Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Binary Masking & Speech Intelligibility

Boldt, Jesper

Publication date:
2010

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Boldt, J. (2010). *Binary Masking & Speech Intelligibility*. Institut for Elektroniske Systemer, Aalborg Universitet.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: February 23, 2025

Binary Masking
&
Speech Intelligibility

Ph.D. Thesis

JESPER BÜNSOW BOLDT

January, 2011

Department of Electronic Systems
Aalborg University
Fredrik Bajers Vej 7
9220 Aalborg Ø, Denmark

Boldt, Jesper Bünsow
Binary masking & Speech Intelligibility
ISBN 978-87-92328-50-2

Copyright ©2010 Jesper Bünsow Boldt, except where otherwise stated.
All rights reserved.

Department of Electronic Systems
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg Ø
Denmark

This thesis was written in LaTeX.

Abstract

The purpose of this thesis is to examine how binary masking can be used to increase intelligibility in situations where hearing impaired listeners have difficulties understanding what is being said. The major part of the experiments carried out in this thesis can be categorized as either experiments under ideal conditions or as experiments under more realistic conditions useful for real-life applications such as hearing aids. In the experiments under ideal conditions, the previously defined ideal binary mask is evaluated using hearing impaired listeners, and a novel binary mask – the target binary mask – is introduced. The target binary mask shows the same substantial increase in intelligibility as the ideal binary mask and is proposed as a new reference for binary masking. In the category of real-life applications, two new methods are proposed: a method for estimation of the ideal binary mask using a directional system and a method for correcting errors in the target binary mask. The last part of the thesis proposes a new method for objective evaluation of speech intelligibility.

This thesis consists of an introduction followed by a collection of papers. The introduction begins with a description of the problem facing a hearing impaired person in difficult listening situations, which is followed by a general introduction to hearing impairment and hearing aids. After this outline, the concept of binary masking is introduced through descriptions of different reference masks (oracle masks), as well as methods for estimation and application of binary masks and comparison to the well-known Wiener filter. Finally, the difference between speech intelligibility and speech quality is considered, and methods for evaluation of speech intelligibility are discussed.

The collection of papers is the main part of the thesis. The first three papers (A–C) evaluate the intelligibility of speech in noise under ideal conditions using the ideal binary mask and the target binary mask. The results presented in the first three papers show the value of the ideal binary mask and the target binary mask for both hearing impaired listeners and normal hearing listeners. Consequently, methods for estimation and error-correction of the ideal binary masks and target binary mask are proposed in Paper D and E. Finally, Paper F proposes a simple method for measuring the intelligibility of binary masked noisy speech.

Abstract (in Danish)

I denne afhandling undersøges, hvordan binære masker kan bruges til øge taleforståeligheden i situationer, hvor hørehæmmede har problemer med at forstå, hvad der bliver sagt. Størstedelen af arbejdet kan kategoriseres som enten lytteforsøg under ideelle betingelser eller som algoritmeudvikling til brug i f.eks. hørerapparater. I lytteforsøgene bliver den ideelle binære mask evalueret med hørehæmmede testpersoner og en ny binær maske bliver defineret. Denne nye binære maske giver de samme forbedringer på taleforståelighed, som den ideelle binære maske og kan derfor ses som en ny reference indenfor binære masker. Under mere realistiske betingelser bliver to nye algoritmer indenfor binære masker evalueret. Den ene algoritme kan bruges til at estimere den ideelle binære maske med to mikrofoner, og den anden kan bruge til at rette fejl i den binære maske. Den sidste del af afhandlingen omhandler en metode til at beregne taleforståelighed vha. en simpel algoritme og uden brug af testpersoner.

Afhandlingen består af en introduktion og en samling af artikler. Introduktion beskriver de problemer, som en hørehæmmet person oplever i vanskelige lydmiljøer og giver en generel introduktion til høretab og hørerapparater. Derefter bliver brugen af forskellige binære masker beskrevet og metoder til at beregne og bruge dem bliver gennemgået. En sammenligning med det klassiske Wiener-filter er også udført. Til sidst i introduktionen er forskellen mellem taleforståelighed og lydkvalitet beskrevet og forskellige metoder til at måle taleforståelighed er gennemgået.

Samlingen af artikler udgør hovedparten af denne afhandling. De første tre artikler (A-C) måler forståeligheden af tale i støj, når den ideelle binære mask og den nye binære maske bruges til at separere talen fra støjen. Resultaterne viser, at de binære masker er brugbare for både normalt hørende og hørehæmmede personer, og det er derfor relevant at forsøge at beregne dem under mere realistiske situationer (artikel D) og rette fejl i de binære masker (artikel E). Den sidste artikel beskriver en simpel måde til at beregne den opnåede taleforståelighed, når de binære masker bruges til at separere tale fra støj.

List of Papers

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

[A] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech Intelligibility in Background Noise with Ideal Binary Time-Frequency Masking", in *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[B] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech Perception of Noise with Binary Gains", in *Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[C] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, D. Wang, "Role of Mask Pattern in Intelligibility of Ideal Binary-Masked Noisy Speech", in *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[D] J. B. Boldt, U. Kjems, M. S. Pedersen, T. Lunner, D. Wang, "Estimation of the Ideal Binary Mask using Directional Systems", in *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[E] J. B. Boldt, M. S. Pedersen, U. Kjems, M. G. Christensen, S. H. Jensen, "Error-Correction of Binary Masks using Hidden Markov Models", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4722–4725.

[F] J. B. Boldt, and D. P. W. Ellis, "A Simple Correlation-based model of Intelligibility for Nonlinear Speech Enhancement and Separation", in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2009, pp. 1849–1853.

In addition, the following patent has been filed:

"A Method for Correcting Errors in Binary Masks", J. B. Boldt, M. S. Pedersen, U. Kjems, M. G. Christensen, S. H. Jensen, EP09168699.8, August 2009.

Preface

This thesis is written as a partial fulfilment of the requirements for the Ph.D. degree at the International Doctoral School of Technology and Science at Aalborg University. The work was carried out during the period from January 2007 to December 2009 under the Industrial PhD programme, which is approved and supported by the Danish Ministry of Science, Technology, and Innovation. The Industrial PhD Programme is conducted in cooperation between a university and a company – in this case the Danish hearing aid company, Oticon A/S.

My workplace has mainly been at Oticon's headquarter in Copenhagen, Denmark, but several other institutions were involved during the past three years: Aalborg University, Columbia University, Eriksholm Research Centre, and the Technical University of Denmark. At all locations, I have been welcomed with openness and sincere interest, and many people have contributed in various ways to my work, progress, and well-being. The list of names would be very long and yet incomplete, but I would like to thank all of you.

In the beginning of the Industrial PhD programme, the travelling between academia and industry was a challenge, but as time evolved it became a strength in the project. At Aalborg University my supervisors Søren Holdt Jensen and Mads Græsbøll Christensen gave me the academic perspective on new ideas and possibilities. At Oticon my supervisors, Michael Syskind Pedersen and Ulrik Kjems provided me with understanding of applied science and industrial research. I would like to thank all of my supervisors for commitment, valuable advices, feedback, and many interesting and fruitful discussions.

Although I intended not to make a list of names, I feel that the following persons deserve, in particular, to be recognized for their participation in my work, and I would like to express my sincere gratitude for their contributions:

DeLiang Wang from the Department of Computer Science and Engineering and the Center for Cognitive Science at the Ohio State University. DeLiang Wang visited Oticon during the first 6 months of my PhD and had the patience and goodwill to share his knowledge with a beginner in the area of binary masking, CASA, and source separation.

Dan Ellis from the Laboratory for Recognition and Organization of Speech and Audio (LabROSA) at Columbia University. Dan welcomed me and gave me a valuable experience during my 6-months visit in his lab. I am grateful for the many hours of supervision and discussions with Dan, from which I learned a lot about science and how to "do science".

Mads Græsbøll Christensen from the Department of Media Technology, Aalborg University. Mads has supported me, challenged me, and given me an insight and understanding of the academic world. In addition, he has taken good care of me during my visits at Aalborg University, and I appreciate our friendship.

I would also like to thank Søren Kamaric Riis at Oticon for giving me the opportunity to pursue my interests and making this PhD programme possible.

Last but not least, I would like to thank my family and friends, but most of all, Cornelia for support, patience, and encouragement in the past three years.

Jesper Boldt

Jesper Bünsow Boldt

Contents

Introduction

1 The Cocktail Party Problem

In situations where signals from various sources are mixed, source separation can be relevant. Examples can be found in astrophysics, geophysics, biology, and medicine, but the term is mainly used when the sources are sound sources such as human speakers or musical instruments. In this case, source separation is the problem of separating one or more sounds from a mixture of multiple sounds, e.g., to separate a single speaker from a large group of talkers or to separate speech from traffic noise.

Source separation will split the sound mixture into one or more target sounds and one or more interferer sounds, and in some systems, the separated sources will be further processed. Examples of this processing are automatic speech recognizers where source separation can be done prior to the recognition, and communication systems, where separation enhances the speech and reduces the background noise before transmission.

In difficult listening situations, some kind of source separation could also be active in the human auditory system: "How do we recognize what one person is saying when others are speaking at the same time?". This question was formulated by Edward Colin Cherry in 1953 [1], but considered as early as 1870 by Hermann von Helmholtz in his book "On the sensations of tone as a physiological basis for the theory of music" [2]. In this book, Hermann von Helmholtz describes the difficult listening situation in the festive ball-room with musical instruments, speaking men and women, rustling garments, gliding feet, clinking glasses, etc. – a mixture of sounds that are "complicated beyond conception", and yet, "... the ear is able to distinguish all the separate constituent parts of the confused whole..." [3]. This capability of the human auditory system, to focus upon and follow one particular speaker in such a sound environment, has been termed "the cocktail party phenomenon" [4, 5], and the problem of replicating this capability is called the "cocktail party problem". One possible explanation of the cocktail party phenomenon is that the human auditory system efficiently identifies and separates the sources prior to recognition at a higher level. However, there is no clear evidence that this is in fact the approach used by the human ear and brain.

The cocktail party phenomenon usually goes unnoticed by people with normal hearing unless the cocktail party takes place in a large room with high reverberation or loud background music. In this situation, even normal hearing listeners can find it difficult to perceive the target speech correctly and intelligibility can be affected. This problem is more pronounced for hearing impaired listeners, who will experience problems in less difficult situations [6, 7, 8] and might give up following the desired conversation. But even for hearing impaired listeners, the cocktail party phenomenon – and the loss of this capability – is usually not noticed, and the hearing impaired listener will often explain the loss of intelligibility as caused by a difficult environment and not by his or her impaired hearing [9].

To enable source separation, the sources in the mixture must be assigned as either target or interferer. This assignment is dependent on the situation and can quickly change, as seen by the following example: When talking to a person at the cocktail party, the speech from that person is the target sound and everything else is the undesired, interfering sound. At some point, a new person enters the conversation and his speech is removed from the group of interfering sources and becomes a target source. The conversation continues with two alternating target sources, but the subject changes into music preferences, and a new source – the music – has changed from interferer to target.

For the purpose of the work in this thesis, the assignment of sources is simplified: Target sound is the speech from the speaker having the listener's attention, and everything else – including reverberation – is interferer sound. This definition makes the distinction between source separation, speech enhancement, and noise reduction less clear, but in this thesis, all three methods are seen as possible solutions to the cocktail party problem: The decreased intelligibility can be compensated either by separating the target speech from the interfering sounds, by enhancement of the target speech, or by reducing the interfering sound.

When source separation, speech enhancement, and noise reduction, are seen as possible solutions to the cocktail party problem, the available methods are numerous and diverse. They include the classic methods such as Wiener filtering and spectral subtraction [10], as well as more recent methods such as independent component analysis [11] and non-negative matrix factorization [12]. Significant results have been obtained from applying these methods, but the cocktail party problem cannot be considered solved, and hearing aid users still have problems understanding speech in noisy conditions [13, 14].

Even though the cocktail party problem is not yet solved, research continues to contribute to a better understanding of the problem and of the human auditory system. A better understanding of the human auditory system could potentially lead to solutions for the cocktail party problem, and as pointed out by B. Edwards: "Nowadays the limiting factor is our basic knowledge pertaining to the functional requirements of what a hearing aid should actually do" [15]. This statement contains a lot of truth, but the limitations within a hearing aid must not be disregarded. Low processing delay, low complexity, small size, and high robustness are requirements that limit the available solutions to a large degree. Furthermore, the more fundamental problem of how to select the target speech has to be solved.

This thesis discusses binary masking as a further possible solution to the cocktail party problem. To say that binary masking fulfils all the above mentioned requirements of a hearing aid is not reasonable, but the simplicity of the method and the results obtained, make the method interesting in hearing aids as a possible solution to the cocktail party problem.

2 Hearing Impairment

Hearing impairment is a broad term referring to different problems related to hearing and hearing loss. The most common hearing loss is the sensorineural hearing loss [16, 17],

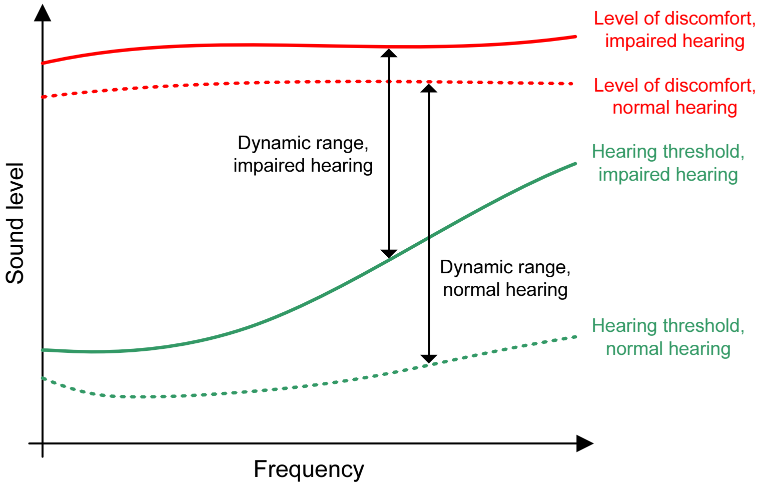

Figure 1: Reduced dynamic range due to sensorineural hearing loss. The hearing threshold is elevated more than the level of discomfort [21, 9]. In the hearing aid, the reduced dynamic range is compensated by compressing and amplifying the sound: Sound levels below the hearing threshold must be amplified, whereas sound levels close to the level of discomfort must not be amplified.

which is caused by damage in the cochlear, or by problems in the neural connections between the cochlear and the auditory nerve. The sensorineural hearing loss is different from, e.g., conductive hearing loss where the transmission of waves in the air to fluids in the cochlea is degraded, and from central hearing loss caused by problems at a higher level of the auditory system.

Sensorineural hearing loss is correlated with age [18], why the name *age-related hearing loss* is often used, but this correlation is not completely understood. Some possible explanations include loss of inner and outer hair cells in the cochlea, a loss of auditory neurons, decreased blood flow in the cochlea, and a stiffening of the basilar membrane [19, 20].

The sensorineural hearing loss is usually characterized by higher hearing thresholds at higher frequencies (a sloping hearing loss [16]), but the dynamic range, frequency resolution, and temporal resolutions are also affected [20]:

- The **dynamic range** is the range between the sound level necessary to detect the sound, the hearing threshold, and the sound level causing discomfort, as seen in Figure 1. The sensorineural hearing loss decreases the dynamic range, because the hearing threshold is elevated more than the threshold of discomfort [9, 21, 20].

- The **frequency resolution** or frequency selectivity is the ability to distinguish spectral components in complex sounds. One reason for the decreased frequency resolution is the broadening of the auditory filters, meaning that different locations on the cochlea become more sensitive to a wider range of frequencies. The broadening of the auditory filters can make the high frequencies difficult to perceive, because of increased masking by low frequencies [20].

- The **temporal resolution** is the ability to detect changes or events as a function of time. The forward and backward masking are the ability to detect a sound before or after another sound, and this non-simultaneous masking is increased by the sensorineural hearing loss [20, 22].

There are large individual differences in the extent to which persons with a sensorineural hearing loss experience the above mentioned difficulties. Further, the method used for measuring, e.g. the non-simultaneous masking, greatly influence the conclusions: If normal hearing listeners and hearing impaired listeners are compared at the same sound level, the effects from the sensorineural hearing loss are evident, but if the persons are tested at the same sensation level – the level relative to the hearing threshold – the effects are less clear [20].

A hearing impaired listener having problems at the cocktail party does not necessarily experience problems in noise free or less difficult conditions, because speech is robust and redundant. Robust and redundant means that part of the speech sound can be lost or modified without negative impact on intelligibility [23, 24, 25]. This redundancy of speech helps the hearing impaired listener to recognize and understand, even though some details of the sound are lost. When multiple sources are present, the listener has the additional task of identifying and separating the target speech. If we assume that both high frequency resolution and high temporal resolution are required to complete this task, a possible explanation to the hearing impaired listener's difficulties at the cocktail party is the reduced frequency and temporal resolution.

An evident effect of the sensorineural hearing loss is the reduced capability to use fluctuations in the interferer sound. Normal hearing listeners can make use of fluctuations in the interfering sound and will obtain a better intelligibility if the sound is modulated [6, 8]. This ability – also known as glimpsing speech [26] – is reduced by the sensorineural hearing loss [6, 8, 17].

3 Hearing Aids

Of the three complications from the sensorineural hearing loss described in the previous section, the reduced dynamic range is the most straightforward to address in the hearing aid by compression of the sound. The reduced frequency and temporal resolution are somewhat more difficult to compensate and the compression in the hearing aid does not necessarily compensate the reduced frequency and temporal resolution [13]. The compression compensates the reduced dynamic range by adjusting the dynamic range of the incoming sound to better correspond to the dynamic range of the hearing impaired listener's damaged cochlear. The compression is a frequency and level dependent gain making weak sounds audible and loud sounds below the level of discomfort [9, 16]. Making sounds audible is a necessary step to compensate for the hearing loss, but many hearing aid users will continue to experience problems with understanding speech in noisy conditions [16, 14].

One further approach towards solving the problem of reduced intelligibility in difficult situations is spatial filtering, where sound arising from certain directions is amplified more than others [27, 28]. The spatial filtering is accomplished using a directional system often based on two closely placed microphones in each hearing aid. This makes it possible to amplify sounds coming from particular directions and thus attenuate interfering sounds. However, this raises the fundamental question of how to decide which source is the target and which sources are interferers. A simple solution – which is acceptable in many situations – is to assume that the target source is located in front of the hearing aid user, though a number of examples can be found where this assumption is not valid, e.g., in a car or in a conversation with several people.

The use of directional systems and other signal processing methods are still limited by the size and placement of the hearing aid, even though the technology in hearing aids has developed fast over the last decade. Complex algorithms take up more of the capacity of the integrated circuit and use more power. In the future, these limitations will hopefully be reduced by the development of new technologies in the hearing aid, but user requirements for the hearing aid such as robustness, reliability, and high performance must continue to be fulfilled.

Most hearing aids do not block the ear canal completely, but allow direct sound through the vent in the hearing aid mould. The result is that the Tympanic membrane in the ear will receive both direct and amplified sound, and this tightens the requirement for low delay through the hearing aid. A large time difference between the direct and the amplified sound can result in negative side effects such as echoes. Thus, the objective is to keep the time delay below the threshold at which degradations of the sound is perceived ($\sim$10 ms [29, 30]). This strongly limits the use of computational expensive algorithms and non-causal methods (batch processing) that require a large number of future samples to be available.

4 Binary Masking

This thesis focuses on binary masking as a solution to the cocktail party problem. In binary masking, sound sources are assigned as either target or interferer in the time-frequency domain. As already defined, the target source is the source having the hearing aid user's attention. Everything else is interferer. When different sources are mixed, the identification of parts belonging to the target and parts belonging to the interferer can be difficult. Fortunately, the use of time-frequency representations can make this assignment easier because the change from time domain to time-frequency domain can help identify which parts of the sound that belong to the target and which parts that belong the interferer. The time-frequency representation can reveal properties of the sound sources that are not visible in the time domain or frequency domain, as seen in Figure 2.

The word **binary** emphasizes the assignment of time-frequency regions as belonging to either the target or the interferer source, but it also suggests what to do with the two sources when they have been identified. In binary masking the target sound is kept by using the value one in the binary mask, whereas the regions with the interferer are removed by using the value zero. In Figure 3 and throughout this thesis, the value one is shown with white, whereas the value zero is shown with black in the binary masks.

The word **mask** is the name for the pattern of values showing which regions in time and frequency that will be kept or removed. This mask can be seen as a matrix which is placed on top of the time-frequency representation of the sound mixture; white areas will be kept and black areas will be removed as seen in Figure 4.

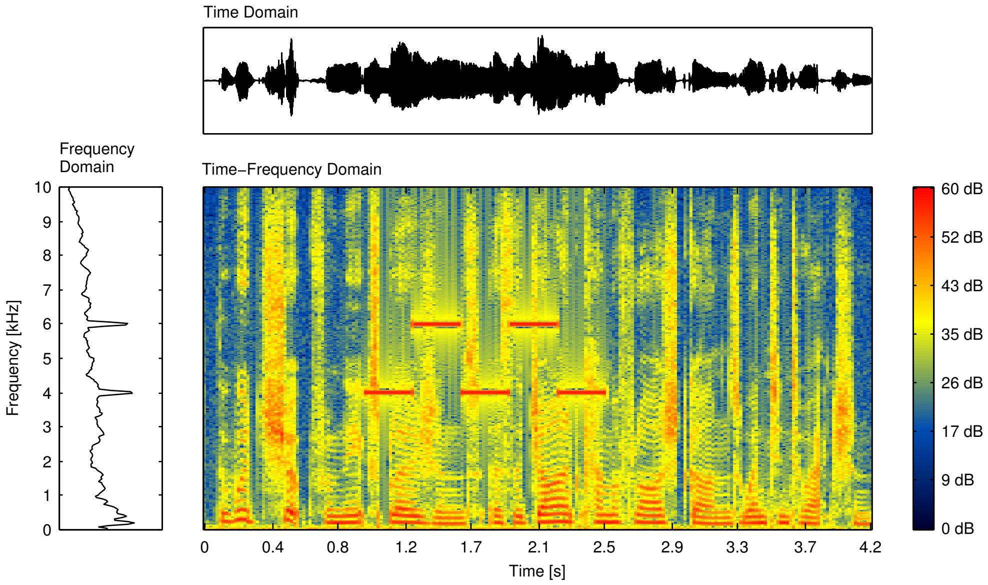

Figure 2: Speech signal mixed with a sinusoidal signal with alternating frequency. The time-frequency representation of the mixture provides more information about the sinewave than the time domain (top) or the frequency domain (left).

In short, binary masking is a method for applying a binary, frequency-dependent, and time-varying gain in a number of frequency channels, and the binary mask defines *what to do when*. This makes it possible to use the binary mask as an intermediate result, which can be used to evaluate and examine the performance of the binary masking algorithm. Binary masking algorithms can be seen as two steps: Estimation of the binary mask and application of the binary mask to carry out the source separation. In the estimation of the binary mask, the time-frequency regions are assigned to either the target or interferer source using, e.g., information about the spatial location of the sources [31] or speech models [32, 33]. In the application of the binary mask, the time-frequency regions assigned to the target source are kept, whereas the regions assigned to the interferer sources are removed or attenuated.

To evaluate the feasibility of an estimated binary mask, it can be compared to an available reference mask. This reference mask makes it possible to compare different estimated binary masks and measure their precision, e.g., by counting the number of errors in the binary mask compared to the reference mask [34, 35]. Two types of reference masks – or oracle masks – are described in Section 6, followed by examples of how to estimate the oracle masks and apply them to the sound. But first, an important property of speech which is a major motivation for the use of binary masking is discussed in the following section.

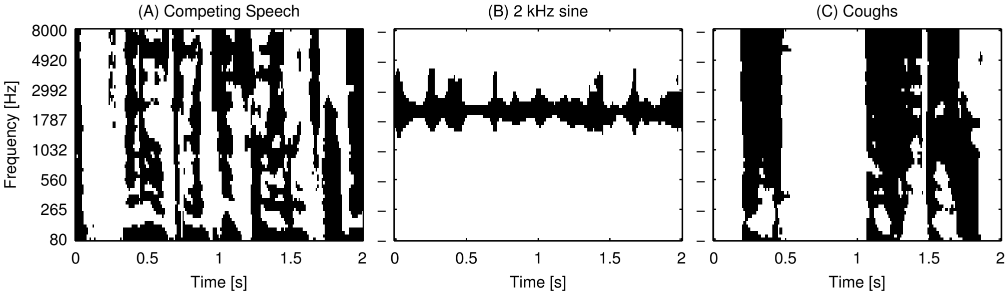

Figure 3: Examples of binary masks. The target sound is speech, and the interferer is (A) speech, (B) a 2 kHz sine, and (C) three coughs. White regions (ones) will be kept, whereas black regions (zeros) will be removed. The binary masks are calculated with the ideal binary mask (Equation (4), Section 6.1) and a 64 channel Gammatone filterbank (Section 7.1).

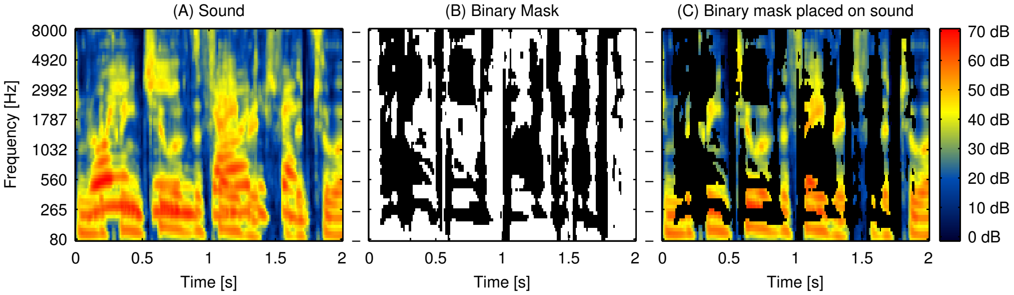

Figure 4: Example of binary masking. When the binary mask (B) is applied to the sound (A), the white regions will be kept and the black regions will be removed. Figure (C) shows the binary mask placed on top of the time-frequency representation (A). The sound in Figure (A) is a mixture of two speakers.

5 Sparsity of Speech

A sparse signal is a signal where the signal energy is found in a small percentage of the samples, and the majority of samples or coefficients are zero or close to zero. Whether a signal has a high or low sparsity – or sparseness – depends on the domain in which the signal is analyzed. In the time domain, when silence intervals between words are ignored, the sparsity of speech is low because energy from the speech signal is found in a large percentage of the samples. If the same speech signal is analyzed in the time-frequency domain, the sparsity is higher, because a small percentage of the units composing the time-frequency representation contains the majority of the energy from the speech signal.

If speech is sparse and different speech signals do not overlap in time and frequency, the following relation is true [31]:

$$\mathbf{S}_1(k,\tau)\mathbf{S}_2(k,\tau) = 0, \ \ \forall\, k, \tau, \tag{1}$$

where $\mathbf{S}_1(k,\tau)$ and $\mathbf{S}_2(k,\tau)$ are the energy of the two sources $s_1(t)$ and $s_2(t)$ at frequency index k and time index τ. A single element in the matrix $\mathbf{S}_j$ is notated as $\mathbf{S}_j(k,\tau)$ and referred to as a time-frequency unit. The assumption in Equation (1) is known as "W-disjoint orthogonality" [31], where W is the window function used in the calculation of the time-frequency representations, e.g., using the short-time Fourier transform or a Gammatone filterbank (see Section 7).

As noted by the authors in [31], Equation (1) is not satisfied for simultaneous speech signals, and this lead to their definition of "approximate W-disjoint orthogonality". In other words, speech is not sparse in the strict sense described by Equation (1). It implies a less strict assumption that only one source is dominant in each time-frequency unit and that the energy in each time-frequency unit mainly belongs to a single source:

$$\mathbf{Y}(k,\tau) \approx \max(\mathbf{S}_j(k,\tau)), \ \ \forall\, \tau, \omega, j, \tag{2}$$

where $\mathbf{Y}(k,\tau)$ is the time-frequency representation of a mixture of J sources:

$$\mathbf{Y}(k,\tau) = \mathbf{S}_1(k,\tau) + \mathbf{S}_2(k,\tau) + \ldots + \mathbf{S}_J(k,\tau), \tag{3}$$

if the sources are uncorrelated. The less strict definition of sparsity in Equation (2), that only one source is dominant in each time-frequency unit, must be accompanied by a definition of dominance. When considering human speech perception, a source can be defined as dominant when the neural response in the human auditory system comes mainly from that source.

The assumption of sparsity provides a simple solution to source separation as long as different sources have most of their energy in different time-frequency units: Keep the time-frequency units where the target source dominates and remove the rest. If sources are disjoint in the time-frequency domain, e.g., two sines at different frequencies, this solution is very efficient, but all real-life sources overlap in the time-frequency domain to some extent. However, if the overlap in time and frequency is limited, the sources can still be separated using binary masking – see [31, 36, 35], and the results in Paper A and Paper C.

The sparsity of speech – or sounds in general – has been examined in a number of studies [37, 31, 38], and measures of sparsity are also proposed and discussed in [39]. If sparsity correlates with the performance of the separation, a reliable measure of sparsity would be valuable. However, there is not a general accepted measure of sparsity that has been used to compare different sounds in different studies. One reason is that sparsity is dependent on the used time-frequency representation as indicated by the expression “W-disjointness” [31].

6 Oracle Masks

In this thesis, the term *oracle mask* is used for binary masks calculated using a priori knowledge which is not available in most real-life applications. In other words, the oracle masks are calculated in ideal situations, where all required information is available and does not need to be estimated. The term *oracle mask* is used instead of *ideal binary mask*, because the latter refers to a particular type of oracle mask described in Section 6.1.

A major objection to the concept of oracle masks is that it is of no use in real-life applications because of the required a priori knowledge. However, the oracle masks establish an upper limit of performance, which makes them useful as references and goals for binary masking algorithms developed for real-life applications such as hearing aids. To use a binary mask as reference or as goal for real-life estimation, it must be optimal in some sense or contribute to the solution of the problem at hand. In this thesis, the problem is the cocktail party problem, and the ideal binary mask and the target binary mask described in the following sections are possible solutions to this problem because of their positive impact on intelligibility. The descriptions of the ideal binary mask and the target binary mask (Section 6.1 and 6.2) are followed by examples of how to estimate these oracle masks (Section 6.3).

6.1 The Ideal Binary Mask

The ideal binary mask (IBM) is an oracle mask, because it requires both the target and interferer to be available as separate sounds. In most real-life situations only the mixture of the two sounds is available. To calculate the ideal binary mask, the energy of the target sound is compared to the energy of the interferer sound within each time-frequency unit:

$$\mathbf{M}_{\mathrm{IBM}}(k,\tau) = \begin{cases} 1, \text{ if } \dfrac{\mathbf{T}(k,\tau)}{\mathbf{N}(k,\tau)} > \mathrm{LC} \\ 0, \text{ otherwise} \end{cases}, \tag{4}$$

where $\mathbf{T}(k,\tau)$ is the energy of the target, $\mathbf{N}(k,\tau)$ is the energy of the interferer, τ the time index, and k the frequency index. The local SNR criterion (LC) is the threshold for classifying the time-frequency unit as dominated by the target or interferer sound and this threshold controls the amount of ones in the ideal binary mask (see Figure 5). If the LC value in Equation (4) is 0 dB, the ideal binary mask will keep all the time-frequency units with a local SNR of more than 0 dB. The local SNR is the SNR within time-frequency units, whereas the global SNR is the overall level difference between two sounds. If sounds from two speakers are mixed at 0 dB global SNR, the local SNR will vary highly from unit to unit because of the sparsity of speech.

Experiments applying the ideal binary mask to noisy speech have documented a substantial improvement in intelligibility. The improvement has been shown for normal hear-

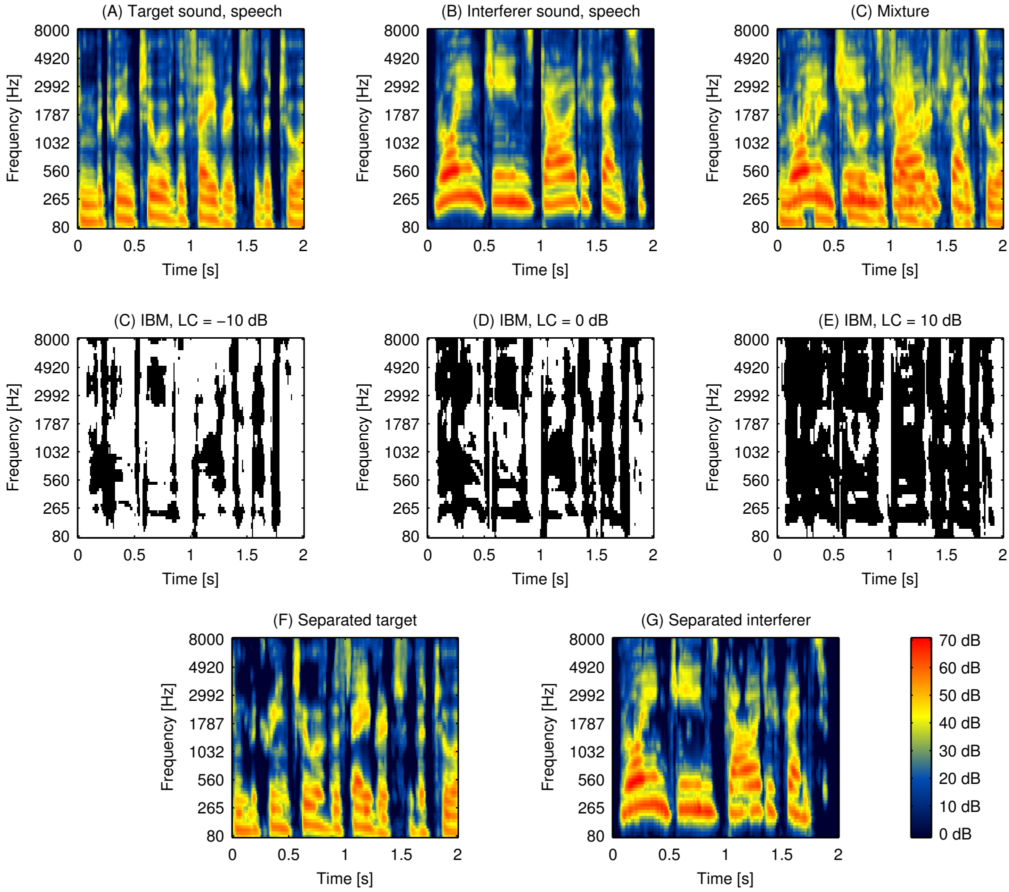

Figure 5: Calculation of the ideal binary mask. If the energy of the target sound (A) is larger than the energy of the interferer sound (B), the ideal binary mask (D) will contain the value one (white color) and otherwise the value zero (black color). The number of ones in the ideal binary mask is determined by the LC value as seen in (C–E). By applying the binary mask (D) to the mixture (C), the target speech is separated as seen in (F). If the target and interferer are swapped the result (G) is obtained. The time-frequency representations are calculated using a 64 channel Gammatone filterbank, and the target and interferer are mixed at 0 dB SNR.

ing listeners [36, 35], cochlear implant users [34], and for hearing impaired listeners in Paper A. The experiments were conducted using various conditions and thus show that the ideal binary mask is able to improve intelligibility for a range of LC values, interferer types, and SNRs.

The results obtained with the ideal binary mask under the various conditions were followed by experiments with ideal binary masking at low SNRs in Paper B and C. If the SNR and the LC value are decreased simultaneously in Equation (4), the ideal binary mask does not change, but the mixture will contain less target sound. Taking this to an extreme and applying the ideal binary mask to speech-shaped noise, creates an intelligible sound as documented in Paper B. This method resembles the method known as vocoding and shows the importance of the temporal speech envelope for intelligibility [40, 41, 42, 43].

6.2 The Target Binary Mask

The results obtained in Paper B – that high intelligibility can be obtained by applying the ideal binary mask to speech-shaped noise – lead to the definition of the target binary mask in Paper C. The speech-shaped noise used in Paper B had the same long-term average spectrum as the target speech, but instead of using speech-shaped noise, the target binary mask can be calculated by comparing the target speech directly with the long-term average spectrum of the target speech:

$$\mathbf{M}_{\text{TBM}}(k,\tau) = \begin{cases} 1, \text{ if } \dfrac{\mathbf{T}(k,\tau)}{\mathbf{r}(k)} > \text{LC} \\ 0, \text{ otherwise} \end{cases}, \tag{5}$$

where the vector $\mathbf{r}(k)$ is the long-term average of the energy in each of the frequency channels in the target sound $\mathbf{T}(k,\tau)$. An important difference between the ideal binary mask (4) and the target binary mask (5) is the presence of the interferer $\mathbf{N}(k,\tau)$. The ideal binary mask requires the interferer to be available and will change depending on the type of interferer, whereas the target binary mask is calculated from the target sound only and therefore is independent of the interferer sound. The ideal binary mask and the target binary mask are compared in Figure 6 using three different types of interferer sound.

The dependency on the target sound only, is not unique to the target binary mask. In [44], an oracle mask was generated by comparing the energy in each band with a global criterion. This criterion was equal in all frequency bands and adjusted to keep 99% of the energy from the target speech. The criterion not being a function of frequency is a major difference between the mask in [44] and the target binary mask. In Equation 5, the criterion $\mathbf{r}(k)$ is a function of frequency, whereas the criterion used in [44] is independent of frequency. The oracle mask in [44] was evaluated using hearing impaired listeners showing a similar impact on intelligibility as the ideal binary mask and target binary mask.

