Computational modeling of speech intelligibility in adverse conditions

Chabot-Leclerc, Alexandre

Publication date:
2016

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

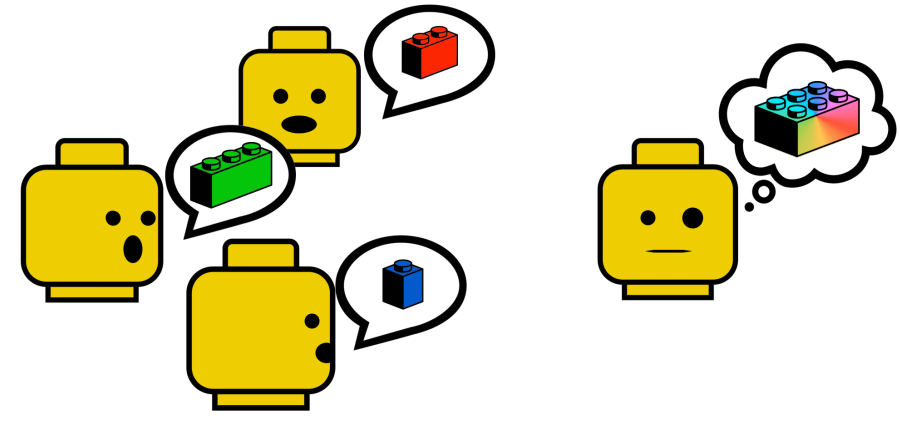
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 21

Alexandre Chabot-Leclerc

Computational modeling of speech intelligibility in adverse conditions



Hearing Systems
Department of Electrical Engineering

Computational modeling of speech intelligibility in adverse conditions

PhD thesis by
Alexandre Chabot-Leclerc

Technical University of Denmark

2018

The defense was held on December 4, 2015.

Illustrations for the front page were taken from The Noun Project and were made by Marco Petrucci, Jon Trillana, Samual Q. Green, and Herbert Spencer.

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark.

The project was partly financed by Phonak (2/3) and by the Technical University of Denmark (1/3).

Supervisors

Prof. Torsten Dau
Assoc. Prof. Ewen N. MacDonald
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

The intelligibility of speech is a measure of how well speech is understood in a given situation. Developing models to predict intelligibility can help develop a better understanding of the essential "features" of speech, how those features are extracted by the auditory system, and how they are combined and used to create understanding. This dissertation expands on a model named the speech-based envelope power spectrum model (sEPSM), which uses the signal-to-noise ratio in the envelope power domain (SNR_{env}) as the decision metric. The sEPSM was analyzed and compared to several other models that either use different front-ends or different decision metrics, such as the audio SNR. The goal was to tease apart the essential components of intelligibility models in a range of conditions known to be challenging.

One condition considered speech that was distorted by a phase jitter process, which destroys its spectral integrity. It was shown that the sEPSM could account for the deleterious effects of phase jitter if an across-channel process was included in the analysis stage, which measures the variability of the envelope power across audio frequencies. In another condition of nonlinear distortion, noise reduction via spectral subtraction, it was shown that across-channel processing was not essential.

Furthermore, a quantitative model was developed in an attempt to predict the speech intelligibility measured in conditions where listeners are known to benefit from using both ears, compared to using either ear alone, such as in a noisy "cocktail party". The model represents a binaural extension of the sEPSM, denoted as B-sEPSM. It consists of realizations of the sEPSM for the monaural pathways, combined with an equalization–cancellation (EC) process to model an across-ear noise reduction mechanism. The sEPSM process also operates at the output of the EC process, such that all pathways are directly comparable. The B-sEPSM was shown to account for intelligibility as a function of the number of maskers, the azimuth of the maskers, the room properties (anechoic or reverberant), the masker types (stationary noise, fluctuating noise, and time-reversed speech), and the interaural time differences of the target and maskers.

Finally, simulation results showed that binaural processing was not always necessary in spatial conditions, and that the SNR_{env} metric could capture aspects of masking that were not considered by models that used the audio SNR as the decision metric. However, none of the models considered could account for

the intelligibility in conditions with so-called “informational masking”, because they did not take into account confusions in the decision-making process experienced by the listeners. A possible method for estimating such confusions was proposed, based on a “distance metric” between the envelope power spectrum representation of the speech estimate and of the noise.

Overall, the results of this thesis support the hypothesis that the SNR_{env} is a powerful metric for intelligibility prediction. Furthermore, the B-sEPSM could be used to investigate the impact on intelligibility of different binaural noise reduction techniques, such as beam-forming, and of various binaural hearing aid compression strategies.

Resumé

Taleforståeligheden er et mål for, hvor godt tale forstås i en given situation. Ved at udvikle modeller, der kan forudsige taleforståeligheden, kan der opnås en bedre forståelse for de essentielle "features" i talen, hvordan det auditive system udtrækker disse features, samt hvordan de kombineres og benyttes til at skabe forståelse for tale. Denne afhandling udvider en model, kaldet den talebaserede modulationseffekt spektrum model (sEPSM), der bruger signal-til-støj modulationseffekt forholdet (SNR_{env}) som beslutningsparameter. sEPSM blev analyseret og sammenlignet med flere andre modeller, der enten benyttede forskellige front-ends eller forskellige beslutningsparametre, såsom signal-til-støj forholdet i audiodomænet (audio SNR). Målet var at adskille de essentielle komponenter i taleforståelighedsmodellerne i en række miljøer, der er kendt for at være vanskelige.

En betingelse betragtede tale forvrænget af en fase-jitter proces, der ødelægger talesignalets spektrale struktur. Det blev vist, at sEPSM kan forklare den skadelige effekt af fase-jitter, såfremt en proces blev inkluderet i analyse stadiet, der måler variationen af modulationseffekten på tværs af de auditive frekvenskanaler. I en anden betingelse med ikke-lineær forvrængning, spektral subtraktion støj reduktion, blev det vist at behandlingen på tværs af kanaler ikke var essentiel.

Endvidere blev en kvantitativ model udviklet i et forsøg på at forudsige taleforståeligheden i miljøer, hvor forsøgspersonen kan drage nytte af to øre fremfor et øre, såsom en støjfyldt "cocktail fest". Modellen repræsenterer en binaural udvidelse af sEPSM, kaldet B-sEPSM. Den består af realiseringer af den monaurale sEPSM model, kombineret med en "equalization-cancellation" (EC) proces, der modellerer på-tværs-af-øret støjreduktion mekanismen. sEPSM processen opererer også ved udgangen af EC processen, således at alle signalveje er direkte kompatible. Det blev vist, at B-sEPSM kunne forklare taleforståeligheden som en funktion af antal støjkilder, azimut vinklen af støjkilderne, egenskaberne af rummet (lyddødt eller med efterklang), støjkildetypen (stationær støjkilde, fluktuerende støjkilde og temporalt inverteret tale) og den interaurale tidsforskel mellem kilden og støjkilderne.

Slutteligt viste resultaterne af en række simuleringer, at den binaurale behandling ikke altid var nødvendig i rummelige miljøer og at SNR_{env} parameteren kunne fange aspekter af maskering, der ikke var fanget af modeller, der benyttede audio SNR som beslutningsparameter. Ingen af de betragtede modeller kunne forudsige taleforståeligheden i miljøer med såkaldt "informationel ma-

skering", fordi de ikke tog højde for forvirringer i den beslutningstagende proces, der blev oplevet af forsøgspersonerne. En mulig metode til at estimere sådanne forvirringer blev foreslået, baseret på en "distance-parameter" mellem modulationseffekt spektrum repræsentationen af taleestimatet og af støjen.

Alt i alt, bakker resultaterne i denne afhandling op om hypotesen, at SNR_{env} er et kraftfuldt mål til at forudsige taleforståelighed. Endvidere kan B-sEPSM benyttes til at undersøge, hvorledes taleforståeligheden påvirkes af forskellige binaurale støj reduktions-teknikker, såsom "beamforming", samt forskellige binaurale kompressionsstrategier, som ofte anvendes i høreapparater.

Resumé

L'intelligibilité de la parole est une mesure de sa compréhension dans une situation donnée. Les modèles de prédiction de l'intelligibilité permettent de mieux comprendre quelles sont les propriétés essentielles des signaux de la parole, comment ces propriétés sont extraites par le système auditif, et comment elles sont combinées, utilisés et comprises. Cette dissertation est basée sur le"modèle multi-résolution de spectre de puissance d'envelope basé sur parole" (multi-resolution envelope power spectrum model, sEPSM). Le sEPSM utilise le rapport signal-sur-bruit dans le domaine de la puissance d'envelope (SNR_{env}) comme métrique de décision. Le sEPSM est analysé et comparé à d'autres modèles d'intelligibilité qui utilisent différents préprocesseurs ou différentes métriques de décision, telle que le rapport signal-sur-bruit dans le domaine audio (S/B). Le but est d'identifier les éléments essentiels de ces modèles pour une sélection de conditions exigeantes.

Dans une de ces conditions, la parole est déformée par un processus de gigue de phase qui détruit l'intégrité spectrale du signal. Il est démontré que le sEPSM peut expliquer l'effet négatif sur l'intelligibilité du processus de gigue de phase en incluant une mesure de la variabilité de la puissance d'envelope en fonction des fréquences audio. Dans une autre condition avec une distortion non-linéaire induite par une réduction de bruit via soustraction spectrale, il est démontré que le processus trans-fréquenciel n'est pas essentiel.

