathbf{w}_\mathrm{H}$ and $\mathbf{w}_\eta^\mathrm{pp}$) and decreasing with increasing right input iSNR for the filters using non-parallel RTF vectors ($\mathbf{w}_\eta^\mathrm{b}$ and $\mathbf{w}_\eta^\mathrm{cw}$), while $\mathbf{w}_\eta^\mathrm{cw}$ outperforms $\mathbf{w}_\eta^\mathrm{b}$.

In conclusion, it appears that even when using anechoic RTFs for the head-mounted microphones, using reverberant RTF estimates between the reference microphones and the external microphone (as in $\mathbf{w}_\eta^\mathrm{b}$ and $\mathbf{w}_\eta^\mathrm{cw}$) may lead to slight binaural cue preservation of the noise without even applying partial noise estimation.

5.3. Experiment 2

In the second experiment we set the input iSNR in the right reference microphone to -5 dB (cf. Table 1) and varied the mixing parameter η in (22) from 0 to 0.2 in steps of 0.05. The results for the second experiment are depicted in Fig. 5. The binaural MVDR beamformer using only the head-mounted microphones $\mathbf{w}_\mathrm{H}$ is obviously not affected by the mixing parameter η but yields a reference of the filter performance without incorporating an eMic.

In terms of iSNR improvement the performance of the binaural MVDR-N beamformer using an external microphone is better than the binaural MVDR beamformer only using the head-mounted microphones for small values of η. This effect decreases with increasing η, i.e., the output iSNR of the binaural MVDR-N beamformer is decreasing with η. For low values of η the filter using the covariance whitening RTF estimate $\mathbf{w}_\eta^\mathrm{cw}$ clearly outperforms all other filters, while for larger η the distance to the other filters decreases. Hence, it appears that η has higher influence on $\mathbf{w}_\eta^\mathrm{cw}$ than on $\mathbf{w}_\eta^\mathrm{b}$ and $\mathbf{w}_\eta^\mathrm{pp}$. The preprocessing done in $\mathbf{w}_\eta^\mathrm{pp}$ proves beneficial for all values of η compared to the filter using the biased estimate $\mathbf{w}_\eta^\mathrm{b}$.

The MSC error of the noise component is decreasing with η for the binaural MVDR-N beamformer, which is intuitively clear because more and more of the noisy reference microphone signal is added to the beamformer output. The filter $\mathbf{w}_\eta^\mathrm{cw}$ clearly outperforms all other filters, while $\mathbf{w}_\eta^\mathrm{b}$ only slightly outperforms $\mathbf{w}_\eta^\mathrm{pp}$ for very small values of η.

The ILD and ITD errors of the speech component are decreasing with increasing η for the binaural MVDR-N beamformer. Please note, that in theory the ILD and ITD errors of the speech component are equal to 0 but due to the use of anechoic RTFs these errors occur. The filter $\mathbf{w}_\eta^\mathrm{cw}$ again outperforms all other filters, while $\mathbf{w}_\eta^\mathrm{pp}$ outperforms $\mathbf{w}_\eta^\mathrm{b}$ in terms of ILD error, and $\mathbf{w}_\eta^\mathrm{b}$ outperforms $\mathbf{w}_\eta^\mathrm{pp}$ in terms of ITD error.

6. Conclusions

In this paper we investigated the influence of three different RTF estimators that estimate the RTFs between the reference microphones of two head-mounted hearing devices and an external microphone on the noise reduction and binaural cue preservation performance of the binaural MVDR-N beamformer using recorded signals. The estimator using so-called covariance whitening outperformed the other estimators. Additionaly, it appeared that using anechoic RTFs for the head-mounted microphones and reverberant RTFs for the external microphone leads to slight binaural cue preservation without even applying partial noise estimation.

7. References

[1] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: Technologies and algorithms," in *Advances in Digital Speech Transmission*. New York, NY, USA: Wiley, 2008, pp. 401–429.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Wiley, 2010, pp. 269–302.