In Paper C, the impact on intelligibility by using the target binary mask and the ideal binary mask was measured in different conditions: 4 different noise types, 3 different SNRs, and 8 different mask densities were evaluated. High intelligibility was found when the ideal binary mask and the target binary mask were applied to noisy speech, and although intelligibility was reduced as a function of decreasing SNR, high intelligibility was still obtained at -60 dB SNR similar to the result in the Paper B. The impact on sound

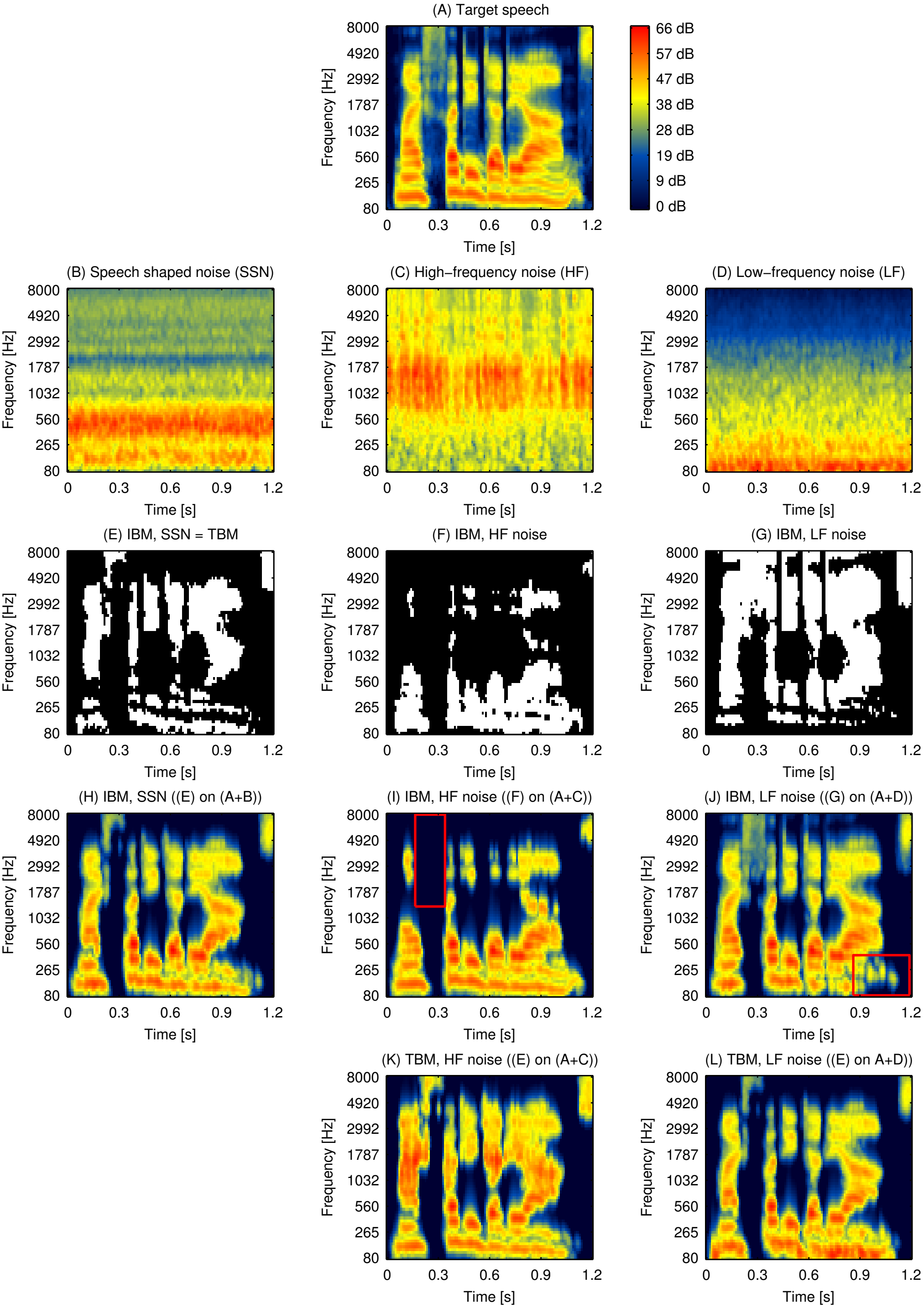
(A) Target speech
(B) Speech shaped noise (SSN)
(C) High−frequency noise (HF)
(D) Low−frequency noise (LF)
(E) IBM, SSN = TBM
(F) IBM, HF noise
(G) IBM, LF noise
(H) IBM, SSN ((E) on (A+B))
(I) IBM, HF noise ((F) on (A+C))
(J) IBM, LF noise ((G) on (A+D))
(K) TBM, HF noise ((E) on (A+C))
(L) TBM, LF noise ((E) on A+D))
Frequency [Hz]
8000
4920
2992
1787
1032
560
265
80
Time [s]
0
0.3
0.6
0.9
1.2
66 dB
57 dB
47 dB
38 dB
28 dB
19 dB
9 dB
0 dB

Figure 6: Comparison of the target binary mask (TBM) to the ideal binary mask (IBM) with three different interferer sounds. The ideal binary masks (E–G) are calculated by comparing the energy in the target speech (A) with speech-shaped noise (B), high-frequency noise from a bottling hall (C), and low-frequency noise from a car (D). If the ideal binary mask is calculated using speech-shaped noise, it becomes the the target binary mask (E). The 4th row (H–J) shows the ideal binary mask applied to the three different sound mixtures: (H) is the ideal binary mask applied to the mixture of target speech and speech-shaped noise. (I) is the ideal binary mask applied to the mixture of target speech and high-frequency noise. (J) is the ideal binary mask applied to the mixture of target speech and low-frequency noise. The 5th row (K–L) shows the target binary mask (E) applied to the high-frequency and low-frequency noise: (K) is the target binary mask applied to the mixture of target speech and the high-frequency noise. (L) is the target binary mask applied to the mixture of target speech and the low-frequency noise. The 3rd row (E–G) shows how the ideal binary mask changes as a function of the interferer: The ideal binary mask (F) has more ones at low-frequencies and fewer ones at high-frequencies. The opposite can be seen for the ideal binary mask calculated using the low-frequency interferer (G). This difference is also apparent when the ideal binary mask is applied to the sound: In (I) and (J) some target speech is missing in the time-frequency regions marked by red boxes when compared to the original signal (A). This loss of sound will not happen if the target binary mask is used: In (K) and (L) energy is found in the areas marked with red boxes in (I) and (J). However, it is important to remember that the high-frequency energy in (K) and low-frequency energy in (L) is a mixture of energy from the target and interferer source. The time-frequency representations are calculated using a 64 channel Gammatone filterbank, and the target and interferers are mixed at -6 dB SNR.

quality from the ideal binary mask or the target binary mask was not measured in Paper B, but a decrease in sound quality can be expected because of the coarse, binary processing. However, the quality of the processed sound using either of the masks is highly dependent on the noise type, SNR, and time-frequency resolution.

Defining the target binary mask establishes a new method for obtaining high intelligibility, and this new oracle mask can be used as reference and as a goal for binary masking. The target binary mask does not change as a function of the interferer sound which makes it easier to build a model of the binary mask. This property is used in Paper E, where a method for error correction of an estimated target binary mask is proposed.

6.3 Estimation of the Binary Mask

The results obtained with the ideal binary mask and the target binary mask in the ideal situations make these oracle masks useful goals when trying to solve the cocktail party problem. To obtain similar results in real-life applications, robust and precise methods for estimating these oracle masks must be found.

In **Computational Auditory Scene Analysis** (CASA) [45, 46], the binary mask is calculated by analyzing the sound mixture using principles from (human) Auditory Scene analysis (ASA) [47, 48]. A major motivation for CASA is the remarkable performance of the human auditory system even in adverse conditions, i.e. the cocktail party phenomenon. A listener is able to follow a single speaker in situations with many competing speakers and interfering sounds, and if this ability could be replicated by an algorithm, this would provide an understanding of how the human auditory system might work and suggest how source separation algorithms could be designed.

In the process of separating the target from the interferer, CASA algorithms often use the two stages from ASA: segmentation and grouping. In the first stage, the sound is decomposed into segments using cues as pitch, onsets, offsets, harmonicity, spatial location, and temporal continuity [48]. A segment is a sound with some inherent similarity, which probably comes from a single source. In a mixture of three speakers, segments are well-defined regions in time and frequency coming from one of the speakers but not assigned to a specific speaker. Assigning a segment to a particular speaker is the purpose of the second stage, where the segments are grouped into streams. A stream is a sound coming from the same source. In the mixture of three speakers, a large number (> 3) of segments can be found but ideally these segments should be grouped into three streams. When the target stream has been identified in time and frequency, it can be separated using the binary mask.

The ideal binary mask was formulated as the goal for CASA [49], but it has also been used as a reference outside the CASA domain [50, 31, 34, 51]. To estimate the oracle masks either multi-channel or single-channel algorithms can be used, but because of the novelty of the target binary mask, only methods for estimating the ideal binary mask has been proposed in the literature as described in the following paragraphs.

In **multi-channel algorithms**, the binary mask can be calculated using time, phase, or level differences from two or more microphone recordings. The microphones are often configured similarly to what is found in a hearing aid, or as a binaural configuration modeling the sound received in the human ears. The first configuration was used in the DUET algorithm [31], where the amplitude and phase differences between two microphones were used to calculate the binary mask. This configuration is also used in the system proposed in Paper D, where the LC value in Equation (4) is calculated from the location of the sources. The binaural configuration was used in [52], where the interaural time difference (ITD) and interaural intensity difference (IIS) [53] are used to calculate the binary mask. A comprehensive review of different multi-channel algorithms useful for hearing aid design can be found in [54].

In **single-channel** algorithms, only a single recording is available, and the binary mask must be calculated from this recording. To do this, the pitch and the harmonic structure can be used for voiced speech [55], and onsets and offsets can be used for unvoiced speech [56]. Outside the CASA domain, the ideal binary mask has been estimated using different "classic" single-channel speech enhancement algorithms [57]. In this study, several gain functions, noise estimation methods, and a voice activity detector were used to estimate the expected local SNR and the ideal binary mask.

Recently, a study proposing a single-channel speech enhancement algorithm using binary masking has reported a substantial improvement in intelligibility [50]. This paper proposes a speaker-dependent algorithm based on a Bayesian classifier that classifies each time-frequency unit as belonging to either the target or the interferer is proposed. The experiments show a significant increase in intelligibility under three different noise conditions and two SNR levels.

The above mentioned methods show that is possible to estimate the ideal binary mask with high precision in certain situations, but also that the binary masks will contain errors in most real-life situations. This problem has been recognized in several papers, where the correlation between errors in the ideal binary mask and intelligibility has been examined [50, 34, 35]. To reduce the number of errors more robust algorithms must be developed or a different approach should be taken as proposed in Paper E. In this paper, errors in the target binary mask are corrected using a hidden Markov model, and the results show that it is possible to build a speaker-independent model of the target binary mask and use this model to reduce the amount of errors.

7 Application of the Binary Mask

To apply the oracle masks – or estimates hereof – to a mixture of sounds, different transforms can be used [58, 59]. In the papers constituting this thesis, a Gammatone filterbank is used, why this is described in detail in Section 7.1. Another widely used method for binary masking is the short-time Fourier transform (see e.g. [60, 61]), which is shortly described in Section 7.2.

7.1 The Gammatone Filterbank

When the binary mask is applied to the sound using a filterbank, the following three steps are taken:

1. Split the time domain signal into subbands using an analysis filterbank.
2. Apply the binary mask by multiplying each subband signal with the binary gain as defined by the binary mask.
3. Transform the modified time-frequency domain signal back to the time domain using a synthesis filterbank.

The three steps are carried out using the setup shown in Figure 7. In all the papers constituting this thesis, the Gammatone filterbank is used to mimic the signal processing in the human auditory system. This filterbank is build of Gammatone filters with frequency dependent bandwidths and non-linear filter spacing as described in the following paragraphs and seen in Figure 8.

The Gammatone filters [62, 63] imitates the auditory filter in the human cochlea, and they are created by multiplying a 4th-order gamma function with a tone (carrier):

$$g(t) = t^{n-1} \exp(-2\pi bt) \cos(2\pi f_c t + \phi), \tag{6}$$

where n is the order ($n = 4$), b is the bandwidth of the filter, f_c the center-frequency, and ϕ the phase. The impulse responses from the Gammatone filters in Figure 9 can be seen as the impulse response at different locations on the basilar membrane. The bandwidths b of the Gammatone filters are determined by the equivalent rectangular bandwidth (ERB) [53]:

$$\begin{aligned} \mathrm{ERB}(f_c) &= 24.7 \cdot (4.37 \cdot 10^{-3} \cdot f_c + 1) && (7) \\ b &= 1.019 \cdot \mathrm{ERB}(f_c) && (8) \end{aligned}$$

where f_c is the center-frequency in Hz, and ERB is the bandwidth in Hz. The equivalent rectangular bandwidth in Equation (7) is a measure of the bandwidths of the human auditory filters [64, 53]. Two filters have the same ERB, if their peak gain is the same and if they retain the same amount of energy from a white noise signal. Equation (8) is a correction to match the bandwidths of the 4th-order Gammatone filter with the bandwidths of the auditory filters in the human auditory system [62, 65].

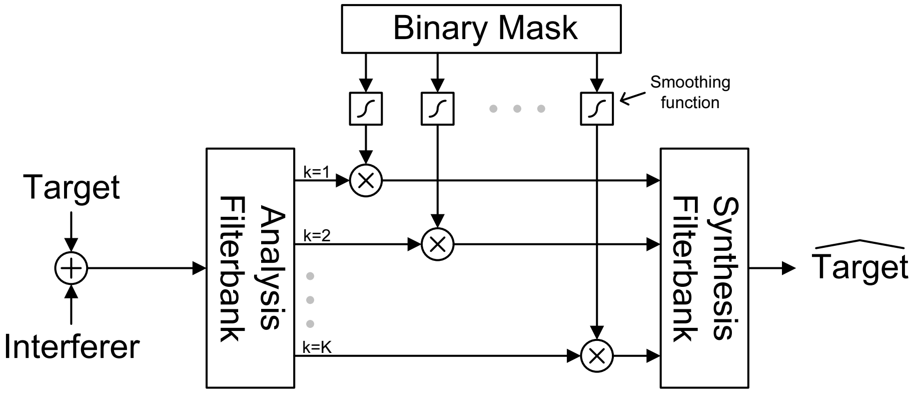

Figure 7: Application of the binary mask using a filterbank setup. The mixture of target and interferer sound is decomposed into K subbands through the analysis filterbank, and each subband signal is multiplied with a smoothed and possibly upsampled binary gain defined by the binary mask. The modified subbands signals are bandpass filtered and combined (summed) in the synthesis filterbank.

The bandwidths of the auditory filters can also be used as a frequency scale [64, 53], where frequency is expressed as the number of auditory filters between 0 Hz and the frequency f:

$$\mathrm{ERB}_n = 21.4 \cdot \log_{10}(4.37 \cdot 10^{-3} \cdot f + 1), \tag{9}$$

This frequency scale (number of ERBs) can be used to calculate the center-frequencies of the filters in the Gammatone filterbank, by distributing the filters linearly between the lowest and highest frequency on the ERB_n scale.

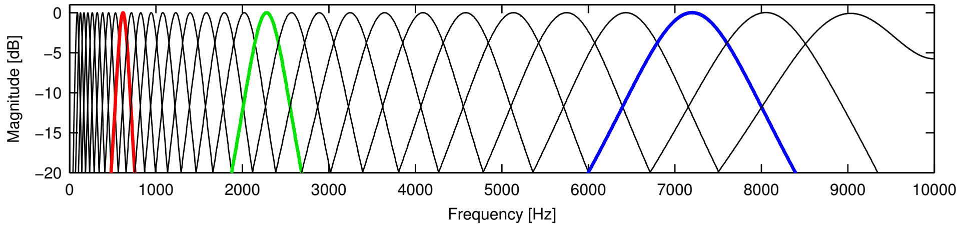

Figure 8: Frequency response of a 32 channel Gammatone filterbank with center-frequencies equally distributed on the ERB_n scale (Equation (9)) between 90 and 9000 Hz. The Gammatone filters are normalized to 0 dB peak gain [63]. The impulse responses of the filters illustrated with red, green, and blue are seen in Figure 9.

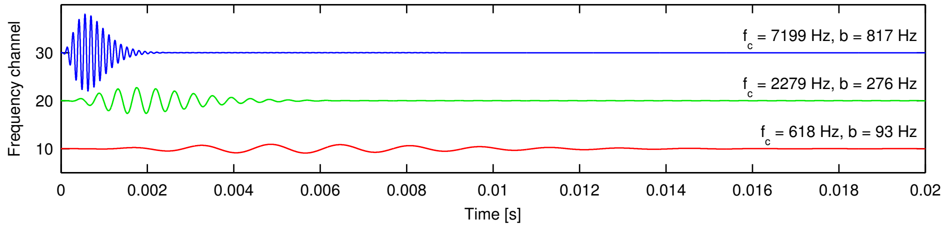

Figure 9: Impulse responses of 4th-order Gammatone filters with center-frequencies f_c and bandwidths b. The filters are normalized to 0 dB peak gain [63], as seen in Figure 8.

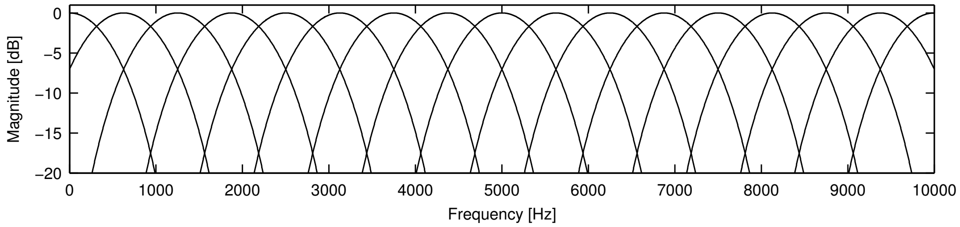

Figure 10: Frequency response of a 32 channel short-time Fourier transform. The analysis window is a 32 point Hamming window normalized to 0 dB peak gain [63].

When the sound has been decomposed into subbands through the analysis filterbank, the binary mask is applied by multiplying each subband signal with the binary gain as shown in Figure 7. Usually, the binary mask is decimated and not calculated on a sample-by-sample basis, although examples of the latter can be found [44]. As an example, the binary masks in Paper E are calculated from time-frequency representations using a frame size of 20 ms with 50% overlap. This decimation makes it necessary to upsample the binary mask before multiplying it with the subband signals, unless decimation has also been used in the analysis filterbank. Furthermore, the binary gain should be smoothed to avoid modulation artifacts. The modulation artifacts are wideband artifacts ("clicks") introduced by abrupt gain changes, see Figure 11. The modulation artifacts are less pronounced, if the binary gain is low-pass filtered (smoothed) before the multiplication with the subband signals and will also be reduced by the bandpass filters in the synthesis filterbank.

Finally, the synthesis filterbank transforms the modified signal from the time-frequency domain back to the time domain. The synthesis filterbank is created by time-reversal of the Gammatone filters in the analysis filterbank. This method compensates for phase shifts introduced by the analysis filterbank and reduces the modulation artifacts from the binary gains.

7.2 Short-Time Fourier Transform

Another useful method for binary masking is the discrete short-time Fourier transform (STFT) [66, 67, 68]. The result of the short-time Fourier transform is frequency channels with equal bandwidths and linearly spaced center-frequencies in Hz as seen in Figure 10.

When the STFT is used for binary masking, the binary mask can be applied by multiplying the binary mask with the magnitudes of the STFT. One example of binary masking using the STFT is [35], where the Fast Fourier transform (FFT) is applied to 20 ms segments with 50% overlap. The binary mask is multiplied with the FFT magnitudes, and the inverse FFT is applied to the modified magnitudes using the phases from the unmodified input signal. Finally, the resulting short time segments from the inverse FFT are combined using the overlap-add-method (OLA).

7.3 Temporal and Spectral Resolution

A main difference between the Gammatone filterbank and the STFT is the spectral resolution. Because the Gammatone filterbank resembles the processing in the human auditory system, it is often used for speech processing and perceptual studies. The STFT can also be used but has the drawback of requiring more frequency channels to obtain the same spectral resolution at low frequencies than the Gammatone filterbank.

A setup for experiments focusing on intelligibility could use 64 Gammatone filters in the filterbank equally spaced between 50 Hz and 10 kHz on the ERB_n frequency scale. A higher number of frequency channels, larger bandwidth of the filterbank, or narrower frequency channels could potentially increase sound quality, but 64 frequency channels are enough to achieve high intelligibility [69] (see also Paper B and Paper C). In many studies, the temporal resolution is 20 ms with 50% overlap [35, 36, 70]. The quasi-stationarity of speech makes a time resolution of 20 ms a reasonable choice, and the

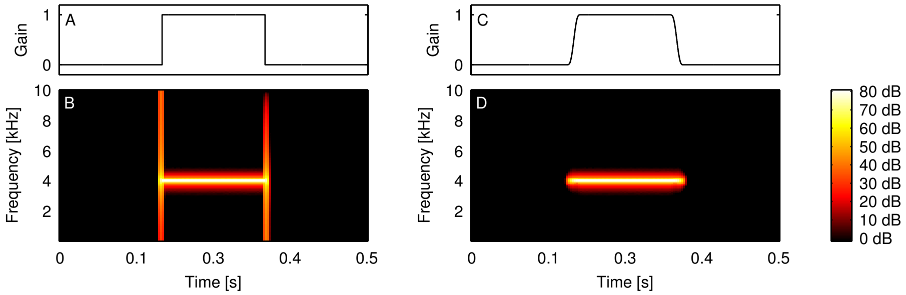

Figure 11: Multiplication of a 4 kHz sine and the binary gain with and without smoothing. If the gain (A) is multiplied with a 4 kHz tone (B), a broadband artifact will be introduced where the gain changes from zero to one or vice versa. A listener will perceive these artifacts as clicks in the sound. If the gain is smoothed by low-pass filtering with a 400 tap (20 ms) Hanning window (C), the artifacts are less pronounced (D).

widespread use of 20 ms temporal resolution makes it easier to compare results between different studies.

8 Time-Frequency Masking

Binary masking can be seen as a subset of a larger category of algorithms which applies a frequency-dependent and time-varying gain to a number of frequency bands, where the gain is not limited to binary values. This type of algorithms can be referred to as time-frequency masking algorithms, or short-time spectral attenuation [71]. When the gain is not limited to binary values, the possibility of attenuation changes the simple, binary decision into a more complex decision of how much each time-frequency unit should be attenuated. In many speech enhancement and noise reduction algorithms, this decision is based on the a priori SNR [72, 73, 74], and the classic algorithms like Wiener filtering, spectral subtraction, and maximum likelihood, can be formulated as a function of this a priori SNR [74]. In real-life applications, the a priori SNR must be estimated, but in the ideal situation the local SNR can be used instead of the a priori SNR. This leads to the following formulation of the Wiener filter [71]:

$$\mathbf{M}_W(k,\tau) = \frac{\mathbf{T}(k,\tau)}{\mathbf{T}(k,\tau)+\mathbf{N}(k,\tau)}, \tag{10}$$

where $\mathbf{T}(k,\tau)$ and $\mathbf{N}(k,\tau)$ is the energy of the target and interferer sounds, respectively. This formulation can be used to compare the gain from the Wiener filter and the ideal binary mask as seen in Figures 12 and 13.

The ideal binary mask produces a mask with values of zero and one, whereas the Wiener filter produces a mask with gain values ranging from zero to one. However, if the overlap between the target and interferer sound in the time-frequency domain is limited, the difference in the applied gain between the ideal binary mask and the Wiener filter is small as illustrated in Figure 14. If each time-frequency unit contains only target or interferer energy, the local SNR will be $-\infty$ dB or $+\infty$ dB resulting in a gain of $-\infty$ dB

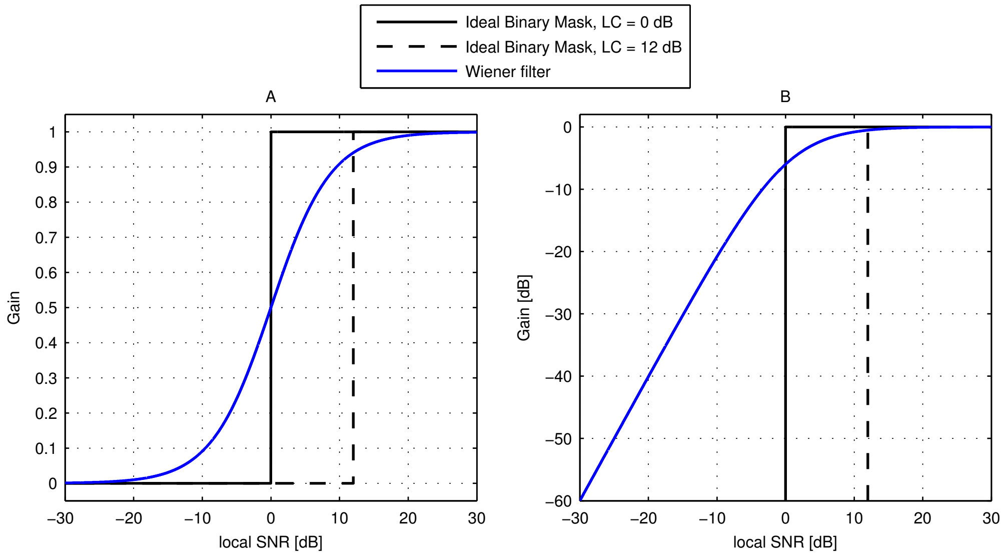

Figure 12: Attenuation curves for the ideal binary mask and the Wiener filter (Equations (4) and (10)). The LC value determines when the gain from the ideal binary mask changes from 0 to 1 ($-\infty$ dB to 0 dB).

or 0 dB using the ideal binary mask or the Wiener filter. The major difference between the Wiener filter and the ideal binary mask is seen when the local SNR is around 0 dB. If the energy of the interferer sound is just above the level of target energy, the ideal binary mask discards everything, whereas half of it will be kept using the Wiener filter.

It is important to emphasize that a comparable evaluation of the ideal binary mask and the Wiener filter with regards to intelligibility and sound quality has not been carried out in the literature. The ideal binary mask has been shown to enable increased intelligibility in the ideal situation, whereas the Wiener filter, when tested under realistic conditions, shows an increase in quality while in most situations only preserving intelligibility [75, 76]. An evaluation of intelligibility and quality using both methods under ideal condition would be interesting, and could also clarify, whether the substantial improvement of intelligibility using the ideal binary mask can be explained by the binary gain, the ideal condition, or both. Looking at the gain curves in Figure 12, it could be interesting to know if the hard binary gain helps emphasize speech cues for the listener, whereas the soft Wiener gain produces a better sound quality because of the smoother gain curve.

9 Speech Intelligibility and Quality

When employing source separation, speech enhancement, or noise reduction, using binary masking or other methods, it is important to realize that there are two different – and sometimes conflicting – goals: To increase speech quality or to increase speech intelligibility. Speech quality is a measure of how clear, natural and free of distortion the speech is, whereas speech intelligibility is a measure of how much of the speech that has

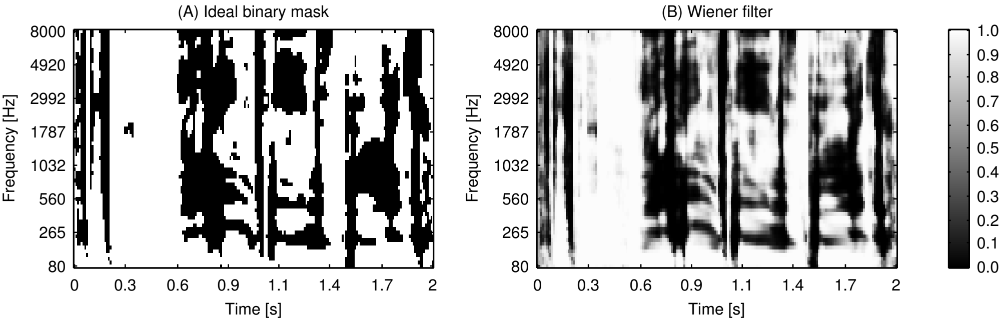

Figure 13: Examples of the ideal binary mask (A) and the Wiener gain (B) calculated from a mixture of male and female speech at 0 dB SNR.

been perceived correctly and recognized. Intelligibility is measured by "recognition" and not by "how much was understood", because some listening tests use nonsense words which cannot be understood, but only recognized correctly [77].

The difference between intelligibility and quality can be illustrated with the following example: If a recording of a sentence is played to a person and the purpose is to measure intelligibility, the person could be asked to repeat the sentence, or by other means reproduce the perceived sentence or words. From the listener's response, intelligibility can be measured, e.g., by the percentage of correctly recognized words, or by the level of noise allowing 50% of the words to be correctly recognized. If, instead, the person is asked his opinion about the quality of the speech without any further instructions, the answer might easily be "compared to what?" If no reference is established when quality is being evaluated, answers like "fine" or "bad" are of very little use unless the listeners are highly trained or expert listeners. Quality is a subjective measure depending on the users' individual reference and experience, whereas intelligibility is an objective measure, because the result of the intelligibility test is not affected by the listeners' subjective judgements. However, this objectivity does not imply that intelligibility is an absolute measure, because the results are dependent on speech material (stimuli), conditions, training, possibility of repeating the stimuli, etc. If the task is to identify words in noise, the results will also depend on the used words, compare for example "house", "bridge", and "airplane" to "cat", "hat", and "bat". The last three words would be more difficult to distinguish because they only differ by the first consonant.

It is important to distinguish between quality and intelligibility to keep a well-defined objective when developing new speech algorithms. Increasing speech intelligibility does not automatically increase speech quality or vice versa [10]. It seems to be a difficult task to obtain both objectives at the same time, and the reason might be that they conflict, e.g. if intelligibility is increased by enhancement of the speech cues like onsets and offsets, and quality is increased by a more smooth sound.

The ideal binary mask is able to increase intelligibility under several conditions, but according to my knowledge a concurrent subjective evaluation of intelligibility and quality using the ideal binary mask has not yet been carried out. If the ideal binary mask is applied to the sound mixture using a small number of frequency channels (e.g. 16 frequency channels), the quality will most likely be affected. If the separated speech is

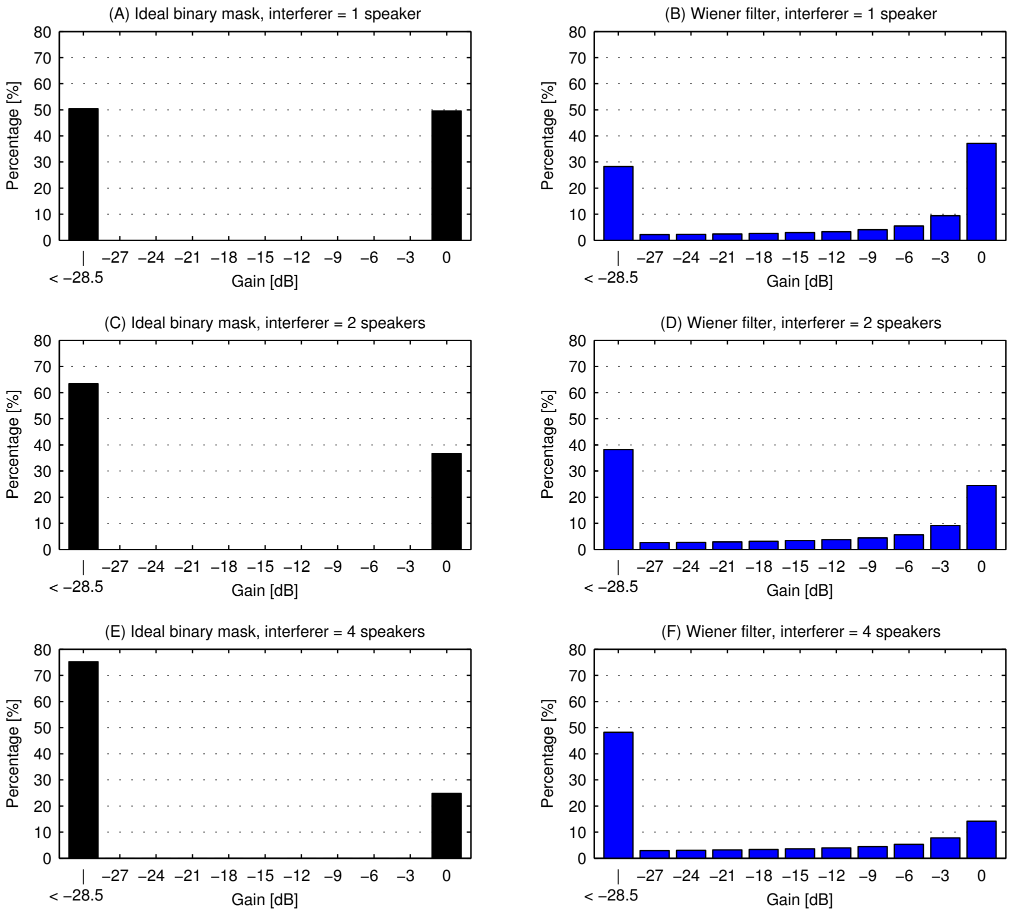

Figure 14: Distribution of gain values with the ideal binary mask and the Wiener filter. The gain using the ideal binary mask is either $-\infty$ or 0 dB, whereas the gain using the Wiener filter is in the range $-\infty$ to 0 dB. However, the highest percentages of gain values using the Wiener filter are found at 0 dB or below -27 dB. Five minutes of target sound (male speaker) and interferer sound is used to calculate the local SNR in all time-frequency units. In (A) and (B) the interferer sound is one female speaker. In (C) and (D) the interferer sound is one female and one male speaker. In (E) and (F) the interferer sound is two female and two male speakers. All speakers are normalized to the same energy, so the global SNR decreases as more speakers are added to the interferer sound. To calculate the local SNR, a 64 channel Gammatone filterbank (80–8000 Hz) is used with center-frequencies equally spaced on the ERB_n scale. The output from the filterbank is divided into 20 ms time-segments with 50% overlap.

compared to the clean target speech, the quality will probably be perceived as lower, whereas, if compared to the original mixture, quality will probably be perceived as being higher. Listeners might perceive higher intelligibility as higher quality.

9.1 What makes Speech Intelligible?

Although intelligibility of speech has been a research area for many decades, the question of "what makes speech intelligible?" has not yet been completely answered. If the question is broadened to "what makes speech intelligible at the cocktail party" the answering becomes even more complex. The purpose of the following section is not to provide a complete answer, but to identify elements within speech and perception that are fundamental to intelligibility or which contribute to the intelligibility of speech. Different answers to the question "what makes speech intelligible?" can be roughly sorted into the following three groups:

1. **Fundamental speech cues** are necessary for intelligibility and to distinguish different phonemes in the language. If the fundamental speech cues are modified or missing, the lexical meaning can change and intelligibility will be reduced.

    - **Formants** are resonances in the oral cavity generated by constrictions using the tongue and the lips. Different vowels are distinguished mainly by the two, lowest formants found between 300 and 2500 Hz [78, 79].
    - **Onsets and offsets** indicate where words begin and end, and they divide the speech into smaller units.
    - **Consonants** are discriminated by place, manner, and voicing [79, 80, 81]. Because consonants have less energy than vowels, they are less robust in noisy conditions [82].

2. **Supplementary speech cues** contribute to the correct perception of speech in noisy conditions. These cues are not fundamental to intelligibility because high intelligibility can be obtained without these cues being available. However, if the complementary speech cues are missing due to noisy conditions, it is more likely that intelligibility is affected since the listener will find it more difficult to identify and separate the target speech.

    - **Pitch** determines the gender of the speaker, and can be used to follow a single speaker in noisy conditions [83, 84, 85, 86]. However, pitch is not crucial for intelligibility in noise free conditions, as experiments with sine-wave-speech [87, 88, 89], vocoding [42, 41], or binary masking (Paper B) show. In some Asian and African languages, pitch is a fundamental speech cue, because pitch changes can change the lexical meaning of a word or sentence [79].
    - **Spatial location** is information about the location of the speaker in the environment, and the binaural cues interaural time difference (ITD), and interaural level difference (ILD), are useful to segregate target from interferer [53, 90, 4].
    - **Harmonicity** can be in simultaneous grouping across frequencies to decide whether a sound segment belongs to the same speaker [55]. If the harmonic structure at low frequencies is different from the harmonic structure at high frequencies, the two sound segments do not originate from the same speaker.

3. **High level processes** can be used to further process the perceived speech in the human auditory system and increase intelligibility. Whereas the previous two groups of contributing factors are characteristics of the perceived sound, the high level processes are located somewhere in the human auditory system.