De plus, un modèle quantitatif est développé pour tenter de prédire l'intelligibilité de la parole mesurée en écoute binaurale et monaurale. Le modèle binaural, appelé le B-sEPSM, consiste en un exemplaire du sEPSM pour chaque voie monaurale, combiné avec un processus d'égalisation–cancellation (EC) pour modéliser le démasquage binaural. Le sEPSM est aussi appliqué à la sortie du processus EC, de sorte que les trois sont soient directement comparables. Il est démontré que le B-sEPSM peut prédire l'intelligibilité en fonction du nombre de sources de bruit, de l'azimuth des sources, des propriété de la pièce (anéchoïque ou réverbérante), du type de bruit (bruit stationnaire, bruit modulé en amplitude, ou parole inversée temporellement), et de la différence de délai interaural entre la cible et les sources de bruit.

Finalement, les résultats d'autres simulations démontrent qu'un processus binaural n'est pas toujours nécessaire dans des conditions où la cible et les sources de bruit sont séparés spatialement, et que le SNR_{env} peut capter certains aspects du masquage qui ne sont pas captés par les modèles qui utilise le S/B audio comme métrique de décision. Toutefois, aucun des modèles

considérés ne peut expliquer l'intelligibilité mesurée dans des conditions où le "masquage informationnel" est présent, parce qu'ils ne prennent pas en compte les confusions subites par les auditeurs dans leur procédé de prise de décision. Une méthode pour estimer ces confusions, basée sur une mesure de "distance" entre le spectre de puissance d'envelope d'une estimation de la parole et du bruit, est suggérée.

Dans l'ensemble, les résultats de cette dissertation supportent l'hypothèse que le SNR_{env} est une métrique fiable pour la prédiction de l'intelligibilité de la parole. De plus, le B-sEPSM peut être utilisé pour investiguer l'effet de techniques binaurales de réduction du bruit et de différentes stratégies de compression binaurales dans les appareils auditifs.

Acknowledgments

I would like to thank my supervisor, Torsten Dau, who never allows a corner to be cut. He's right: it's the people who matter. Ewen N. MacDonald for his always insightful comments and his invaluable help when I was lost. He never mistakes the tree for the forest. Søren Jørgensen for his brilliant, albeit sometimes frustrating (!), work on which a lot of this thesis is based. I'm not sure I would have finished that first paper without him. Anna Dabrowska for doubling my happiness and halving my sorrows. And finally, everyone at the Hearing Systems group for their support and sense of humor. Without *y'all*, this whole experience would not have been memorable and life changing.

Related publications

Journal papers

- Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (**2014**). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction," J. Acoust. Soc. Am. **135**, 3502–3512
- Chabot-Leclerc, A., MacDonal, E. N., and Dau, T. (**2016**). "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain," J. Acoust. Soc. Am., **140**, 192–205.

Conference papers

- Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (**2013**). "The role of across-frequency envelope processing for speech intelligibility," Proc. of Meetings on Acoust., International Congress on Acoustics. 060128.
- Chabot-Leclerc, A. and Dau, T. (**2014**). "Predicting speech release from masking through spatial separation in distance," Proc. 7th Forum Acust. SS16_17.
- Chabot-Leclerc, A., MacDonald, E. N., and Dau, T. (**2015**). "Predicting masking release of lateralized speech," Proc. of the 5th International Symposium on Auditory and Audiological Research, ISAAR, Submitted.

Contents

1

General introduction

"People don't talk like this, theytalklikethis. Syllables, words, sentences run together like a watercolor left in the rain. To understand what anyone is saying to us we must separate these noises into words and the words into sentences so that we might in our turn issue a stream of mixed sounds in response. If what we say is suitably apt and amusing, the listener will show his delight by emitting a series of uncontrolled high-pitched noises, accompanied by sharp intakes of breath of the sort normally associated with a seizure or heart failure. And by these means we converse. Talking, when you think about it, is a very strange business indeed." — Bill Bryson[1]

1.1 Motivation and background

Research in human speech communication often studies speech sounds as a function of their intelligibility in different conditions. In this case, speech intelligibility is a measure of how well a speech sound is understood by a listener. For some tasks, using human listeners to test intelligibility can be costly or impossible. In these cases, computational models of speech intelligibility can be used as design aids that attempt to predict human speech understanding. For example, they can be used as validation tools when designing signal processing algorithms—validating an algorithm against a model is both cheaper and faster than running an experiment with listeners. They are also used in room design, where they allow acousticians and architects to predict what intelligibility will be in the real room, given a computer model of the room. Models are an essential part of the scientific method. They are used to validate hypotheses about the functioning of the auditory system, to represent complex processes, and to generate ideas of how the hearing system actually performs certain actions

[1] From *The Mother Tongue: English and How It Got That Way* (1991), William Morrow Paperbacks.

(Dau, 2008). For example, Jeffress (1948) used a model to suggest that acoustic timing differences between binaural signals could be captured using delay lines in the neural auditory pathway. Signals coming from both ears would branch out along paths of different lengths and meet at an array of neurons. Only the neurons where both signals arrived at the same time would fire, which therefore maps the interaural delay along one dimension of neurons. Later psychophysical and physiological work, summarized by Joris et al. (1998), showed that neurons in the auditory brainstem actually performed in a similar way as suggested by Jeffress. In this case, real-life systems were identified *after* a model suggested their existence. The models developed and evaluated in this thesis are built as research tools. They aim to analyze and validate hypotheses about the functioning of the hearing system. Speech intelligibility models evolved over time, both in complexity and in accuracy. An overview of this evolution is presented here, as well as a framework with which to reason about the structure of the models.

1.2 Speech intelligibility frameworks

Intelligibility models can be used to predict speech performance at different levels of precision. Models that predict confusions are labelled as "microscopic" intelligibility models. For example, the models of Zaar and Dau (2015) or of Cooke (2006) attempt to predict consonant confusions. In contrast, the models considered in this thesis are qualified as "macroscopic". They attempt to predict human performance in more general terms, usually as a proportion of correctly understood words or sentences.

In order to make their prediction, models evaluate features extracted from the input signals. Figure 1.1 shows the basic structure of the intelligibility models considered in this thesis. Input signals are processed by a "front-end", or "preprocessing", which extracts a set of "features" from the input signals, and yields an internal representation. This internal representation is then analyzed by a "back-end" which is often split between a "decision metric" that evaluates the intelligibility of the processed signals, and a conversion to intelligibility. In many cases, this conversion takes the form of a psychometric function-like mapping of the model output to intelligibility.

The decision metric dictates the inputs required by the model. In some rare cases, such as in Falk et al. (2010), the model requires only access to the

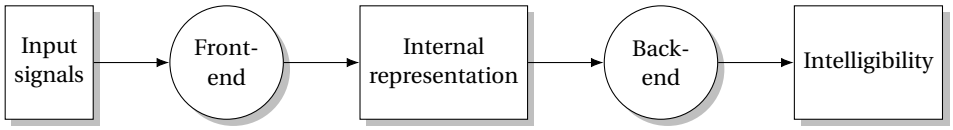


Figure 1.1: Structure of a general speech intelligibility model. Rectangular nodes represent input and output signals and circular nodes represent processes.

processed speech signal as input, in which case it is said to be "non-intrusive". In most cases, the models are "intrusive" and the decision metric requires two different inputs to make a comparison. The pairs of inputs can be the clean speech and the masker alone (S & N), the clean speech and the noisy speech (S & $(S+N)$), or the noisy speech and the masker alone ($(S+N)$ & N). Models that require speech and masker separately typically use a metric based on audibility, measured as the signal-to-noise ratio (SNR). The SNR is a metric designed to quantify the amount of masking produced by one signal against another (Fletcher, 1923). The masking effect captured by the SNR is typically described as energetic masking (EM), i.e., when the neural response produced by a noisy speech signal is nearly indistinguishable from the response of the masker alone (Moore, 2012). French and Steinberg (1947) formalized the idea of using the audio SNR as the decision metric under a model named the articulation index (AI). The AI is calculated as the (weighted) SNR at the output of a filterbank mimicking the frequency selectivity of the peripheral auditory system. This approach was very successful in predicting the intelligibility of band-limited speech (Kryter, 1962) and was instrumental in the development of the telephone system. The AI was later extended as the speech intelligibility index (SII; ANSI, 1997), which introduced additional parameters to account for different types of speech materials, as well as updated weights for the relative importance of various frequencies to speech intelligibility.

Although SNR-based models were very successful, they failed in reverberant conditions, e.g., in a quiet church where the SNR is positive but speech understanding is nonetheless impaired. To address this limitation, Houtgast and Steeneken (1973a) suggested a metric named the modulation transfer function (MTF), which requires both the clean and the noisy signal inputs. The MTF evaluates the reduction in modulation depth of a probe signal (or of speech, Payton and Braida, 1999) after processing, at the output of a modulation filterbank. The MTF concept required the addition of a dimension beyond the audio

spectrum. It was used in the speech transmission index (STI; Steeneken and Houtgast, 1980; IEC, 2003) framework, where the reduction in the modulation depth is converted to an equivalent SNR; the idea being that reverberation fills the gaps in the probe signal in the same way stationary noise does. Elhilali et al. (2003) proposed a model based on the same MTF back-end but replaced the temporal modulation front-end with a model that represents the spectral *and* temporal modulations in the auditory system (Chi et al., 1999). Their spectro-temporal modulation index (STMI) model captures these modulations using two-dimensional filters (spectro-temporal receptive fields, STRF) that have the form of a Gabor function. The STMI model was shown to account for the intelligibility of speech corrupted by a phase jitter process, as well as by linear phase shifts, and reverberation. The STI and STMI models are analyzed and discussed further in Chap. 2.