[3] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[4] B. Cornelis, S. Doclo, T. Van den Bogaert, J. Wouters, and

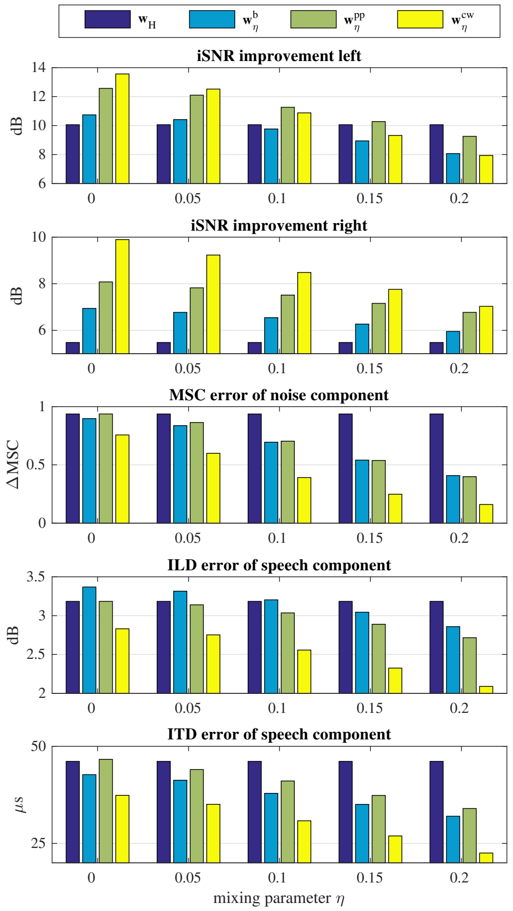


Figure 5: *Results of the second experiment where the mixing parameter* η *has been changed and the right input* iSNR *has been set to* -5 dB

M. Moonen, “Theoretical analysis of binaural multi-microphone noise reduction techniques,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[5] T. Klasen, T. van den Bogaert, M. Moonen, and J. Wouters, “Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,” *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[6] D. Marquardt, “Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques,” Ph.D. dissertation, Carl von Ossietzky Universität Oldenburg, 2015.

[7] A. Bertrand and M. Moonen, “Robust Distributed Noise Reduction in Hearing Aids with External Acoustic Sensor Nodes,” *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 14 pages, Jan. 2009.

[8] N. Cvijanovic, O. Sadiq, and S. Srinivasan, “Speech enhancement using a remote wireless microphone,” *IEEE Transactions on Consumer Electronics*, vol. 59, no. 1, pp. 167–174, Feb. 2013.

[9] D. Yee, H. Kamkar-Parsi, H. Puder, and R. Martin, “A speech enhancement system using binaural hearing aids and an external microphone,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 246–250.

[10] J. Szurley, A. Bertrand, B. Van Dijk, and M. Moonen, “Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 5, pp. 952–966, May 2016.

[11] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, “Informed sound source localization using relative transfer functions for hearing aid applications,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 3, pp. 611–623, Jan. 2017.

[12] N. Gößling, D. Marquardt, and S. Doclo, “Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field,” in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, USA, Mar. 2017, pp. 1–5.

[13] R. Ali, T. van Waterschoot, and M. Moonen, “A noise reduction strategy for hearing devices using an external microphone,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos island, Greece, Aug. 2017, (submitted).

[14] S. Markovich, S. Gannot, and I. Cohen, “Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[15] S. Markovich-Golan and S. Gannot, “Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015.

[16] I. Cohen, “Relative transfer function identification using speech signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[17] I. Kodrasi and S. Doclo, “EVD-Based Multi-Channel Dereverberation of a Moving Speaker Using Different RETF Estimation Methods,” in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, USA, Mar. 2017, pp. 116–120.

[18] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Mass. MIT Press, 1997.

[19] C. Faller and J. Merimaa, “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,” *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[20] M. Dietz, S. D. Ewert, and V. Hohmann, “Auditory model based direction estimation of concurrent speakers from binaural signals,” *Speech Communication*, vol. 53, pp. 592–605, 2011.

[21] D. Marquardt, V. Hohmann, and S. Doclo, “Interaural Coherence Preservation in MWF-based Binaural Noise Reduction Algorithms using Partial Noise Estimation,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 654–658.

[22] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, “Theoretical Analysis of Binaural Transfer Function MVDR Beamformers with Interference Cue Preservation Constraints,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[23] Y. Huang and J. Benesty, “A class of frequency-domain adaptive approaches to blind multichannel identification,” *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[24] H. Kayser, S. Ewert, J. Annemüller, 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, p. 10 pages, 2009.

[25] D. P. Jarret, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, “Rigid sphere room impulse response simulation: Algorithm and application,” *Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[26] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, “Intelligibility-weighted measures of speech-to-interference ratio and speech system performance,” *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

List of Authors