    - **Redundancy** and phonetic restoration of speech is important for intelligibility in noisy conditions. If some parts of the speech is inaudible because of noise, the remaining parts can be used to perceive what was being said (see e.g. [23, 25, 91, 24]). The recognition of a word is not dependent on one, unique realization of the word – many sounds or acoustic patterns can lead to the same perception. It has been shown that a few unobstructed glimpses in time and frequency of the target speech can be enough to achieve a high intelligibility [26], and that the size and amount of these glimpses have a high correlation with intelligibility [92].
    - **Context** helps to determine the correct words, if the speech was not perceived correctly. Context works on many levels by limiting the number of possible words that can be chosen to substitute the wrongly perceived word. Knowledge about, e.g., the spoken language, the speaker, and the subject of the conversation reduces the context entropy [77].
    - **Auditory-visual integration** is the use of visual information to help recognize the speech correctly [93, 8]. The visual impact on speech perception is strong, and conflicting auditory and visual cues has been shown to create a different perception – the McGurk effect [94]. For hearing impaired listeners, auditory-visual integration in terms of lip-reading can be a fundamental speech cue.
    - **Continuity** and illusion of continuity is the auditory equivalent of the Gestalt principle of closure [47]. If a sound is masked by a louder sound, we expect the masked sound to continue "behind" the louder sound [24], e.g., if a tone is masked by a noise burst. This illusion or principle helps to restore missing speech sound that was masked by the interferer.

9.2 Evaluation of Speech Intelligibility

Having defined speech intelligibility and considered the elements contributing to it, it is of interest to consider how intelligibility can be evaluated and measured. Basically, this can be done in two ways; either by subjective evaluation using human listeners or by objective evaluation by means of an algorithm. The advantage of using listeners is the reliability of the result, but it comes at the price of a more demanding process. On the other hand, the objective evaluations are quicker to carry out, but less reliable.

Objective Evaluation of Speech Intelligibility

Objective evaluations are based on models of the human perception and calibrated using results from listening tests, but since the understanding of the human auditory system is not yet complete, these models are also incomplete and to some extent unreliable. This is not to say that objective evaluations should be avoided – they can be of great help in the development of new algorithms and point out problems or wrong directions in this process – but it is important to interpret the results correctly and know the limitations of the model. Some objective evaluations are based on simplified models and calibrated using a limited amount of data, e.g., the nSec measure in Paper F or the HIT-FA [50]. With these types of simple, objective evaluations it is very important to consider and discuss the result and the possible generalization of the method, although this type of criticism must be applied to all objective evaluations.

A well-known objective measure of intelligibility is the Articulation Index (AI) [95] which is based on calculation of the SNR in a number of independent frequency bands. The AI has been further developed into the speech intelligibility index (SII) [96], the speech transmission index (STI) [97], and a number of variants of these methods [98, 99]. These objective measures have a reliable performance in additive noise, but precision is reduced when non-linear processing like binary masking is used [100, 101]. Recently, an objective measure based on cross-correlation between the pre-processed test signal and the reference signal [100] showed a fine concordance with the results from the listening test in paper C. The pre-processing was implemented using the auditory model from [102]. A similar, but simpler approach, is used in the nSec measure described in Paper E.

The above mentioned objective evaluations are used to measure the performance of a particular algorithm or concept, but may also serve a different purpose: to provide insight into and understanding of what makes speech intelligible. If the models of speech perception correlate well with results obtained in subjective listening test, these models can be used as explanations of how speech intelligibility is obtained [84, 92].

Subjective Evaluation of Speech Intelligibility

A subjective evaluation is more demanding than an objective evaluation, because of the use of human listeners. Subjective evaluations are carried out by playing a stimuli to the listener and measure intelligibility from his response. The stimuli can be various sets of speech material (corpus) from short nonsense words to long sentences. In the subjective listening tests in the Papers A–C, the Dantale II corpus was used [103]. This corpus consists of syntactically fixed but semantically unpredictable sentences. This means that each sentence in the Dantale II corpus has the same structure with *name, verb, numeral, adjective*, and *object*, e.g., "Anders receives six fine cars" or "Linda owns ten old jackets". It is not possible to guess any particular word from the preceding or following words.

An important consideration when creating and using a corpus for subjective listening tests is the context entropy [77]. High context entropy means that listeners have a small probability of being able to predict a word from previous or following words in a sentence, and high context entropy is preferable since it is not the listeners' ability to predict which is to be measured in a subjective listening test. To maximize the context entropy, nonsense words or syllables can be used. In the Dantale II corpus, context entropy is maximized by carefully selecting 10 different words in each category (name, verb, etc.). The limited number of words in each place in the sentence makes the Dantale II corpus a closed set, and even though the number of different sentences is large, over time listeners will learn the different words occurring in the sentences as well as the structure of the sentences. Thus some training effect exists when using the Dantale II corpus, and this training effect must be taken into consideration when planning the listening test, e.g., by allowing the test person to listen to a number of test sequences before the measurement of intelligibility is initialized.

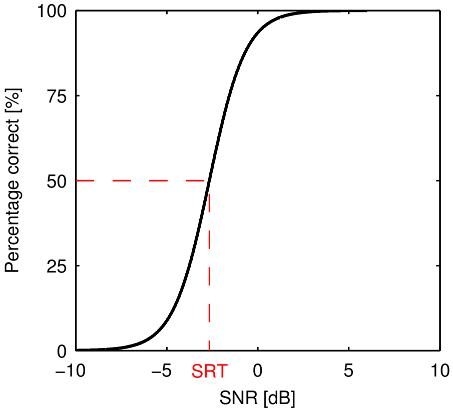

Figure 15: The speech reception threshold (SRT) shown on the psychometric function. The speech reception threshold is the SNR at which 50% of the stimuli (e.g. words) are correctly recognized. Please note, that higher SRT is lower performance [17].

To measure intelligibility, the number of correctly identified words can be counted, or an adaptive test can be used, in which the condition, e.g., the SNR, changes as a function of the response from the subject. In Paper B and Paper C, the number of correctly recognized words in each Dantale II sentence was used to measure the intelligibility. In Paper A and Paper C, the speech reception threshold (SRT) was used. The SRT is a measure of the required SNR allowing 50% correct recognition. The SRT is a point on a psychometric curve showing intelligibility as a function of SNR (see Figure 15). If the slope of the psychometric curve is a necessary parameter, at least two points on the curve must be measured as done in Paper C, where the 20% and 80% correct recognition were used to establish the psychometric curve.

10 Contributions

The papers that form the main body of this thesis fall into two groups. The first group comprising papers A–C is based on subjective listening tests, and contributes to the knowledge and understanding of speech intelligibility and how this can be improved using binary masks. In the second group of papers (D and E) the knowledge obtained in papers A–C is used in real-life applications for estimation and error-correction of the oracle masks. In Paper F, a method for objective evaluation of intelligibility is proposed based on the results obtained with the oracle masks.

Paper A: In this paper, the intelligibility of ideal binary masked noisy speech is evaluated using hearing impaired and normal hearing listeners. The results confirm previous results [36] and further show that hearing impaired listeners also achieve a large increase in intelligibility from the ideal binary mask. The increase in intelligibility for the hearing impaired listeners is higher than for normal hearing listeners in the two noise conditions tested, and this results in a similar performance of the two groups when the ideal binary mask is used on noisy speech.

Paper B: Based on the finding that applying the ideal binary mask to pure noise will generate highly intelligible speech, this paper investigates how large a number of channels is required using normal hearing listeners. When the ideal binary mask is applied to speech-shaped noise, the noise is modulated with a very coarse envelope from the target speech. This method generates a new type of artificial, but intelligible, speech with highly reduced temporal and spectral information similar to vocoding and sine-wave-speech [40, 41, 42, 43, 87].

Paper C: In this paper, the method from paper B is formalized by the definition of the target binary mask (Equation (5)). The impact on speech intelligibility from the ideal binary mask and target binary mask is examined and shows high intelligibility for a range of mask densities, different noise types, and SNR levels. These results make the method from paper B even more interesting by showing that high intelligibility can be obtained by modulating different noise types with either the target binary mask or the ideal binary mask.

Paper D: The substantial increase in intelligibility obtained by using the ideal binary mask makes it interesting to try to estimate the ideal binary mask in real-life applications. This paper introduces a simple method for estimating the ideal binary mask using a directional system with two microphones in a configuration similar to what is found in a hearing aid. The results show that the ideal binary mask can be estimated with high precision in the evaluated conditions, measured by the number of correct time-frequency units.

Paper E: Realizing that most real-life estimates of the ideal binary mask or the target binary mask will contain errors, a method for correcting these errors is proposed and evaluated. The focus in this paper is error-correction of the target binary mask using hidden Markov models, and it is shown how a model of the target binary mask can be build and used to reduce errors in the target binary mask.

Paper F: This paper introduces a simple method for objective evaluation of speech intelligibility and a model of how intelligibility is obtained. The method is based on the results in Paper C showing that intelligible speech can be created by modulating noise sounds with the target binary mask or the ideal binary mask. This knowledge is used in the intelligibility model which is based upon the correlation in time and frequency between the target and the processed speech.

11 Conclusion

The diversity of the contributions in this thesis allows only for a very general overall conclusion, but the obtained results attempt to narrow the gap between binary masking under ideal conditions and binary masking under more realistic conditions. The results obtained under ideal conditions show that high intelligibility can be obtained with binary masking in a variety of mixtures of speech and interferers for both normal hearing listeners and hearing impaired listeners. The results under more realistic conditions show that it is indeed possible to obtain reasonable estimates of the ideal binary mask and the target binary mask under more realistic conditions.

In the first group of papers (A–C), it was confirmed that binary masking could be useful for the hearing impaired, and the requirements for obtaining high intelligibility have been relaxed: High intelligibility using binary masking does not require a large number of frequency channels, target sound does not need to be present in the mixture, and prior knowledge of the interferer sound is not required. These relaxed requirements for high intelligibility using binary masking have increased the potential for the use of binary masking in hearing aids or other devices where intelligibility is a major concern.

On the opposite side of the gap, the papers in the second group (D–E) have focused on how to obtain the ideal and target binary mask in more realistic conditions and with limited resources. It was shown that the ideal binary mask can be estimated with high precision using a directional system with low complexity, and that a speaker-independent model of the target binary mask can be build and used to correct errors.

Although the gap between results under ideal conditions and more realistic conditions has been narrowed, it has not been closed. Hearing aids with true user benefits obtained through binary masking remain to be seen. To get to that point, more knowledge about the impact on intelligibility and quality from binary masking must be obtained – in particular for hearing impaired listeners. Binary masking can negatively influence sound quality, but it is not evident from existing studies how hearing impaired listeners will perceive and judge the sound quality. It is also important to deepen the understanding of how intelligibility is affected by the binary mask. As shown, the ideal binary mask and the target binary mask are able to increase intelligibility by a large amount, but situations can be found, where the method could fail, e.g., when more difficult recognition tasks such as consonant identification are considered. To enable the use of binary masking in hearing aids, algorithms must meet the fundamental requirements of low delay, low complexity, and high robustness. This is not easy to achieve, but if the human auditory system were understood to a larger extent, it might be possible to generalize and simplify the used methods into efficient algorithms.

References

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

[2] H. von Helmholtz, *Tonempfindungen, als physiologische grundlage für die theorie der musik*, 3rd ed. Braunschweig: Verlag von Friedrich Vieweg u. Sohn, 1870.

[3] ——, *On the sensations of tone*, 2nd ed. New York, NY: Dover Publications Inc., 1954.

[4] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[5] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, 2005.

[6] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, 1990.

[7] R. W. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 577–587, January 1998.

[8] J. G. W. Bernstein and K. W. Grant, "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3358–3372, 2009.

[9] H. Dillon, *Hearing aids*, 1st ed. Boomertang Press/Thieme, 2001.

[10] P. C. Loizou, *Speech enhancement: theory and practice*, 1st ed. CRC Press, 2007.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. Wiley, 2001.

[12] M. N. Schmidt, "Single-channel source separation using non-negative matrix factorization," Ph.D. dissertation, Technical University of Denmark, 2008.

[13] B. Edwards, "Beyond amplification: Signal processing techniques for improving speech intelligibility in noise with hearing aids," *Seminars In Hearing*, vol. 21, no. 2, pp. 137–156, 2000.

[14] S. Kochin, "MarkeTrak VII: Customer satisfaction with hearing instruments in the digital age," *The Hearing Journal*, vol. 58, no. 9, pp. 30–43, 2005.

[15] B. Edwards, "Hearing aids and hearing impairment," in *Speech Processing in the Auditory System*, S. Greenberg, W. Ainsworth, A. N. Popper, and R. R. Fay, Eds. New York, NY: Springer, 2004, pp. 339–421.

[16] A. Schaub, *Digital hearing aids*, 1st ed. New York, NY: Thieme, 2008.

[17] B. C. J. Moore, “Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms,” *Speech Communication*, vol. 41, no. 1, pp. 81–91, 2003.

[18] D. W. Robinson and G. J. Sutton, “Age effect in hearing - a comparative analysis of published threshold data,” *International Journal of Audiology*, vol. 18, no. 4, pp. 320–334, 1979.

[19] R. R. Gacek and H. F. Schuknecht, “Pathology of presbycusis,” *International Journal of Audiology*, vol. 8, no. 2-3, pp. 199–209, 1969.

[20] B. C. J. Moore, *Cochlear hearing loss*, 2nd ed. Wiley, 2007.

[21] I. Shapiro, “Evaluation of Relationship Between Hearing Threshold and Loudness Discomfort Level in Sensorineural Hearing Loss,” *Journal of Speech and Hearing Disorders*, vol. 44, pp. 31–36, 1979.

[22] E. Zwicker and K. Schorn, “Temporal resolution in hard-of-hearing patients,” *International Journal on Audiology*, vol. 21, no. 6, pp. 474–492, 1982.

[23] G. A. Miller and J. C. R. Licklider, “The intelligibility of interrupted speech,” *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 167–173, 1950.

[24] R. M. Warren, “Perceptual restoration of missing speech sounds.” *Science*, vol. 167, no. 3917, pp. 392–393, 1970.

[25] P. Howard-Jones and S. Rosen, “Uncomodulated glimpsing in ’checkerboard’ noise,” *Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2915–2922, 1993.

[26] M. Cooke, “Glimpsing speech,” *Journal of Phonetics*, vol. 31, no. 3-4, pp. 579–584, 2003.

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

[28] J. M. Kates and M. R. Weiss, “A comparison of hearing-aid array-processing techniques,” *Journal of the Acoustical Society of America*, vol. 99, no. 5, pp. 3138–3148, 1996.

[29] 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.

[30] H. Dillon, G. Keidser, A. O’Brien, and H. Silberstein, “Sound quality comparisons of advanced hearing aids,” *Hearing Journal*, vol. 56, no. 4, pp. 1–6, 2003.

[31] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[32] Y. Shao and D. Wang, “Sequential organization of speech in computational auditory scene analysis,” *Speech Communication*, vol. 51, no. 8, pp. 657–667, 2009.

[33] G. Hu and D. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[34] Y. Hu and P. C. Loizou, “A new sound coding strategy for suppressing noise in cochlear implants,” *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 498–509, 2008.

[35] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction,” *Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[36] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[37] J. J. Burred and T. Sikora, “On the use of auditory representations for sparsity-based sound source separation,” *2005 5th International Conference on Information Communications and Signal Processing*, pp. 1466–1470, 2005.

[38] S. T. Roweis, “Factorial models and refiltering for speech separation and denoising,” in *Proceedings of Eurospeech*, September 2003, pp. 1009–1012.

[39] N. Hurley and S. Rickard, “Comparing measures of sparsity,” *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[40] H. Dudley, “Remaking speech,” *Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[41] 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.

[42] M. F. Dorman, P. C. Loizou, and D. Rainey, “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,” *Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2403–2411, 1997.

[43] P. C. Loizou, M. Dorman, and Z. Tu, “On the number of channels needed to understand speech,” *Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2097–2103, 1999.

[44] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, “Determination of the potential benefit of time-frequency gain manipulation,” *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[45] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[46] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis*. Hoboken, New Jersey: Wiley & IEEE Press, 2006.

[47] A. S. Bregman, *Auditory scene analysis*, 2nd ed. MIT Press, 1990.

[48] M. Cooke and D. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, no. 3-4, pp. 141–177, 2001.

[49] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Kluwer Academic, 2005, pp. 181–197.

[50] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[51] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, no. 3, pp. 267–285, 2001.

[52] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[53] B. C. J. Moore, *An introduction to the psychology of hearing*, 5th ed. Academic Press, 2004.

[54] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[55] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[56] ——, "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1306–1319, 2008.

[57] Y. Hu and P. Loizou, "Techniques for estimating the ideal binary mask," in *Proceedings of the 11th International Workshop on Acoustic Echo, and Noise Control*, 2008.

[58] Q. Li, "An auditory-based transform for audio signal processing," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

[59] M. K. I. Molla, K. Hirose, and N. Minematsu, "Separation of mixed audio signals by source localization and binary masking with hilbert spectrum," in *Independent Component Analysis and Blind Signal Separation*. Springer, 2006, vol. 3889, pp. 641–648.

[60] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[61] N. Li and P. C. Loizou, "Factors influencing glimpsing of speech in noise," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1165–1172, 2007.

[62] J. Holdsworth, R. D. Patterson, I. Nimmo-Smith, and P. Rice, "SVOS final report, annex C: Implementing a GammaTone filterbank," *Rep. 2341, MRC Applied Psychology Unit.*, 1988.

[63] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer, Perception Group – Advanced Technology Group, Tech. Rep. 35, 1993.

[64] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[65] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[66] T. F. Quatieri, *Speech signal processing*. Prentice Hall, 2002.

[67] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[68] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[69] N. Li and P. C. Loizou, "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. EL59–EL64, 2008.

[70] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[71] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003.

[72] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," in *Cochlear and Brainstem Implants*, A. R. Møller, Ed. Basel, Switzerland: Advances in Oto-Rhino-Laryngology, Karger, 2006, vol. 64, pp. 109–143.

[73] R. C. Hendriks, "Advances in DFT-based single-microphone speech enhancement," Ph.D. dissertation, Delft University of Technology, 2008.

[74] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2. IEEE, 1996, pp. 629–632.

[75] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

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

[77] J. B. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[78] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, 1st ed. Wiley-Interscience, 2000.

[79] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing*, 1st ed. Prentice Hall, 2001.

[80] C. Christiansen, "Speech intelligibility prediction of linearly and nonlinearly processed speech in noise," Master's thesis, Technical University of Denmark, 2008.

[81] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[82] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Human Auditory System*, S. Greenberg, W. Ainsworth, A. N. Popper, and R. R. Fay, Eds. New York, NY: Springer-Verlag, 2004, pp. 231–308.

[83] C. J. Darwin, D. S. Brungart, B. D. Simpson, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2913–2922, 2003.

[84] S. Srinivasan and D. Wang, "A model for multitalker speech perception," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3213–3224, 2008.

[85] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, 2001.

[86] D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2527–2538, 2001.

[87] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, no. 4497, pp. 947–950, 1981.

[88] J. Barker, "The relationship between auditory organisation and speech perception: Studies with spectrally reduced speech," Ph.D. dissertation, University of Sheffield, 1998.

[89] J. Barker and M. Cooke, "Is the sine-wave speech cocktail party worth attending?" *Speech Communication*, vol. 27, no. 3-4, pp. 159–74, 1999.

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

[91] M. Kashino, “Phonemic restoration: The brain creates missing speech sounds,” *Acoustical Science and Technology*, vol. 27, no. 6, pp. 318–321, 2006.

[92] M. Cooke, “A glimpsing model of speech perception in noise,” *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[93] K. W. Grant, J. B. Tufts, and S. Greenberg, “Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals,” *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1164–1176, 2007.

[94] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[95] N. R. French and J. C. Steinberg, “Factors governing the intelligibility of speech sounds,” *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[96] ANSI S3.5-1997, “American national standard: Methods for the calculation of the speech intelligibility index,” 1997.

[97] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech-transmission quality,” *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[98] J. M. Kates and K. H. Arehart, “Coherence and the speech intelligibility index,” *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[99] R. L. Goldsworthy and J. E. Greenberg, “Analysis of speech-based speech transmission index methods with implications for nonlinear operations,” *Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[100] C. Christiansen, M. S. Pedersen, and T. Dau, “Prediction of speech intelligibility based on an auditory preprocessing model,” *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, 2010.

[101] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, “An evaluation of objective quality measures for speech intelligibility prediction,” in *Proceedings of Interspeech*, 2009, pp. 1947–1950.

[102] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. I. Modulation detection and masking with narrowband carriers,” *Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[103] K. Wagener, J. L. Josvassen, and R. Ardenkjaer, “Design, optimization and evaluation of a danish sentence test in noise.” *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

Paper A

Speech Intelligibility in Background Noise with Ideal Binary Time-Frequency Masking

DeLiang Wang, Ulrik Kjems, Michael S. Pedersen,
Jesper B. Boldt, and Thomas Lunner

Journal of the Acoustical Society of America
Vol. 125, no. 4, pp. 2336–2347, April 2009.

©2009 Acoustical Society of America

Speech intelligibility in background noise with ideal binary time-frequency masking

DeLiang Wang[a)]
Department of Computer Science & Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

Ulrik Kjems, Michael S. Pedersen, and Jesper B. Boldt
Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark

Thomas Lunner
Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark and Department of Clinical and Experimental Medicine and Technical Audiology, Linköping University, S-58183 Linköping, Sweden

(Received 4 March 2008; revised 21 January 2009; accepted 27 January 2009)

Ideal binary time-frequency masking is a signal separation technique that retains mixture energy in time-frequency units where local signal-to-noise ratio exceeds a certain threshold and rejects mixture energy in other time-frequency units. Two experiments were designed to assess the effects of ideal binary masking on speech intelligibility of both normal-hearing (NH) and hearing-impaired (HI) listeners in different kinds of background interference. The results from Experiment 1 demonstrate that ideal binary masking leads to substantial reductions in speech-reception threshold for both NH and HI listeners, and the reduction is greater in a cafeteria background than in a speech-shaped noise. Furthermore, listeners with hearing loss benefit more than listeners with normal hearing, particularly for cafeteria noise, and ideal masking nearly equalizes the speech intelligibility performances of NH and HI listeners in noisy backgrounds. The results from Experiment 2 suggest that ideal binary masking in the low-frequency range yields larger intelligibility improvements than in the high-frequency range, especially for listeners with hearing loss. The findings from the two experiments have major implications for understanding speech perception in noise, computational auditory scene analysis, speech enhancement, and hearing aid design. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3083233]

PACS number(s): 43.71.Gv, 43.66.Dc [KWG] Pages: 2336–2347

I. INTRODUCTION

Human speech communication typically takes place in complex acoustic backgrounds with environmental sound sources, competing voices, and ambient noise. It is remarkable that human speech understanding remains robust in the presence of such interference. This perceptual ability is thought to involve the process of auditory scene analysis (Bregman, 1990), by which the auditory system first analyzes a noisy input into a collection of sensory elements in time and frequency, also known as segments (Wang and Brown, 2006), and then selectively groups segments into auditory streams which correspond to sound sources.

It is well known that listeners with hearing loss have greater difficulty in speech perception in background noise. A standard way to quantify speech intelligibility in noise is a speech-reception threshold (SRT), which is the mixture signal to noise ratio (SNR) required to achieve a certain intelligibility score, typically 50%. Hearing-impaired (HI) listeners need 3–6 dB higher SNR than normal-hearing (NH) listeners in order to perform at the same level in typical noisy backgrounds (Plomp, 1994; Alcantara *et al.*, 2003). For speech-shaped noise (SSN) which is a steady noise with a long-term spectrum matching that of natural speech, the SRT increase for HI listeners is from 2.5 to 7 dB (Plomp, 1994). For fluctuating noise or competing speech, the increase is considerably higher (Festen and Plomp, 1990; Hygge *et al.*, 1992; Eisenberg *et al.*, 1995; Peters *et al.*, 1998); for a single competing talker, the increase is as much as 10–15 dB (Carhart and Tillman, 1970; Festen and Plomp, 1990; Peters *et al.*, 1998). Note that, for typical speech materials, a 1 dB increase in SRT leads to a 7%–19% reduction in the percent correct score, and a 2–3 dB elevation creates a significant handicap for understanding speech in noisy listening conditions (Moore, 2007).

Although modern hearing aids improve the audibility and comfort of noisy speech, their ability to improve the intelligibility of noisy speech is unfortunately very limited (Dillon, 2001; Alcantara *et al.*, 2003). Extensive research has been made to develop noise reduction algorithms in order to close the SRT gap between HI and NH listeners. Monaural speech enhancement algorithms, such as Wiener filtering and spectral subtraction, perform statistical analysis of speech and noise and then estimate clean speech from noisy speech (Lim, 1983; Benesty *et al.*, 2005). Although these algorithms produce SNR improvements, they have not led to increased speech intelligibility for human subjects (Levitt, 2001;

[a)] Author to whom correspondence should be addressed. Electronic mail: dwang@cse.ohio-state.edu

Moore, 2003b; Edwards, 2004). Attempts have also been made to directly enhance speech cues, especially formants which are spectral peaks of speech (Bunnell, 1990; Simpson *et al.*, 1990). This processing results in clearer formant structure; however, listening tests with both NH and HI listeners show little improvement in speech intelligibility (Baer *et al.*, 1993; Alcantara *et al.*, 1994; Dillon, 2001). Unlike monaural speech enhancement, beamforming (spatial filtering) with a microphone array has been demonstrated to achieve significant speech intelligibility improvements, particularly with large arrays (Kates and Weiss, 1996; Levitt, 2001; Schum, 2003). On the other hand, practical considerations of hearing aid design often limit the size of an array to two microphones, and the effectiveness of beamforming degrades in the presence of room reverberation (Greenberg and Zurek, 1992; Levitt, 2001; Ricketts and Hornsby, 2003). Additionally, to benefit from spatial filtering target speech and interfering sounds must originate from different directions.

Recent research in computational auditory scene analysis (CASA) has led to the notion of an ideal binary time-frequency mask as a performance upper bound to measure how well CASA algorithms perform (Wang and Brown, 2006). With a two-dimensional time-frequency (*T-F*) representation or decomposition of the mixture of target and interference, where elements in the representation are called *T-F* units, an ideal binary mask (IBM) is defined as a binary matrix within which 1 denotes that the target energy in the corresponding *T-F* unit exceeds the interference energy by a predefined threshold and 0 denotes otherwise. The threshold is called the local SNR criterion (*LC*), measured in decibels. More specifically, IBM is defined as

$$\mathrm{IBM}(t,f) = \begin{cases} 1 & \text{if } s(t,f) - n(t,f) > \mathrm{LC} \\ 0 & \text{otherwise,} \end{cases}$$

where $s(t,f)$ denotes the target energy within the unit of time t and frequency f and $n(t,f)$ the noise energy in the *T-F* unit, with both $s(t,f)$ and $n(t,f)$ measured in decibels. The mask is considered ideal because its construction requires access to the target and masker signals prior to mixing, and under certain conditions the IBM with $LC=0$ dB has the optimal SNR gain among all the binary masks (Wang, 2005; Li and Wang, 2009). As a separation technique, applying the IBM with $LC=0$ dB to the mixture input retains the *T-F* regions of the mixture where target energy is stronger than interference energy while removing the *T-F* regions where target energy is weaker than interference energy.

Varying *LC* results in different IBMs. Recently, Brungart *et al.* (2006) tested the effects of IBM with different *LC* values on speech mixtures with one target utterance and 1–3 competing utterances of the same talker, where the sound levels of all the utterances are set to be equal. Their experimental results show that, when $0 \text{ dB} \geqslant LC \geqslant -12$ dB, IBM produces nearly perfect intelligibility scores, which are dramatically higher than in a control condition where speech mixtures are presented to listeners without processing. They suggest that the choice of $LC=-6$ dB, which lies near the center of the performance plateau, may be better than the commonly used 0 dB *LC* for intelligibility improvement. Furthermore, they attribute the intelligibility improvement to the removal of informational masking which occurs when the listener is unable to successfully extract or segregate acoustically detectable target information from the mixture. Anzalone *et al.* (2006) investigated the intelligibility improvements of a related version of IBM, defined by a comparison between target energy and a threshold rather than a comparison between target energy and interference energy. Using mixtures of speech and SSN, they found that IBM leads to substantial SRT reductions: more than 7 dB for NH listeners and more than 9 dB for HI listeners. In addition they reported that, while NH listeners benefit from ideal masking in both the low-frequency (LF) and high-frequency (HF) ranges, HI listeners benefit from ideal masking only at LFs (up to 1.5 kHz). Li and Loizou (2007) used the IBM to generate “glimpses,” or *T-F* regions with stronger target energy, to study several factors that influence glimpsing of speech mixed with babble noise. Their results show that it is important to generate glimpses in the LF to mid-frequency range (up to 3 kHz) that includes the first and the second formant of speech, but not necessary to glimpse a whole utterance; high intelligibility is achieved when the listener can obtain glimpses in a majority of time frames. More recently, Li and Loizou (2008b) extended the findings of Brungart *et al.* (2006) to different types of background interference, including speech babble and modulated SSN. Moreover, they evaluated the impact of deviations from the IBM on intelligibility performance and found that there is a gradual drop as the amount of mask errors increases. A subsequent study by Li and Loizou (2008a) shows that NH listeners obtain significant intelligibility improvements from IBM processing with as few as 12 frequency channels, and IBM processing in the LF to mid-frequency range that includes the first and the second formant appears sufficient.

In this paper, we evaluate the effects of IBM processing on speech intelligibility with two kinds of background noise: SSN and cafeteria noise, using both NH and HI listeners. While SSN is commonly used in the literature, the cafeteria noise we use contains a conversation between two speakers in a cafeteria background and it resembles the kind of noise typically encountered in everyday life. Our study adopts the standard IBM definition with a comparison between target and interference and measures speech intelligibility by SRT at the 50% level. As suggested by the findings of Brungart *et al.* (2006), we set *LC* to −6 dB in IBM construction. Intrigued by the observation of Anzalone *et al.* (2006) that HI listeners derive little benefit from IBM in the HF range, we conduct an experiment to test whether ideal masking in the HF range is indeed not important for HI subjects. Unlike Anzalone *et al.* (2006) who applied a constant gain to compensate for the hearing loss of their HI subjects, we apply gain prescriptions to fit individual HI listeners.

In what follows, Sec. II details IBM processing. Section III describes an experiment that tests the effects of ideal masking on mixtures of speech with SSN or cafeteria noise. Section IV describes an experiment that compares the effects of ideal masking in LF, HF, and all-frequency (AF) ranges. Further discussion is given in Sec. V. Finally, Sec. VI concludes the paper.

II. IDEAL BINARY MASKING

The concept of IBM in CASA is directly motivated by the auditory masking phenomenon which, roughly speaking, refers to the perceptual effect that a louder sound renders a weaker sound inaudible within a critical band (Moore, 2003a). So keeping noise in *T-F* units with stronger target energy as done in the standard IBM definition with 0 dB *LC* should not reduce speech intelligibility, and this is indeed what was found by Drullman (1995). On the other hand, IBM processing removes all the *T-F* units with stronger interference energy as the target energy in these units is assumed to be masked by the interference. Removing these masker-dominated units also serves to remove informational masking, which is a dominant factor for reduced speech intelligibility in speech and other modulated maskers (Brungart, 2001). Hence IBM processing, as a form of ideal time-frequency segregation, is expected to yield larger speech intelligibility improvement in a modulated noise condition than in a steady noise condition (Brungart *et al.*, 2006).

Like earlier studies (Brungart *et al.*, 2006; Anzalone *et al.*, 2006), we use a gammatone filterbank to process a stimulus and then time windowing to produce a cochleagram which is a two-dimensional *T-F* presentation (Wang and Brown, 2006). Specifically, we use a 64-channel filterbank that is equally spaced on the equivalent rectangular bandwidth (ERB) rate scale with center frequencies distributed from 2 to 33 ERBs (corresponding to 55–7743 Hz). The bandwidth of each filter is 1 ERB. We note that this filterbank is similar to the one used in Anzalone *et al.* (2006) whereas Brungart *et al.* (2006) used a 128-channel filterbank covering the frequency range of 80–5000 Hz. The response of each filter is divided into 20 ms frames with a frame shift of 10 ms, hence generating a two-dimensional matrix of *T-F* units. The cochleagram of a stimulus is simply the two-dimensional graph of response energy within all the *T-F* units. For a given mixture of target signal and background noise, the IBM is calculated by comparing whether the local SNR within a *T-F* unit is greater than *LC*. As mentioned in Sec. I, we fix $LC=-6$ dB in this study as suggested by Brungart *et al.* (2006). Such a choice of negative *LC* retains certain *T-F* units where the target energy is weaker but not much weaker than the interference energy, in accordance with Drullman's observation that weaker speech energy below the noise level still makes some contribution to speech intelligibility (Drullman, 1995). Indeed, a pilot test with 0 dB *LC* indicates that SRT improvements are not as high as those produced with $LC=-6$ dB. More generally, in order to produce large auditory masking, the masker needs to be stronger than the masked signal (Moore, 2003a).

Given an IBM, the waveform output of IBM can be resynthesized from the mixture input by weighting the mixture cochleagram by the IBM and correcting phase shifts introduced during gammatone filtering (see Wang and Brown, 2006). Such an output can then be played to a listener as a stimulus in our experiments. Figure 1 illustrates IBM for a mixture of a speech utterance and a cafeteria background. The SNR of the mixture is 0 dB. Figure 1(a) shows the cochleagram of the target speech, Fig. 1(b) that of the background noise, and Fig. 1(c) that of the mixture. Figure 1(d) displays the IBM with $LC=-6$ dB, and Fig. 1(e) the cochleagram of the resynthesized result of ideal masking with the IBM in Fig. 1(d). The ideally masked mixture in Fig. 1(d) is clearly more similar to the target speech shown in Fig. 1(a) than the original mixture shown in Fig. 1(c) is. As a comparison, Fig. 1(f) shows the IBM with $LC=0$ dB, and Fig. 1(g) the cochleagram of the corresponding ideal masking output. With the increased *LC*, the IBM has fewer 1's and retains less mixture energy.

III. EXPERIMENT 1: EFFECTS OF IDEAL BINARY MASKING ON SPEECH-RECEPTION THRESHOLD

This experiment was designed to quantify the SRT effects of IBM for both NH and HI listeners. Sentences from the Dantale II corpus (Wagener *et al.*, 2003) were used as target speech, and tests were conducted with two different backgrounds: SSN and cafeteria noise.

A. Methods

1. Stimuli

The Dantale II corpus (Wagener *et al.*, 2003) comprises sentences recorded by a female Danish speaker. Each sentence has five words with a fixed grammar (name, verb, numeral, adjective, and object), for example, "Linda bought three lovely presents" (English translation). Each word in a sentence is randomly chosen from ten equally meaningful words. As a result, recognizing a subset of words in a sentence does not help with the recognition of the remaining words. There are a total of 15 test lists, and each list has ten sentences with no repeating word. There are a few seconds of silence between sentences within each list to allow a listener time to report what has been heard. Similar to the Swedish sentence test (Hagerman, 1982), the closed set corpus was designed for repeated use, and training effects are minimal after familiarization with a few lists (Wagener *et al.*, 2003). We use the speech-shaped noise included with the Dantale II corpus, which is produced by superimposing the speech material in the corpus. The cafeteria noise employed is a recorded conversation in Danish between a male and female speaker that took place in a cafeteria background (Vestergaard, 1998). To emphasize temporal modulation effects, the long-term spectrum of this noise was shaped to match that of the Dantale II speech material (Johannesson, 2006). Target speech and background noises are all digitized at 20 kHz sampling frequency.

A speech utterance and a background noise are first processed separately by a 64-channel gammatone filterbank (see Sec. II), which produces a flat frequency response within the frequency range of the filterbank. Filter responses are then windowed into 20 ms rectangular frames with a 50% overlap between consecutive frames, resulting in a two-dimensional cochleagram. This 100 Hz frame rate is frequently used in speech processing (Rabiner and Juang, 1993). For a given mixture of a Dantale II list and a background noise, the mixture SNR is calculated during the intervals that contain speech energy. To account for the forward masking of the continuously present noise that occurs between two consecu-

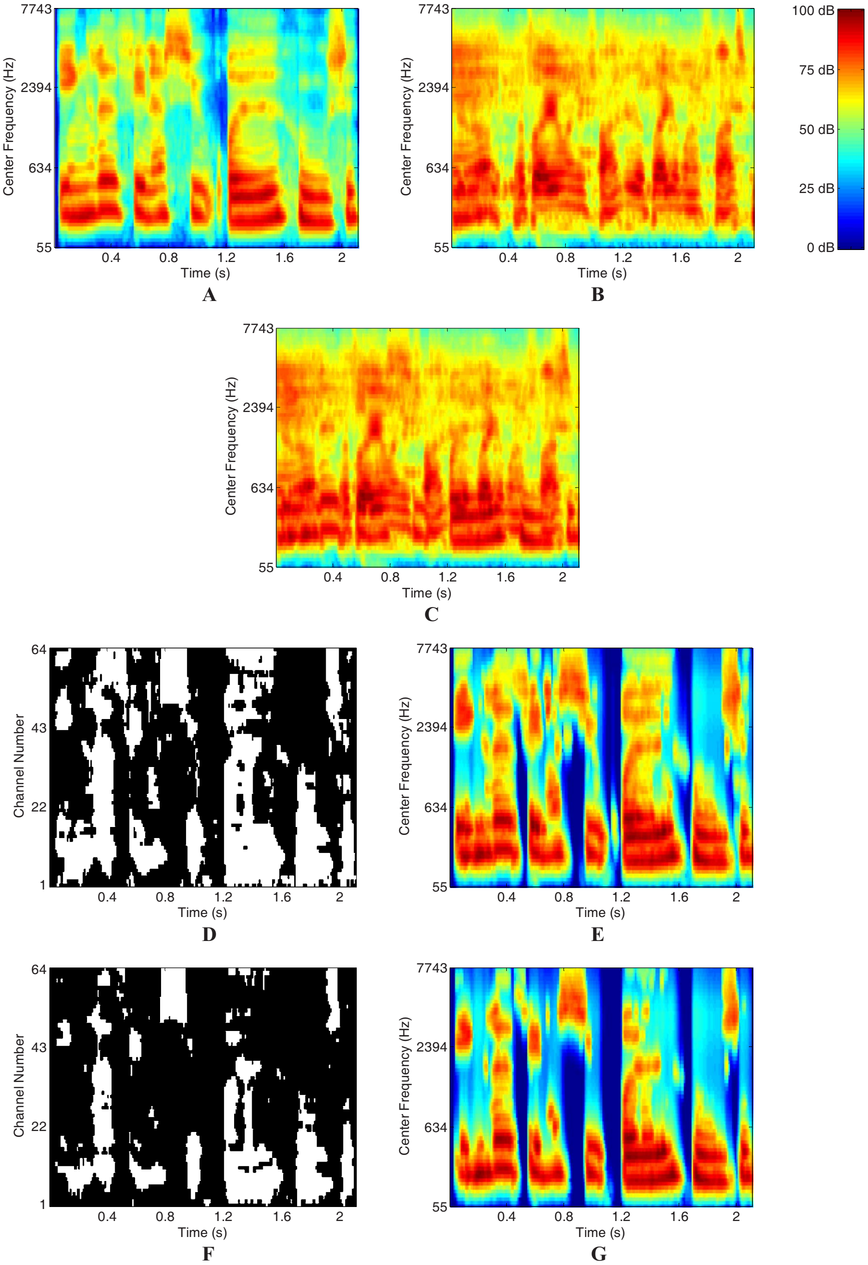

FIG. 1. (Color online) Illustration of IBM (A) Cochleagram of a target speech utterance. (B) Cochleagram of a cafeteria background. (C) Cochleagram of a 0 dB mixture of the speech and the background shown in A and B. (D) IBM with $LC=-6$ dB, where 1 is indicated by white and 0 by black. (E) Cochleagram of the segregated mixture by the IBM in D. (F) IBM with $LC=0$ dB. (G) Cochleagram of the segregated mixture by the IBM in (F).