Jørgensen and Dau (2011) introduced the speech-based envelope power spectrum model (sEPSM) framework, based on the work of Dubbelboer and Houtgast (2008) and Ewert and Dau (2000), who suggested the concept of SNR in the modulation domain. The sEPSM captures the SNR in the envelope power domain ($\mathrm{SNR}_{\mathrm{env}}$) at the output of a modulation filterbank. The $\mathrm{SNR}_{\mathrm{env}}$ metric assumes that an estimate of the speech-alone modulations, $\hat{S}$, can be made based on the modulation spectrum of the noisy speech, $S+N$, and the noise alone, N, i.e., $\hat{S}=(S+N)-N$. The sEPSM was shown to account for the intelligibility of speech mixed with speech-shaped noise (SSN) as well as for the effects of reverberation and of noise reduction via spectral subtraction, unlike the SII and STI (Dubbelboer and Houtgast, 2007; Hilkhuysen et al., 2014). However, the sEPSM fails in conditions where the spectral integrity of the speech is compromised, such as in the case of a phase jitter process (Chabot-Leclerc et al., 2014, Chap. 2). In a series of studies, Stone et al. demonstrated that noises typically considered as "steady", such as SSN or white noise, provide more modulation masking (MM) than EM (Stone et al., 2011; Stone et al., 2012; Stone and Moore, 2014). This demonstration highlights the relevance of the $\mathrm{SNR}_{\mathrm{env}}$ metric for intelligibility prediction, which considers MM rather than EM. Additionally, Stone et al. (2012) showed that masking release produced by fluctuations of the masker's amplitude reflects a release from MM rather than a release from EM. Aspects of EM and MM are covered in Chap. 3 and Chap. 4.

Although all the models mentioned above perform well in the conditions in which they were validated, none of them can account for the masking release

(MR) due to fluctuations in the masker. In the case of the STI, the model is limited by the fact that all modulation information must come from the probe signal. In the case of the sEPSM, a modulated masker would yield a smaller SNR_{env} and thus a lower intelligibility. AI-based models fail to capture the gaps present in fluctuating maskers which provide "glimpses" to the listener. The approach taken in all modeling frameworks has been to introduce a higher temporal sensitivity by applying each model's metric on a short-time basis. Rhebergen and Versfeld (2005) proposed the extended SII (ESII), and Jørgensen et al. (2013) proposed the multi-resolution sEPSM (mr-sEPSM). Both models share the feature that the duration of the short-term window is not fixed. In the ESII, the duration of the windows depends on the center frequency of the peripheral filters and ranges from 35 ms for the filter centered at 150 Hz to 9.4 ms for the filter centered at 8 kHz. In the mr-sEPSM, the duration of the windows depends on the center frequency of the modulation filters and is equal to the inverse of the modulation channel's center frequency.

The AI, SII, STI, STMI, and sEPSM mentioned above all use energy, either in the audio or in the modulation domain, as part of their decision metric. In contrast, there is another class of models that considers the similarity between a test signal and a reference signal (often clean speech) as their decision metric. Kates and Arehart (2005) proposed an extension of the SII, the CSII, which replaces the SNR calculation with a calculation of the long-term coherence between the clean speech and the noisy speech. The CSII could account for the effect of additive noise, peak clipping, and center clipping. The model of Christiansen et al. (2010) calculates a cross-correlation operation at the output of a model of auditory preprocessing (Dau et al., 1997a) in 20 ms windows. Their model could account for the intelligibility of time-frequency weighted noisy speech, i.e., ideal binary mask (IBM) processing, as well as noise vocoding. Taal et al. (2011) proposed the short-time objective intelligibility measure (STOI) to account for the intelligibility of IBM processing. STOI also considers the cross-correlation between the clean and degraded speech, but at the output of a discrete Fourier transform (DFT) front-end and using longer time windows of about 400 ms. Although the CSII, STOI, and the model of Christiansen et al. (2010) account well for different types of non-linear processing, they fail to account for reverberation or have never been tested in this condition.

1.3 Binaural models of speech intelligibility

All intelligibility models mentioned above only considered monaural processing. However, listening with two ears can have an effect as large as 13 dB in some conditions (Marrone et al., 2008; Bronkhorst, 2000). The source of this advantage is typically split into two components, one denoted as a "better-ear" (BE) and an other denoted as "binaural unmasking" (BU). The BE concept assumes that the hearing systems can select only the signal at the ear with the most favorable SNR in a given situation. This advantageous SNR is characterized by the interaural level difference (ILD) between the ears caused by the shadow cast by the head. However, a binaural advantage can also be observed in conditions without ILDs (Licklider, 1948), in which case the benefit is attributed to the BU process and its use of the interaural time differences (ITDs) to "cancel" some of the noise (Durlach, 1963). A prominent example of the existence of the BU process in speech intelligibility is the 6 dB binaural intelligibility level difference (BILD; Levitt and Rabiner, 1967a; Johansson and Arlinger, 2002) measured in a condition where speech is presented out of phase between the two ears together with an in-phase noise masker, compared to the condition where both signals are presented in phase. A similar and even larger effect, of up to 15 dB, is observed when the target is a tone, in which case the difference is denoted as binaural masking level difference (BMLD; Licklider, 1948; Durlach, 1963). In other words, the BE process is a *selection* process, whereas the BU process involves an *interaction* between the two ear signals and is sometimes denoted as "pure" binaural processing. Even though multiple experiments (Bronkhorst and Plomp, 1988; Culling et al., 2004; Lavandier and Culling, 2010; Best et al., 2013) have demonstrated the presence and supported the explanation of those two mechanism, there is no clear explanation as whether both of them are continually in use, or if one of the two is necessary or dominant in a given condition (See Chap. 4).

Binaural intelligibility models typically consist of a combination of a monaural model and a model used to predict BMLDs. Beutelmann and Brand (2006), for example, proposed a model that combines the SII with the equalization–cancellation (EC) model of Durlach (1963) that was later simplified and extended in Beutelmann et al. (2010). The extended model, the binaural speech intelligibility model (BSIM), provided an analytical expression to predict intelligibility and was shown to account for the intelligibility obtained with interferers pre-

sented in a reverberant environment and at different azimuths. Lavandier and Culling (2010), in contrast, considered the SII and an analytical prediction of the BMLD (Culling et al., 2005), in which case the predicted BMLD is *added* to the BE benefit. Their model could account for the intelligibility of speech mixed with noise in a reverberant environment for different target and masker locations. Wan et al. (2010) proposed the steady-state EC (SSEC) model. It is very similar to the model of Beutelmann et al. (2010), but differs in the fact that the efficiency of the EC process is limited by a temporal and level jitter, rather than by the addition of a Gaussian noise. The SSEC could account for intelligibility in anechoic conditions for different masker types, different number of maskers, and for different spatial location of maskers (Hawley et al., 2004; Marrone et al., 2008). Van Wijngaarden and Drullman (2008) proposed a binaural extension of the STI which included binaural interactions in the form of interaural cross correlation (Jeffress, 1948).

Similarly to monaural models, binaural models have been extended to predict the intelligibility in conditions with one or multiple fluctuating maskers. In the short-time BSIM (stBSIM; Beutelmann et al., 2010), the (long-time) BSIM is applied in 20 ms windows, and in the short-time EC model (STEC; Wan et al., 2014), the model of Wan et al. (2010) is also applied in 20 ms windows. In this thesis, a binaural extension of the mr-sEPSM is proposed (Chap. 4), which incorporates the EC model of Durlach (1963), as used in the STEC (Wan et al., 2014) to account for binaural intelligibility benefits.

1.4 Overview of the thesis

In this thesis, several models are compared according to the modeling framework described in Fig. 1.1, with the aim of identifying the contributions of the different front-ends, back-ends, and of short- *vs.* long-term processing to speech intelligibility prediction. This thesis is composed of four chapters, some of which are based on already published manuscripts, as indicated in each chapter.

Chapter 2 analyzes the role of across-channel processing and of the decision metric for speech intelligibility prediction. Predictions by the sEPSM are compared to two modified versions of the sEPSM, as well as to the spectro-temporal modulation index (STMI) model by Elhilali et al. (2003), which uses the MTF as the decision metric. One version uses a similar two-dimensional

modulation filtering stage as the STMI and keeps the SNR_{env} as the decision metric. The second version keeps the one-dimensional modulation filtering of the original sEPSM, but introduces an across (peripheral) audio-frequency mechanism, inspired by models of comodulation masking release (CMR; Dau et al., 2013; Piechowiak et al., 2007). The models are evaluated using data measured in conditions where speech is presented against speech-shaped noise (SSN) and further processed by either reverberation, noise reduction via spectral subtraction, or a phase jitter process.

In *Chapter 3*, the SRT data measured by Westermann and Buchholz (2015a) are used to study the role of binaural processing, the decision metric and short- *vs.* long-term processing in a condition where spatial release from masking is created by moving a masker, on axis, away from a listener. Predictions by the (long-term) sEPSM and by the mr-sEPSM (Jørgensen et al., 2013), which both use the SNR_{env} metric, are compared to predictions by the binaural model of Jelfs et al. (2011) (a reimplementation of the model of Lavandier and Culling, 2010), the binaural speech intelligibility model (BSIM; Beutelmann et al., 2010), and the (short-time) ESII (Rhebergen et al., 2005), all three of which use audibility as the decision metric.

Chapter 4 introduces a binaural extension of the mr-sEPSM, named the B-sEPSM. The "better-ear" process is modeled as two realizations of the mr-sEPSM and the binaural unmasking consists of an EC process as suggested by Wan et al. (2014). A selection mechanism then selects the best of the BE and BU SNR_{env} in the multi-resolution time-scale. The model is evaluated against three experiments from the literature that systematically investigate intelligibility as a function of the number of maskers, the azimuth of the maskers, the room properties (anechoic and reverberant), the masker types, and the ITD of the target and masker.

Chapter 5 further tests the binaural model presented in Chap. 4 using data measured by Lőcsei et al. (2015). The model is tested in ITD-only conditions with masker types that produce different amounts of energetic, modulation, and so-called "informational masking". This chapter investigates the possibility of estimating the confusions that cause parts of the informational masking using the multi-resolution internal representations in the mr-sEPSM.

Finally, *Chapter 6* summarizes the main findings and discusses the limitations and perspectives of the proposed modeling approaches.