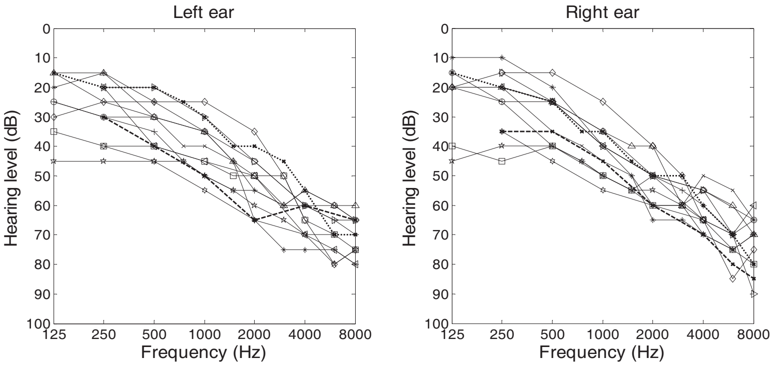

FIG. 2. Audiograms of the 13 HI listeners who participated in the experiments. The dashed line indicates the subject who only participated in Experiment 1, and the dotted line the subject who only participated in Experiment 2.

tive sentences (Moore, 2003a), a period of 100 ms is added before the onset of a sentence for mixture SNR calculation. For a mixture input with a specified SNR, IBM is constructed from the cochleagrams of the target speech and the background noise with *LC* fixed at −6 dB. The IBM is then used to resynthesize a waveform stimulus from the mixture cochleagram. Note that, as a result, the masker signals in between sentences are removed by IBM processing because during such intervals there is only masker energy.

As control conditions, mixtures of speech and background noise were also presented to listeners without segregation. To incorporate filtering effects and any distortions that might be introduced during cochleagram analysis and synthesis, a mixture in an unsegregated condition is processed through an all-1 binary mask or the IBM with the *LC* of negative infinity, therefore including all the *T-F* units in the resynthesis.

2. Listeners

A total of 12 NH listeners and a total of 12 listeners with sensorineural hearing loss participated in this experiment. All subjects were native Danish speakers. The NH listeners had hearing thresholds at or below 20 dB HL from 125 Hz to 8 kHz, and their ages ranged from 26 to 51 with the average age of 37. The NH listeners had little prior experience with auditory experiments, and were not informed of the purpose or design of the experiment.

The 12 HI listeners had a symmetric, mild-to-moderate, sloping hearing loss. The audiograms of these listeners are shown in Fig. 2. They had an age range from 33 to 80 with the average age of 67. All the HI listeners were experienced hearing aid wearers. The tests were performed with their hearing aids taken off, and compensation was applied to each HI subject individually. Specifically, a gain prescription was computed from an individual's audiogram using the NAL-RP procedure (Dillon, 2001), and then used to produce amplification with appropriate frequency-dependent shaping. The hearing losses in the left ear and the right ear were compensated for separately. The subjects had participated in Dantale II listening tasks before, but were not told of the purpose and design of this experiment.

3. Procedure

There are a total of four test conditions in this experiment: two ideal masking conditions with SSN and cafeteria noise and two control conditions with unsegregated mixtures. Three Dantale II lists with a total of 30 sentences were randomly selected from the corpus for each test condition. Subjects were instructed to repeat as many words as they could after listening to each stimulus that corresponded to one sentence, and they were not given any feedback as to whether their responses were correct or not. To familiarize them with the test procedure, subjects were given a training session at the beginning of the experiment by listening to and reporting on three lists of clean sentences. The order of the four conditions was randomized but balanced among the listeners (Beck and Zacharov, 2006). A subject test with the four conditions and a training session together took less than 1 h, and a short break was given roughly halfway through the test.

The Dantale II test employs an adaptive procedure in order to find the 50% SRT. The procedure is to present test sentences at SNR that is continuously adapted according to the number of correctly reported words in the previous sentence (Hansen and Ludvigsen, 2001). In a test condition with 30 sentences, the first 10 sentences are used to reach a steady 50% SRT level and the final SRT is determined by averaging the SNR levels for the last 20 sentences.

Speech and noise were both set to the same initial sound pressure level (SPL) for NH listeners. For HI listeners, the initial SPL of speech was set to 5 dB higher than the noise SPL in Experiment 1, and to the same SPL of noise in Experiment 2. In unsegregated conditions, the noise level was fixed while the speech level was adjusted during the adaptive procedure. In ideal masking conditions, as input SNR drops IBM becomes sparser with fewer 1's. To ensure that ideally masked stimuli remain audible at very low SNRs, the speech

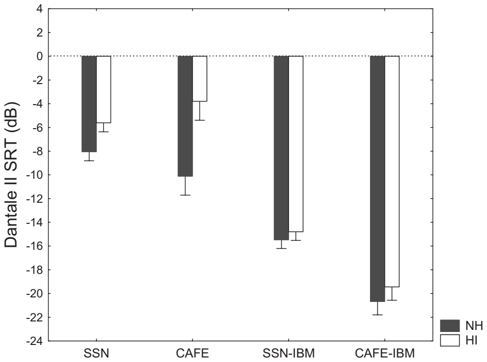

FIG. 3. SRTs for different conditions of Experiment 1 for NH and HI listeners. A more negative SRT corresponds to better performance. Error bars indicate 95% confidence intervals of the means.

level was fixed while the noise level was adjusted in all IBM conditions. As a result, with fewer retained *T-F* units their sound levels became higher even though the levels of the speech signals within these units were unchanged, and the loudness of a processed mixture was thus kept within a small range. This way of adjusting input SNR ensured that the stimuli in all four conditions were comfortably audible.

During a test, a subject was seated in a sound attenuating booth. Test stimuli were generated using the built-in sound card (SoundMAX) in a control computer (IBM ThinkCenter S50) and then presented diotically to a listener through headphones (Sennheiser HD 280 Pro). For HI listeners, an external amplifier (Behring Powerplay HA4000) was used to increase the sound level so that the stimuli within the test range were all audible and yet not uncomfortably loud. The amplification level was adjusted once for each HI listener before the test began.

4. Statistical analysis and power

During the planning phase of the study, the experiment was statistically powered to detect a within-subject between-condition difference of 1.0 dB on mean scores across conditions on the Dantale II test described subsequently for $p < 0.05$ at 80% power. This required at least ten complete data sets per condition. Analysis of variance (ANOVA) was performed on all of the data from NH and HI subjects, with within-subject factors of type of processing (IBM or unsegregated) and of type of noise (SSN or cafeteria noise), and a between-subject factor of subject type (NH and HI). *Post hoc* tests were the Bonferroni test and/or the Fisher least-significant-difference (LSD) test, applied where appropriate. The Bonferroni test was used as the most conservative test to indicate differences between means, while the Fisher LSD test was used as the most conservative test for a null result. All statistics were performed using STATISTICA version 7 (StatSoft, 2007).

B. Results and discussion

Figure 3 shows the SRT results of all four test conditions: SSN, CAFE, SSN-IBM, and CAFE-IBM, for both NH and HI listeners. For NH listeners, the mean SRT for unsegregated mixtures with SSN (SSN) is −8.15 dB, for unsegregated mixtures with cafeteria noise (CAFE) is −10.25 dB, for ideal masking with SSN (SSN-IBM) is −15.56 dB, and for ideal masking with cafeteria noise (CAFE-IBM) is −20.70 dB. The ANOVA for NH subjects showed that the main effects of processing type and noise type were significant [$F(1,11)=606.1$, $p<0.001$, and $F(1,11)=78.1$, $p<0.001$, respectively], and there was also a significant interaction between processing type and noise type [$F(1,11)=32.3$, $p<0.001$]. The Bonferroni *post hoc* tests indicated that all means were significantly different [$p<0.005$] from one another. The results show that ideal masking leads to lower (better) SRT compared to unsegregated mixtures regardless of background noise, that the cafeteria background yields a lower SRT than the SSN, and that ideal masking has a greater effect on the cafeteria background. The SRT for the unsegregated SSN condition is comparable to the reference level of −8.43 dB for the Dantale II task (Wagener *et al.*, 2003). The lower SRT for the cafeteria background is consistent with previous studies showing that NH listeners exhibit higher intelligibility in fluctuating backgrounds (Festen and Plomp, 1990; Peters *et al.*, 1998).

For the SSN background, IBM produces an average SRT improvement of 7.4 dB. This level of improvement is consistent with what was found by Anzalone *et al.* (2006) using the HINT test (Nilsson *et al.*, 1994), but higher than the 5 dB improvement reported by Brungart *et al.* (2006) using the CRM task (Bolia *et al.*, 2000). The main difference between our experiment and Brungart *et al.* (2006) lies in different *LC* values: their test uses 0 dB *LC* whereas *LC* is set to −6 dB in our study. As reported in Brungart *et al.* (2006) the choice of $LC=-6$ dB seems better than $LC=0$ dB in terms of speech intelligibility (see also Sec. II).

For the cafeteria background, ideal masking lowers SRT by 10.5 dB on average, which represents a larger gain than for the SSN background. Unlike SSN, the cafeteria background contains significant spectral and temporal modulations which contribute to better intelligibility in the unsegregated condition. We stress that the larger SRT improvement for this background is achieved on top of the better performance of listening to unsegregated mixtures.

For HI listeners, the mean SRTs are −5.61, −3.80, −14.79, and −19.44 dB for the SSN, CAFE, SSN-IBM, and SSN-CAFE conditions, respectively. The ANOVA where both NH and HI subjects were included showed that the main effects of subject type, processing type, and noise type were significant [$F(1,22)=17.2$, $p<0.001$; $F(1,22)=1959.0$, $p<0.001$; and $F(1,22)=100.6$, $p<0.001$, respectively], and there were also significant interaction effects between-subject type and processing type, subject type and noise type, and processing type and noise type [$F(1,22)=49.9$, $p<0.001$; $F(1,22)=19.2$, $p<0.001$; and $F(1,22)=163.9$, $p<0.001$ respectively], as well as a three-way interaction between subject type, processing type, and noise type [$F(1,11)=19.7$, $p<0.001$]. The Bonferroni as well as the Fisher LSD *post hoc* tests on the three-way interaction indicated that all means were significantly different ($p<0.006$)

except for the SSN-IBM and CAFE-IBM conditions where the differences between NH and HI listeners were insignificant ($p > 0.05$). The *post hoc* results show that ideal masking produces lower SRT compared to unsegregated mixtures regardless of noise type, and has a greater effect for the cafeteria background. No difference, however, was revealed between the NH subjects and the HI subjects in the two IBM conditions by either the more conservative Bonferroni test or the less conservative Fisher LSD test. The elevated levels of SRT in the two unsegregated conditions show that HI listeners perform worse in speech recognition in noisy environments, and the levels of SRT increment, 2.5 dB for the SSN condition and 6.5 dB for the CAFE condition, are broadly compatible with previous findings of HI listeners' increased difficulty in speech understanding in noise (see Sec. I). IBM lowers SRT substantially. The SRT gain resulting from ideal masking is 9.2 dB for the SSN background, and this level of improvement is compatible with that reported in Anzalone *et al.* (2006). For the cafeteria background, ideal masking produces a very large SRT improvement of 15.6 dB.

By comparing NH and HI results in Fig. 3, it is clear that HI listeners benefit from ideal masking even more than NH listeners, particularly for the cafeteria background. The results suggest that, after IBM processing, the intelligibility performance is comparable for HI and NH listeners in both SSN and cafeteria backgrounds. It is remarkable that the speech intelligibility of HI listeners becomes statistically indistinguishable from that of NH listeners after ideal masking.

IV. EXPERIMENT 2: EFFECTS OF BAND-LIMITED IDEAL BINARY MASKING ON SPEECH-RECEPTION THRESHOLD

The results of Experiment 1 show large SRT improvements resulting from IBM processing. As mentioned in Sec. I, a main finding reported by Anzalone *et al.* (2006) is that, while NH listeners benefit from IBM in both the LF and HF ranges, HI listeners benefit from ideal masking only in the LF range. This finding is significant because it suggests that, to alleviate the hearing loss of HI listeners, one need not worry about performing *T-F* masking in the HF range; speech segregation at HFs tends to be more difficult than at LFs (Wang and Brown, 2006). Although their interpretation based on the upward spread of masking is reasonable, the fact that they apply constant amplification with no spectral shaping to compensate for the sloping hearing loss of their subjects may suggest a simpler interpretation: the lack of the IBM benefit in the HF range may be partially accounted for by the potentially less compensated hearing loss at HFs. Experiment 2 was primarily designed to assess the importance of IBM processing at HFs for HI listeners as compared to NH listeners. In this experiment, we compensated for the hearing loss of individual listeners based on their audiograms. We compare the intelligibility performance in three setups: IBM in the LF range only, ideal masking in the HF range only, and ideal masking in the AF range. Both SSN and cafeteria backgrounds are used. Consequently, there are a total of six test conditions in this experiment.

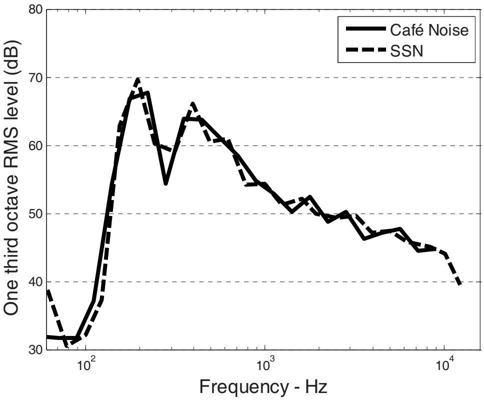

FIG. 4. Long-term spectrum of the SSN in Dantale II (redrawn from Wagener *et al.*, 2003). The spectrum is expressed as root mean square levels in one-third octave bands. Also shown is the long-term spectrum of the cafeteria noise.

A. Methods

1. Stimuli

As in Experiment 1, Dantale II sentences were used as target speech, and SSN and cafeteria noise were used as two different backgrounds. The IBM processing in the AF condition was the same as in Experiment 1. For the LF condition, the same IBM processing as in Experiment 1 was used in the lower 32 frequency channels while an all-1 mask was applied to the higher 32 frequency channels. This way of processing produces no segregation in the HF range. In the HF condition, the reverse was done: IBM was applied to the higher 32 channels while an all-1 mask was applied to the lower 32 channels (hence no segregation in the LF range). This equal division of the 64-channel gammatone filterbank yields a frequency separation boundary approximately at 1.35 kHz. Note that this boundary separating LF and HF ranges is a little lower than the 1.5 kHz boundary used in Anzalone *et al.* (2006). Our choice was partly motivated by the consideration that both the speech material and the SSN in the Dantale II corpus have energy distribution heavily tilted toward LFs so that IBM processing below 1 kHz likely removes significantly more noise than IBM processing above 1 kHz. The long-term spectrum of the SSN (Wagener *et al.*, 2003) is shown in Fig. 4, along with the long-term spectrum of the cafeteria noise. With the 1.5 kHz boundary, the NH results from Anzalone *et al.* (2006) show that the SRT in their LF condition is a little lower than the SRT in their HF condition.

2. Listeners

12 NH listeners and 12 HI listeners participated in this experiment. The pool of NH listeners was the same as that participated in Experiment 1 except for one. This substitution lowered the average age from 37 to 36 without altering the age range. The pool of HI listeners also remained the same as in Experiment 1 except for one. This substitution (see Fig.

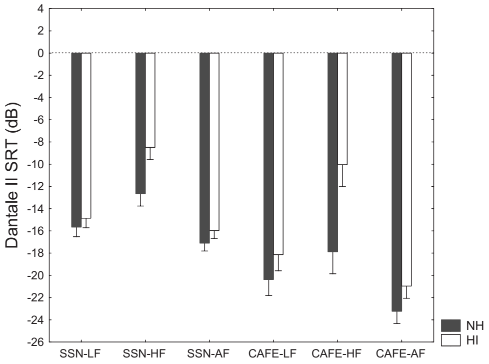

FIG. 5. SRTs for different conditions of Experiment 2 for NH and HI listeners. Error bars indicate 95% confidence intervals of the means.

2), plus a listener whose birthday occurred between the two experiments, changed the average age from 67 to 66 without changing the age range. Again, subjects were naïve to the purpose and design of the experiment. NH listeners were familiar with the Dantale II sentences by virtue of participating in Experiment 1, and as noted in Sec. III A 2, HI listeners had experience listening to Dantale II sentences prior to Experiment 1. Due to the limited number of test lists (15) available in the Dantale II corpus, the same lists used in Experiment 1 were also employed in Experiment 2. It is worth mentioning that the corpus was designed for repeated usage (Wagener *et al.*, 2003; see also Sec. III A 1).

3. Procedure and statistical analysis

The procedure of this experiment is the same as in Experiment 1 except for the following. To vary the input SNR, the noise level was adjusted while the speech level was fixed as in the ideal masking conditions of Experiment 1. In the LF and HF conditions, there is no segregation in half of the frequency channels. As the input SNR decreases in the negative range, the sound level of a stimulus in these conditions is dominated by the background noise in the unsegregated frequency range and hence becomes increasingly louder. To ensure that LF and HF stimuli are not too loud for NH listeners who have very low SRTs, the speech level was fixed at a lower volume than in Experiment 1. Despite this change of sound level, all test stimuli were still comfortably audible for NH listeners. Note that this change did not impact HI listeners as the amount of amplification was individually set for them. ANOVA was performed similarly on all the data from NH and HI subjects as in Experiment 1, with within-subject factors of type of processing (LF, HF, or AF) and of type of noise (SSN or CAFE), and a between-subject factor of subject type (NH or HI).

B. Results and discussion

Figure 5 shows the SRT results of all six test conditions in Experiment 2: SSN-LF, SSN-HF, SSN-AF, CAFE-LF, CAFE-HF, and CAFE-AF, for both NH and HI listeners. The ANOVA for NH subjects showed that the main effects of processing type and noise type were significant [$F(2,22)=255.5$, $p<0.001$, and $F(1,11)=231.2$, $p<0.001$, respectively], and there was also a significant interaction between processing type and noise type [$F(2,22)=4.4$, $p<0.05$]. The Bonferroni tests indicated that all NH means were significantly different ($p<0.006$) from one another, except between the SSN-AF and the CAFE-HF condition. For the SSN background, the mean SRT is −15.66 dB in the LF condition, −12.65 dB in the HF condition, and −17.10 dB in the AF condition. The results show that NH listeners perform better when IBM processing is applied in the LF range than in the HF range, and the difference in SRT is approximately 3 dB. This SRT difference is larger than the SRT difference of slightly more than 1 dB reported by Anzalone *et al.* (2006), despite the fact that the boundary separating LFs and HFs is 1.35 kHz in our processing and 1.5 kHz in their processing. Even with the lower frequency boundary we find that, with the same input SNR, the HF condition leaves more noise than the LF condition since the noise energy is distributed mostly in the LF range (see Fig. 4). The discrepancy is likely due to different ways of IBM processing used in the two studies. The AF condition yields the lowest SRT, which is about 1.6 dB lower than in the LF condition.

For the cafeteria background, the mean SRT is −20.37 dB in the LF condition, −17.88 dB in the HF condition, and −23.24 dB in the AF condition. Clearly NH subjects perform better in this background than in SSN, consistent with the results of Experiment 1. Again, NH listeners benefit more from IBM processing at LFs than at HFs and the relative benefit is 2.5 dB. The AF condition also gives the lowest SRT, which is about 2.9 dB lower than in the LF condition. That NH subjects performed better in the AF condition than in the LF condition for both the SSN and cafeteria backgrounds suggest that they do benefit from IBM in the HF range, even though the benefit is not as high as from the LF range.

The ANOVA where both HI and NH subjects were included showed that the main effects of subject type, processing type, and noise type were significant [$F(1,22)=19.1$, $p<0.001$; $F(2,44)=255.4$, $p<0.001$; and $F(1,22)=317.2$, $p<0.001$, respectively], and there were also significant interaction effects between subject type and processing type, subject type and noise type, and processing type and noise type [$F(2,44)=31.2$, $p<0.001$; $F(1,22)=18.3$, $p<0.001$; and $F(2,44)=14.2$, $p<0.001$, respectively], as well as a three-way interaction between subject type, processing type, and noise type [$F(2,44)=5.4$, $p<0.01$]. Table I shows the Fisher LSD *post hoc* tests. As seen in the table, all conditions were significantly different ($p<0.05$) from one another within the NH subjects (conditions {1}–{6} contrasted against each other) and within the HI subjects (conditions {7}–{12} contrasted against each other). However, the differences between NH and HI were insignificant for the conditions of SSN-LF and SSN-AF.

For HI listeners, the mean SRTs for the SSN background are −14.85, −8.49, and −15.96 dB for the LF, HF, and AF conditions, respectively. The SRT advantage of the LF condition over the HF condition is 6.4 dB, whereas the advantage of the AF condition over the LF condition is only

TABLE I. Fisher LSD *post hoc* significance tests for the three-way interaction of subject type, processing type, and noise type. Significance levels above $p > 0.05$ are given in boldface.

Subject type	Processing type	Test condition	{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}	{9}	{10}	{11}
NH	SSN-LF	{1} −15.66											
	SSN-HF	{2} −12.65	0.00										
	SSN-AF	{3} −17.10	0.00	0.00									
	CAFE-LF	{4} −20.37	0.00	0.00	0.00								
	CAFE-HF	{5} −17.88	0.00	0.00	**0.07**	0.00							
	CAFE-AF	{6} −23.24	0.00	0.00	0.00	0.00	0.00						
HI	SSN-LF	{7} −14.85	**0.45**	0.05	0.04	0.00	0.01	0.00					
	SSN-HF	{8} −8.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
	SSN-AF	{9} −15.96	**0.77**	0.00	**0.29**	0.00	**0.08**	0.00	0.01	0.00			
	CAFE-LF	{10} −18.13	0.03	0.00	**0.34**	0.04	**0.81**	0.00	0.00	0.00	0.00		
	CAFE-HF	{11} −10.05	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	CAFE-AF	{12} −20.96	0.00	0.00	0.00	**0.58**	0.01	0.04	0.00	0.00	0.00	0.00	0.00

1.1 dB. These data are generally comparable with those in Anzalone *et al.* (2006). The results suggest that HI listeners derive considerably more benefit from ideal masking at LFs than at HFs, and the SRT difference is much larger than for NH listeners (see Fig. 5). Although part of the larger gap may be caused by a larger SRT gain (9.2 dB) in HI listeners than that (7.4 dB) in NH listeners due to IBM processing, the fact that the relative advantage of the AF condition over the LF condition for HI listeners is even a little smaller than for NH listeners (1.1 dB versus 1.6 dB) strongly indicates that IBM processing in LF is to a greater extent responsible for the SRT improvement of ideal masking in HI listeners than in NH listeners. In other words, almost all the benefit of IBM can be obtained by IBM only in the LF range. This, of course, is not to say that ideal masking in HF does not improve speech intelligibility compared to no segregation. As illustrated in Fig. 3, IBM processing at all frequencies results in a 9.2 dB SRT improvement compared to no segregation, and the AF condition produces a 7.5 dB relative advantage over the HF condition. This comparison suggests that ideal masking at HFs produces some improvement in speech intelligibility.

For the cafeteria background, the SRTs in the LF, HF, and AF conditions are −18.13, −10.05, and −20.96 dB, respectively (see Fig. 5). The SRT advantage of LF processing over HF processing is 8.1 dB and that of AF over LF is 2.8 dB. These results show a similar pattern as for the SSN background, even though the SRT difference of 2.8 dB between the LF and AF conditions clearly reaches statistical significance (see Table I), and HF processing yields a significant SRT improvement over no segregation as suggested by comparing with the data in Experiment 1. The use of the fluctuating cafeteria background reinforces the conclusion that ideal masking in LF produces a much stronger benefit than that in HF, and this effect is greater in HI listeners than in NH listeners.

The two AF conditions for the SSN and cafeteria backgrounds are the same as the corresponding ideal masking conditions in Experiment 1. The NH performances in Experiment 2 are somewhat better than in Experiment 1. A comparison between Fig. 5 and Fig. 3 shows that the discrepancies are 1.5 dB for SSN and 2.5 dB for cafeteria noise. The only difference in stimuli is the sound level; as pointed out in Sec. IV A 3, the sound level is softer in Experiment 2 than in Experiment 1. For example, at the input SNR of −10 dB, the sound level in Experiment 1 is about 63 dB(A) SPL for the SSN background and 75 dB(A) for the cafeteria background, while the corresponding levels in Experiment 2 are 51 and 51 dB(A), respectively. Studies suggest that softer sound can produce better recognition under certain conditions (Hagerman, 1982; Studebaker *et al.*, 1999). To examine whether the sound volume was a factor in the performance differences, we performed a follow-up experiment with the same pool of the NH listeners who participated in Experiment 2. The follow-up experiment was to simply check subjects' percent correct scores at the sound levels used in the two experiments when the input SNR was fixed at one of the SRTs (alternating between subjects) already obtained in the experiments. The cafeteria background noise was used. The scores are 50.6% with the louder level of Experiment 1 and 58.6% with the softer level of Experiment 2. The 8% difference is statistically significant [$t(11)=3.31$, $p<0.01$], but unlikely large enough to explain the 2.5 dB SRT difference. Perhaps more important is a learning effect. Unlike HI listeners who were experienced with the Dantale II task, NH listeners used in this investigation had little prior experience with auditory experiments before participating in Experiment 1. When they participated in the second experiment, the familiarity with the Dantale II task acquired during Experiment 1 likely contributed to their better performance. In the predecessor to Dantale II—the Hageman sentence test—Hagerman and Kinnefors (1995) found a training effect of about 0.07 dB per ten sentences, which may explain the differences between Experiments 1 and 2. This interpretation is consistent with the observation that the corresponding performance differences between Experiment 1 and Experiment 2 are smaller for HI listeners; one-third of the mean performance differences is accounted for by the replacement of one HI listener from Experiment 1 to Experiment 2 (see Sec. IV A 2).

V. GENERAL DISCUSSION

The robustness of speech recognition in noise by NH listeners is commonly attributed to the perceptual process of glimpsing, or "listening in the dips," which detects and gathers *T-F* regions of a sound mixture where target speech is relatively stronger compared to interference (Miller and Licklider, 1950; Howard-Jones and Rosen, 1993; Assmann and Summerfield, 2004; Li and Loizou, 2007). As glimpsing involves grouping, this account is closely related to the ASA account that applies to both speech and nonspeech signals (Bregman, 1990). Poorer performance of listeners with hearing loss in fluctuating backgrounds is generally explained as their inability to take advantage of temporal and spectral dips, perhaps caused by reduced frequency selectivity and temporal resolution (Moore, 2007). IBM could be understood as producing glimpses or performing ASA for the listener. The fact that ideal masking also improves intelligibility of NH listeners suggests that even listeners without hearing loss can fail to make full use of the speech information available in a noisy input. The less-than-ideal performance in noisy environments is probably caused by the failure in detecting a glimpse—a *T-F* region with relatively strong target energy—or grouping detected glimpses. This failure becomes more acute with hearing loss. Because ideal masking does an "ideal" job of glimpsing for the auditory system, it helps to nearly equalize the performances of HI and NH listeners (see Fig. 3).

The results of Experiment 1 demonstrate that listeners with or without hearing loss benefit more from IBM processing in the cafeteria background than in the SSN background. The cafeteria background has temporal and spectral modulations, and as a result the amount of informational masking caused by target-masker similarity is expected to be higher than that in SSN. Indeed, some listeners voluntarily commented after the experiment that the conversation in the background distracted their attention, making it harder to concentrate on target utterances. The larger SRT improvement observed for the cafeteria background is thus consistent with the interpretation that ideal masking removes or largely attenuates informational masking (Brungart *et al.*, 2006). In a situation extremely conducive to informational masking, namely, the mixtures of speech utterances of the same talker, Brungart *et al.* (2006) found that the effect of ideal masking is tantamount to a 22–25 dB improvement in input SNR. The 10.5 dB SRT improvement obtained through ideal masking in the cafeteria background, although greater than that obtained in the SSN background, is much smaller than that obtained in mixtures of same-talker utterances. The improvement is also smaller than those obtained in mixtures of different-talker utterances (Chang, 2004), although the gap is not quite as big as in same-talker mixtures. One can therefore expect even larger SRT improvements when interference is one or several competing talkers, a kind of background that produces very large performance gaps between NH and HI listeners as reviewed in Sec. I.

The results of Experiment 2 are on the whole consistent with the related findings of Anzalone *et al.* (2006) even though we used individual gain prescriptions to compensate for listeners' hearing loss. The results are also qualitatively consistent with the findings of Li and Loizou (2007) illustrating that glimpses in the LF to mid-frequency range are more beneficial for speech intelligibility than those in the HF range. However, a few differences between our results and the results of Anzalone *et al.* (2006) are worth noting. First, although considerably smaller than LF processing, there is a benefit from ideal masking in the HF range for HI listeners in our study whereas their study did not show a significant benefit. A possible reason is the individual gain prescription employed in our study that makes segregated speech relatively louder in the HF range than the constant gain applied in their study. Second, we find a relatively greater LF benefit in NH listeners than in their study. The main reason, we believe, is that LF processing removes more background noise than HF processing for a given input SNR. With negative input SNRs (see Fig. 5), the residual noise in the HF condition is in the LF range while that in the LF condition is in the HF range, and the background noises used in our experiments have energy distributed mostly in the LF range, as shown in Fig. 4. This explanation, not considered by Anzalone *et al.*, gives a partial account for the larger LF benefit for listeners with hearing loss. The large SRT gap between LF and HF processing for HI listeners (see Fig. 5), however, cannot be fully explained this way as the gap is substantially larger—to the extent that the SRT performance in LF processing is almost the same as in AF processing. Another likely reason is upward spread of masking (Anzalone *et al.*, 2006) which listeners with sensorineural hearing loss are especially susceptible to (Jerger *et al.*, 1960; Gagne, 1988; Klein *et al.*, 1990). Upward spread of masking is a more prominent factor in the HF condition because of no segregation in the LF range. Also, with more hearing loss at HFs (see Fig. 2), HI listeners are less able to utilize audible HF speech information in recognition compared to NH listeners (Dubno *et al.*, 1989; Ching *et al.*, 1998; Hogan and Turner, 1998). This could also contribute to a steeper performance decline of HF processing relative to AF processing for HI listeners than for NH listeners.

Despite different definitions of IBM, the SRT improvements observed in our study and in Anzalone *et al.* (2006) are very close for the SSN background. It is all the more remarkable considering that their IBM is generated on a sample-by-sample basis while ours is generated on a frame-by-frame basis, which has a drastically lower temporal resolution, and that, in their experiments, IBM-determined gains take the values of 1 and 0.2 while the gains take the values of 1 and 0 in our experiments. The use of two-valued gains is a key similarity between the studies. The most important difference is, of course, that our definition is based on a comparison between target and interference energy and theirs is between target energy and a fixed threshold. The local SNR based IBM is arguably easier to estimate computationally, as many speech segregation algorithms compute binary time-frequency masks by exploring local SNR explicitly or implicitly (Divenyi, 2005; Wang and Brown, 2006). Also, there is little basis in a noisy signal to identify those *T-F* regions of significant target energy where interference is much stronger.

The results from our experiments have major implications for CASA and speech enhancement research aiming to improve speech intelligibility in noisy environments. In addition to affirming the general effectiveness of IBM as a computational goal, our data provide direct evidence that a choice of *LC* at −6 dB for IBM construction, first suggested by Brungart *et al.* (2006), is effective for improving human speech recognition. A comparison between the data of Brungart *et al.* (2006) and ours for the SSN background indicates that the IBM with −6 dB *LC* yields larger SRT improvement than commonly used 0 dB *LC*. Compared to 0 dB *LC*, the choice of −6 dB *LC* retains those *T-F* units where local SNR falls between 0 and −6 dB in ideal masking (see Fig. 1). From the standpoint of SNR, such inclusion will lower the overall SNR of the segregated signal. In other words, if the objective is to improve the SNR of the output signal, the choice of −6 dB *LC* is a poorer one compared to that of 0 dB *LC*. This discussion casts further doubt on the suitability of traditional SNR as a performance metric to evaluate sound separation systems, and at the same time, could shed light on why monaural speech enhancement algorithms often improve SNR but not speech intelligibility (see Sec. I). Another strong implication of our results (see also Anzalone *et al.*, 2006) is that performing speech separation in the LF range is a great deal more important than in the HF range, particularly for improving speech intelligibility of HI listeners.

Our results point to a very promising direction for hearing aid design to improve speech intelligibility in noise of listeners with hearing loss, that is, by designing hearing aids that function in similar ways to IBM. IBM processing improves SRT by a large margin, and HI listeners derive larger benefit than NH listeners. Equally important, the profile of improvement with respect to different kinds of background noise seems to match that of typical hearing impairment. We consider it a highly significant result that ideal masking almost equalizes the intelligibility performances of HI and NH listeners (see Fig. 3). Of course, facing a noisy input IBM cannot be directly constructed and algorithms must be developed to estimate IBM. Encouraging effort has been made in CASA with the explicit goal of IBM estimation (Wang and Brown, 2006), and in limited conditions high-quality estimates are obtainable (see, e.g., Roman *et al.*, 2003). However, computing binary masks close to the IBM in unconstrained acoustic environments remains a major challenge. On the other hand, the extent of intelligibility gain for HI listeners produced by IBM processing much more than fills the SRT gap from NH listeners; Experiment 1 shows a gap of 2.5 dB for the SSN background and a gap of 6.5 dB for the cafeteria background while the ideal masking improvements for HI listeners are 9.2 and 13.8 dB for the two backgrounds, respectively. Hence, perfect IBM estimation is not necessary to bring the performance of HI listeners to the same level as that of NH listeners.

VI. CONCLUSION

The present study was designed to evaluate the impact of IBM on speech intelligibility in noisy backgrounds for both NH and HI listeners. Two experiments were conducted and the main results are summarized below.

- For NH listeners, IBM processing resulted in 7.4 dB SRT reduction for SSN and 10.5 dB reduction for cafeteria noise.
- For HI listeners, IBM processing resulted in 9.2 dB SRT reduction for SSN and 15.6 dB reduction for cafeteria noise.
- After IBM processing, the intelligibility performances for HI listeners and NH listeners were comparable.
- For NH listeners, IBM processing at LFs produced greater SRT reduction than at HFs. The differences were 3 dB for SSN and 2.5 dB for cafeteria noise.
- For HI listeners, IBM processing at LFs produced greater SRT reduction than at HFs. The differences were 5.5 dB for SSN and almost 8 dB for cafeteria noise.

ACKNOWLEDGMENTS

We thank the Associate Editor Ken Grant, and two anonymous reviewers for their helpful comments. The work was conducted while D.W. was a visiting scholar at Oticon A/S. The authors are grateful to M. Schlaikjer, L. Bramsløw, and M. Hartvig, for their assistance in the experiments, and Y. Li for his assistance in figure preparation. D.W. was supported in part by an AFOSR grant (F49620-04-01-0027) and an NSF grant (IIS-0534707).

Alcantara, J. I., Dooley, G., Blamey, P., and Seligman, P. (**1994**). "Preliminary evaluation of a formant enhancement algorithm on the perception of speech in noise for normally hearing listeners," Audiology **33**, 15–27.

Alcantara, J. I., Moore, B. C. J., Kuhnel, V., and Launer, S. (**2003**). "Evaluation of the noise reduction system in a commercial digital hearing aid," Int. J. Audiol. **42**, 34–42.

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Assmann, P., and Summerfield, A. Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York) pp. 231–308.

Baer, T., Moore, B. C. J., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev. **30**, 49–72.

Beck, S., and Zacharov, N. (**2006**). *Perceptual Audio Evaluation: Theory, Method and Application* (Wiley, Chichester, NY).

Benesty, J., Makino, S., and Chen, J., eds. (**2005**). *Speech Enhancement* (Springer, New York).

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Brungart, D. S. (**2001**). "Information and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Bunnell, H. T. (**1990**). "On enhancement of spectral contrast in speech for hearing-impaired listeners," J. Acoust. Soc. Am. **88**, 2546–2556.

Carhart, R. C., and Tillman, T. W. (**1970**). "Interaction of competing speech signals with hearing losses," Arch. Otolaryngol. **91**, 273–279.

Chang, P. (**2004**). "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," M.S. thesis, The Ohio State University Department of Computer Science and Engineering, Columbus, OH; http://www.cse.ohio-state.edu/pnl/theses.html (Last viewed September 2008).

Ching, T. Y. C., Dillon, H., and Byrne, D. (**1998**). "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, 1128–1140.

Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York).

Divenyi, P., ed. (**2005**). *Speech Separation by Humans and Machines* (Kluwer Academic, Norwell, MA).

Drullman, R. (**1995**). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.

Dubno, J. R., Dirks, D. D., and Ellison, D. E. (**1989**). "Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. I: The contribution of selected frequency regions," J. Acoust. Soc. Am. **85**, 347–354.

Edwards, B. (**2004**). "Hearing aids and hearing impairment," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York).

Eisenberg, L. S., Dirks, D. D., and Bell, T. S. (**1995**). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," J. Speech Hear. Res. **38**, 222–233.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Gagne, J.-P. (**1988**). "Excess masking among listeners with a sensorineural hearing loss," J. Acoust. Soc. Am. **83**, 2311–2321.

Greenberg, J. E., and Zurek, P. M. (**1992**). "Evaluation of an adaptive beamforming method for hearing aids," J. Acoust. Soc. Am. **91**, 1662–1676.

Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.

Hagerman, B., and Kinnefors, C. (**1995**). "Efficient adaptive methods for measurements of speech reception thresholds in quiet and in noise," Scand. Audiol. **24**, 71–77.

Hansen, M., and Ludvigsen, C. (**2001**). "Dantale II—Danske Hagermann sætninger (Dantale II—Danish Hagermann sentences)," Danish Speech Audiometry Materials (Danske Taleaudiomaterialer), Værløse, Denmark.

Hogan, C. A., and Turner, C. W. (**1998**). "High-frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432–441.

Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomudulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**, 2915–2922.

Hygge, S., Ronnberg, J., Larsby, B., and Arlinger, S. (**1992**). "Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds," J. Speech Hear. Res. **35**, 208–215.

Jerger, J. F., Tillman, T. W., and Peterson, J. L. (**1960**). "Masking by octave bands of noise in normal and impaired ears," J. Acoust. Soc. Am. **32**, 385–390.

Johannesson, R. B. (**2006**). "Output SNR measurement method," Report No. 052-08-04, Oticon Research Centre Eriksholm, Snekkersten, Denmark.

Kates, J. M., and Weiss, M. R. (**1996**). "A comparison of hearing-aid array-processing techniques," J. Acoust. Soc. Am. **99**, 3138–3148.

Klein, A. J., Mills, J. H., and Adkins, W. Y. (**1990**). "Upward spread of masking, hearing loss, and speech recognition in young and elderly listeners," J. Acoust. Soc. Am. **87**, 1266–1271.

Levitt, H. (**2001**). "Noise reduction in hearing aids: A review," J. Rehabil. Res. Dev. **38**, 111–121.

Li, N., and Loizou, P. C. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.

Li, N., and Loizou, P. C. (**2008a**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. **123**, EL59–EL64.

Li, N., and Loizou, P. C. (**2008b**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.

Lim, J., ed. (**1983**). *Speech Enhancement* (Prentice-Hall, Englewood Cliffs, NJ).

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.

Moore, B. C. J. (**2003a**). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic, San Diego, CA).

Moore, B. C. J. (**2003b**). "Speech processing for the hearing-impaired: Successes, failures, and implications for speech mechanisms," Speech Commun. **41**, 81–91.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK).

Nilsson, M., Soli, S., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.

Plomp, R. (**1994**). "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," Ear Hear. **15**, 2–12.

Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Ricketts, T., and Hornsby, B. W. (**2003**). "Distance and reverberation effects on directional benefit," Ear Hear. **24**, 472–484.

Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.

Schum, D. J. (**2003**). "Noise-reduction circuitry in hearing aids, II: Goals and current strategies," Hear. J. **56**, 32–41.

Simpson, A. M., Moore, B. C. J., and Glasberg, B. R. (**1990**). "Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners," Acta Oto-Laryngol. **469**, 101–107.

StatSoft, Inc. (**2007**). STATISTICA (data analysis software system), version 7, http://www.statsoft.com (Last viewed February 2008).

Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., and Gwaltney, C. A. (**1999**). "Monosyllabic word recognition at higher-than-normal speech and noise levels," J. Acoust. Soc. Am. **105**, 2431–2444.

Vestergaard, M. (**1998**). "The Eriksholm CD 01: Speech signals in various acoustical environments," Report No. 050-08-01, Oticon Research Centre Eriksholm, Snekkersten, Denmark.

Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Wang, D. L., and Brown, G. J., eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ/IEEE, New York).

Paper B

Speech Perception of Noise with Binary Gains

DeLiang Wang, Ulrik Kjems, Michael S. Pedersen,
Jesper B. Boldt, and Thomas Lunner

Journal of the Acoustical Society of America
Vol. 124, no. 4, pp. 2303–2307, October 2008.

©2008 Acoustical Society of America

Speech perception of noise with binary gains

DeLiang Wang[a)]
Department of Computer Science & Engineering, and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

Ulrik Kjems, Michael S. Pedersen, and Jesper B. Boldt
Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark

Thomas Lunner
Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark and Department of Clinical and Experimental Medicine, and Technical Audiology, Linköping University, S-58183 Linköping, Sweden

(Received 9 December 2007; revised 7 July 2008; accepted 8 July 2008)

For a given mixture of speech and noise, an ideal binary time-frequency mask is constructed by comparing speech energy and noise energy within local time-frequency units. It is observed that listeners achieve nearly perfect speech recognition from gated noise with binary gains prescribed by the ideal binary mask. Only 16 filter channels and a frame rate of 100 Hz are sufficient for high intelligibility. The results show that, despite a dramatic reduction of speech information, a pattern of binary gains provides an adequate basis for speech perception.

© *2008 Acoustical Society of America.* [DOI: 10.1121/1.2967865]

PACS number(s): 43.71.Gv, 43.66.Dc [KWG] Pages: 2303–2307

I. INTRODUCTION

Human speech recognition shows remarkable robustness in a variety of listening conditions, including competing talkers, environmental sounds, and background noise. Understanding how speech is recognized under these conditions is important not only for auditory perception but also for automatic speech recognition, where robustness to acoustic interference remains elusive (Lippmann, 1997; Allen, 2005). Related research in computational auditory scene analysis (CASA) and blind source separation makes use of a binary time–frequency (T–F) masking technique (Roman *et al.*, 2003; Hu and Wang, 2004; Yilmaz and Rickard, 2004). Time–frequency masking operates on a T–F representation or decomposition of the input into a two-dimensional matrix of T–F units. Such a representation can be readily generated by passing the input signal through a filterbank and then time windowing the response of each filter. Then binary masking as a means of separation amounts to identifying a binary mask where 1 indicates that the acoustic energy in the corresponding T–F unit is retained and 0 indicates that the energy is removed. In other words, binary masking applies a pattern of binary gains to the mixture signal. It should be noted that the term “masking” here means weighting the mixture, which is different from the same term used in psychoacoustics where it means blocking the target sound by acoustic interference.

Among T–F masks, the so-called ideal binary mask (IBM) has been suggested to be a goal of CASA (Wang, 2005). The IBM is a matrix where 1 indicates that the signal-to-noise ratio (SNR) within the corresponding T–F unit exceeds a threshold LC (local SNR criterion) and 0 otherwise. The mask is “ideal” because its construction requires that speech and noise be available before they are mixed, and the mask possesses certain optimality in terms of overall SNR gain when LC is set to 0 dB (Li and Wang, 2008).

Recent studies in speech perception show that applying IBM to noisy speech leads to large speech intelligibility improvements (Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Li and Loizou, 2008). In particular, Brungart *et al.* (2006) and Li and Loizou (2008) have shown that, with fixed levels of input SNR (between −10 and 0 dB), a range of LC values produces nearly 100% correct scores. The large intelligibility gain has been attributed to ideal segregation (or detection) that directs the listener’s attention to the T–F regions of noisy speech where the target speech is relatively strong. This explanation emphasizes the importance of the target signal contained in the T–F units labeled 1 for intelligibility. How important is the binary pattern of an ideal mask itself? This investigation was designed to isolate the intelligibility contribution of an IBM by removing the target speech signal from all T–F units.

Specifically, with linear filters, including gammatone filters (Patterson *et al.*, 1988; Wang and Brown, 2006), increasing or decreasing the SNR of a mixture while changing LC by the same amount does not alter the IBM. On the other hand, although co-reducing input SNR and LC does not change the IBM, the masked mixture becomes progressively noisier or contains less target signal. Taking this manipulation to the limit, i.e., setting both mixture SNR and LC to $-\infty$ dB, leads to an output that contains only noise with no target speech at all. This particular output corresponds to turning on or off the filtered noise according to a pattern prescribed by the IBM. Our study evaluates speech intelligibility of noise gated by the IBM obtained in this way.

[a)]Author to whom correspondence should be addressed. Electronic mail: dwang@cse.ohio-state.edu

II. METHOD

A. Stimuli

Our tests use sentences from the Dantale II data set as target speech and a speech-shaped noise as interference (Wagener *et al.*, 2003). The speech material in the Dantale II corpus consists of sentences recorded by a female Danish speaker. Each sentence has five words with a fixed grammar (name, verb, numeral, adjective and object), e.g., "Michael had five new plants" (English translation). Each position in a sentence takes a randomly chosen word from ten equally meaningful words. The speech-shaped noise included in the Dantale II corpus is produced by adding repeated utterances of each of the 250 test sentences in the corpus (see Wagener *et al.*, 2003). Both speech and noise materials were digitized at 20 kHz sampling frequency.

A speech utterance and the noise are first processed by a gammatone filterbank with varying numbers of filter channels. With 32 filters equally spaced on the equivalent rectangular bandwidth (ERB) rate scale with center frequencies distributed in the range of 2–33 ERBs (or 55–7743 Hz), the frequency response of the filterbank is nearly flat. In addition to a 32-channel gammatone filterbank, we also tested 16-, 8-, and 4-channel filterbanks. Each of the filterbanks spans the same frequency range with filters equally spaced on the ERB-rate scale, and in all cases each filter has the bandwidth of 1 ERB. With reduced channels, the frequency response of a filterbank is no longer flat and information in certain frequency bands is lost in comparison to the 32-channel filterbank case. A filter response is then windowed into time frames using 20 ms rectangular windows and a frame shift of 10 ms. This 100 Hz frame rate is commonly used in speech processing (Rabiner and Juang, 1993). The resulting $T-F$ representation has been called a cochleagram (Wang and Brown, 2006). The IBM is constructed from the cochleagrams of the target speech and the speech-shaped noise with both the mixture SNR (calculated during the period of a sentence) and LC set to 0 dB. The IBM is then applied to the noise cochleagram alone in a synthesis step to generate a waveform stimulus [see Wang and Brown (2006) for details of cochleagram analysis and synthesis]. Figure 1 illustrates the signal processing scheme using a Dantale II sentence. Take, for example, the 8-channel filterbank case. Figure 1(G) shows the IBM for this case, which is produced by comparing the 8-channel cochleagram of the Dantale II sentence and the 8-channel cochleagram of the speech-shaped noise. Applying the binary mask in Fig. 1(G) to gate the noise results in a waveform signal, which is represented in the cochleagram format in Fig. 1(H). Note that Fig. 1 represents the waveform signals from different channel numbers using the same 32-channel cochleagram representation in order to facilitate comparison. In other words, all the cochleagrams in Fig. 1 serve the purpose of signal representation and do not indicate the size of the filterbank used in IBM construction.

B. Subjects

Twelve normal-hearing, native Danish-speaking listeners participated in the experiment. All listeners had normal hearing, i.e., their hearing thresholds did not exceed 20 dB HL, and their ages ranged from 26 to 51 with the average age of 36.

C. Procedure

In each condition of the experiment, two lists, each with ten sentences, were randomly selected from the Dantale II corpus for IBM construction. Speech-shaped noise gated by the IBM was then presented to a listener. The subjects were instructed to repeat as many words as they could after listening to each stimulus corresponding to one sentence, and no feedback was provided to them regarding whether their responses were correct or not. A stimulus was presented only once. Subjects were given a training session by listening to two lists of clean (or unprocessed) sentences, which were not included in the subsequent test. Each subject test had four conditions corresponding to the filterbanks with 4, 8, 16, and 32 channels. The four test conditions plus training took less than 30 min. The presentation order of the four conditions was randomized and balanced among the 12 listeners.

Speech and noise were both set to the sound pressure level of 70 dB initially. To account for level differences caused by the use of different-sized filterbanks, stimuli were scaled by factors of two, four, and eight, for the 16-channel, the 8-channel, and the 4-channel filterbank, respectively. This level calibration resulted in stimuli with approximately the same loudness. On each trial, a stimulus was generated by the built-in sound card (SoundMAX) in a control computer (IBM ThinkCenter S50) and then presented diotically to a listener through headphones (Sennheiser HD 280 Pro) in a sound treated hearing test room.

III. RESULTS

Figure 2 shows the word recognition performance for all four conditions. The mean speech intelligibility scores for the four conditions are: 7.75%, 54.25%, 92.92%, and 97.08%, with increasing number of filter channels. The results show that nearly perfect speech recognition is obtained with 32 channels, and a high recognition rate is obtained with 16 channels. The subjects recognized more than half of the words when the number of channels was set to 8, but were unable to perform the recognition task when the number of channels was 4. A repeated measures analysis of variance (ANOVA) was conducted and the effect of number of channels was significant, $F(3,33)=179.05$, $p<0.00001$. The Tukey honest significant difference (HSD) procedure revealed that all pairwise differences among the means were significant, $p<0.001$, except for the difference between 16 and 32 channels, which was not significant. Both ANOVA and post hoc Tukey HSD tests were conducted on the rationalized arcsine-transformed percentage scores (Studebaker, 1985).

The performance variability across different listeners was small for the 32-channel and the 16-channel cases, suggesting that the acoustic information was sufficient for them to perform the recognition task. On the other hand, the individual variability for the 8-channel case was significantly larger than the 16-channel case, $F(1,11)=5.50$, $p<0.01$, sug-

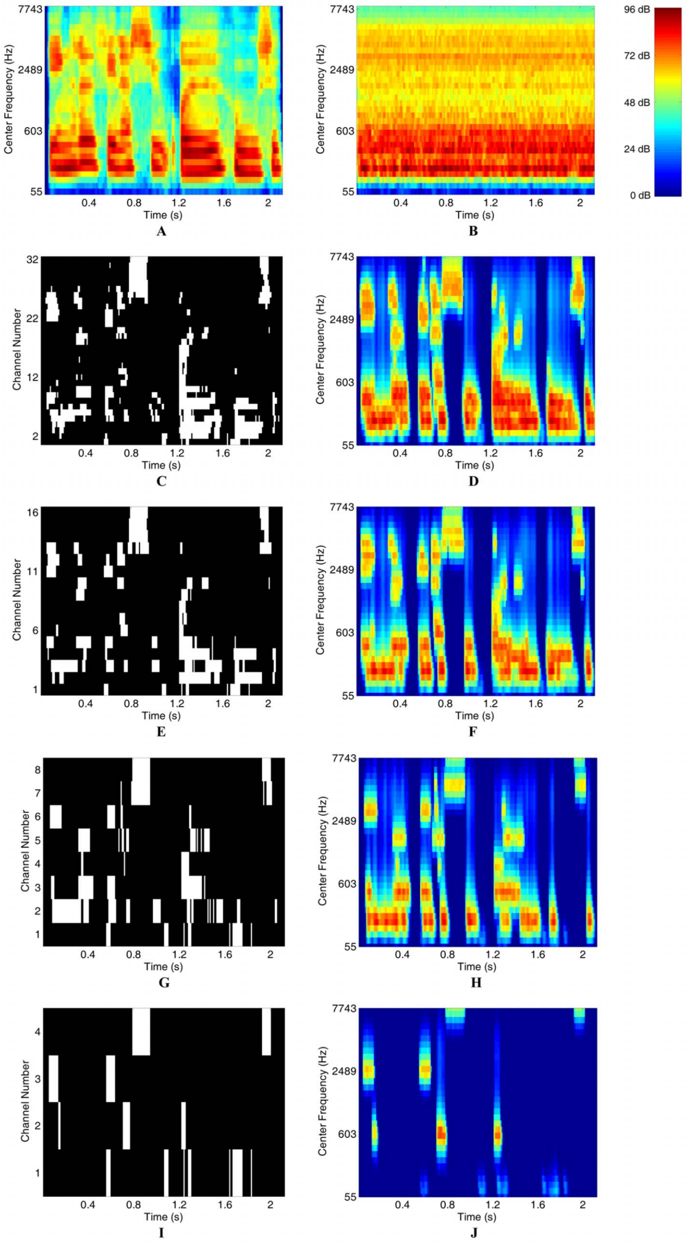

FIG. 1. (Color online) Illustration of gated noise by binary gains. (A) 32-channel cochleagram of a Dantale II sentence. (B) 32-channel cochleagram of speech-shaped noise. (C) IBM with 32 channels, where 1 is indicated by white and 0 by black. (D) 32-channel cochleagram of gated noise by the IBM in (C). (E) IBM with 16 channels. (F) 32-channel cochleagram of gated noise by the IBM in (E). (G) IBM with 8 channels. (H) 32-channel cochleagram of gated noise by the IBM in (G). (I) IBM with 4 channels. (J) 32-channel cochleagram of gated noise by the IBM in (I).

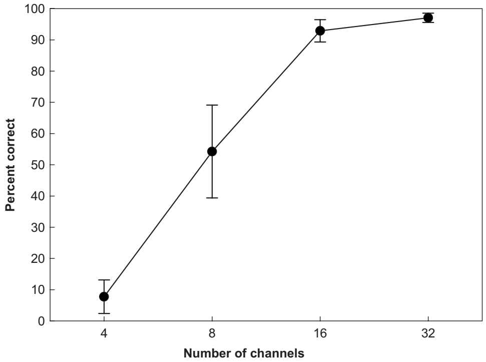

FIG. 2. Word intelligibility scores for 12 normal-hearing listeners with respect to different filterbank sizes. Dots denote the mean scores and vertical bars indicate 95% confidence intervals.

gesting that factors such as the ability and tendency to guess, concentration, and prior experience with corrupted speech, come into play.

The results in Fig. 2 clearly demonstrate that very high recognition can be obtained by turning on and off 16 bands of noise at a rate of 100 Hz following a specific pattern. The speech signal plays the sole role of determining the IBM. Such a stimulus contains little speech-specific information. The spectral shape of speech is drastically reduced to a binary variation, and so is the temporal envelope. The harmonic structure of voiced speech is absent, and the temporal fine structure (the carrier signal underlying the temporal envelope) of the stimulus reflects that of noise, not speech. Despite this dramatic reduction of speech information, listeners are capable of speech perception.

So what cues enable listeners to perceive speech from IBM-gated noise? The binary pattern encodes a general outline of spectrotemporal energy variations of speech relative to noise. Binary-gated noise crudely reflects the formant structure of speech; as shown in Fig. 1, IBM-gated noise appears to "carve out" regions of noise energy that roughly match the spectrotemporal peaks of speech. Our results indicate that such a pattern of energy variations is sufficient for recognition purposes.

IV. DISCUSSION AND CONCLUSION

Our study bears resemblance to the well-known study by Shannon *et al.* (1995) demonstrating that only a few bands of noise modulated by the corresponding speech envelopes suffice for speech intelligibility [for a much earlier study using more bands see Dudley (1939)]. There are, however, several differences between our binary-gated noise and the vocoded noise of Shannon *et al.* Perhaps the most important and obvious difference is that, within a frequency channel, noise modulation uses a binary envelope in our study and a full envelope in vocoded noise. Second, the IBM is derived by a comparison between target speech and speech-shaped noise, while temporal envelopes in vocoded noise are obtained from target speech alone. We note that many speech separation algorithms compute a binary mask by implicitly or explicitly exploiting local SNR (Divenyi, 2005; Wang and Brown, 2006), making the ideal mask amenable to computational estimation. Third, the bandwidths of noise bands in Shannon *et al.* change as the number of the bands varies in order to cover the entire frequency range of interest; in IBM-gated noise, the bandwidth of each frequency channel is fixed to 1 ERB regardless of the number of filtbank channels. It is also worth mentioning that recognizing vocoded noise takes hours of training, while no training on binary-gated noise was given in our experiment.

Like vocoded noise, the type of noise used in binary gating likely has an effect on speech intelligibility. The speech-shaped noise used in this study is a steady noise with a long-term spectrum matching that of the utterances in the Dantale II corpus, and may be particularly effective for IBM gating, although our informal listening indicates that other types of steady noise, such as pink noise, can also produce intelligible speech. Our experiment was conducted using Danish utterances. Byrne *et al.* (1994) reported that the long-term average speech spectra of a group of languages, including Danish and English, are quite similar, suggesting that, though there are likely some language effects, the main observations of our experiment may hold for English and other languages. Also, the IBM used in this study is constructed when input SNR and LC are set to be equal ($-\infty$ dB). Fixing one of them while varying the other produces different IBMs. For example, when input SNR is set to 0 dB, increasing LC results in ideal masks with fewer and fewer 1's, whereas decreasing LC leads to more and more 1's. Is equating input SNR and LC most effective for intelligibility of IBM-gated noise? Further investigation is required to address the issues regarding noise type, language, and input SNR and LC levels.

That a pattern of binary gains is apparently sufficient for human speech recognition, like previous work on vocoded noise, raises intriguing questions on the nature of human speech recognition. What speech information is truly indispensable for intelligibility? Could the IBM itself be what is being recognized? Almost perfect speech recognition from such drastically reduced speech information has broad implications for CASA, automatic speech recognition, hearing prosthesis, and coding and compression in speech communication.

ACKNOWLEDGMENTS

The authors thank Associate Editor Ken Grant, and two anonymous reviewers for their constructive criticisms and suggestions. The work was performed while D.L.W. was a visiting scholar at Oticon A/S. The authors thank M. Schlaikjer, L. Bramsløw, and M. Hartvig for their assistance in the experiments, and Y. Li for his assistance in figure preparation. D.L.W. was supported in part by AFOSR Grant No. F49620-04-01-0027 and NSF Grant No. IIS-0534707.

Allen, J. B. (**2005**). *Articulation and Intelligibility* (Morgan & Claypool, San Rafael, CA).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Byrne, D., Dillion, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., and Ludvigsen, C. (**1994**). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108–2120.

Divenyi, P., ed. (**2005**). *Speech Separation by Humans and Machines* (Kluwer Academic, Norwell, MA).

Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.

Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Li, Y., and Wang, D. L. (**2008**). "On the optimality of ideal binary time-frequency masks," in *Proceedings of IEEE ICASSP*, pp. 3501–3504.

Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–16.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS final report, part B: Implementing a gammatone filterbank," Rep. No. 2341, MRC Applied Psychology Unit.

Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Wang, D. L., and Brown, G. J., eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ).

Yilmaz, O., and Rickard, S. (**2004**). "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**, 1830–1847.

Paper C

Role of Mask Pattern in Intelligibility of Ideal Binary-masked Noisy Speech

Ulrik Kjems, Jesper B. Boldt, Michael S. Pedersen,
Thomas Lunner, and DeLiang Wang

Journal of the Acoustical Society of America
Vol. 126, no. 3, pp. 1415–1426, September 2009.

©2009 Acoustical Society of America

Role of mask pattern in intelligibility of ideal binary-masked noisy speech

Ulrik Kjems,[a] Jesper B. Boldt, and Michael S. Pedersen
Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark

Thomas Lunner
Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark and Department of Clinical and Experimental Medicine, and Technical Audiology, Linköping University, S-58183 Linköping, Sweden

DeLiang Wang
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 11 September 2008; revised 9 May 2009; accepted 22 June 2009)

Intelligibility of ideal binary masked noisy speech was measured on a group of normal hearing individuals across mixture signal to noise ratio (SNR) levels, masker types, and local criteria for forming the binary mask. The binary mask is computed from time-frequency decompositions of target and masker signals using two different schemes: an ideal binary mask computed by thresholding the local SNR within time-frequency units and a target binary mask computed by comparing the local target energy against the long-term average speech spectrum. By depicting intelligibility scores as a function of the difference between mixture SNR and local SNR threshold, alignment of the performance curves is obtained for a large range of mixture SNR levels. Large intelligibility benefits are obtained for both sparse and dense binary masks. When an ideal mask is dense with many ones, the effect of changing mixture SNR level while fixing the mask is significant, whereas for more sparse masks the effect is small or insignificant.
© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3179673]

PACS number(s): 43.71.An, 43.71.Gv, 43.66.Ba, 43.66.Dc [MSS] Pages: 1415–1426

I. INTRODUCTION

The human ability to understand speech in a variety of adverse conditions is remarkable, and the underlying processes are not well understood. According to Bregman's auditory scene analysis account, the auditory system processes the acoustic input in two stages: an analysis and segmentation stage where the sound is decomposed into distinct time-frequency (T-F) segments followed by a grouping stage (Bregman, 1990; Wang and Brown, 2006). The grouping stage is divided into primitive grouping and schema driven grouping that represent bottom-up and top-down processes, respectively. Hence, in order to recognize speech in background noise, the auditory system would employ a combination of bottom-up processing of available cues, and top-down application of schemas, which represent learned patterns.

In this paper these processes are studied using the technique of ideal T-F segregation (ITFS), which was proposed by Brungart *et al.* (2006) to induce idealized grouping when listening to a mixture of target speech and noise. ITFS is based on the use of ideal binary mask (IBM), which was originally proposed as a benchmark for measuring the segregation performance of computational auditory scene analysis systems (Wang, 2005). The ITFS technique applies an IBM to the mixture, and several recent studies have utilized the technique for revealing important factors for speech intelligibility in noise (Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009).

A binary mask is defined in the T-F domain as a matrix of binary numbers. We refer to the basic elements of the T-F representation of a signal as T-F units. A frequency decomposition similar to the human ear can be achieved using a bank of gammatone filters (Patterson *et al.*, 1988), and signal energies are computed in time frames (Wang and Brown, 2006). The IBM is defined by comparing the signal-to-noise ratio within each T-F unit against a local criterion (LC) or threshold measured in units of decibels. Only the T-F units with local signal to noise ratio (SNR) exceeding LC are assigned 1 in the binary mask. Let $T(t,f)$ and $M(t,f)$ denote target and masker signal power measured in decibels, at time t and frequency f, respectively, the IBM is defined as

$$\mathrm{IBM}(t,f) = \begin{cases} 1 & \text{if } T(t,f) - M(t,f) > \mathrm{LC}, \\ 0 & \text{otherwise}. \end{cases} \quad (1)$$

An IBM segregated signal can be synthesized from the mixture by deriving a gain from the binary mask, and applying it to the mixture before recombination in a synthesis filter bank. However, not all studies follow the same procedure—sometimes the short-time Fourier transform is used (for instance Li and Loizou, 2008) which typically yields lower frequency resolution at low frequencies, but much higher resolution at high frequencies.

In Brungart *et al.*, 2006, the IBM was used as a means to retain the effect of energetic masking, thereby separating the

[a] Author to whom correspondence should be addressed. Electronic mail: uk@oticon.dk

energetic masking and informational masking effects. They argued that since the IBM removes those T-F units dominated by the masker, ITFS can be said to retain the effect of energetic masking, while removing informational masking caused by the excluded units with relatively significant masker energy. Informational masking refers to the inability to correctly segregate audible target information from the mixture. Their study showed a plateau of nearly perfect intelligibility of ITFS processed mixtures when varying the value of LC from −12 to 0 dB. Meanwhile, the IBM with 0 dB LC is considered to be the theoretically optimal mask out of all possible binary masks in terms of SNR gain (Li and Wang, 2009). Brungart *et al.* (2006) noted that lowering the mixture SNR by 1 dB while fixing LC causes the exact same T-F units to be left out as increasing the LC by 1 dB while fixing the mixture SNR; in other words, the IBM remains the same in these two scenarios. They demonstrated remarkably similar performance curves by altering the test conditions in the two ways described, which they interpret as rough equivalence in the effect of energetic masking.

Anzalone *et al.* (2006) showed large intelligibility benefits of IBM segregation and reported positive results on hearing impaired subjects, although their IBM definition is different from the previously outlined ITFS procedure. They computed the IBM by comparing the target signal to a fixed threshold adjusted to retain a certain percentage of the total target energy. Furthermore they attenuated the T-F units designated as non-target by 14 dB, in contrast to the total elimination described above. Their results showed more than 7 dB improvement in speech reception threshold (SRT) for normal hearing and more than 9 dB improvement for hearing impaired subjects.

In a study comparing impaired and normal-hearing subjects, Wang *et al.* (2009) demonstrated large improvements in SRT for both normal-hearing and hearing impaired groups due to ITFS processing of speech mixtures. Their study of the normal-hearing group shows an 11 dB improvement in SRT with a cafeteria noise masker containing conversational speech and an improvement of 7 dB for speech-shaped noise (SSN). For the hearing impaired group, the SRT improvement was 16 dB in cafeteria noise and 9 dB in SSN. As a surprising result, the SRTs obtained from the normal-hearing and hearing impaired groups on the ITFS processed mixtures were comparable.

Li and Loizou (2008) used short time Fourier transforms to apply ideal binary masking to mixtures with a two-talker masker, as well as modulated and unmodulated SSN maskers. They found large intelligibility benefits similar to Brungart *et al.* (2006) when varying the LC parameter, although they reported wider plateaus of LC values with almost perfect intelligibility (−20 to +5 dB compared to −12 to 0 dB in Brungart *et al.*, 2006), which they attributed to differences in speech material and filterbank setup. They further suggested that it may be the pattern of the binary mask itself that matters for intelligibility, rather than the local SNR of each T-F unit.

Wang *et al.* (2008) demonstrated that applying a binary pattern of gains obtained from an IBM with a SSN masker to the masker signal alone produces high intelligibility scores, a type of process related to noise vocoding (Dudley, 1939; Shannon *et al.*, 1995). Using different numbers of filterbank bands, they showed that intelligibility is lost when the number of channels is 8 or smaller, a result which differs from that reported by Shannon *et al.* (1995) who used continuous, rather than binary, values for envelope manipulation. There, high intelligibility was reported using noise vocoded in just four channels.

A. Motivation

The large benefits in intelligibility outlined previously could make the IBM a candidate for applications such as hearing aids, provided that the IBM can be approximated sufficiently well. In this paper we will not consider how such estimation might be done. However, to devise such applications it is important to understand the mechanisms by which the IBM enhances intelligibility. In the above described literature, much attention has been given to explaining intelligibility of IBM segregated mixtures by considering audibility of the target signal. By focusing on absolute regions of LC (Brungart *et al.*, 2006), emphasis is put on the interpretation that the IBM reduces informational masking by directing listeners' attention to the T-F units containing target information (Li and Loizou, 2008). This view is basically related to models of intelligibility based on target audibility in additive noise, such as the speech intelligibility index (ANSI, 1997), where intelligibility is described as a function of the proportion of target signal that is audible in different frequency bands. Cooke (2006) and Srinivasan and Wang (2008) proposed related computational models that operate on mixture input directly and produce recognition results from automatic speech recognition that are compatible with human intelligibility performance.

However, some of the previous published results seem inconsistent with this view. In particular, the observation of Wang *et al.* (2008) that IBM-processed noise is intelligible suggests that the resulting temporal envelope of the processed mixture is important. The speech transmission index (Houtgast and Steeneken, 1971) considers how distortions to the envelope affect speech intelligibility. Recent extensions have been made to improve the model predictions of nonlinearly processed speech (Goldsworthy and Greenberg, 2004). While the speech intelligibility index model cannot explain the noise gating results of Wang *et al.* (2008), a model based on speech transmission index described by Goldsworthy and Greenberg (2004) may perform better. This means that the target modulation carried by the IBM may play a key role in intelligibility of processed mixtures.

Based on the observation that the IBM is insensitive to the covariation of LC and mixture SNR, we propose to focus on the *difference* between the LC and the mixture SNR levels when comparing performance across mixture SNR levels. We therefore introduce a *relative criterion* (RC), defined as RC=LC−SNR in units of decibels.

By focusing on RC and varying the mixture SNR, it is possible to vary the effects of the target component of the IBM processed mixture relative to that of the masker. For example, by taking the mixture SNR to a large negative

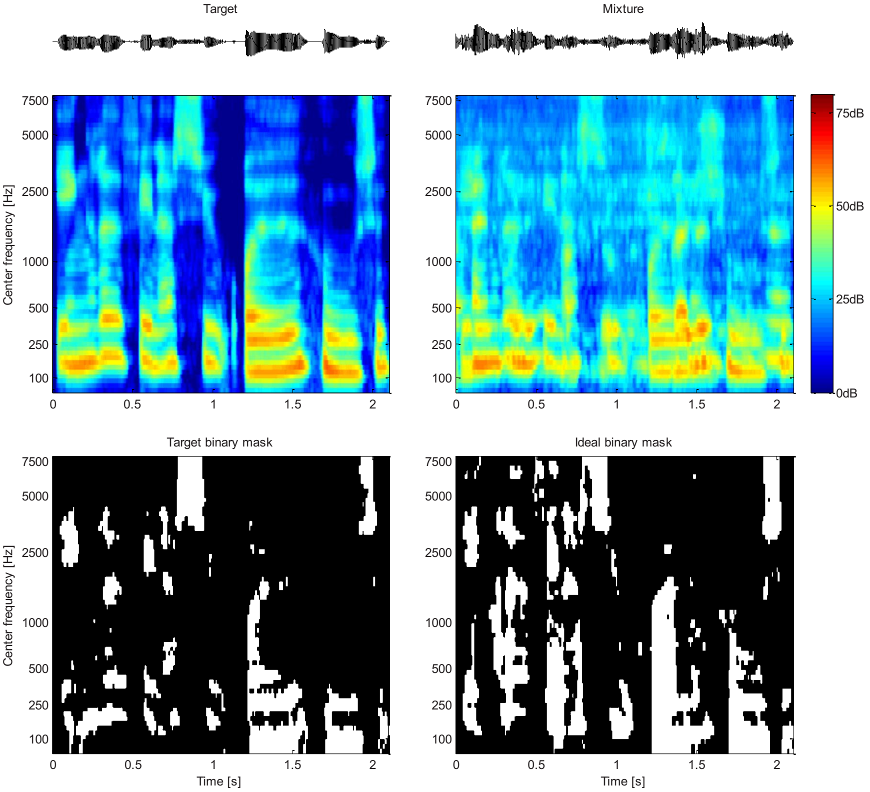

FIG. 1. (Color online) Illustration of IBM and TBM. Upper row shows waveform signal for a clean target sentence (left) and the sentence corrupted with cafeteria noise (right). Middle row shows the cochleagram representation of the two signals. Bottom left and right show the TBM and IBM, respectively, with white indicating the value of 1.

value, we can measure intelligibility of IBM-gated noise similar to Wang *et al.* (2008). On the other hand, by taking the mixture SNR to a level near the SRT we may measure how processing with the exact same binary mask affects intelligibility near the SRT.

B. Aims of the experiments

The change in focus from LC to RC brings up several research questions, which we will address in this paper. One aim is to investigate how the range of RC for optimal intelligibility depends on mixture SNR. Are there regions of RC where mixture SNR level has little or no effect on intelligibility? This question is directly addressed by the experiments in this paper. If under some circumstances mixture SNR level plays a minor role, the masker signal type may play a major role. So the second aim is to investigate the effects of masker type. We know that the plateau of optimal LC values is narrower for same-talker speech maskers (Brungart *et al.*, 2006) compared to a SSN masker. So far, intelligibility of IBM-processed noise has only been reported for stationary noise (Wang *et al.*, 2008).

Third, we wanted to compare the effects of alternative ideal mask definitions. The mask used by Anzalone *et al.* (2006) was computed based on the target signal alone; yet large intelligibility improvements were obtained. We(1) define the target binary mask (TBM) as the one obtained by comparing, in each T-F unit, the target energy to that of a SSN reference signal matching the long-term spectrum of the target speaker. This comparison still uses the LC parameter as a SNR threshold. The binary mask that results from this process can then be applied to a mixture of the target and a *different* masker. Figure 1 illustrates an example TBM and IBM computed from a target sentence in cafeteria noise, and shows the differences between the resulting masks. The top row shows the time domain waveforms of the clean and noisy target sentences. The middle row shows cochleagrams of the clean and noisy target sentence using a filterbank of 1 ERB (equivalent rectangular bandwidth) wide gammatone filters with center frequencies from 55 Hz to 7.7 kHz. The bottom row shows the TBM (left) and IBM (right). The two masks are noticeably different. The TBM pattern resembles the target sentence and is unaffected by the specific masker.