2

The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction[a]

Abstract

Speech intelligibility models typically consist of a preprocessing part that transforms stimuli into some internal (auditory) representation and a decision metric that relates the internal representation to speech intelligibility. The present study analyzed the role of modulation filtering in the preprocessing of different speech intelligibility models by comparing predictions from models that either assume a spectro-temporal (i.e., two-dimensional) or a temporal-only (i.e., one-dimensional) modulation filterbank. Furthermore, the role of the decision metric for speech intelligibility was investigated by comparing predictions from models based on the signal-to-noise envelope power ratio, SNR_{env}, and the modulation transfer function, MTF. The models were evaluated in conditions of noisy speech (1) subjected to reverberation, (2) distorted by phase jitter, or (3) processed by noise reduction via spectral subtraction. The results suggested that a decision metric based on the SNR_{env} may provide a more general basis for predicting speech intelligibility than a metric based on the MTF. Moreover, the one-dimensional modulation filtering process was found to be sufficient to account for the data when combined with a measure of across (audio) frequency variability at the output of the auditory preprocessing. A complex spectro-temporal modulation filterbank might therefore

[a] This chapter is based on Chabot-Leclerc et al. (2014).

not be required for speech intelligibility prediction.

2.1 Introduction

Early models of speech intelligibility, such as the articulation index (AI; French and Steinberg, 1947), consider the effects of energetic masking as the main factor influencing the intelligibility of speech presented in background noise. The decision metric employed by the AI, i.e., the measure used to quantify the effects of the transmission channel on speech intelligibility, mainly considers the *audibility* of the speech, quantified by a weighted average of the signal-to-noise ratios (SNRs) measured in frequency bands covering the speech spectrum.The AI has been demonstrated to account well for conditions with static interferers, like additive noise (French and Steinberg, 1947), and for conditions with spectrally filtered speech (Kryter, 1962). However, it fails in conditions with temporal distortions, such as reverberation, because it does not consider the modifications to the temporal envelope of the (speech) signal.

In contrast, the speech transmission index (STI; Houtgast et al., 1980; Steeneken and Houtgast, 1980; IEC, 2003) considers the integrity of the temporal envelope fluctuations of a reference signal in the decision metric, quantified by the modulation transfer function (MTF), which was included in a revised version of the AI, the speech intelligibility index (SII; Pavlovic, 1987; ANSI, 1997). The MTF measures the reduction of the envelope fluctuations of a target reference signal as the ratio between the modulation magnitude spectrum of the processed reference signal and that of the clean reference signal, for a number of audio frequencies. The MTF therefore captures the effects of the distortions on the envelope of the reference signal, and leads the STI to account for speech intelligibility in reverberant conditions as well as when the speech is presented in a stationary background noise (Houtgast et al., 1980; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). However, the STI fails in conditions with nonlinear processing, such as envelope compression (Rhebergen and Versfeld, 2005), phase jitter, phase shifts (Elhilali et al., 2003), or spectral subtraction (Ludvigsen et al., 1993; Dubbelboer and Houtgast, 2007).

To overcome this limitation, Payton and Braida (1999) as well as Goldsworthy and Greenberg (2004) introduced modifications of the STI, generally referred to as speech-based speech transmission index (sSTI) methods. The main difference from the original STI method is that speech is used as the reference

signal rather than a modulated wideband noise, and that the integrity of the temporal envelope is quantified by other metrics than the MTF. Although the sSTI methods seemed promising, they have never been evaluated with quantitative comparisons between measured and predicted speech intelligibility data. Moreover, Dubbelboer and Houtgast (2008) proposed that, in the case of noise reduction via spectral subtraction, the MTF-concept was inherently limited because it compares the clean reference signal to the processed signal and thereby neglects the effects of the intrinsic modulations in the noise itself on speech intelligibility. Jørgensen and Dau (2011) supported this view by showing that the inherent modulations in the noise-alone envelope can, in some conditions, be enhanced to a larger degree by the spectral subtraction process than the modulations in the noisy speech, which decreases intelligibility.

An alternative approach was taken by Elhilali et al. (2003), who predicted intelligibility based on the spectro-temporal modulation index (STMI). The STMI measures the integrity of the spectral and temporal modulation energy of a signal, inspired by neural responses to spectro-temporally varying stimuli in the auditory cortex of ferrets (Depireux et al., 2001; Kowalski et al., 1996). This concept thus differs from the STI, which considers the energy of the modulations only in the temporal domain. The STMI considers a two-dimensional (2D) (spectro-temporal) MTF as the decision metric, effectively assuming a spectro-temporal modulation bandpass filterbank. Elhilali et al. (2003) defined two versions of the STMI. One version used a spectro-temporally modulated noise as the reference signal, denoted as a ripple, analogous to the temporally modulated noise in the case of the STI. The second version used clean speech as the reference signal, as in the sSTI methods. The ripple-based and speech-based STMI, respectively denoted as STMI^{R} and STMI^{T}, were shown to be consistent with the STI in conditions with additive noise and reverberation. Furthermore, both STMI versions could account for the nonlinear distortion effects due to phase jittering and phase shifts, to which the STI is insensitive. The key component in the STMI to account for the phase distortions was assumed to be the processing across the frequency axis, i.e., the evaluation of the integrity of the spectral modulations in the speech signal. However, since the STMI is still based on the MTF concept, it should have the same limitations as the STI when noisy speech is processed by spectral subtraction, because the MTF does not capture the effect of the processing on the intrinsic noise modulations (Dubbelboer and Houtgast, 2008).

Recently, Jørgensen and Dau (2011) proposed the signal-to-noise envelope power ratio (SNR_{env}) as an alternative decision metric, inspired by the work of Dubbelboer and Houtgast (2007). Instead of measuring the reduction of the clean speech envelope modulation power, as done by the MTF, the SNR_{env} estimates the ratio of speech and noise envelope modulation power. The SNR_{env} was implemented in the speech-based envelope power spectrum model (sEPSM) and is estimated at the output of a temporal modulation filterbank. The SNR_{env} was shown to be consistent with the STI in conditions with additive noise and reverberation, while also accounting for the effect of spectral subtraction. The key component allowing the SNR_{env} to account for spectral subtraction is the consideration of the intrinsic modulations of the noise (alone). The power of these modulations is typically *increased* as a consequence of the noise reduction processing which leads to a masking effect on speech in the modulation domain. This effect is neglected in the MTF concept. However, the sEPSM can be expected to fail in conditions with distortions that affect the *spectral* structure of the signal (e.g., the spectral peaks representing the speech formants) since the model does not assume any explicit across-frequency processing besides simple information integration.

Thus, conceptually, the STMI and the sEPSM introduced different modifications to the STI: The STMI introduced an across-frequency mechanism via a *spectro-temporal* modulation filterbank that seems important for the prediction of phase jitter effects, but kept the MTF-based decision metric. The sEPSM introduced another decision metric, based on the SNR_{env}, which seems important for the prediction of effects of spectral subtraction, but kept the analysis of only *temporal* modulations while neglecting across-frequency effects of a given distortion on the modulation pattern of the stimuli.

The present study investigated if the combination of the two models would provide a more general, and thus more powerful, modeling framework for predicting speech intelligibility. Two model realizations were considered, both based on the sEPSM structure from Jørgensen and Dau (2011) and thus employing the SNR_{env} metric. One realization replaced the temporal modulation filterbank by a two-dimensional spectro-temporal modulation filterbank, as in the STMI, denoted in the following as "2D-sEPSM." The other realization kept the purely temporal [one-dimensional (1D)] modulation filterbank and introduced a mechanism that measured the variance of the outputs of this modulation filterbank *across* peripheral channels, denoted in the following as

"sEPSM^X." In this model, the contribution to intelligibility from a given modulation channel was assumed to be proportional to the amount of the variance across peripheral channels for that particular modulation channel. Such a mechanism was inspired by models of co-modulation masking release (CMR; e.g., van de Par and Kohlrausch, 1998; Piechowiak et al., 2007; Dau et al., 2013). CMR refers to the greater detectability of a tone centered in a narrow band noise, surrounded by one or more flanking noise bands with co-modulated waveforms, compared to the same situation with uncorrelated flanking noise bands. The addition of a tone in the co-modulated noise bands introduces a decorrelation of the waveforms across frequency, which has been suggested to be a cue for detection (van de Par and Kohlrausch, 1998). An across-channel decorrelation corresponds to an increase in the variation across frequency bands. However, in contrast to the synthetic stimuli considered in CMR experiments, natural speech contains a highly variable pattern of spectro-temporal fluctuations in an auditory spectrogram-like representation, reflected by a large across-channel variance. Distortions that would decrease the across-frequency variance would thus reflect a degradation of the speech representation.

The two models were evaluated in conditions of reverberation, spectral subtraction, and phase jitter processing and compared to predictions obtained with the STMI^T (Elhilali et al., 2003) and the original sEPSM (Jørgensen and Dau, 2011).

2.2 Model descriptions

Figure 2.1 shows a sketch of the overall structure of the model(s) considered in the present study. The first three stages represent the auditory "preprocessing," consisting of a gammatone filterbank, an envelope extraction process, and a modulation filtering process, which are specific to each of the two model realizations. An absolute sensitivity threshold is included in both models, such that only peripheral filters with output energy above the normal hearing threshold are considered. The two final stages indicated in Fig. 2.1 represent the decision module, consisting of the SNR_{env} calculation and an "ideal observer," as defined in Jørgensen and Dau (2011).