(1) In the published paper, "We" was erroneously corrected to "They". "We define" is the correct phrase.

On the other hand, the IBM pattern depends on the masker signal as well.

The TBM has several useful properties. The mask is, by this definition, identical to the IBM when SSN is the masker, so the TBM can be used as a measure of how general the IBM generated with the SSN masker is. Furthermore, relating to schema-based auditory scene analysis, the TBM could be interpreted as a simplified template of a learned pattern, indicating where in time and frequency to expect target energy. We therefore expect that the TBM leads to comparable benefits in intelligibility compared to the IBM. In some speech enhancement applications, it may be easier to estimate a TBM rather than an IBM, and it is therefore useful to know the extent to which the TBM results in intelligibility improvements.

The remainder of this paper is organized as follows. A listening experiment is described in Sec. II, and the results are reported and discussed in Sec. III. Section IV concludes the paper.

II. EXPERIMENTAL SETUP

A listening experiment was conducted to measure speech intelligibility of ITFS processed mixtures. The aim was to measure the influence of mixture SNR level, RC value, masker type, and to compare mask construction schemes: IBM and TBM.

A. Stimuli

The target phrases were from the Dantale II corpus (Wagener *et al.*, 2003) which is the Danish version of the Swedish Hagerman sentence (Hagerman, 1982) test and the German Oldenburg sentence test (Kollmeier and Wesselkamp, 1997). The corpus consists of 150 sentences designed to have low redundancy. The phrases were all spoken by the same female Danish speaker. The sentences were five words long following the same grammatical structure: name-verb-numeral-adjective-noun. An English translated example is "Michael had five new plants." Each word was randomly selected out of ten possibilities in each position of a sentence, taking coarticulation into account (Wagener *et al.*, 2003). Since long-term spectral characteristics are quite similar among different languages (Byrne *et al.*, 1994), the main observations of the present experiment could hold for English and other languages, though there are likely some language effects.

The target sentences were presented in nine second intervals, allowing the subjects time to repeat the words they recognize as well as guess. An operator recorded the number of correctly recognized words for each sentence.

Four masker signals were used: SSN, cafeteria noise, car interior noise, and noise from a bottling hall. We use the SSN included with the Dantale II corpus, which is produced by superimposing the speech material in the corpus. The cafeteria masker was a recording of an uninterrupted conversation between a male and a female Danish speaker in a cafeteria background (Vestergaard, 1998). The signal was equalized to match the long-term spectrum of the target sentences. This was done to isolate the effects of masker modulation and long-term average spectrum. The car interior noise was a recording during highway driving and was chosen to represent a quasi-stationary noise with strong low-frequency content. The fourth noise used was a recording of bottles rattling on a conveyor belt in a bottling hall (Vestergaard, 1998), and was chosen to represent a signal with strong high-frequency content. All stimuli were diotically presented through headphones.

TABLE I. Seven combinations of masker type and mask type. Note that TBM and IBM with SSN masker are identical.

	Speech shaped noise	Cafeteria noise	Car interior	Bottling noise
IBM	1	2	3	4
TBM		5	6	7

For each masker type, three mixture SNR levels were selected along with eight values of RC. Given that the IBM and the TBM are identical with the SSN masker, there were seven combinations of masker type and mask type, as shown in Table I.

Mixture SNR levels were set to match measured 20% and 50% SRTs for each masker type. The third SNR level was fixed at −60 dB to create IBM-gated noise similar to Wang *et al.* (2008).

B. Sessions

The experiment was divided into two sessions. In Session I, the slope and SRT of each subject's psychometric curve of the unprocessed mixtures and each of the four maskers were measured using the adaptive Dantale II procedure, and the mixture SNR levels for 20% and 80% correct word identification were derived (Brand and Kollmeier, 2002; Wagener *et al.*, 2003). In Session II, intelligibility was measured on a grid of three different mixture SNR levels and eight different RC values (including an "unprocessed" condition, see later) for each of the seven conditions in Table I. This generated a total of 3 SNR levels, 8 RC values, and 7 conditions of Table I, resulting in $3 \times 8 \times 7 = 168$ points, where intelligibility was measured. Each combination was tested on each subject using two sentences. Hence, each subject listened to a total of $2 \times 168 = 336$ sentences, which required reuse of sentences. To prevent memorization, order of the sentences was balanced as much as possible within and across subjects, and appeared random to the subjects.

From Session I measurements, logistic functions

$$P(\mathrm{SNR}) = (1 + \exp(4s_{50}(L_{50} - \mathrm{SNR})))^{-1} \qquad (2)$$

were fitted by means of the maximum likelihood method, assuming a binomial distribution of individual sentence scores (Brand and Kollmeier, 2002) yielding the 50% SRT (L_{50}) and slope (s_{50}) parameters for each subject and each masker type. The two initial sentences of each adaptation were discarded, and to reduce the effects of outliers, the data from the three best and three worst performing subjects were left out before averaging in order to derive the 20% and 50% SRT values. Pilot experiments revealed an effect of a princi-

TABLE II. SRT at 50% correct L_{50} and slope s_{50} parameters of the logistic function, Eq. (2), estimated from Session I measurements, using maximum likelihood with correction for gated noise (see text, Sec. II B). The next column shows the derived 20% SRT for average subject performance. The last two columns show the upper and lower RC values for the four masker types. Offline simulations were used to determine the RC values for obtaining IBM sparseness of 1.5% and 80% ones in the mask. The three TBM conditions 5–7 of Table I all used RC values corresponding to IBM/SSN with mixture SNR corresponding to masker type.

Masker type	50% SRT mixture SNR (L_{50}) (dB)	Slope at SRT (s_{50}) (%/dB)	20% SRT mixture SNR (dB)	RC for 1.5% ones in mask (dB)	RC for 80% ones in mask (dB)
Speech shaped noise	−7.3	15.1	−9.8	12.7	−30.3
Cafeteria	−8.8	7.5	−13.8	24.6	−27.4
Car interior	−20.3	12.7	−23.0	27.5	−25.2
Bottling noise	−12.2	5.7	−18.4	23.1	−34.9

pal difference between the continuous masker used in Session I, and the binary gated masker used with the ITFS signals in Session II. The effect caused a slightly decreased performance in the latter case. This effect has previously been described by Wagener (2003, Chap. 5) where a comparison of continuous versus gated noise indicated a 1.4 dB increase in SRT (L_{50}) and a decrease in slope (s_{50}) from 21%/dB to 18%/dB. Accordingly, we adjusted the measured SRT from Session I by adding 1.4 dB and slope by multiplying 18/21, resulting in the values listed in the first two columns of Table II. The third column shows 20% SRT derived from the adjusted parameters. The measured SRTs and slopes for speech-shaped and cafeteria noise all agree with previous results on the same material (Wagener, 2003; Wang *et al.*, 2009).

In order to determine the range of RC values to use, offline simulations were carried out to identify the RC values that yielded mask densities of 1.5% and 80% measured as percent ones in the mask within speech intervals (see Sec. II D for signal processing details). For each masker type seven RC values were then identified by equidistant sampling (in decibels) between these two points. For the three TBM conditions, the set of RC values equaled the set for IBM/SSN (condition 1 in Table I) since the binary masks are identical by definition. An eighth additional unprocessed condition was added, where the mask was set to 1 in all frequency bands within the speech intervals, and 0 outside these intervals, creating essentially a gated masker.

Speech intervals were derived from the target sentences alone and were used for all mixture SNR computations by averaging target and masker energy within speech intervals only. A speech interval was defined by low-pass filtering the absolute target sample values using a first-order IIR low-pass filter with the time constant of 1 ms (for 20 kHz sample rate the transfer function was $H(z)=\lambda/(1-(1-\lambda)z^{-1})$, $\lambda=0.04877$), thresholding the result at 60 dB below the maximum value, and further designating all non-speech intervals less than 2 s as speech to include inter-word intervals in all sentences. All detected speech onsets were shifted 100 ms backward to account for forward masking effects (Wang *et al.*, 2009).

C. Subjects

A total of 15 normal-hearing, native Danish speaking subjects participated in the experiment. The subjects volunteered for the experiment and were not paid for their participation. Their age ranged from 25 to 52 with a mean age of 35. The audiograms of all subjects indicated normal hearing with hearing thresholds below 20 dB HL in the measured range of 250 Hz–8 kHz.

D. Signal processing

All target and masker signals were resampled from 44.1 to 20 kHz sampling rate. Gain factors for target and masker were computed in order to achieve a given mixture SNR and fixed mixture power. This was done by computing the signal energies of target and masker within the speech intervals previously defined. The target and masker signals were processed separately by means of a gammatone filterbank, consisting of 64 channels of 2048-tap FIR filters; each channel has the bandwidth of 1 ERB and channel center frequencies range from 2 to 33 ERBs (corresponding to 55–7743 Hz) linearly distributed on the ERB-rate scale (Patterson *et al.*, 1988; see also Wang and Brown, 2006). The filterbank response was divided into 20 ms frames with 10 ms overlap, and the total signal energy was computed within each T-F unit.

For IBM processing, a binary mask was formed by comparing the local SNR within a T-F unit against LC, assigning 1 if the local SNR was greater than LC and 0 otherwise. For TBM processing, the reference masker (i.e. the SSN masker) was processed through the filterbank, with a gain set to achieve a 0 dB mixture SNR. The TBM was formed by comparing the local SNR within a T-F unit using the reference masker against the RC threshold, assigning 1 if the local SNR was greater than RC.

The binary mask signal was then upsampled to the full 20 kHz sampling rate by means of a sample-hold scheme followed by low-pass filtering using a 10 ms Hanning filter. In each band, the target-masker mixture was delayed 20 ms in time, accounting for the total delay from the T-F unit energy summation, sample-hold, and low-pass filtering, before the upsampled mask was multiplied with the mixture. Finally, the ITFS processed waveform was synthesized using time reversed gammatone filters.

The target and masker stimuli for Session I were processed through the filterbank analysis and synthesis procedure (no binary mask was applied), reducing the signal bandwidth to 55 Hz–7.74 kHz in order to match processed signals in Session II.

E. Procedure

1. Session I: SRT and slope measurements

The first session consisted of an adaptive Dantale procedure for each of the four masker types. Prior to this the subjects were given a short training session consisting of 30 randomly chosen sentences using speech-shaped and cafeteria noise maskers. These maskers were chosen to let listeners familiarize themselves with the task under stationary and non-stationary noise conditions.

In the adaptive Dantale procedure, the mixture SNR was varied after each sentence according to the number of correctly identified words, and the 20% and 80% SRTs were tracked in an interleaved manner (Brand and Kollmeier, 2002). The 20% and 80% points were chosen since they were proposed by Brand and Kollmeier (2002) to be optimal for the simultaneous measurement of the logistic function parameters L_{50} and s_{50} of Eq. (2). A total of 30 sentences were presented for each masker type in the adaptive procedure. To account for learning effects, the order of masker types was balanced across subjects (Beck and Zacharov, 2006).

2. Session II: ITFS mixtures

In the second session, each subject listened to 336 offline computed ITFS sentences. The stimuli alone lasted approximately 51 min so the subjects were allowed two breaks in the middle.

Prior to the main experiment, subjects were exposed to 60 sentences of training using all four noise types. First, for each masker type ten sentences corresponded to the unprocessed condition with increasingly lower mixture SNRs. The remaining 20 training sentences corresponded to various ITFS conditions, randomly selected but increasing difficulty. We found from pilot experiments that an extended training procedure was required to reduce learning effects and subject variability.

Learning and other temporal effects were accounted for by using a balanced design: for each subject the ordering of the seven conditions was changed and for each condition the ordering of SNR levels and RC values were balanced as much as possible.

Subjects were seated in a sound treated room where sounds were presented using Sennheiser HD280 Pro headphones connected to a SoundBlaster SB0300 sound card, using a PC running MATLAB.

3. Level of presentation

All mixtures were normalized to have same broadband long-term signal power before ITFS processing, both across mixture SNR and across noise types. The SSN condition was used to calibrate the presentation level to 65 dB(A) sound pressure level, and the volume control settings were then held fixed. The calibration was done using a sound level meter coupled to an earpiece of the headphones. The resulting presentation levels were measured to 62 dB(A) for cafeteria noise, 60 dB(A) for car interior noise, and 68 dB(A) for bottling hall noise.

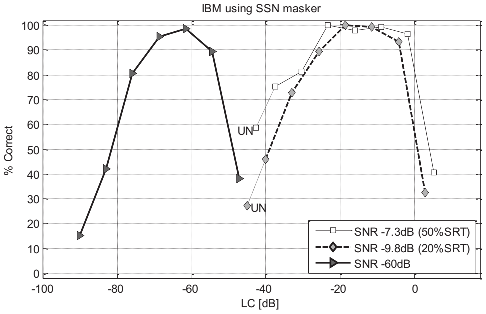

FIG. 2. Percentage of correctly identified words for IBM-processed mixtures with SSN masker as function of LC used for generating the IBM. Three mixture SNR levels are shown. The unprocessed conditions do not correspond to a particular LC value, but are inserted to the left of the respective curves, marked as "UN" and connected with dotted lines. Chance performance level is 10%.

III. RESULTS AND DISCUSSION

Figure 2 shows the percentage of correctly identified words as a function of LC for IBM segregated mixtures with the SSN masker in the three mixture SNR settings, averaged over all subjects. The unprocessed conditions do not correspond to a particular LC value, and are inserted as the leftmost points of the respective curves (marked as "UN") and connected with dotted lines to the curves.

The unprocessed data points resulted in higher performance than expected; across conditions the average scores are 25.7% and 59.5% out of 600 answers, which are larger than the 20% and 50% expected scores. This could be explained by the training that was encountered during Session I and during the training session introduced between Session I and Session II, as described in Sec. II E 2.

Each of the three curves shows a plateau or peak of very high intelligibility; for the 50% SRT (SNR of −7.3 dB), the interpolated average performance was above 95% in the interval −25 dB < LC < −2 dB, a 23 dB wide region. For 20% SRT (SNR of −9.8 dB) the interval was −22 dB < LC < −6 dB and 16 dB wide, while for the −60 dB case the interval was −69 dB < LC < −59 dB and 10 dB wide. The results for 20% and 50% SRT have similar profiles as those reported by Brungart *et al.* (2006) and Li and Loizou (2008). In Brungart *et al.*, 2006, the range is −12 dB < LC < 0 dB using a multi-talker task and similar ITFS processing. The plateaus in the present study are wider than those of Brungart *et al.* (2006), due to higher scores at lower LC values, while plateau upper bounds are similar. Li and Loizou (2008) reported plateaus from −20 to +5 dB at −5 dB SNR and −20 to 0 dB at −10 dB SNR using a sentence test with a SSN masker and a T-F representation with linear frequency. The observed differences are probably due to differences in sentence material and mixture SNR.

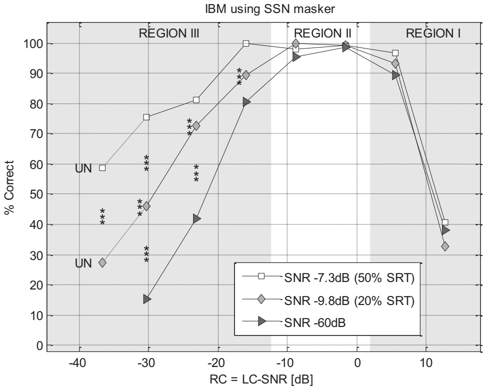

FIG. 3. Percentage of correctly identified words for IBM-processed mixtures with SSN masker as function of RC=LC−SNR. This figure gives a different plot of the same data in Fig. 3. Asterisks (*) indicate significant differences between intelligibility scores at adjacent mixture SNR levels, or when placed to the left of a diamond, between the scores at the lowest and highest mixture SNR levels, according to a Tukey HSD test. In the figure, * corresponds to $p<0.05$, ** to $p<0.01$, and *** to $p<0.001$.

The −60 dB SNR curve, however, is different. First of all, since the mask here was applied to essentially pure noise, this is consistent with the results of Wang *et al.* (2008) who demonstrated that listeners achieve nearly perfect recognition from IBM-gated noise where the mask is obtained from speech and SSN. This process of producing intelligible speech from noise may be viewed as a form of noise gating. Our results extend their findings by showing that the vocoding ability of the IBM applies to a range of LC values. This range is not much smaller than those of the performance plateaus at much higher mixture SNR levels, a finding that has not previously been reported.

Secondly, the shape of the −60 dB curve is similar to but narrower than the curves at higher SNR levels, but its position on the LC axis is very much shifted. As pointed out by Brungart *et al.*, (2006), the IBM is insensitive to covariations of LC and mixture SNR. This means that the mask pattern is a function of the difference LC-SNR, which was termed RC in Sec. I A.

A. Performance versus RC

Depicting the performance curves versus RC rather than LC brings the curves together, as shown in Fig. 3. Most notably the decline in performance at high RC values seems to be aligned well. Recall that the IBMs for the three SNR levels are equal for a fixed RC regardless of mixture SNR.

A two-way analysis of variance (ANOVA) with repeated measures was performed on the rationalized arcsine transformed subject mean percentage scores (Studebaker, 1985). The ANOVA revealed significant effect of mixture SNR, RC, and of interaction terms, as indicated in Table III. To further investigate the interaction effect, a *post hoc* Tukey HSD test was performed comparing all pairwise differences across SNR. In Fig. 3, asterisks are used to indicate significant pairwise differences, where the significance level is indicated by their number: * indicates $p<0.05$, ** indicates $p<0.01$, and *** indicates $p<0.001$. In this case, all pairwise comparisons that were significant were at the level of $p<0.001$. The significance of the difference between the upper and lower SNR performance is indicated to the left of the corresponding data point of the middle SNR curve (diamond).

In Fig. 4, plots similar to Fig. 3 are shown for the remaining conditions tested. The two rows of the plots show IBM and TBM processing, respectively. The three columns correspond to the three remaining masker types: cafeteria, car interior, and bottle noise. As shown in Table III, a two-way ANOVA in all conditions revealed significant effects of mixture SNR, RC, and of interaction terms.

The results in Fig. 4 show patterns similar to that of Fig. 3. Tukey HSD tests revealed significant differences across mixture SNR for low RC values just as was the case for the IBM/SSN condition.

1. Interpretation using regions in RC

In a manner similar to Brungart *et al.* (2006) we divide the performance curves into three distinct regions. The main difference in our analysis is that our regions are defined in terms of RC instead of LC. The purpose is to interpret the intelligibility improvement in terms of RC (Fig. 3), instead of LC (Fig. 2). While the aim of the analysis by Brungart *et al.* (2006) was to separate effects of informational and energetic masking, our analysis highlights the importance of the binary mask pattern.

TABLE III. Two way ANOVA test results using rationalized arcsine transformed mean subject scores (Studebaker, 1985) revealed significance of effects of mixture SNR, RC, and interaction terms for the measurement data shown in Figs. 3 and 4.

	Effect of mixture SNR	Effect of RC	Effect of interaction
Test statistic	$F(2,28)$	$F(7,98)$	$F(14,196)$
IBM/SSN (Fig. 3)	136.1, $p<0.000\,01$	153.1, $p<0.000\,01$	13.8, $p<0.000\,01$
IBM/cafeteria	340.5, $p<0.000\,01$	149.7, $p<0.000\,01$	17.4, $p<0.000\,01$
IBM/car noise	172.4, $p<0.000\,01$	295.5, $p<0.000\,01$	12.0, $p<0.000\,01$
IBM/bottling noise	173.0, $p<0.000\,01$	126.0, $p<0.000\,01$	12.2, $p<0.000\,01$
TBM/cafeteria	253.1, $p<0.000\,01$	95.1, $p<0.000\,01$	11.8, $p<0.000\,01$
TBM/car noise	133.1, $p<0.000\,01$	156.8, $p<0.000\,01$	12.3, $p<0.000\,01$
TBM/bottling noise	234.3, $p<0.000\,01$	146.7, $p<0.000\,01$	15.2, $p<0.000\,01$

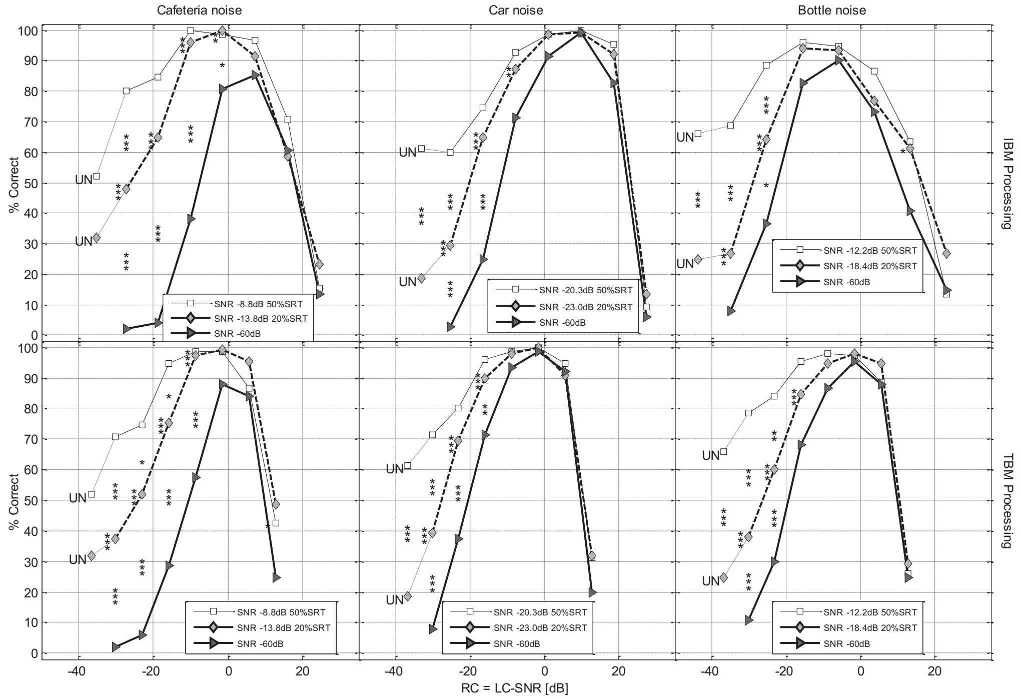

FIG. 4. Percentage of correctly identified words versus RC for IBM-processed mixtures (upper row) and TBM processed mixtures (lower row). Each column corresponds to a masker type. The three curves in each plot correspond to the mixture SNR levels (squares: 50% SRT, diamonds: 20% SRT, and triangles: −60 dB mixture SNR). Asterisks (*) indicate significant difference between adjacent mixture SNR levels, or when placed to the left of a diamond, between the lowest and highest mixture SNR levels (* corresponds to $p<0.05$, ** to $p<0.01$, and *** to $p<0.001$), according to Tukey HSD tests.

Region I corresponds to large RC values, where intelligibility decreases with increasing RC due to increasing sparseness of the ideal mask. In our results from the IBM/SSN condition, performance decreased for RC>−2 dB.

Region II corresponds to an intermediate range of RC values, with nearly perfect performance. For the IBM/SSN condition this occurred as a plateau at RC values between −8.8 and −1.6 dB, where intelligibility was above 95%.

Region III ranges below approximately RC=−10 dB in the IBM/SSN case. In this region performance decreases as RC decreases and the number of T-F units included in the IBM increases, until the performance of the unprocessed mixture is reached.

A general pattern in our data is that the influence of mixture SNR on the recognition performance decreases with increasing RC: In Regions I and II the effect was small or insignificant, while in Region III there was significant influence.

The fact that the performance in Region I (high RC values) showed only a negligible or small effect of mixture SNR level suggests that the target component of the processed mixture plays a relatively small role. Our results seem to indicate that some of the traditional cues for speech perception, such as F0, periodicity, and other temporal fine structure cues, are less important in Region II than in Region III and of even smaller importance in Region I. Otherwise one would have expected a difference in performance across mixture SNRs. So the application of the IBM seems, on the one hand, to improve the intelligibility relative to the unprocessed condition and, on the other hand, to reduce or eliminate the listener's ability to make use of speech cues other than what is carried in the binary mask. This result is of particular interest for the design of hearing aids, since reports suggest that the ability of hearing impaired subjects to make use of temporal fine structure cues is limited compared to normal listeners (Lorenzi *et al.* 2006; Hopkins *et al.* 2008), making the trade-off more favorable for the hearing impaired.

In Region III, there was an overall significant effect of mixture SNR (indicated with asterisks in Figs. 3 and 4). We further note that across all seven mask scheme/masker conditions, the increase in performance at the mixture SNR corresponding to 20% SRT from Region III to Region II is accompanied by an increasing vocoding ability at −60 dB mixture SNR.

2. Influence of masker type

The results in Fig. 4 show that the RC values beneficial to intelligibility varied across the seven mask scheme/masker

TABLE IV. Measured peak intelligibility score (in percentage) for noise gating data (at a mixture SNR of −60 dB) together with average width (in RC) of performance plateau where the interpolated performance was within 95% of the peak value, for the four masker types and two mask computation schemes.

	Speech shaped noise		Cafeteria		Car interior		Bottling noise	
IBM	98.7%	23.6 dB	85.3%	20.7 dB	99.3%	23.0 dB	90.0%	19.0 dB
TBM			88.0%	16.9 dB	98.7%	21.5 dB	95.3%	18.4 dB

conditions. While the plateau became narrower at lower mixture SNR levels, its position shifts across the seven conditions tested. As already described, mixture SNR, which factors in the definition of RC, is not a good indicator of intelligibility across masker types. For instance, in the IBM/bottle noise curve at the mixture SNR corresponding to 50% SRT, the performance plateau—the region of RC values where intelligibility is within 95% of the maximum score—ranged from −22 to −3 dB (measured on interpolated mean data), while in the IBM/car noise curve the corresponding plateau occurs in the RC range of −4 to 19 dB.

Table IV shows the average plateau width for the three mixture SNR levels for each of the seven mask scheme/masker conditions. The IBM/SSN condition produced the broadest plateau, 23.6 dB on average, and the TBM/cafeteria the narrowest plateau of 16.9 dB. Comparing mask schemes within masker signals, the IBM showed slightly wider average plateaus for all masker types. The table also gives the peak intelligibility scores of various noise gating curves.

B. Discussion of binary noise gating results

The noise gating performance curves (SNR −60 dB) form a performance lower bound for each masker type: in no case was the noise gating performance significantly greater than that for any other mixture SNR level. The measured peak value of the noise gating performance curves varied across masker type and mask computation scheme as indicated in Table IV. The effect of masker type was greater than the effect of mask computation scheme (from 85.3% for IBM/cafeteria to 99.3% for IBM/car noise).

The cafeteria noise was a relatively poor signal for vocoding, yielding maximum scores of 85% correct using IBM and 88% using TBMs, a result which may be explained by the sparse energy distribution in retained T-F units: The presence of 1 in the binary mask may coincide with a dip in the noise signal. In our data, the performance in the TBM/cafeteria condition with the −60 dB SNR was significantly lower at RC=15 dB than those with higher SNR levels. The modulation dips of the cafeteria masker made the distribution of T-F energy in the processed signal relatively sparse, a likely reason for reduced intelligibility performance.

Figure 5 shows the density of the binary mask measured as percentage ones in the mask averaged over all speech intervals (see Sec. II B) as function of channel center frequency for different masker types. The bold lines correspond to the RC value with the highest noise gating intelligibility (at mixture SNR of −60 dB). The figure shows that when the target and masker signal spectra are matched (speech-shaped and cafeteria noise) the result is a more uniform mask density compared to when the signals are not matched (bottle noise and car noise).

It should be noted that, for stationary maskers, the TBM is similar to the IBM with a LC parameter made frequency dependent in such a way that the resulting distribution of mask sparseness resembles that of the TBM (i.e. IBM with SSN masker). Since the TBM in the bottle noise case brings some intelligibility benefits over the IBM, it is possible that speech separation algorithms that estimate the IBM would also benefit from making the LC parameter frequency dependent, to ensure that enough ones are present in frequency bands relevant for speech.

C. Results from TBM

In Fig. 6, the results of applying the TBM to mixtures of the four masker types are compared. From left to right the mixture SNR level corresponds to 50% SRT, 20% SRT, and −60 dB. The curves corresponding to the four different maskers appear to align well. This is further reflected in Table V, showing the results from a two-way ANOVA with repeated measures performed on the rationalized arcsine mean subject scores, for each of the three mixture SNR levels. Compared to the previous analysis, the effects are not as strong; in fact, the noise type influence was not above the standard 5% significance level for the 20% SRT data and the interaction term for the 50% SRT data was also not significant. Tukey HSD tests revealed significance in the pairwise differences across masker type only for cafeteria noise in −60 dB SNR against all three other noise types, and only for RC values of −23.1, −15.9, and −8.7 dB as indicated with asterisks in Fig. 6.

D. Performance versus mask density

Given the importance of mask density for resulting intelligibility, the performance scores versus resulting overall mask density are plotted in Fig. 7. The mask density was measured as resulting percentage of ones in all frequency bands within speech intervals. The unprocessed condition is indicated as having 100% ones in the mask. The IBM results are connected with solid lines, and the TBM results are connected with dashed lines. Note that a nonlinear abscissa is used to better illustrate the performance differences at low percentages.

All curves show maximum performance between 15% and 60% ones in the masks. The curves all show a sharp decline toward zero at low percentages, a plateau in the middle which is wider for higher mixture SNRs and a gradual drop to the level of unprocessed mixtures, from 40%

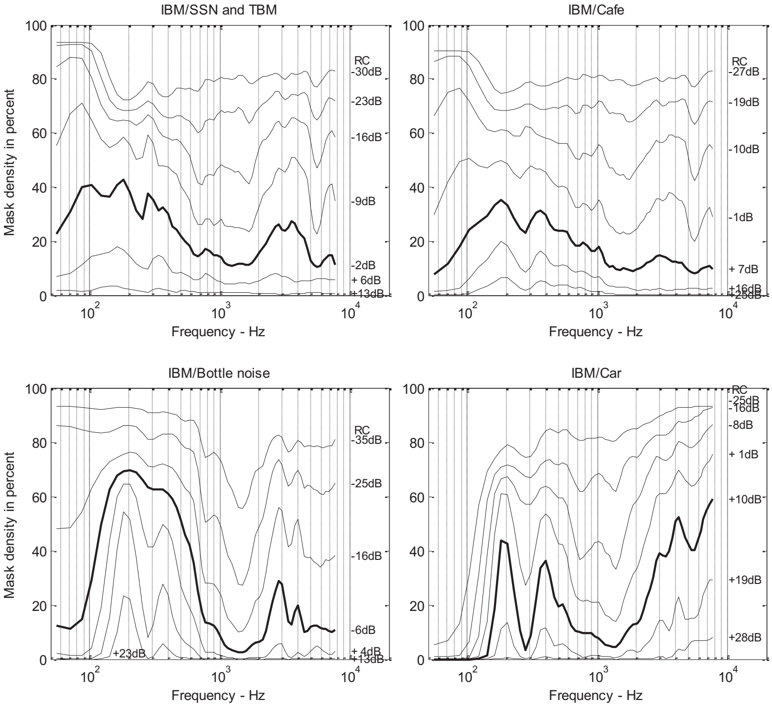

FIG. 5. Mask density (in percentage of mask value 1) as function of channel center frequency averaged over entire sentence material. The corresponding RC value used for computing the mask is indicated to the right of each curve. The bold line corresponds to the RC value with the highest intelligibility for IBM-gated noise (at mixture SNR of −60 dB). The mask densities of the TBM masks equals that of the IBM/SSN by definition.

to 100% ones in mask. The TBM and IBM curves are generally similar, with slightly larger scores for the target binary mask except for the cafeteria masker at high percentage of ones. Below 5%–10% ones, the TBM scores were higher than for the IBM for all masker types. For the exceptional case of the cafeteria noise, the IBM strategy based on mixture SNR was apparently better than the TBM scheme according to the target energy. Overall, it is rather remarkable how well the TBM and IBM results are aligned, considering their differences with respect to RC in Fig. 4.

IV. CONCLUSION

By measuring intelligibility of ideal binary-masked noisy speech, we have shown that intelligibility performance

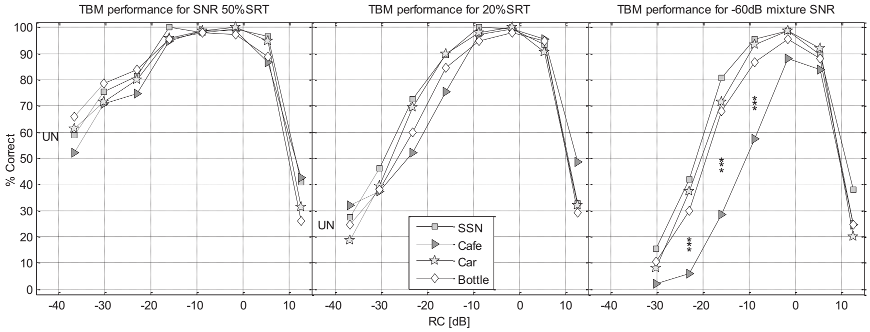

FIG. 6. Percentage of correctly identified words versus RC for TBM processed mixtures comparing the effect of noise types. Note that all curves use the same mask for a given RC. The three plots correspond to the three mixture SNR levels. The individual curves correspond to masker types. Asterisks (*) indicate significant difference between adjacent noise types (* corresponds to $p<0.05$, ** to $p<0.01$, and *** to $p<0.001$), according to a Tukey HSD test.

TABLE V. Two way ANOVA test was performed on rationalized arcsine transformed mean subject scores revealing significance of effects of noise type, RC, and interaction terms for the measurement data shown in Fig. 6.

	Effect of noise type	Effect of RC	Effect of interaction
Test statistic	$F(3,42)$	$F(7,98)$	$F(21,294)$
50% SRT data	3.80, $p<0.017$	92.3, $p<0.000\,01$	1.54, $p<0.063$
20% SRT data	2.78, $p<0.053$	147.4, $p<0.000\,01$	2.25, $p<0.001\,7$
−60 dB SNR data	87.9, $p<0.000\,01$	297.1, $p<0.000\,01$	6.19, $p<0.000\,01$

curves became aligned across a large range of mixture SNR levels when using the RC defined as the difference of LC and SNR. This alignment was demonstrated for four masker types, using the IBM as well as the proposed TBM. By fixing RC and varying the mixture SNR level, we identified three regions in RC, differentiated by intelligibility and influence of the mixture SNR level. In Regions I and II, weak or insignificant influence was found, whereas in Region III the influence was large and significant. The size and location of the regions varied with masker type.

By applying IBM processing to mixtures of low negative SNR levels, we have extended the findings of Wang *et al.* (2008) showing that the processing acts as binary noise gating and produces intelligible speech at a range of sparseness configurations parametrized by RC. We further showed that the proposed TBM based on the target signal alone was comparable to the IBM in terms of intelligibility improvements. For a given level of mask sparseness, the mean measured TBM intelligibility scores were even slightly higher than those of the IBM in some conditions.

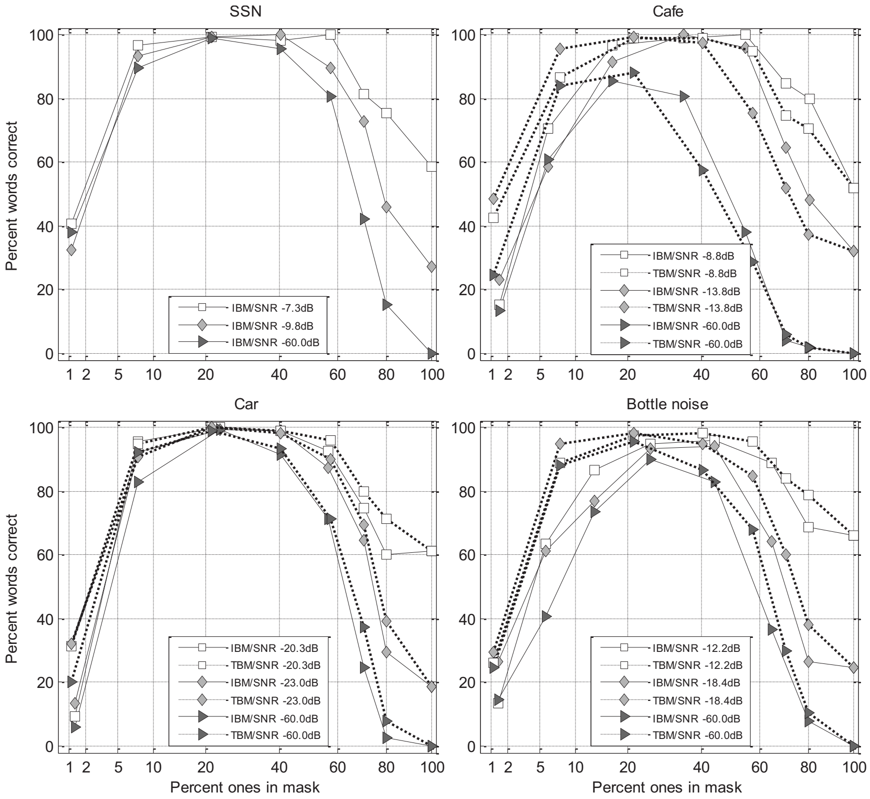

FIG. 7. Percentage of correctly identified words as function of mask density. The four plots show the four masker types: SSN, cafeteria, car noise, and bottle noise. Each plot corresponds and the three mixture SNR levels and the two mask computation schemes. The IBM results are connected with solid lines, and the TBM results with dotted lines. The unprocessed condition is marked as 100% ones in mask.

ACKNOWLEDGMENTS

The authors would like to thank Tayyib Arshad for assistance in performing the experiments, volunteering subjects for participating, and colleagues for discussions and proofreading. The research of D.W. was supported in part by an AFOSR grant (FA9550-08-1-0155) and a NSF grant (IIS-0534707).

ANSI S3.5-1997 (**1997**). "American National Standard: Methods for the calculation of the speech intelligibility index" (American National Standards Institute, New York).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Beck, S., and Zacharov, N. (**2006**). *Perceptual Audio Evaluation: Theory, Method and Application* (Wiley, Chichester, UK).

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge MA).

Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Byrne, D., Dillon, H., and Tran, K. (**1994**). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108–2120.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.

Goldsworthy, R. L., and Greenberg, J. E. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2008**). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," J. Acoust. Soc. Am. **123**, 1140–1153.

Houtgast, T., and Steeneken, H. J. M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367.

Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am. **102**, 2412–2421.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "An efficient auditory filterbank based on the gammatone function," Report No. 2341, MRC Applied Psychology Unit, Cambridge.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Srinivasan, S., and Wang, D. L. (**2008**). "A model for multitalker speech perception," J. Acoust. Soc. Am. **124**, 3213–3224.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Vestergaard, M. (**1998**). "The Eriksholm CD 01: Speech signals in various acoustical environments," Report No. 050-08-01, Oticon Research Centre Eriksholm, Snekkersten.

Wagener, K. (**2003**). "Factors Influencing Sentence Intelligibility in Noise," Ph.D. thesis, Oldenburg University, Oldenburg, Germany.

Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken NJ).

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**, 2303–2307.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.

Paper D

Estimation of the Ideal Binary Mask using Directional Systems

Jesper B. Boldt, Ulrik Kjems, Michael S. Pedersen,
Thomas Lunner, and DeLiang Wang

*Proceedings of the 11th International Workshop on
Acoustic Echo and Noise Control 2008.*

ESTIMATION OF THE IDEAL BINARY MASK USING DIRECTIONAL SYSTEMS

Jesper Bünsow Boldt[1,2], *Ulrik Kjems*[2], *Michael Syskind Pedersen*[2], *Thomas Lunner*[3], *DeLiang Wang*[4]

[1]Department of Electronic Systems, Aalborg University, DK-9220 Aalborg Øst, Denmark
[2]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[3]Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark
[4]Department of Computer Science and Engineering & Center for Cognitive Science,
The Ohio State University, Columbus, OH 43210-1277, USA
email: {jeb, uk, msp, tlu}@oticon.dk, dwang@cse.ohio-state.edu

ABSTRACT

The ideal binary mask is often seen as a goal for time-frequency masking algorithms trying to increase speech intelligibility, but the required availability of the unmixed signals makes it difficult to calculate the ideal binary mask in any real-life applications. In this paper we derive the theory and the requirements to enable calculations of the ideal binary mask using a directional system without the availability of the unmixed signals. The proposed method has a low complexity and is verified using computer simulation in both ideal and non-ideal setups showing promising results.

Index Terms— Time-Frequency Masking, Directional systems, Ideal Binary Mask, Speech Intelligibility, Sound separation

1. INTRODUCTION

Time-frequency masking is a widely used technique for speech and signal processing used in automatic speech recognition [1], computational auditory scene analysis [2], noise reduction [3, 4], and source separation [5, 6, 7, 8]. The technique is based on time-frequency (T-F) representation of signals and makes it possible to utilize the temporal and spectral properties of speech and the assumption of sparseness of speech. An important quality of T-F masking is the availability of a reference mask, which defines the maximum obtainable speech intelligibility for a given mixture. This *ideal binary mask* (IBM) [9] has recently been demonstrated to have large potential for improving speech intelligibility in difficult listening conditions [10, 4, 3]. To calculate the IBM, the unmixed signals must be available, which is a requirement rarely met in any real-life application. However, the significant increase in speech intelligibility by the IBM makes it a valuable goal for T-F algorithms trying to increase speech intelligibility. The T-F representation is obtained using e.g. the short-time Fourier transform or a Gammatone filterbank [11], and the IBM is calculated by comparing the power of the target signal to the power of the masker (interfering) signal for each unit in the T-F representations:

$$\mathrm{IBM}(\tau,k)=\begin{cases}1, \text{ if } \dfrac{\mathbf{T}(\tau,k)}{\mathbf{M}(\tau,k)}>\mathrm{LC}\\ 0, \text{ otherwise}\end{cases}, \tag{1}$$

where $\mathbf{T}(\tau,k)$ is the power of the target signal, $\mathbf{M}(\tau,k)$ is the power of the masker signal, LC is a local SNR criterion, τ the time index, and k the frequency index. The LC value is the threshold for classifying the T-F unit as target or masker and determines the amount of target and masker signal in the processed signal, if the binary mask is applied to the mixture. In computational auditory scene analysis (CASA), an LC value of 0 dB is commonly used, but recent studies have shown that a certain range of LC values different from zero provides the same major improvement in speech intelligibility [10, 3].

In this paper we show that it is indeed possible to calculate the IBM without the availability of the unmixed signals. This is made possible with the proposed method and the required theory and constraints are derived. The proposed method has a very low complexity and is based on a first-order differential array. To verify the method and document the theory, computer simulations are performed: First, in the ideal situation where all constraints are met, and subsequently in situations where one or more constraints are not met. These simulations verify the precision of the method in the ideal situations, and the robustness of the method in non-ideal situations.

2. IBM ESTIMATION

The proposed method is based on two first-order differential arrays (cardioids) pointing in opposite directions. One target source and one masker source are present and separated in space as shown in Figure 1. We assume that the directional patterns and the azimuths of the two sources are known. If the spacing between the two microphones in the first-order differential array is much smaller than the acoustic wavelength, the output can be approximated by [12]:

$$\begin{aligned} C_T(f) &\approx G(f)\left(a_0T(f)+a_1M(f)\right) \qquad (2)\\ C_M(f) &\approx G(f)\left(b_0T(f)+b_1M(f)\right), \qquad (3)\end{aligned}$$

where f is the frequency, $G(f)$ is a high-pass system, $T(f)$ is the target signal, $M(f)$ is the masker signal, and a_0, a_1, b_0, b_1 are directional gains for the target and masker signal as shown in Figure 1. To obtain the T-F representations of $C_T(f)$ and $C_M(f)$ the two signals are further processed as shown in Figure 2: Filtering through a K-point filterbank, squaring the absolute value, low-pass filtering, and downsampling by a factor P. Assuming that $T(f)$ and $M(f)$ are uncorrelated, the four steps result in the two directional power signals:

$$\begin{aligned} \mathbf{D}_T(\tau,k) &= |G(k)|^2\left(a_0^2\mathbf{T}(\tau,k)+a_1^2\mathbf{M}(\tau,k)\right) \qquad (4)\\ \mathbf{D}_M(\tau,k) &= |G(k)|^2\left(b_0^2\mathbf{T}(\tau,k)+b_1^2\mathbf{M}(\tau,k)\right), \qquad (5)\end{aligned}$$

where $\mathbf{T}(\tau,k)$ and $\mathbf{M}(\tau,k)$ are the powers of the target and masker signals, respectively. To estimate the IBM using the two directional

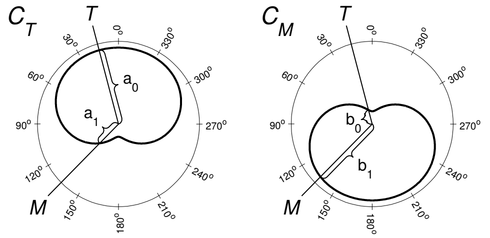

Fig. 1. The directional patterns of the two first-order differential arrays. C_T points towards the target signal T, and C_M points towards the masker signal M. The directional gains a_0, a_1, b_0, and b_1 are functions of the azimuths of the two sources T and M.

power signals (4, 5), we change (1) to

$$\widehat{\mathrm{IBM}}(\tau,k)=\begin{cases}1, \text{ if } \dfrac{\mathbf{D}_T(\tau,k)}{\mathbf{D}_M(\tau,k)}>\mathrm{LC}' \\ 0, \text{ otherwise}\end{cases}, \tag{6}$$

where LC′ is the applied local SNR criterion derived in the next section, and $\widehat{\mathrm{IBM}}$ is the estimate of the IBM.

2.1. The relation between LC and LC′

To estimate the IBM with the directional system using (6), the LC′ value must be derived from the LC value used in the definition of the IBM (1). Leaving out the time and frequency indices in the directional signals from (4, 5) we get, using (6):

$$\frac{a_0^2\mathbf{T}+a_1^2\mathbf{M}}{b_0^2\mathbf{T}+b_1^2\mathbf{M}}>\mathrm{LC}' \;\Leftrightarrow\; \frac{\mathbf{T}}{\mathbf{M}}>\frac{b_1^2\mathrm{LC}'-a_1^2}{a_0^2-b_0^2\mathrm{LC}'}. \tag{7}$$

To allow this rearrangement, we introduce the constraints

$$a_0^2-b_0^2\mathrm{LC}'>0 \quad\text{and}\quad b_1^2\mathrm{LC}'-a_1^2>0, \tag{8}$$

which guarantee that $\mathbf{T}/\mathbf{M}>0$ and prevent the target and masker from being interchanged. A prerequisite for estimating the IBM is that C_T captures more target signal than masker signal, and C_M captures more masker signal than target signal. Otherwise, the binary mask will be inverted. Using the definition of the IBM from (1) in combination with (7) we obtain

$$\mathrm{LC} = \frac{b_1^2\mathrm{LC}'-a_1^2}{a_0^2-b_0^2\mathrm{LC}'} \Leftrightarrow \tag{9}$$

$$\mathrm{LC}' = \frac{a_0^2\mathrm{LC}+a_1^2}{b_0^2\mathrm{LC}+b_1^2}. \tag{10}$$

Since we can express LC′ in terms of LC, we can actually estimate the IBM without having the unmixed sounds available, if the directional gains are known.

2.2. The asymptotes of LC′

If the directional gains are known, the LC′ value can be calculated from the wanted LC value using (10). If the directional gains are unknown, a fixed LC′ must be used in (6), and the LC value will change depending on the location of the sources (9). Combining the two constraints from (8) we get that

$$\frac{a_1^2}{b_1^2}<\mathrm{LC}'<\frac{a_0^2}{b_0^2}, \tag{11}$$

which are the two asymptotes of LC′ as shown in Figure 3. The asymptotes are determined by the amount of target and masker signal captured by C_T compared to C_M. If no target signal is found in C_M, the high asymptote will be at $+\infty$ dB, and if no masker signal is found in C_T, the low asymptote will be at $-\infty$ dB. In the interval bounded by the two asymptotes we find a region where the relation between LC and LC′ becomes approximately linear. In this region, changes of LC′ produce an equal change of LC. However, changes of LC′ near the asymptotes produce very large changes of LC. We refer to this relation as the *sensitivity* of the method. If the sensitivity is high, errors on $\mathbf{D}_T$, $\mathbf{D}_M$, or the directional gains, can have a significant impact on the LC value. The minimum sensitivity is found in the approximately linear regions which should be as large as possible. The asymptotes makes the LC′ be defined for all LC values, whereas the opposite is not true. If the LC′ value used in (6) is below the low asymptote, the mask becomes an all-one mask. If the LC′ is above the high asymptote the mask becomes an all-zero mask.

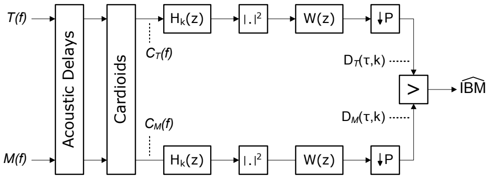

Fig. 2. Blockdiagram for estimation of the ideal binary mask. The acoustic delays model the delay from sources to the microphones in the first-order differential array. $\mathrm{H_k(z)}$ is the k'th analysis filter in the filterbank, $\mathrm{W(z)}$ is a low-pass filter, and ↓P is a decimation. The block labeled $>$ is the implementation of Equation (6).

3. SIMULATIONS

To verify that it is possible to estimate the IBM with the proposed method, a computer simulation was performed showing the precision of the estimate. Furthermore, simulations were done in non-ideal situations to illustrate the robustness of the method. The precision were measured by the number of correct T-F units in the $\widehat{\mathrm{IBM}}$ with respect to the IBM. Two instances of the system shown in Figure 2 were used: The first instance was used to calculate the $\widehat{\mathrm{IBM}}$ and was configured as follows: The acoustic delays were calculated from the azimuth of the two sources using a free-field model [13] with no reverberation. Two microphones were placed with a distance of 1 cm on the line through 0° and 180°, and the distance from the microphones to the sources was 1 m. Two cardioid signals were derived from the microphone signals, and each of the cardioid signals was processed by a 128 band Gammatone filterbank [11] with center frequencies linearly distributed on the ERB frequency scale from 100 Hz to 8000 Hz, each filter having a bandwidth of 1 ERB. The LP filter $W(z)$ was a 20 ms rectangular window followed by a 100 fold decimation corresponding to a 10 ms shift at the used sampling frequency of 20 kHz. The second instance of the system from Figure

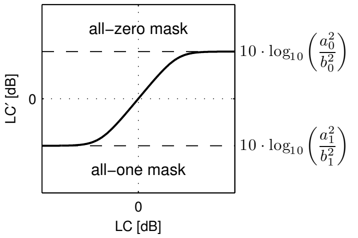
all−zero mask
all−one mask
LC′ [dB]
0
0
LC [dB]
$10 \cdot \log_{10}\left(\frac{a_0^2}{b_0^2}\right)$
$10 \cdot \log_{10}\left(\frac{a_1^2}{b_1^2}\right)$

Fig. 3. LC′ as a function of LC. The asymptotes are defined by the directional gains. Using LC′ values outside the region bound by the two asymptotes produce all-one or all-zero masks.

2 was used to calculate the IBM. This instance was equal to the previous without the cardioids. Instead, the target and masker sound were recorded separately by a single microphone located between the microphones used in the previous instance.

In the first simulation, the free-field model was used to calculate the acoustic delays, while the masker source was moved from $180 - 0°$, and the target source was fixed at $30°$. The two sources were male and female speech with 0 dB SNR and a duration of 11 seconds. A fixed LC′ value of 0 dB was compared to an adaptive LC′ value calculated using (10) and an LC value of 0 dB.

3.1. Simulation 1

The results from the first simulation are shown in Figure 4. The solid line is the percentage of correct T-F units using an adaptive LC′ value, and the dashed line is LC′ fixed at 0 dB. In both situations we see a high percentage of correct T-F units when the masker azimuth is in the range $180° - 150°$, and the small number of wrong T-F units ($< 2\%$) can be explained by the cardioid filters only used to calculate the $\widehat{\text{IBM}}$.

As the masker source is moved towards the target source, the percentage of correct T-F units decreases faster for the fixed LC′ than the adaptive LC′. At $90°$ the fixed LC′ has decreased to almost 50% whereas the adaptive LC′ remains above 95%. This decrease is explained by the $\widehat{\text{IBM}}$ becoming an all-one mask which in this case has around 50% correct T-F units. When the masker azimuth is $90°$ an equal amount of masker signal is captured by C_T and C_M, and the low asymptote in Figure 3 will be at 0 dB. In this situation the 0 dB fixed LC′ value is equal to an LC value of $-\infty$ dB. Moving the masker source further, we see a rapid decrease in correct T-F units for the adaptive LC′, when the masker source passes the target source at $30°$. The decrease from above 90% to below 10% correct T-F units is explained by the interchange of target and masker because (11) is not satisfied anymore. If C_T captures more masker than target sound or C_M captures more target than masker sound, the $\widehat{\text{IBM}}$ is the inverse of the IBM with a very low number of correct T-F units.

The small decrease in correct T-F units for the adaptive LC′ value between $180°$ to $45°$ can be explained by increased sensitivity of the system. As the masker and target get closer, the two asymptotes from Figure 3 get closer which leads to amplification of the errors introduced by the cardioid filters used for calculating the $\widehat{\text{IBM}}$.

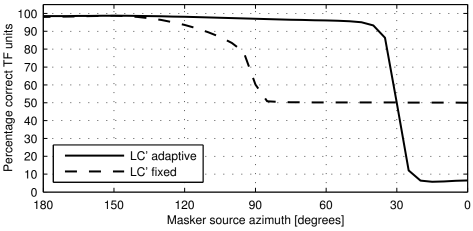

Fig. 4. The percentage of correct T-F units in the $\widehat{\text{IBM}}$ with respect to the IBM. The target was fixed at $30°$ while the masker was moved from $180°$ to $0°$. The adaptive LC′ value was calculated from the directional gains using an LC value of 0 dB, whereas the fixed LC′ was kept at 0 dB.

3.2. Simulation 2

To further examine the precision and robustness of the proposed method in a non-ideal setup a second simulation was carried out. The setup was identical to simulation 1, except the number of sources and the acoustical delays. One target and three masker sources were present: A male target speaker at $0°$, a female masker speaker moving from $180°$ to $0°$, a female masker speaker at $135°$, and a male masker speaker at $180°$. The speakers were located 2 m from the microphones and the sounds have a duration of 15 seconds. The acoustical delays were the free-field model from simulation 1 and impulse responses from a behind-the-ear (BTE) hearing aid shell on a Head and Torso Simulator (HATS) in three different acoustical environments: Anechoic, low reverberation time (RT_{60}=400 ms), and high reverberation time (RT_{60}=1000 ms). The reverberation time is defined as the time before the room impulse response is decreased by 60 dB.

As in the previous simulation, it is evident from Figure 5 that the percentage of correct T-F units decreases when the moving masker passes $90°$. In Figure 4 the fixed LC drops to 50% whereas in Figure 5 the free-field simulation drops to around 72% correct unit. This difference is explained by the two masker sources at $135°$ and $180°$ in simulation 2, which prevent the mask from becoming an all-one mask. Compared to simulation 1, where the all-one mask has 50% correct T-F units, the all-one mask in simulation 2 has 34% correct T-F units. Using impulse responses from a hearing aid on a HATS in an anechoic room, the percentage of correct T-F units between $95°$ and $40°$ is increased compared to the free-field simulation. This increase is explained by the cardioids being non-ideal and attenuating the moving masker more at these angles. As soon as reverberation is present, the precision of the $\widehat{\text{IBM}}$ decreases. Using impulse responses from the low reverberant room we get around 83% correct units when the moving masker is located at $180°$. If the wrong T-F units at this point are divided into wrong ones and wrong zeros with respect to the IBM we find 14% wrong zeros and 19% wrong ones. In other words, the $\widehat{\text{IBM}}$ will remove 14% of the target signal and will retain 19% of the masker signals compared to the IBM if applied to the mixture signal.

4. DISCUSSION

In this paper an important connection between the ideal binary mask and a realizable computation of the binary mask has been estab-

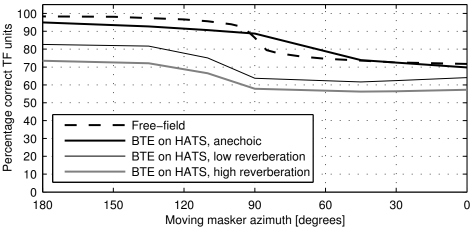

Fig. 5. The percentage of correct T-F units in the $\widehat{\text{IBM}}$ with respect to the IBM. Free-field and impulse responses from a hearing aid shell (BTE) on a HATS in three different acoustical environments were used, and four sources were present: Target at $0°$, a moving masker from $180°$ to $0°$, and two fixed maskers at $135°$ and $180°$. The LC$'$ value was 0 dB in all simulations.

lished. To calculate the IBM, the target and masker signals must be available prior to being mixed. This requirement can be relaxed by using a directional system to estimate the IBM, and from (6), we see that the $\widehat{\text{IBM}}$ can be equal to the IBM if only two sources are present, and their directional gains are known. The directional gains are used to calculate the LC$'$ value from the LC value and requires that the directional patterns of the cardioids and the target and masker azimuth are known.

From the first simulation, we find that the proposed method makes it possible to obtain an estimate of the IBM with a very high precission. When the two sources are spatially well separated, the setup with fixed LC$'$ and adaptive LC$'$ both provide a high number of correct T-F units. But as the two sources become closer, the setup with the adaptive LC$'$ shows a significant advantage compared to the fixed LC. The simulation illustrates what happens when the masker source is captured equally by the target and masker cardioid. The binary mask becomes an all-one mask with 50% correct T-F units. The same situation occurs when the target source is captured equally by the two cardioids, and the result is an all-zero mask. The method of varying the LC$'$ value has an advantage over fixating the LC$'$ value, and the target and masker source can become closer before the estimate is degraded significantly.

In the second simulation, we examine the robustness of the proposed method, when conditions are changed from the ideal ones. Introducing more sources and impulse responses from a BTE shell on a HATS in an anechoic room does not undermine the method and a significant increase in speech intelligibility can still be expected from the proposed method. However, a significant decrease in the percentage of correct T-F units is seen when reverberation is introduced, which are agreeable with the results reported using the DUET algorithm in echoic environments [7]. The errors introduced in the estimated binary mask can be divided into two types of errors, and in [3] the wrong ones and wrong zeros are referred to as type I and type II errors, respectively. In their paper, the impact on speech intelligibility of the two types of errors are measured showing that type II errors have a larger impact on speech intelligibility compared to type I errors. This interesting result should be taken into consideration when further developing the proposed method, but the results from [3] can not be used directly to predict speech intelligibility of the method proposed in the present paper. One reason is the difference in setup: We use a Gammatone filterbank whereas a linear filterbank is used in [3]. Another reason is the distribution of errors: It is expected that type II errors scattered uniformly as in [3] will have less impact on speech intelligibility compared to e.g. type II errors placed at onsets in the target sound.

5. CONCLUSION

In this paper we have proposed a method that makes it possible to estimate the ideal binary mask without having the unmixed signals available. If certain constraints are met, the precision of the estimated binary mask is very high, and even if the constraints are not met the proposed method shows promising results having the low complexity of the method in mind. These results establish an important connection between the ideal binary mask and a realizable system for T-F masking, and the precision and robustness of the proposed method in non-ideal conditions makes it very promising for further research and development.

6. REFERENCES

[1] M. Cooke and D.P.W. Ellis, “The auditory organization of speech and other sources in listeners and computational models,” *Speech Comm.*, vol. 35, no. 3, pp. 141–177, 2001.

[2] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis*, Wiley & IEEE Press, Hoboken, New Jersey, 2006.

[3] Ning Li and Philipos C. Loizou, “Factors influencing glimpsing of speech in noise,” *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1165–1172, 2007.

[4] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, “Determination of the potential benefit of time-frequency gain manipulation,” *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[5] N. Roman, D. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[6] D. Kolossa and R. Orglmeister, “Nonlinear postprocessing for blind speech separation,” in *Proc. ICA 2004*, Granada, Spain, September 22-24. 2004, pp. 832–839.

[7] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[8] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, “Two-microphone separation of speech mixtures,” *IEEE Trans. Neural Netw.*, vol. 19, no. 3, pp. 475–492, 2008.

[9] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, Pierre Divenyi, Ed., pp. 181–197. Kluwer Academic, 2005.

[10] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[11] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, “SVOS final report, part b: Implementing a gammatone filterbank,” *Rep. 2341, MRC Applied Psychology Unit.*, 1988.

[12] G. W. Elko, “Superdirectional Microphone Arrays,” in *Acoustic Signal Processing for Telecommunication*, Steven L. Gay and Jacob Benesty, Eds., chapter 10, pp. 181–237. Kluwer Academic Publishers, 2000.

[13] J. Blauert, *Spatial hearing. The Psychophysics of human sound localization*, MIT Press, Cambridge, USA, 1999.

In the following five figures, the results from paper D are elaborated by showing the hit and false alarm rates. The hit rate (1 - type II errors) is the percentage of correct ones with respect to the IBM. The false alarm rate (type I errors) is the percentage of zeros that have falsely been estimated as being one.

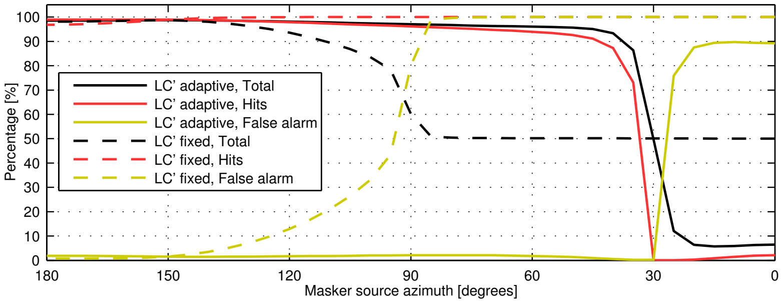

Figure 1: The results from Figure 4 split into hits and false alarms.

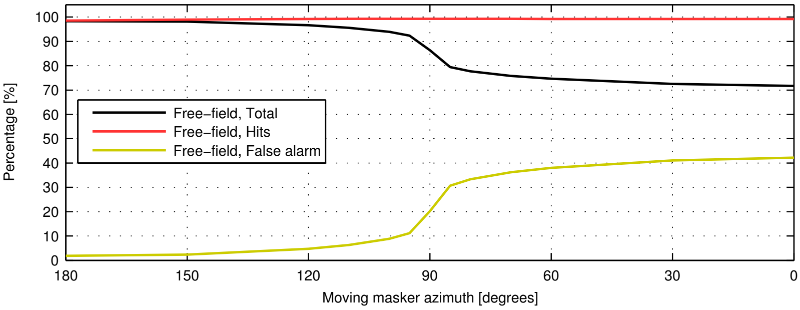

Figure 2: The Free-field result from Figure 5 split into hits and false alarms.

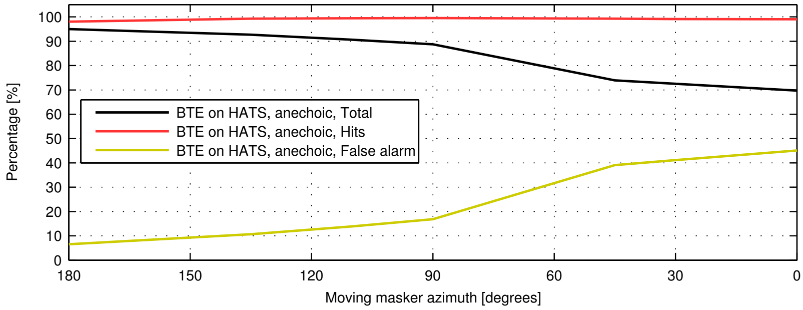

Figure 3: The BTE on HATS, anechoic from Figure 5 split into hits and false alarms.

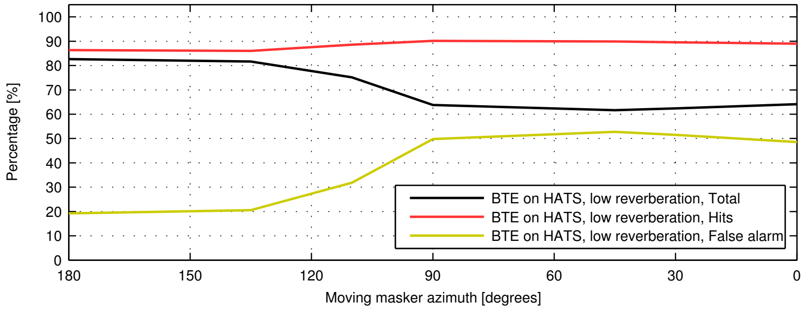

Figure 4: The BTE on HATS, low reverberation from Figure 5 split into hits and false alarms.

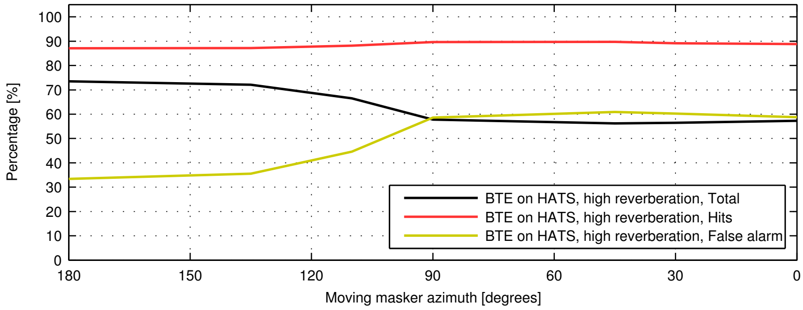

Figure 5: The BTE on HATS, high reverberation from Figure 5 split into hits and false alarms.

Paper E

Error-Correction of Binary Masks using Hidden Markov Models

Jesper B. Boldt, Michael S. Pedersen, Ulrik Kjems,
Mads G. Christensen, and Søren H. Jensen

*Proceedings of the International Conference on Acoustics,
Speech and Signal Processing 2010, pp. 4722–4725.*

©2010 IEEE

ERROR-CORRECTION OF BINARY MASKS USING HIDDEN MARKOV MODELS

Jesper Bünsow Boldt[1,2], *Michael Syskind Pedersen*[1], *Ulrik Kjems*[1],
Mads Græsbøll Christensen[2], *Søren Holdt Jensen*[2]

[1]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[2]Department of Electronic Systems, Aalborg University, DK-9220 Aalborg Øst, Denmark
{jeb,msp,uk}@oticon.dk,{mgc,shj}@es.aau.dk

ABSTRACT

Binary masking is a simple and efficient method for source separation, and a high increase in intelligibility can be obtained by applying the target binary mask to noisy speech. The target binary mask can only be calculated under ideal conditions and will contain errors when estimated in real-life applications. This paper proposes a method for correcting these errors. The error-correction is based on a hidden Markov model and uses the Viterbi algorithm to calculate the most probable error-free target binary mask from a target binary mask containing errors. The results demonstrate that it is possible to correct errors in the target binary mask and reduce the noise energy. However, speech energy is also reduced by the error-correction, but the impact on speech intelligibility and speech quality are not established or evaluated in the present study.

Index Terms— Binary masking, target binary mask, hidden Markov model, speech intelligibility, error-correction.

1. INTRODUCTION

Time-frequency masking has been widely used for source separation [1] and in experiments on intelligibility of noisy speech [2]. Basically, time-frequency masking is a general method of applying a time-varying and frequency-dependent gain to a signal. This gain is called the mask, and when the mask only contains the values zero and one, the method is referred to as binary masking. Using binary masking to separate speech from noise, in e.g. hearing aids and cochlear implants, is interesting because binary masking is a simple method and has a substantial impact on intelligibility. In [2, 3] the ideal binary mask increased intelligibility for normal hearing listeners and in [4] for hearing impaired listeners. The ideal binary mask was further studied in [5] together with the target binary mask, and when used to separate speech from noise both binary masks caused a large increase in intelligibility .

The target binary mask (TBM) is calculated by comparing the energy of the target speech with the long-term average energy of speech from the same speaker. If the energy of the target speech $\mathbf{T}(k,\tau)$ exceeds this long-term average energy $\mathbf{r}(k)$ by a certain amount, the value one will be assigned to the TBM at that particular time τ and frequency k. If not, the value zero is assigned:

$$\text{TBM}(k,\tau) = \begin{cases} 1, \text{ if } \dfrac{\mathbf{T}(k,\tau)}{\mathbf{r}(k)} > \text{LC} \\ 0, \text{ otherwise} \end{cases}, \tag{1}$$

where LC is the local SNR criterion. The LC value controls the amount of ones in the TBM. High intelligibility was obtained in [5] within the range of 20% - 60% ones in the TBM.

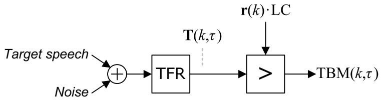

Fig. 1. Setup for estimating the target binary mask. The TFR block calculates the time-frequency representation of the mixture of target speech and noise. The time-frequency representation $\mathbf{T}(k,\tau)$ is compared to $\mathbf{r}(k)\cdot\text{LC}$ to get the target binary mask TBM(k,τ).

The positive impact on intelligibility, makes the TBM interesting in situations where intelligibility is reduced, e.g., for hearing aid users in difficult listening environments. The classic approaches to this problem have been evaluated in [6]. In this study, different speech enhancement algorithms are evaluated on normal hearing listeners. However, only a single algorithm in a single noise condition is able to increase intelligibility significantly.

It is possible to increase intelligibility with the TBM, but the method has the obvious drawback of requiring the target speech to be available. In most real-life situations, the target speech is not available, and the TBM must be estimated from the available sound. This estimate will contain errors, and the hypothesis of this paper is that these errors can be corrected. To our knowledge, no methods for estimating or correcting errors in the TBM have been proposed in the literature – a major reason for this being the novelty of the TBM. In this study, we focus on the error-correction of the TBM, well aware that the estimation of the TBM is not a trivial problem.

The error-correction employs a model of the error-free TBM. This model is a hidden Markov model (HMM) build from the TBM calculated using speech from multiple speakers to test the generality of the model. If the TBM from different speakers do not share some common characteristics, it is difficult to build a model of the TBM and use it for error-correction. To make the error-correction independent of the speaker, the long-term average energy $\mathbf{r}(k)$ in Equation (1) will not be adjusted to the individual speaker, but calculated as the long-term average energy of the speech used in the experiments.

The setup shown in Figure 1 is used to evaluate the proposed method. In this setup, the TBM will be error-free, if no noise sound is present and $\mathbf{r}(k)$ is known. When noise is added to the target speech, two types of errors will be found in the noisy TBM: False ones, if the noise sound causes the energy in the individual time-frequency units to exceed the threshold $\mathbf{r}(k)\cdot\text{LC}$. False zeros, if the speech and noise cancel each other in certain time-frequency units. At high signal-to-noise ratios (SNR), no errors will be found in the

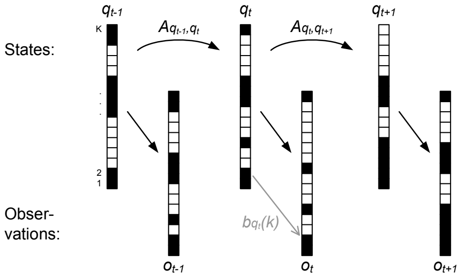

Fig. 2. Structure of the hidden Markov model used in this study. At time t, the state q_t generates an observation o_t and changes state with probability $\mathbf{A}_{q_t,q_{t+1}}$. The probability of being in state q_t and observing a one at frequency k is defined by $b_{q_t}(k)$ as shown with gray color.

TBM, but as the SNR decreases the amount of errors increases. The majority of these errors will be false ones, and ultimately, if the SNR is further reduced, the binary mask will become an all-one mask. In this situation the error-correction is of no use, even though the total number of errors could be reduced by forcing some time-frequency to zero in the target binary mask. Error-correction is expected to be possible when the TBM contains a comparable amount of correct and false ones.

2. BINARY MASK MODEL

The error-correction is based on a hidden Markov model [7] of the TBM. The HMM is a widely used statistical model for pattern recognition and speech processing, and it is particularly well suited to model time-series with time-varying statistical properties. In the HMM, the hidden layer contains multiple states, which, at each time increment, can change and generate an observation. From a sequence of observations, the most probable sequence of states can be calculated using the Viterbi algorithm [7].

In the HMM used in this study, the observations are the noisy TBM, the states are the error-free TBM, and the error-correction is the step of calculating the most probable error-free TBM from the noisy TBM using the Viterbi algorithm. The observations and states in the HMM are binary vectors of size K, as seen in Figure 2. K is the number of frequency channels. Each state in the HMM represents the TBM at a single time τ, and this approach assumes that the TBM can be build from a small number of states. However, when a limited number of states is used to build the TBM errors will be introduced, as seen in Figure 3. We refer to the process of building the TBM with a limited number of states as quantization of the TBM.

In the HMM, the probability of changing state is determined by the state-transition probability matrix $\mathbf{A}$, where the elements $a_{i,j}$ are the probability of changing from state i to state j [7]. In each state j, the observation probability $b_j(k)$ determines the probability of a one at frequency k. If the TBM could be build from N states without quantization error, the observation probabilities $b_j(k)$ would be binary and identical to the states. However, when the TBM is build from, e.g. 512 states, quantization error will be introduced because the TBM contains more than 512 different states. This means that the observation probabilities will have values between zero and ones: If d binary vectors from the training data are quantized to the same state j, and c out of the d states have a one at frequency k, we find that $b_j^T(k) = c/d$.

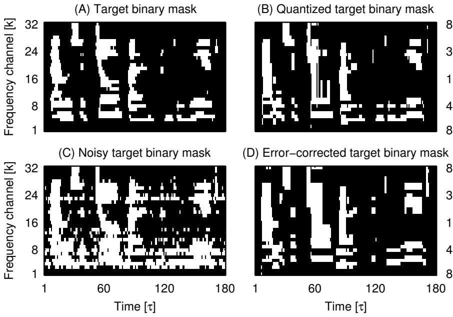

Fig. 3. Error-correction of the noisy target binary mask. (A) is the target binary mask from 1.8 s of speech. (B) is (A) quantized using 256 states. (C) is the speech from (A) mixed with speech shaped noise at 0 dB SNR. (D) is (C) after error-correction using a 256 state hidden Markov model.

The probability of a false one in the TBM generated by the noise sound is given by $b^N(k)$. This probability is independent on the state in the HMM but dependent on the type and level of the noise sound. If we assume that the target speech and noise sound are independent and do not overlap in time and frequency, the observation probability in state j at frequency k is given by

$$b_j(k) = b_j^T(k) + b^N(k) - b_j^T(k)b^N(k), \tag{2}$$

where $b_j^T(k)$ is the probability of a one generated by the target speech. In the experiments, $b^N(k)$ will be estimated from a short segment of the noise sound in the beginning of each experiment.