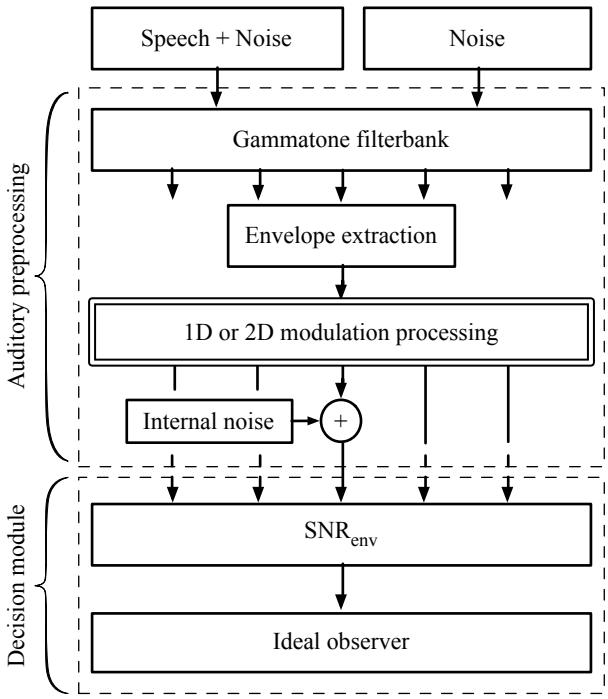


Figure 2.1: Block diagram of the overall structure of the modeling framework. The model consists of a gammatone bandpass filterbank followed by envelope extraction via the Hilbert transform, producing the "auditory spectrogram" of the incoming signal. The modulation filtering differs in the two considered model versions, the 2D-sEPSM and the (1D-) sEPSM^X. The SNR_{env} is calculated from the envelope power spectra at the output of either the two-dimension modulation filtering (2D-sEPSM) or the 1D modulation filtering (sEPSM^X). The resulting values are combined across modulation filters and audio filters. The overall SNR_{env} is converted to a percentage of correctly recognized speech items using an ideal observer process.

2.2.1 Model 1: 2D envelope power spectrum model (2D-sEPSM)

In the 2D-sEPSM, the acoustic signal is filtered using a bandpass filterbank consisting of 128 fourth-order gammatone filters equally spaced on a logarithmic scale between 90Hz and 3.5kHz (24 filters/octave over a 5.3 octave range). The envelope of the output of each gammatone filter is extracted using the Hilbert transform, low-pass filtered using a first-order Butterworth filter with a cutoff frequency of 150 Hz (Ewert and Dau, 2000; Kohlrausch et al., 2000), and short-term averaged in blocks of 8 ms to form an "auditory spectrogram"-like representation.[a] Next, the joint spectral and temporal modulation content is ex-

[a] Short-term averaging is used as a single operation to low-pass filter and down-sample the signal to reduce computation time. The resulting sampling rate of 125 Hz is sufficiently large to not affect the highest temporal modulation filter centered at 32 Hz.

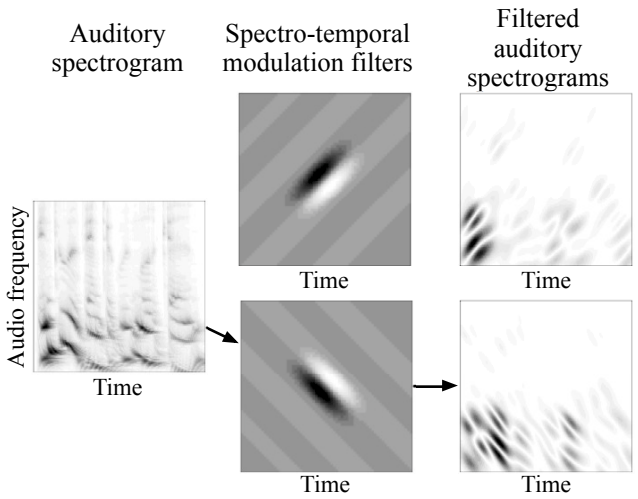


Figure 2.2: Illustration of the spectro-temporal modulation filtering of the auditory spectrogram. The first column shows an auditory spectrogram. The center column shows two examples of spectro-temporal modulation filters tuned to different temporal modulation frequencies, spectral modulation frequencies, and directions. The third column represents the output of the filtering of the auditory spectro-temporal by the respective spectro-temporal modulation filter.

tracted from the auditory spectrogram using a bank of spectrally and temporally selective modulation filters (Chi et al., 1999). The outputs of this processing, the filtered spectro-temporal envelopes, have the four dimensions of time, cochlear frequency, temporal modulation frequency, and spectral modulation frequency. Figure 2.2 shows two examples of spectro-temporal modulation filters with their respective filtered auditory spectrograms. The 2D modulation filters are third-octave wide, octave-spaced, and tuned ($Q = 1$) to a range of temporal modulations frequencies (ω) between 2 and 32 Hz and spectral modulations frequencies (Ω) between 0.25 and 8 cycles/octave. The impulse responses of the 2D modulation filters have the form of a Gabor function. Detailed information on the 2D filtering stage can be found in Chi et al. (1999) and Chi et al. (2005).

In the decision device, the long-term envelope power of the filtered spectro-temporal envelope is calculated as the variance[b] across time, leaving a three-dimensional internal representation of the noisy speech mixture, $T_{\text{mix}}(f, \omega, \Omega)$,

[b] The variance across time is a measure of the "ac-coupled" power, i.e., the power related to the fluctuations around the mean amplitude. The power of signals with finite energy is expressed as the mean-squared operation, which is mathematically different from the variance. However, in the special case of bandpass filters, as in the modulation bandpass filterbank, the output signals have zero mean, which means that the variance and the mean-squared operations are identical.

and of the noise alone, $N(f,\omega,\Omega)$.[c] The internal representation is then normalized with the squared time-average of the auditory spectrogram to assert that the envelope power is independent of overall stimulus level. The $\mathrm{SNR}_{\mathrm{env}}$ can thereafter be expressed as

$$\mathrm{SNR}_{\mathrm{env}} = \frac{T_{\mathrm{mix}} - N}{N}. \tag{2.1}$$

It is assumed that the envelope power of the mixture does not exceed the envelope power of the clean speech, as performed in the $\mathrm{STMI}^{\mathrm{T}}$,

$$T_{\mathrm{mix}} = \min(T_{\mathrm{mix}}, T_{\mathrm{clean}}), \tag{2.2}$$

and that the envelope power of the noise is never larger than the envelope power of the mixture:

$$N = \min(N, T_{\mathrm{mix}}). \tag{2.3}$$

The lower limits of T_{mix} and N are represented by a small positive value ϵ reflecting an internal noise threshold and correspond to the internal noise term indicated in Fig. 2.1:

$$\begin{aligned} T_{\mathrm{mix}} &= \max(T_{\mathrm{mix}}, \epsilon) \\ N &= \max(N, \epsilon), \end{aligned} \tag{2.4}$$

which prevents the numerator of Eq. (2.1) to be zero when $T_{\mathrm{mix}} = N$. ϵ was set to -40 dB.[d]

2.2.2 Model 2: 1D envelope power spectrum model with variance weighting across frequency ($\mathrm{sEPSM}^{\mathrm{X}}$)

The $\mathrm{sEPSM}^{\mathrm{X}}$ assumes 22 gammatone filters with $1/3$-octave spacing of the center frequencies, covering the range from 63 Hz to 8 kHz, as in the original sEPSM. The envelope of the output of each gammatone filter is extracted via the Hilbert

[c] The spectro-temporal envelope power spectra, T and N, are considered equivalent to the temporal envelope power spectrum of the noisy speech, $P_{\mathrm{env,S+N}}$ and of the noise alone, $P_{\mathrm{env,N}}$, as defined in Jørgensen and Dau (2011). It is assumed that the $\mathrm{SNR}_{\mathrm{env}}$ concept is applicable to envelope power spectra, indistinguishably of the number of dimensions.

[d] This threshold corresponds to the value of -20 dB in the original model of Jørgensen and Dau (2011) that considered only temporal modulation channels. The assumption that the internal noise is independent in all (spectral and temporal) modulation channels considered in the 2D-sEPSM leads to the lower value of ϵ.

transform and low-pass filtered using a first-order Butterworth filter with a cutoff frequency of 150 Hz (Kohlrausch et al., 2000). The envelope is analyzed by a filterbank consisting of a third-order low-pass filter in parallel with six overlapping second-order bandpass filters. The cutoff frequency of the low-pass filter is 1 Hz and the bandpass filters have center frequencies from 2 to 64 Hz with octave spacing and a constant Q-factor of 1. Thus, for the sEPSM$^{\mathrm{X}}$, the filtered temporal envelope is a three-dimensional function of time, audio-filter center frequency, and modulation-filter center frequency.

The long-term envelope power, P_{env}, is calculated from the temporal output of each modulation filter as the variance of the filter output across time. The $\mathrm{SNR}_{\mathrm{env}}$ is then calculated from the *normalized* envelope power of the noisy speech, $P_{\mathrm{env,S+N}}$, and the noise alone, $P_{\mathrm{env,N}}$,

$$\mathrm{SNR}_{\mathrm{env}} = \frac{P_{\mathrm{env,S+N}} - P_{\mathrm{env,N}}}{P_{\mathrm{env,N}}}, \tag{2.5}$$

where the normalization factor was the squared time-average of the unfiltered envelopes of the respective stimuli. Similar to the 2D-sEPSM, the model's sensitivity is limited and the envelope powers below −20 dB are set to −20 dB (Jørgensen and Dau, 2011). This sensitivity is expressed as ϵ in Eq. (2.4) and corresponds to the internal noise indicated in Fig. 2.1.

2.2.3 Transformation from $\mathrm{SNR}_{\mathrm{env}}$ to probability of being correct

In both models, the $\mathrm{SNR}_{\mathrm{env}}$ contributions from all G modulation filters and L audio filters are integrated according to

$$\mathrm{SNR}_{\mathrm{env}} = \left[\sum_{g=1}^{G} \sum_{l=1}^{L} (\mathrm{SNR}_{\mathrm{env},g,l})^2 \right]^{1/2}. \tag{2.6}$$

In the case of the sEPSM$^{\mathrm{X}}$, the $\mathrm{SNR}_{\mathrm{env}}$ contribution from the modulation filter g is weighted as follows:

$$\mathrm{SNR}_{\mathrm{env},g,l} = [\sigma_g^2]^{\beta} \cdot \mathrm{SNR}_{\mathrm{env},g,l}, \tag{2.7}$$

where σ_g^2 represents the across-channel variance for modulation filter g, evaluated across all 22 audio filters, and β is a free parameter with a value determined by an optimal fit of the model predictions to the conditions with phase jitter. The

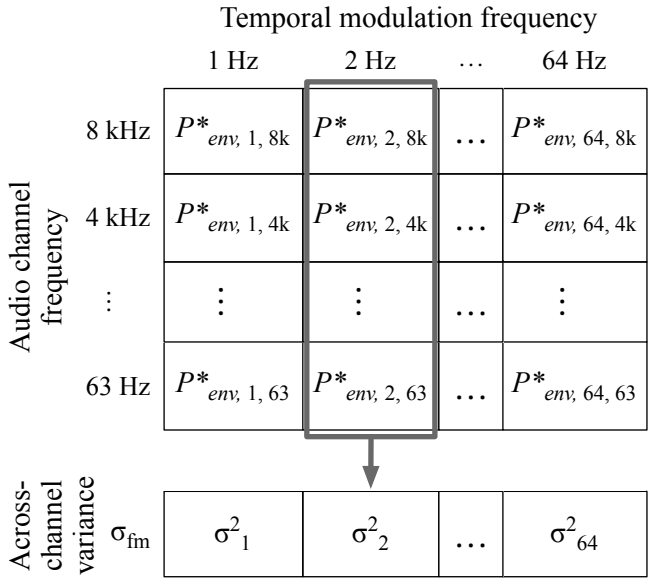


Figure 2.3: Illustration of the calculation of the across-channel variance of the envelope power for a given time frame. Each cell represents the normalized envelope power, $P^*_{\text{env},g,l}$, of the noisy speech at the output of a temporal modulation filter g, and audio filter, l. The across-channel variance, σ_g^2, for a given modulation filter center frequency corresponds to the variance across rows for a given column of the matrix.

value of σ_g^2 was based on several computational steps: First, the long-term envelope power of the noisy speech mixture, $P_{\text{env},g,l}$, was computed at the output of modulation-filter g and audio-filter l and normalized with a factor proportional to the bandwidth in hertz of the audio filter. The proportionality factor was the root-mean-square level of the noisy speech mixture. The normalization ensured that σ_g^2 did not reflect differences in the overall level across peripheral channels that might arise due to greater energy contained in the audio filters with larger bandwidths. Finally, the variance (σ_g^2) of the *normalized* $P^*_{\text{env},g,l}$ was computed across the 22 peripheral filters. Figure 2.3 illustrates the across-channel variance computation using a matrix representation; each row corresponds to a different audio channel and each column represents a modulation channel. In each cell of the matrix, the indices g and l of $P^*_{\text{env},g,l}$ represent the center frequencies of the filters. The across-channel variance is calculated across rows in a given column.

For both models, the overall combined SNR_{env} is converted to a sensitivity index, d', of an "ideal observer" using the relation:

$$d' = k(\text{SNR}_{\text{env}})^q, \tag{2.8}$$

where k and q are parameters independent of the experimental condition. d' is converted to a percentage of correct responses using an m-alternative forced choice (mAFC) decision model (Green and Swets, 1988) combined with an unequal-variance Gaussian model. The ideal observer is assumed to compare the input speech item with m stored alternatives and select the item, x_S, that yields the largest similarity. The $m-1$ remaining items are assumed to be noise, one of which, $x_{N,\max}$, has the largest similarity with the input speech item. The value of x_S is a random variable with a mean of d' and variance σ_S^2. Similarly, the value of $x_{N,\max}$ is a random variable with mean μ_N and variance σ_N^2. The selected item is considered correct if the value of x_S is larger than $x_{N,\max}$. The corresponding probability of being correct is estimated from the difference distributions of x_S and $x_{N,\max}$:

$$P_{\text{correct}}(d') = \Phi\left(\frac{d'-\mu_N}{\sqrt{\sigma_S^2+\sigma_N^2}}\right), \tag{2.9}$$

where Φ designates the cumulative normal distribution. The values of σ_N and μ_N are determined by the number of response alternatives, m. For the open-set paradigm used in the present study the value of m reflects the number of words in a normal listener's active vocabulary, here set to 8000 inspired by Müsch and Buus (2001). σ_S is inversely proportional to the slope of the ideal observer's psychometric function, and reflects different degrees of redundancy in different speech materials. The value of σ_S is a free parameter determined by fitting model predictions to speech intelligibility data in conditions with a stationary speech-shaped noise (see Jørgensen and Dau, 2011).

2.3 Method

2.3.1 Speech and noise material

The target speech was Danish sentences from the Conversation Language Understanding Evaluation (CLUE) test, consisting of unique meaningful five-word sentences (Nielsen and Dau, 2009). The CLUE test is similar to the hearing in noise test (HINT; Nilsson et al., 1994). For all test materials, the sentences were mixed with speech-shaped stationary noise (SSN). The noise was created to match the long-term frequency spectrum of the sentence material. The sentence sound files were randomly concatenated and saved as initial noise file.

The files were then randomized in a new order and added to the noise file. The final noise was the results of 150 superpositions.

2.3.2 Stimuli and experimental conditions

Three conditions of processed noisy speech were considered: Reverberant speech, speech processed by spectral subtraction, which is a form of (single microphone) noise reduction used often in hearing aids and mobile phones, and speech subjected to a phase-jitter distortion, which is a distortion that can be found in telephone channels and is caused by the fluctuations of the power supply voltage (Lee and Messerschmitt, 1994; Bellamy, 2000). In all conditions, sentences were mixed with the stationary speech-shaped noise at a given SNR before processing. In the phase jitter condition, the measured data presented here were collected in the present study. The data in the phase-jitter condition are represented as percent correct because of their non-monotonic shape as a function of the phase jitter parameter. In the reverberation and spectral subtraction conditions, data were taken from Jørgensen and Dau (2011). The data in the reverberation and spectral subtraction conditions are represented as speech reception thresholds (SRTs), representing the SNR corresponding to 50% intelligibility.

Reverberation

The noisy sentences were convolved with impulse responses corresponding to particular reverberation times. The impulse responses were created using the ODEON room acoustic software version 10 (Christensen, 2009). The simulated room was shaped like a rectangular auditorium with maximal dimensions of $28\times 16\times 10$ m (length-width-height). The source and the receiver were horizontally aligned with a fixed distance of 5 m, and placed approximately in the center of the room. All surfaces had the same absorption coefficient, which was adjusted individually across frequency such that the room had similar reverberation times (T_{30}) in the octave bands from 63 to 8000 Hz. Four different values of absorption were considered, such that the impulse responses corresponded to four different values of T_{30}: 0.4, 0.7, 1.3, and 2.3 s. The corresponding acoustic clarity (C50), defined as the ratio of the energy of the first 50 ms of the impulse response to the energy of the remaining part, was 0.6, −2.9, −6.6, and −8.0 dB, respectively.

Spectral subtraction

Sentences were mixed with noise and processed by a spectral subtraction algorithm as described in Jørgensen and Dau (2011), which was similar to that presented by Berouti et al. (1979). An estimate of the noise power spectrum was multiplied with an over-subtraction factor, κ, and subtracted from the power spectrum of the noisy speech in 24-ms time frames with 50% overlap. The estimate of the noise power spectrum was calculated as the mean value of the noise power spectral density. After subtraction, negative values of the noisy speech spectrum were set to zero. Finally, the spectrum was combined with the phase of the original noisy speech and transformed back to the time domain using an overlap-add method. The details on the algorithm can be found in Jørgensen and Dau (2011). Six different over-subtraction factors were considered: 0, 0.5, 1, 2, 4 and 8, where $\kappa = 0$ corresponded to the reference condition with no spectral subtraction.

Phase jitter

Noisy speech distorted by phase-jitter was obtained by multiplying noisy speech with an SNR of 5 dB, $s(t)$, with a cosine function with a random phase, as described in Elhilali et al. (2003):

$$r(t) = R\{s(t)e^{j\Theta(t)}\} = s(t)\cos(\Theta(t)), \tag{2.10}$$

where $\Theta(t)$ is a random process uniformly distributed over $[0, 2\pi\alpha](0 \leq \alpha \leq 1)$, and α is the parameter controlling the amount of jitter. The α-values used covered the range 0 to 1 in steps of 0.125. For $\alpha = 0.5$ and 1, the signal becomes a temporally modulated white noise because each signal sample is multiplied by a random value in the interval -1 to 1. The value of 5 dB SNR was chosen such that noisy unprocessed speech had an intelligibility of 100% (Nielsen and Dau, 2009, Fig. 7).

2.3.3 Apparatus and procedure

For the conditions with phase jitter, the stimuli were stored digitally at a sampling frequency of 44.1 kHz and presented diotically through a pair of calibrated Sennheiser HD580 headphones (Wedemark, Germany) driven by a high-quality soundcard in a double-walled sound-attenuating booth. The setups was cali-

brated using the SSN signal measured with a 1/2 inch microphone and a Brüel & Kjær artificial ear of type 4152 (Nærum, Denmark). The speech had a constant sound pressure level of 65 dB and noise was added to achieve the desired SNR before further processing. Each sentence was presented once with the noise starting 1 s before and ending 600 ms after the sentence; the noise was ramped on and off using 400 ms cosine ramps. Eighteen ten-sentence lists were presented to each listener: Two lists were used for each α value and two sentences per list for each SNR, resulting in 100 data points per condition per listener (2 lists × 10 sentences × 5 words). The lists and SNRs were presented in random order. The training consisted of three lists using α-values of 0, 0.25, and 0.5, presented in a random order. The listeners were asked to repeat the sentence heard and were allowed to guess. No feedback was provided.