3. TRAINING

To train the HMM and evaluate the error-correction, the EUROM corpus was used [8]. The training data was generated by calculating the TBM from 36 minutes of speech spoken by 4 male and 4 female speakers normalized to equal energy. This speech was processed using a 32 band Gammatone filterbank [9] with centerfrequencies between 80 Hz to 8000 Hz equally spaced on the ERB scale (equivalent rectangular bandwidth). To obtain $\mathbf{T}(k,\tau)$, each subband signal from the Gammatone filterbank was divided into 20 ms frames with 10 ms overlap, and the energy was calculated from each frame in each subband. The long-term average energy $\mathbf{r}(k)$ was calculated as the average of each frequency channel in $\mathbf{T}(k,\tau)$.

The 36 minutes of speech produced a TBM with 216000 columns and 32 rows from which N states were found while minimizing the quantization error as measured by the total amount of false ones and false zeros. This quantization was done using the K-mode algorithm which is similar to the well-known K-means algorithm but useable for clustering binary data [10]. From the quantized TBM, the state-transition probability matrix $\mathbf{A}_{i,j}$ was calculated by counting the number of state changes from state i to state j and divide by the total number of visits in state i. To find $b_j^T(k)$, the columns in the TBM quantized to the same state were identified, and the probability of a one at each frequency was calculated. To

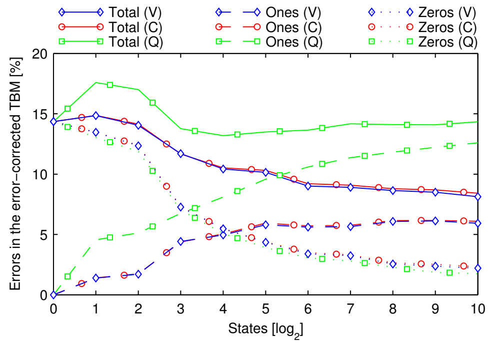

Fig. 4. Errors in the error-corrected target binary mask as a function of number of states in the hidden Markov model. (V) is error-correction using the Viterbi algorithm, (C) is error-correction using a causal Viterbi algorithm, and (Q) is simple quantization of the noisy target binary mask. The percentages of errors before error-correction were 0.2% false zeros, 15.9% false ones, and 16.1% in total.

find $b^N(k)$, the probabilities of a one in each of the frequency channels was obtained from the TBM as described in Section 2 using 5 seconds of noise.

4. EVALUATION

To evaluate the proposed method, two simulations were carried out. The first simulation examines the relation between the number of states and the performance of the error-correction as measured by the percentage of errors. The second simulation examines the loss of target energy and the remaining noise energy before and after the error-correction under different conditions. In both simulations, 10 sentences from a male and a female speaker were used. These two speakers were not part of the training data.

In the first simulation, the sentences were mixed with speech shaped noise at 0 dB SNR using the setup in Figure 1. The HMM was trained as described in Section 3 with a varying number of states between 1 and 1024. Figure 4 shows the percentage of errors after the error-correction. For comparison, the percentage of errors in the quantized noisy TBM and error-correction using a causal Viterbi algorithm are also shown.

All percentages are calculated relative to the total number of time-frequency units in the binary mask. The percentages of errors in the TBM before error-correction were 0.2% false zeros and 15.9% false ones giving 16.1% in total. As more states are used in the HMM, the amount of false ones increases, whereas the amount of false zeros decreases, as seen Figure 4. Using a single state in the HMM, this single state will be the all-zero column vector, making it impossible to have false ones in the error-corrected TBM. When the number of states is between 1 and 32, these states will contain few ones which limits the amount of false ones. Using 1024 states, the percentage of errors is 8.1% - a reduction of 8 percentage points compared to the noisy TBM. However, the reduction of false ones has the drawback of increasing the amount of false zeros relative to the noisy TBM.

The Viterbi algorithm uses previous, current and future observations from the noisy TBM to calculate the most probable state sequence. In low-delay applications this dependency on the future is critical and for comparison a causal Viterbi algorithm was implemented. This algorithm finds the most probable sequence based upon the previous and current observations only, and, as seen in Figure 4, this modification does not significantly reduce performance.

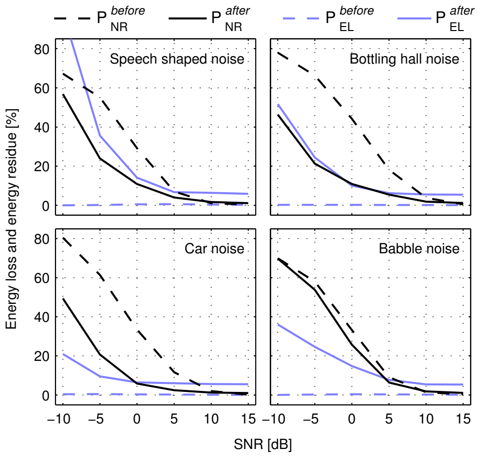

Fig. 5. The noise residue P_{NR} and the loss of target energy P_{EL} before and after the error-correction. The hidden Markov model used for error-correction has 256 states. P_{NR} and P_{EL} is calculated relative to the error-free target binary mask.

In the second simulation, an HMM with 256 states was trained as described in Section 3 and used to evaluate the performance between -10 dB and 15 dB SNR. The sentences were mixed with four different noise types: speech shaped noise, a high-frequency sound from a bottling hall, a low-frequency sound from the interior of a car, and babble noise. Performance was measured using the percentage of energy loss and the percentage of noise residue [11]:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \qquad (3) \qquad\qquad P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \qquad (4)$$

where $I(n)$ is the resynthesized sound using the TBM, $O(n)$ is the resynthesized output before or after the error-correction, $e_1(n)$ is the sound found in $I(n)$ but not in $O(n)$, and $e_2(n)$ is the sound found in $O(n)$ but not in $I(n)$. As seen in Figure 5, a similar performance for the different noise types is seen when changing SNR. At low SNR, the percentage of noise energy P_{NR}^{before} is high for the noisy TBM before error-correction. As the SNR increases, the amount of false ones in the TBM decreases resulting in a lower percentage of noise energy. Ultimately, when SNR is further increased, the P_{NR}^{before} is reduced to 0% because no false ones are found in the noisy TBM. The noise energy after error-correction P_{NR}^{after} shows a reduction at SNRs below 10 dB but a very small increase at SNRs around 15 dB. This increase shows that error-correction of an error-free TBM can introduce false ones due to the limited number of states in the HMM. The percentage of energy loss P_{EL}^{before} shows that loss of target energy using the noisy TBM is close to 0%, because very few false zeros are found in the noisy TBM before error-correction. When error-correction is introduced, the loss of target energy increases as shown by P_{EL}^{after}: At low SNRs the loss of target energy is significant, but as the SNR increases this loss is reduced and levels off at around 8%. The lower limit of P_{EL}^{after} at 8% is explained by the limited number of states in the HMM. If few errors are found in the TBM, the error-correction will increase both false ones and false zeros. For all four noise types, except the babble noise, the best performance is found around $0 - 5$ dB SNR, when the error-correction reduces the noise energy more than the target energy. Listening to the processed sound before and after error-correction confirms this finding.

5. DISCUSSION

The results confirm that a model of the TBM can be build and used to correct errors in the noisy TBM, but the reduction of false ones has the drawback of increasing the amount of false zeros. Even though the relation between errors and intelligibility has been examined in [12], it is difficult to use these results to determine the intelligibility of the TBM before and after the error-correction. In [12], the errors are uniformly distributed in time and frequency and the frequency resolution is different. The authors in [12] find that false ones reduce intelligibility more than false zeros. However, the location of errors and the noise type must have a significant impact on intelligibility, e.g. if the false zeros are found at onsets in the target speech.

If the relation between errors in the TBM and intelligibility was well-established, this would change the training and use of the HMM. If false ones reduce intelligibility more than false zeros, the model could be modified to allow more false zeros than false ones. Such a weighting would make it possible to adjust the level of lost target energy and remaining noise energy. Furthermore, the impact from errors on intelligibility is probably frequency dependent and thus it might be useful to reduce errors at some frequencies at the prize of more errors at other frequencies.

An interesting question to consider, is if the performance of the error-correction will continue to improve with an increasing number of states. Using more states will reduce the errors, but errors will be difficult to avoid. Errors in the TBM can make a wrong sequence of states more probable than the correct sequence. This limitation is a drawback of working in the binary domain, because the amount of information about the target speech and the noise sound is greatly reduced compared to the time-frequency domain.

Another limitation of the proposed method is the speaker dependency of the TBM. The TBM changes with different speakers, so the model used in the error-correction has to model different speech sounds as well as different speakers. This makes the model more general, but also less precise for the individual speakers. If $\mathbf{r}(k)$ was adjusted to each speaker, the TBM from different speakers would probably be more similar and less complex to model. This could reduce the number of required states in the hidden Markov model without affecting performance. The Viterbi algorithm has a complexity of $O(N^2T)$, where T is the number of observations [13], why decreasing the number of states will make the method more usable in hearing aids.

More complex models, e.g., factorial HMMs, could also be used to make the error-correction more efficient. However, the complexity of the model should be considered with respect to the complexity of the domain. The binary domain is a simplified domain compared to the time-frequency domain, and applying a very complex model in a simple domain might not be optimal. Instead, models in the time-frequency domain should be used. This also applies to the present study, because the large number of states can be a problem in applications like hearing aids even though the complexity of the model itself is low.

6. CONCLUSION

In this study, a method for error-correction of the TBM has been presented. The method is based on a HMM and trained on the TBM calculated under ideal conditions. The results of this study demonstrate that errors can be reduced, although the reduction of false ones has the drawback of increasing the amount of false zeros. The possibility of correcting errors in the TBM makes algorithms for estimating the TBM in real-life applications like hearing aids and cochlear implants more interesting and useful. The method used in this study can be further improved, e.g. using a speaker dependent $\mathbf{r}(k)$ or by weighting of different frequencies, but the model could also be useful for similar problems involving erroneous binary patterns.

7. REFERENCES

[1] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[2] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[3] N. Li and P. C. Loizou, "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. EL59–EL64, 2008.

[4] Deliang Wang, Ulrik Kjems, Michael S. Pedersen, Jesper B. Boldt, and Thomas Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2336–2347, 2009.

[5] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 1 edition, 2007.

[7] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[8] D. Chan et al., "EUROM - a spoken language resource for the EU," in *Proc. Eurospeech*, 1995, vol. 1, pp. 867–870.

[9] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS final report, part b: Implementing a gammatone filterbank," *Rep. 2341, MRC Applied Psychology Unit.*, 1988.

[10] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[11] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[12] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, 2008.

[13] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall, 1 edition, 2001.

Paper F

A Simple Correlation-based Model of Intelligibility for Nonlinear Speech Enhancement and Separation

Jesper B. Boldt, Daniel P. W. Ellis

Proceedings of the European Signal Processing Conference 2009,
pp. 1849–1853.

©2009 European Association for Signal Processing

A SIMPLE CORRELATION-BASED MODEL OF INTELLIGIBILITY FOR NONLINEAR SPEECH ENHANCEMENT AND SEPARATION

Jesper B. Boldt[1], *Daniel P. W. Ellis*[2]

[1]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[2]LabROSA, Dept. of Electrical Engineering, Columbia University
New York, NY 10027, USA
jeb@oticon.dk, dpwe@ee.columbia.edu

ABSTRACT

Applying a binary mask to a pure noise signal can result in speech that is highly intelligible, despite the absence of any of the target speech signal. Therefore, to estimate the intelligibility benefit of highly nonlinear speech enhancement techniques, we contend that SNR is not useful; instead we propose a measure based on the similarity between the time-varying spectral envelopes of target speech and system output, as measured by correlation. As with previous correlation-based intelligibility measures, our system can broadly match subjective intelligibility for a range of enhanced signals. Our system, however, is notably simpler and we explain the practical motivation behind each stage. This measure, freely available as a small Matlab implementation, can provide a more meaningful evaluation measure for nonlinear speech enhancement systems, as well as providing a transparent objective function for the optimization of such systems.

1. INTRODUCTION

Speech enhancement concerns taking a target speech signal that has been corrupted, by the addition of interfering sources and transmission through an acoustic channel, and mitigating the impact of these corruptions. Enhancement can have two, distinct goals: improving *quality*, which relates to how "clear" or "natural" the enhanced speech sounds, and improving *intelligibility*, which focuses on the more practical problem of whether a listener can understand the message in the original target speech. Although we might expect that quality and intelligibility are strongly correlated, there are ample situations in which speech of relatively low quality can nonetheless achieve high intelligibility [17, 22], and where improving quality does not necessarily improve intelligibility [10].

In this paper we ignore quality (and related effects such as listener fatigue) and concentrate on intelligibility. We focus specifically on time-frequency masking algorithms, which have been widely used in automatic speech recognition [6], computational auditory scene analysis [21], noise reduction [15, 2], and source separation [23, 16]. In this type of algorithm, a time-varying and frequency-dependent gain is applied across a number of frequency channels. In some variants, the gains are quantized to zero or one, giving a *binary masking* algorithm where the pattern of gains is referred to as the binary mask. One type of binary mask – the ideal binary mask (IBM) – has shown to be able to increase speech intelligibility significantly [3, 2, 14]. This mask is 'ideal' in that it relies on perfect knowledge of both clean target and interference prior to mixing, and is constructed to pass only those time-frequency cells in which the target energy exceeds the interference. An intriguing property of the IBM is that applying such a mask to a sound consisting only of noise results in high intelligibility for the speech upon which the mask was based [22, 13], even though the perceived quality of the reconstructed speech is very poor: depending on the resolution of the time-frequency distribution, it will have no pitch or other fine structure, and fine nuances of energy modulation are lost. Similar characteristics are found to those of noise-excited channel-vocoded speech [17]. An attempt to measure the signal to noise ratio (SNR) in such signals would find no trace of the original target in the final output, so SNR-based measures will not be a useful basis for accounting for this intelligibility. What is preserved, however, is the broad envelope in time and frequency. This suggests that an intelligibility estimate could be developed based on the similarity of this envelope between target speech and system output.

In this paper, we use correlation as a measure of similarity between time-frequency envelopes of target and enhanced speech. Given this basic principle, we make a number of design choices and system enhancements with a view to matching the general properties of observed subjective intelligibility of nonlinearly-enhanced signals. At each stage, we strive for the simplest and most transparent processing that can effectively match the subjective results. Our outcome is a simple correlation-based measure that can predict intelligibility with approximately the same fidelity as more complex models based on far more detailed models of auditory processing [4]. We feel this simplicity and transparency is a considerable advantage as a guide for developing enhancement systems.

2. NORMALIZED SUBBAND ENVELOPE CORRELATION

To estimate intelligibility, the correlation between the time-frequency representations of the target (reference) and the output of the time-frequency masking algorithm is calculated:

$$\sum_{\tau}\sum_{k}\mathbf{T}(\tau,k)\cdot\mathbf{Y}(\tau,k), \tag{1}$$

where τ the time index, k the frequency index, $\mathbf{T}(\tau,k)$ is the energy envelope of the target signal, and $\mathbf{Y}(\tau,k)$ is the energy envelope of the output. This correlation will not have an upper bound, and in low energy regions of $\mathbf{T}(\tau,k)$ the inclusion of potential unwanted energy in $\mathbf{Y}(\tau,k)$ will have a very small impact on the correlation. To improve this behavior, we normalize with the Frobenius norm of $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$ and refer to this measure as the normalized subband envelope correlation ($\mathbf{nSec}$):

$$\mathbf{nSec} = \sum_{\tau}\sum_{k}\frac{\mathbf{T}(\tau,k)\cdot\mathbf{Y}(\tau,k)}{||\mathbf{T}(\tau,k)||||\mathbf{Y}(\tau,k)||} \tag{2}$$

The $\mathbf{nSec}$ is bounded between zero and one. The lower bound is reached if no energy is found in the same regions of $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$. The upper bound is reached if the two signals are identical or only differ by a scale factor. Geometrically interpreted, $\mathbf{nSec}$ is the angle between $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$ if calculated using a single time or frequency index.

3. EXPERIMENTAL DATA

To verify that $\mathbf{nSec}$ is a useful measure of speech intelligibility, we use the results from Kjems et al. [13], where speech intelligibility of

This work was supported by Oticon A/S, The Danish Agency for Science, Technology and Innovation, and by the NSF under grant no. IIS-0535168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

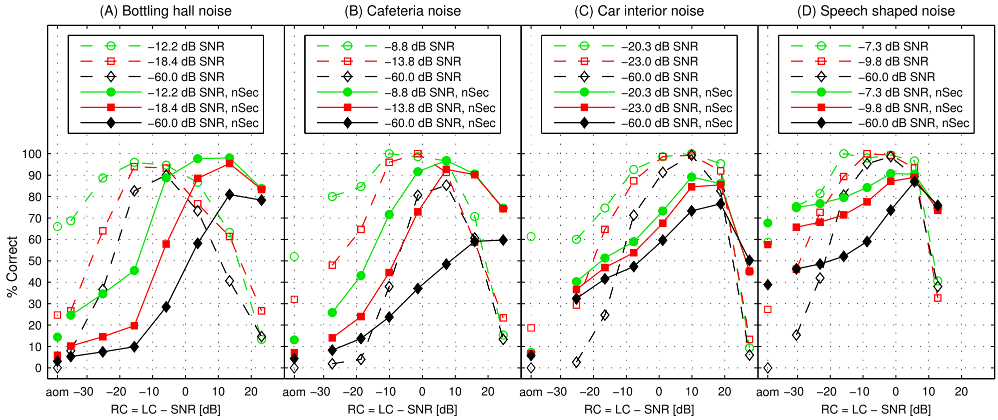

Figure 1: Estimated intelligibility by **nSec** compared to subjective listening tests in four different noise conditions and three SNR levels. **nSec** is shown with solid lines/filled symbols, and subjective listening tests are shown with dotted lines/hollow symbols. The results are plotted as a function of the RC value which determines the sparseness of the binary mask - higher RC values imply fewer ones in the binary mask. The all-one mask (aom) is the unprocessed condition and does does not correspond to a specific RC value.

IBM-masked noisy speech is measured using normal hearing subjects, three SNR levels, four noise types, and different local SNR criterions (LC). LC is the threshold used to construct the IBM; a larger LC results in an IBM with proportionally fewer nonzero values:

$$\mathrm{IBM}(\tau,k) = \begin{cases} 1, \text{ if } \dfrac{\mathbf{T}(\tau,k)}{\mathbf{N}(\tau,k)} > \mathrm{LC} \\ 0, \text{ otherwise} \end{cases} , \tag{3}$$

where $\mathbf{N}(\tau,k)$ is the energy envelope of the noise signal.

Two of the three SNR levels used in the experiments by Kjems et al. were set to 20% and 50% intelligibility of the speech and noise mixtures with no binary masking. The third SNR level was fixed at -60 dB to examine the effect of applying the IBM to pure noise. Four different noise conditions were used: speech shaped noise (SSN), cafeteria noise, car interior noise with mainly low-frequency energy, and noise from a bottling hall with mainly high-frequency energy. The LC values resulted in IBMs consisting of between 1.5% and 80% of nonzero cells, and an all-one mask (aom) was used to measure the intelligibility of the unprocessed mixture with no binary masking. A 64 channel Gammatone filterbank with centerfrequencies from 55 Hz to 7742 Hz equally spaced on the ERB (equivalent rectangular bandwidth) scale was used, and the output was divided into 20 ms frames with 10 ms overlap. The results are shown with dotted lines and hollow symbols in Figure 1 (and are repeated in subsequent figures). To align the results, they are plotted as a function of the RC value defined as $\mathrm{RC} = \mathrm{LC} - \mathrm{SNR}$ in units of dB. Using this x-coordinate, the binary masks will be identical at the same RC value and independent of the SNR levels.

To compare the **nSec** with the results by Kjems et al., we use 10 sentences from their experiment which have been mixed with noise and processed with the IBM. Silence between the sentences are removed from the waveforms, and $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$ are calculated using a 16 channel Gammatone filterbank with center frequencies from 80 Hz to 8000 Hz equally spaced on the ERB scale. The energy from each frequency channel in the filterbank is divided into segments of 80 ms with 40 ms overlap. All processing is done at 20 kHz. The calculated time-frequency representations $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$ are inserted in Equation 2, and the **nSec** scaled by a factor of 100 is shown with solid lines and filled symbols in Figure 1.

4. MODIFICATIONS TO THE nSec

Looking at Figure 1, it can be seen that even though the **nSec** is not aligned with the subjective listening tests, the overall shape and behavior is encouraging: Increasing SNR gives a better or similar **nSec**, and a distinct peak in correlation as a function of RC value is seen at all curves expect for the -60 dB SNR cafeteria noise (Fig.1.B). If this curve had been continued to higher RC values, it would have made a peak at some point, because higher RC values makes the binary mask more sparse with fewer ones, and, ultimately, $\mathbf{Y}(\tau,k)$ will be zero. At the other extreme, at low RC values, the **nSec** levels off which is most evident from Figure 1.A and 1.D. The reason is that at some RC value, the time-frequency units added to $\mathbf{Y}(\tau,k)$ by lowering the RC value will not change the numerator of Equation 2 because no energy is found at these time-frequency units in $\mathbf{T}(\tau,k)$. At the same time, the denominator will continue to increase as the RC value decreases, due only to the added energy in $||\mathbf{Y}(\tau,k)||$; $||\mathbf{T}(\tau,k)||$ is a fixed value independent of SNR and RC value.

Comparing the three SNR levels, it can be seen that the peak of the **nSec** shifts towards lower RC values for higher SNRs – a reasonable property, if we recognize that the IBM for a certain target and noise sound is a function of the RC value only, and that increasing SNR level implies that the RC value can be lowered without increasing the number of noise-dominated time-frequency units in the binary masked mixture. At increasing SNR levels, the RC value is lowered by increasing the LC value with less than the increase in SNR level.

The **nSec** for the speech shaped noise (Fig.1.D) with an all-one mask is considerable higher at all three SNR levels compared to other noise types. The **nSec** of the -60 dB SNR mixture with an all-one mask is approximately 0.4, despite the fact that practically no target sound is found in the mixture. Two random signals will always give a positive correlation as long as they contain energy in some of the same time-frequency regions, and the speech shaped noise do, since it was made by superimposing 30 sequences of the speech from the corpus with random silence durations and starting times [20]

The last observation we make of the unmodified **nSec** is that the location of the peaks are at higher RC values compared to the subjective listening tests. This property is caused by the fact that

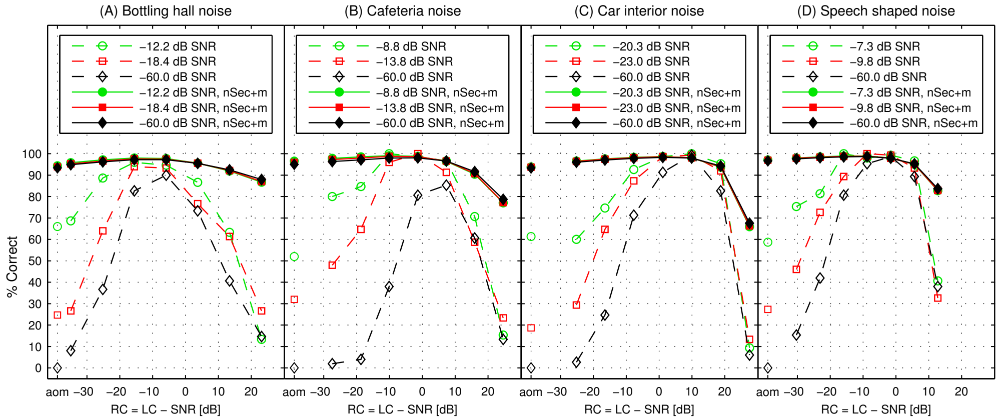

Figure 2: The modified $\mathbf{nSec}$ with frequency normalization, compression, and DC removal compared to the subjective listening tests.

$\mathbf{nSec}$ is decreased when $\mathbf{Y}(\tau,k)$ contains energy in the low energy regions of $\mathbf{T}(\tau,k)$; to get a high $\mathbf{nSec}$, the binary mask should only present the high energy regions of $\mathbf{T}(\tau,k)$.

To improve the alignment between the subjective listening tests and the $\mathbf{nSec}$, the following modifications are introduced:

4.1 Frequency Normalization

In speech signals, high frequencies have less energy than low frequencies, but this difference does not reflect the frequencies' importance to intelligibility. Using the $\mathbf{nSec}$ without any frequency normalization will make the low frequencies dominate the result. Furthermore, the auditory system can to some degree adapt to the listening situation and a minor fixed coloration of the speech spectrum is not expected to affect intelligibility. To compensate for the difference in energy and any fixed colorations, we normalize the frequency channels to equal energy. This normalization has the drawback that at increasing RC values, when the binary mask becomes more sparse, some frequency channels will contain few non-zero elements, which would become very large because of the normalization. To avoid these high level time-frequency units, amplitude compression should follow the normalization (although in frequency channels with no non-zero elements, no normalization should be applied).

4.2 Compression

To decrease the relative importance of high level time-frequency units mainly produced by the frequency normalization, compression can be applied to $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$. Compression will move the peaks of the $\mathbf{nSec}$ curves towards lower RC values, but also reduces the difference between the three SNR levels. To align the $\mathbf{nSec}$ peaks with the subjective listening tests, $\mathbf{T}(\tau,k)$ and $\mathbf{Y}(\tau,k)$ are raised to the power of 0.15.

4.3 DC removal

As previously stated, the $\mathbf{nSec}$ will be positive even if two random signals are used because their energy is always positive. To reduce this offset in the time-frequency representations, each frequency channel should be high-pass filtered. This high-pass filtering will push the values down to zero in the case where we have flat, but nonzero, energy and emphasize changes in energy instead of absolute levels. The used high-pass filter has a single zero at 1 and a single pole at 0.95.

5. RESULT

As seen in Figure 2, the modifications improve the correspondence between the subjective listening test and the $\mathbf{nSec}$. The differences are most pronounced at low and high RC values where the slope of the modified $\mathbf{nSec}$ is too shallow, and in the unprocessed condition (aom) the results are too low and too closely placed in the bottling hall and cafeteria noise condition. Ideally, the three SNR levels should give a intelligibility of 50%, 20% and 0%, but the compression, which was introduced to shift the peaks of the $\mathbf{nSec}$ towards lower RC values, also compresses the results at low RC values, making them more equal. At high RC values the shallow slope of the modified $\mathbf{nSec}$ is also a outcome from using compression. Compression increases the impact of low-amplitude time-frequency units and a more sparse mask is needed to reduce the $\mathbf{nSec}$.

To allow some nonlinearity in the relationship between the $\mathbf{nSec}$ and speech intelligibility, a logistic function can be applied:

$$p(c) = \frac{1}{1+e^{(o-c)/s}}, \tag{4}$$

where o is the offset, and s is the slope of the logistic function [4]. To find the offset and the slope we use the unconstrained nonlinear minimization function `fminsarch` in Matlab to minimize the squared error between $\mathbf{nSec}$ and the results from the subjective listening test using speech shaped noise. The found offset and slope of $o = 0.62$ and $s = 0.09$ are used to transform the $\mathbf{nSec}$ results from Figure 2 into the results seen in Figure 3. The overall performance is improved: a better correspondence between the subjective listening tests and the $\mathbf{nSec}$ is seen, but this is achieved at the expense of the match in the situation with no binary masking (aom).

6. DISCUSSION

Our proposed method uses a different approach compared to intelligibility measures as AI, SII, and STI [1, 8, 19] by using the correlation as the fundamental function for measuring intelligibility. In the AI, SII, and STI, the intelligibility is measured as a sum and weighting of SNR in a number of frequency channel. A more similar approach to ours is used in [11] for measuring speech quality and in [4] to measure intelligibility. In both works, the cross-correlation coefficient is used to measure the similarity between *internal representations* of the target and test signal. The internal representations are the expected patterns of neural activity from the auditory periphery calculated using the model by Dau et. al. [7]. In [4] the modu-

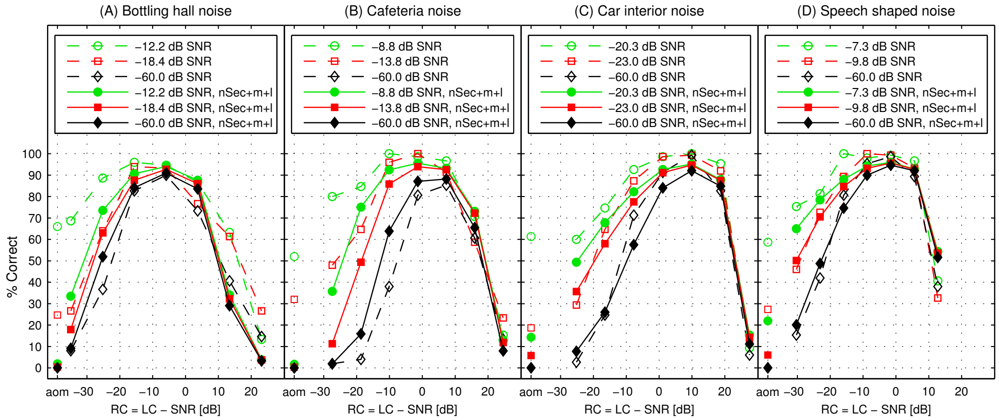

Figure 3: The modified $\mathbf{nSec}$ transformed with a logistic function (Eq. 4) and compared to the subjective listening tests.

lation filterbank is replaced by a modulation low-pass filter, and the cross-correlation coefficient is calculated using short time frames of 20 ms with 50% overlap. The cross-correlation coefficients are grouped into low, medium, and high level correlation frames as in [12], but only the average of the high level correlation frames is used in the model output which is mapped to intelligibility using a logistic function equal to Equation 4.

The model by Christiansen et al. shows significant improvements compared to the speech-based coherence SII method [12] and the speech-based STI method [9], when using the same subjective results as in this study. The predicted intelligibility, using the same 10 sentences as for the $\mathbf{nSec}$, is shown in Figure 4, and a substantial reason for the promising results is, as explained by the authors, the use of 20 ms frames, which is an interesting difference from the $\mathbf{nSec}$. The main deviation between the model and the subjective results is found using the bottling hall noise (Fig.4.A), which is explained by Christiansen et al. to be caused by a too high influence from the low frequencies on the final result.

It is of interest to compare our approach and results with the model by Christiansen et al., but concluding which one is better is not appropriate from the results shown in Figure 3 and 4. Mainly because the logistic function used in Figure 3 was fitted directly to the subjective results using the speech shaped noise condition, whereas the logistic function used by Christiansen et al. was fitted to the psychometric curves from subjective listening tests of unprocessed mixtures at different SNR levels. The consequence of this difference is evident using the all-one mask, where the results from the $\mathbf{nSec}$ are too close and too low, which is not the case for the model by Christiansen et al. An interesting difference between the two methods is the bottling hall noise, where the $\mathbf{nSec}$, although very similar at the three SNR levels, has a better alignment of the peaks, which is caused by the frequency normalization as explained in section 4.1.

We might question whether the proposed modifications of the $\mathbf{nSec}$ are the correct ones to use, and if they appear in the correct order. The modifications could be compared to processing steps in the auditory system, but in this case we have selected and ordered them purely to adjust the $\mathbf{nSec}$ to the subjective results and not to simulate specific aspects of the auditory system. Similarly, the use of the correlation as underlying basis was supported by the preliminary results seen in Figure 1, and not by assumptions about correlation being used at some level in human perception. Introducing additional steps – simple or complicated – could potentially improve the precision of the method, but would also introduce additional processing and parameters that would make the system less transparent for the user.

Another approach to measure intelligibility is the use of automatic speech recognition systems, where the number of correctly identified words or phonemes are used as a measure of speech intelligibility. This method has shown promising results [18, 5], but it is vulnerable to peculiarities of speech recognition systems that can make them differ widely from the perception of listeners. Trivial mismatch between the processed signals and the training data used by the recognizer can result in misleading low results.

A straightforward approach to evaluate time-frequency masking algorithms is to count the number of errors in the binary mask. Although we believe that the binary mask itself can explain a large amount of the intelligibility, this approach has various drawbacks e.g. the type of errors can have widely differing impact [15], the location of errors is important, and it is not certain which type of binary mask should be used as reference. Furthermore, this approach will not show the difference between applying the same binary mask to mixtures at different SNR levels.

The $\mathbf{nSec}$ has shown a fine agreement with subjective listening test of the IBM applied to different mixtures and SNR levels, but this is only one of many methods of time-frequency masking. In the present work, we have not examined how the $\mathbf{nSec}$ will behave using e.g. non-binary masks – the general case of applying a time-varying gain in a number of frequency bands – but we are hopeful that it will continue to agree with human performance. We note that the $\mathbf{nSec}$ can fail if the target and system output become misaligned e.g. if the processed mixture is delayed compared to the target, however this could be accommodated by searching over a timing skew parameter (full cross-correlation).

7. CONCLUSION

By focusing on the correlation between the broad spectral envelope of target and system output, while completely ignoring the fine structure, we arrive at an intelligibility measure able to match a range of subjective results that would be very difficult to explain by SNR measures. We therefore suggest that future work on nonlinear speech enhancement, if it is concerned with intelligibility, should use measures based on correlation in place of SNR. To this end, we have released a simple drop-in implementation of our measure, written in Matlab[1].

[1] See http://labrosa.ee.columbia.edu/projects/intelligibility/

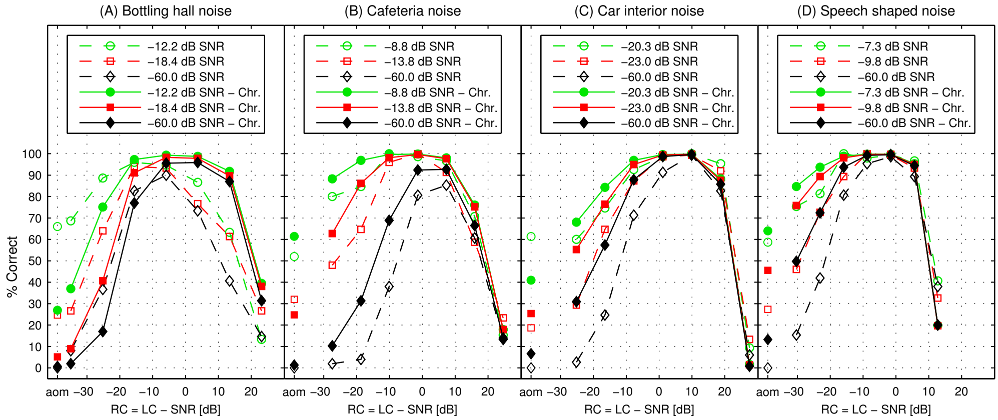

Figure 4: The predicted intelligibility using the model by Christiansen et al. [4] compared to the subjective listening test.

Although there are other, existing intelligibility measures that are able to match subjective data as closely as ours, our measure is constructed to be as simple as possible, with a consequent benefits in terms of transparency and diagnosis: when a system performs poorly under this measure, it is relatively easy to look at the processed envelopes going into the final correlation to see in which regions they are most different, thereby suggesting where to look for improvements. We hope that measures of this kind can help to focus and promote progress in speech intelligibility enhancement systems.

REFERENCES

[1] ANSI S3.5-1997. American national standard: Methods for the calculation of the speech intelligibility index, 1997.

[2] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney. Determination of the potential benefit of time-frequency gain manipulation. *Ear and Hearing*, 27(5):480–492, 2006.

[3] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. 120(6):4007–4018, 2006.

[4] C. Christiansen, T. Dau, and M. S. Pedersen. Prediction of speech intelligibility based on an auditory preprocessing model. *submitted to Speech Communication*, —-.

[5] M. Cooke. A glimpsing model of speech perception in noise. 119(3):1562–1573, 2006.

[6] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Comm.*, 34(3):267–285, 2001.

[7] T. Dau, B. Kollmeier, and A. Kohlrausch. Modeling auditory processing of amplitude modulation. I. Modulation detection and masking with narrowband carriers. 102:2892–2905, 1997.

[8] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. 19(1):90–119, 1947.

[9] I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. 100(3):1703–1716, 1996.

[10] Y. Hu and P. C. Loizou. A comparative intelligibility study of speech enhancement algorithms. pages IV–561–564, Hawaii, 2007.

[11] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. 14(6):1902–1911, 2006.

[12] J. M. Kates and K. H. Arehart. Coherence and the speech intelligibility index. 117(4 I):2224–2237, 2005.

[13] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. 126(3):1415–1426, 2009.

[14] N. Li and P. C. Loizou. Effect of spectral resolution on the intelligibility of ideal binary masked speech. 123(4):EL59–EL64, 2008.

[15] N. Li and P. C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. 123(3):1673–1682, 2008.

[16] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems. Two-microphone separation of speech mixtures. 19(3):475–492, 2008.

[17] R. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–4, 1995.

[18] S. Srinivasan and D. Wang. A model for multitalker speech perception. 124(5):3213–3224, 2008.

[19] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. 67(1):318–326, 1980.

[20] K. Wagener, J. L. Josvassen, and R. Ardenkjaer. Design, optimization and evaluation of a danish sentence test in noise. *International Journal of Audiology*, 42(1):10–17, 2003.

[21] D. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis*. Wiley & IEEE Press, Hoboken, New Jersey, 2006.

[22] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, and T. Lunner. Speech perception of noise with binary gains. 124(4):2303–2307, 2008.

[23] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. 52(7):1830–1847, 2004.