2.3.4 Listeners

Measurements were obtained with five normal-hearing listeners (four men, one woman), aged from 21 to 27 years. Their pure-tone thresholds were of 20 dB hearing level or better in the frequency range 0.25 to 8 kHz. All of them were native Danish speakers and students at the Technical University of Denmark. They were paid for their participation.

2.3.5 Model setup and parameters

Predictions were generated using 150 sentences from the CLUE material. The sentences were down-sampled to 8192 Hz for the 2D-sEPSM and to 22050 Hz for the sEPSM^{X} to reduce computation time. The duration of the noise samples was matched to the duration of each sentence and mixed at five SNRs, ranging from −9 to 9 dB in 3 dB steps, except in the phase jitter condition where the SNR was 5 dB only. In all cases, the processing (reverberation, spectral subtraction, or phase jitter process) was applied to both the noisy speech and the noise alone. The percentage of correct responses was obtained for each sentence and for each combination of SNR and distortion parameter (T_{30}, κ or α). The final predictions were calculated as the average across all 150 sentences at a given combination of SNR and distortion parameter. A predicted psychometric function was obtained by connecting predicted responses with straight lines, and the SRT for a specific condition was obtained as the SNR corresponding to 50% intelligibility.

For the sEPSM^X, the values of the parameters k and β were adjusted to minimize the root-mean-square error (RMSE) between the prediction and the measured SRT in the reference condition (speech-shaped noise only), and between the prediction and the data in the condition with phase jitter when α is 1.[e] The values of m, q, and σ_S were taken from Jørgensen and Dau (2011). For the 2D-sEPSM, the parameters k and q were adjusted in the same way. All parameters were then kept fixed in all other experimental conditions and the values are given in Table 2.1.

Table 2.1: Calibrated values of the parameters k, q, β, and of the constants σ_s, m of the ideal observer for the CLUE speech material.

Model	k	q	σ_s	m	β
2D-sEPSM	0.70	0.28	0.6	8000	–
sEPSM^X	0.79	0.5	0.6	8000	0.28
sEPSM	1.1	0.5	0.6	8000	–

2.4 Results

2.4.1 Reverberant speech

Figure 2.4 shows the obtained SRTs as a function of the reverberation time. The open squares represent the measured data from Jørgensen and Dau (2011). The mean SRT in the reference condition without reverberation ($T_{30} = 0$) was obtained at an SNR of −3.5 dB. The vertical bars denote one standard deviation of the listeners' SRT. The measured SRT increased with increasing reverberation time, reflecting a decrease in intelligibility, because of the increasing low-pass effect of reverberation on the temporal modulations (Houtgast and Steeneken, 1985). These data are consistent with Duquesnoy and Plomp (1980). The different filled symbols represent model predictions. The filled black squares show the results obtained with the 2D-sEPSM and the black triangles represent the

[e] The sEPSM^X introduced the across-channel weighting, which included the parameter β, in the model's integration stage to account for the effect of phase jitter on the across-channel structure of the signals. While the weighting coefficient itself is determined from the stimuli as the variance of the envelope power across filters, the value of β had to be determined globally by comparing predictions to the data in a condition of phase jitter. Thus, the parameters k and β were adjusted to obtain the best fit to the reference condition, i.e., with SSN only, and to the condition where the phase jitter parameter equaled 1, i.e., where the speech was completely corrupted by the phase jitter process.

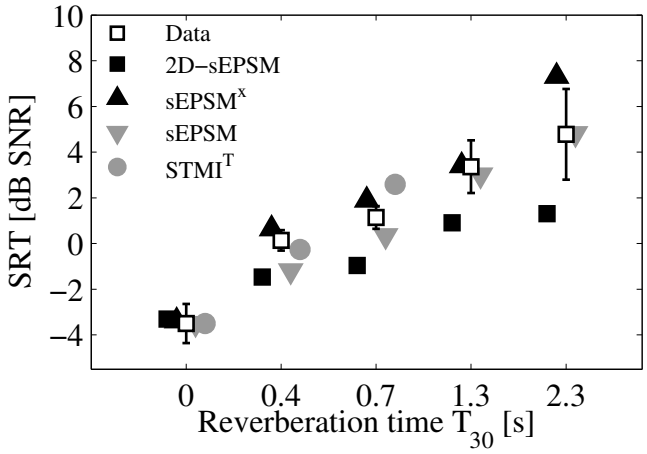


Figure 2.4: SRTs as a function of the reverberation time, T_{30}. The open squares represent a replot of the data in Jørgensen and Dau (2011), with the vertical bars indicating one standard deviation. The filled symbols show predictions obtained with the different models. The black squares and the black triangles indicate predictions obtained with the 2D-sEPSM and the sEPSM^{X}, respectively. In addition, for comparison, the gray triangles show predictions obtained with the original sEPSM without an across-channel process (Jørgensen and Dau, 2011). The gray filled circles represent the predictions obtained with the STMI^{T} (Elhilali et al., 2003). The STMI^{T} prediction did not reach 50% intelligibility when the reverberation time was 1.3 and 2.3 s, therefore the SRT could not be calculated and is not shown.

predictions obtained with the sEPSM^{X}. The Pearson correlation coefficient between the 2D-sEPSM predictions (filled squares) and the measured data was 0.99 and the RMSE was 2.2 dB. In the case of the sEPSM^{X} (black triangles), the Pearson coefficient was 0.98 and the RMSE amounted to 1.2 dB. For comparison, the predictions obtained with the original sEPSM (without any across-frequency process) from Jørgensen and Dau (2011) are also shown, indicated by the filled gray triangles. Furthermore, predictions using the STMI^{T} based on Elhilali et al. (2003) are shown as the gray filled circles. Table 2.2 shows the summary of the RMSE and Pearson correlation coefficients in the three conditions for all models presented. All models could account for the increase of SRT with increasing reverberation. However, the 2D-sEPSM generally underestimated the effect of reverberation by about 2 to 3 dB while the sEPSM^{X} overestimates it by a similar amount for the largest T_{30}.

2.4.2 Spectral subtraction

Figure 2.5 shows the results for the condition with spectral subtraction. SRTs are shown as a function of the over-subtraction factor κ. The measured SRTs, replot-

Table 2.2: Summary of the Pearson correlation coefficients (ρ) and RMSEs between the data and the predictions for the four models presented here, and for the three types of processing applied to noisy speech. The correlation coefficient and RMSE are unavailable for the STMI$^{\mathrm{T}}$ in the reverberation condition because no SRT can be calculated for T_{30} larger than 0.7 s. The correlation coefficient for the sEPSM in the phase jitter condition cannot be calculated.

Model	Reverberation ρ/RMSE	Phase jitter ρ/RMSE	Spectral subtraction ρ/RMSE
2D-sEPSM	0.99/2.2 dB	0.95/24.5%	0.93/1.4 dB
sEPSM$^{\mathrm{X}}$	0.98/1.2 dB	0.99/7.2%	0.99/0.3 dB
sEPSM	0.99/0.7 dB	NA/65.5%	0.99/0.5 dB
STMI$^{\mathrm{T}}$	NA/NA	0.94/19.6%	−0.94/3.6 dB

ted from Jørgensen and Dau (2011), increased with increasing over-subtraction factor. The predicted SRTs obtained with the 2D-sEPSM (black filled squares) and those using the sEPSM$^{\mathrm{X}}$ (black triangles) also increased with κ. The Pearson correlation coefficient between the data and the 2D-sEPSM was 0.93 and the RMSE was 1.4 dB. The sEPSM$^{\mathrm{X}}$ predictions had a correlation with the data of 0.99 and a RMSE of 0.4 dB. For comparison, the predictions using the original sEPSM were replotted from Jørgensen and Dau (2011) and are indicated by the gray triangles. Furthermore, the gray filled circles show the predictions obtained with the STMI$^{\mathrm{T}}$. This model predicted a *decrease* of SRT, i.e., increasing speech intelligibility with increasing κ, in contrast to the measured data.

2.4.3 Phase jitter

The open symbols in Fig. 2.6 show the measured speech intelligibility data collected in the present study, expressed as the percentage of correct words as a function of the phase jitter parameter, α, at a fixed SNR of 5 dB. The vertical bars represent one standard deviation. The intelligibility score showed a characteristic trend as a function of α, with 100% intelligibility for α close to 0, a steep drop of intelligibility down to 0% for $\alpha = 0.5$, followed by a local maximum of about 45% for $\alpha = 0.75$ and, finally, 0% intelligibility for $\alpha = 1$. This trend in the data is consistent with the data presented in Elhilali et al. (2003), although their results did not show a local maximum for $\alpha = 0.75$, most likely because the α-values used were different from the ones used in the present study. The percentage of correctly understood words drops to zero when α is 0.5 or 1 because for those values, the phase jitter process is equivalent to multiplying the speech signal with white noise, yielding speech-modulated white noise. A two-way analysis

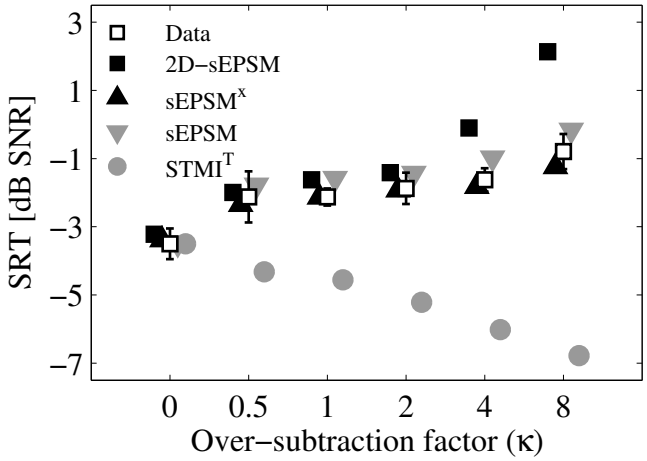


Figure 2.5: SRTs as a function of the over-subtraction factor κ in conditions of spectral subtraction. The open squares represent measured data from Jørgensen and Dau (2011), with the vertical bars indicating one standard deviation. The filled black squares show the predictions obtained with the 2D-sEPSM and the upward triangles represent the results using the sEPSM$^{\mathrm{X}}$. For direct comparison, the filled gray downward triangles show predictions with the original sEPSM (re-plotted from Jørgensen and Dau (2011)). The filled gray circles show predictions obtained with the STMI$^{\mathrm{T}}$ as proposed by Elhilali et al. (2003).

of variance (ANOVA) of the data showed a significant effect of α ($F_{8,44} = 228.7$, $p < 0.001$) but no significant difference between listeners ($F_{4,252} = 3.3$, $p = 0.023$). A *post-hoc* test with Bonferroni correction and with 95% confidence intervals showed that intelligibility percentages $\alpha = 0.375$ and $\alpha = 0.75$ were different from all other values. Two data points are significantly different from each other if they are labeled by different letters indicated the figure.

The filled symbols represent predictions obtained with the different models. The 2D-sEPSM accounted for the main characteristics in the data, with 100% intelligibility below $\alpha = 0.25$, minima at $\alpha = 0.5$ and 1, and a local maximum at $\alpha = 0.75$. However, the predicted intelligibility scores never reached values below 39%. The Pearson correlation coefficient between the data and the 2D-sEPSM was 0.95 and the RMSE was 24.5%. The predictions obtained with the sEPSM$^{\mathrm{X}}$ followed the data more closely than the 2D-sEPSM, with minima of about 1.5% correct responses for $\alpha = 0.5$ and 1 and a local maximum of 35% for $\alpha = 0.75$. The correlation between the sEPSM$^{\mathrm{X}}$ predictions and the data was $\rho = 0.99$, with an RMSE value of 7%. The original sEPSM (downward triangles) without across-channel processing was insensitive to the effects of the phase jitter (similar to the STI as demonstrated in Elhilali et al. (2003)), thus predicting constant speech intelligibility independent of α, in strong contrast to the data. The predictions

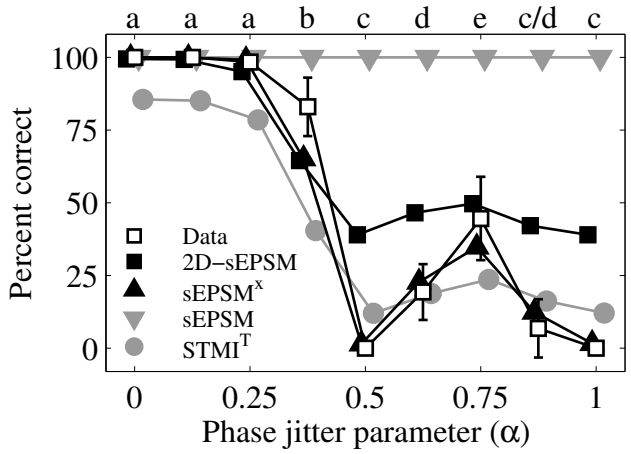


Figure 2.6: Word intelligibility as a function of the phase-jitter parameter α, for a fixed SNR of 5 dB. The open symbols indicate average measured data collected in the present study. Vertical bars show one standard deviation of the average listeners' percentage of word intelligibility. The filled symbols show predictions obtained with the different models. 2D-sEPSM predictions are shown as filled black squares; sEPSM$^{\text{X}}$ predictions are indicated as filled black triangles. For comparison, predictions obtained with the original sEPSM without across-channel processing are shown as gray triangles. Predictions from the STMI$^{\text{T}}$ are represented as filled gray circles. Data points that differ significantly from each other are labeled by different letters above the figure (2-way ANOVA, Bonferroni *post-hoc* correction, 95% confidence interval).

obtained with the STMI$^{\text{T}}$ showed the correct trend, but the dynamic range of intelligibility values was smaller than in the measured data, with values between 12% and 85%. Overall, the results suggest that all models except the original sEPSM could account for the main effects on speech intelligibility caused by the phase jitter distortion.

2.5 Discussion

2.5.1 The role of the decision metric

All considered models could account for the main effect of reverberation on the intelligibility of noisy speech, with correlation coefficients of 0.99, 0.98, and 0.98, for the 2D-sEPSM, sEPSM$^{\text{X}}$ and sEPSM, respectively. The MTF-based STMI$^{\text{T}}$ showed the correct trend of a decrease in intelligibility with increasing reverberation time, although an SRT could not be obtained for reverberation times above 0.7 s. In contrast, only the models considering the decision metric based on the SNR_{env} could account for the detrimental effect of spectral subtraction, with correlation coefficients of 0.93, 0.99, and 0.99 for the 2D-sEPSM, the sEPSM$^{\text{X}}$ and

the sEPSM, respectively. The MTF-like metric of the STMI^{T} could not account for the spectral subtraction data, exhibiting a correlation coefficient of −0.93, possibly because it does not consider the effects of the nonlinear processing on the noise modulations alone. This is consistent with the results from Jørgensen and Dau (2011). The two new models presented here, the 2D-sEPSM and the sEPSM^{X}, employed the SNR_{env} metric but applied it to two different internal representations (a three-dimensional versus a two-dimensional representation of the modulation power), and provided reasonable results across the three different conditions considered in the present study, with average correlation coefficients across all conditions of 0.95 and 0.99, for the 2D-sEPSM and the sEPSM^{X}, respectively. This suggests that the SNR_{env} is a powerful metric for speech intelligibility prediction that is robust with respect to specific assumptions made in the auditory preprocessing; the details of the implementation of the envelope extraction, [e.g. with or without a lateral inhibitory network (LIN) (Elhilali et al., 2003)], the particular shape of the peripheral filters, and the number of modulation dimensions, did not affect the predictions in a crucial way.

2.5.2 The role of across-frequency modulation processing

In the case of the phase-jitter distortion, which mainly affects the spectral structure of the speech, the results demonstrated that the original sEPSM (Jørgensen and Dau, 2011) fails. The results obtained with the two new models considered here, the 2D-sEPSM and the sEPSM^{X}, showed that the failure of the original sEPSM was caused by the lack of an across (audio-) frequency mechanism. The across-frequency process in the 2D-sEPSM is reflected in the spectro-temporal modulation filtering stage because it evaluates the *joint* modulation, i.e., the combined modulations across time and audio frequency. Such a stage, inspired by physiology, has been proposed as the basis for extracting relevant information in various modeling tasks, such as speech segregation (Mesgarani et al., 2006) and discrimination of natural sounds (Woolley et al., 2005), and as a feature extraction mechanism for speech recognition (Kleinschmidt, 2002; Nemala et al., 2013). However, the "2D" modulation filtering does not assume any information reduction in the processing and may represent a rather complex internal representation of the stimuli for modeling speech perception. The sEPSM^{X}, in contrast, applies a temporal-only modulation-filtering process (as in the original sEPSM), also motivated by physiological data in the auditory brainstem

and cortex in cats (Langner and Schreiner, 1988; Schreiner and Urbas, 1988), recent imaging studies in humans (Xiang et al., 2013), as well as computational modeling results from behavioral signal detection and modulation masking studies in humans (e.g., Dau et al., 1997a; Dau et al., 1997b; Verhey et al., 1999; Derleth and Dau, 2000; Jepsen et al., 2008). The approach to measure the variation of modulation activity across frequency after the preprocessing in the model is also consistent with recent concepts in computational auditory scene analysis (Elhilali et al., 2009a), comodulation masking release (CMR; Piechowiak et al., 2007; Dau et al., 2013) and sound texture synthesis (e.g., McDermott and Simoncelli, 2011). Using the across-channel variance as the measure of coherent across-frequency activity has been a pragmatic choice in the present study. Other across-channel operations, such as those based on cross-correlation or the temporal coherence of the temporal envelope of neighboring channels (Richards, 1987; Par and Kohlrausch, 1998), may represent alternative measures.

The sEPSM^{X} performed slightly better than the 2D-sEPSM in the conditions considered in the present study. The sEPSM^{X} showed an average Pearson correlation of 0.98 with the data across all conditions and an average RMSE of 0.79 dB for the reverberation and spectral subtraction conditions and a RMSE of 9% for the phase jitter condition. The 2D-sEPSM showed an average Pearson correlation of 0.96 across all conditions and average RMSEs of 1.81 dB and 24.5%, respectively, for the same conditions. Compared to the original sEPSM, both models showed a slightly worse performance in conditions with reverberation. However, it should be noted that the two models presented here were not optimized in terms of best fits with the data in all conditions. The parameters of the 2D-sEPSM and of the sEPSM^{X} were optimized using the data in the phase-jitter conditions, but all parameters were then kept constant for the conditions with additional processing. The main focus of this study was a comparison of two conceptual across-frequency processes in connection with different types of decision metric for speech-intelligibility prediction. The sEPSM^{X} appears conceptually simpler than the 2D-sEPSM. However, more work is needed to clarify which approach may be more powerful and plausible when applied to a broader range of experimental conditions.

2.5.3 The role of the auditory preprocessing in the models

The similarity of the predictions obtained with the 2D-sEPSM from the present study and the STMI^{T} (from Elhilali et al., 2003) in the phase jitter conditions

suggests that the sharp tuning of the auditory filters assumed in the STMI framework (Elhilali et al., 2003) may not be critical for the simulation results. The preprocessing of the STMI includes peripheral filters with a quality factor, Q, of 4, followed by a lateral inhibitory network (LIN), which effectively sharpens the auditory filters to a Q-factor of 12 (Shamma et al., 1986; Wang and Shamma, 1994; Lyon and Shamma, 1996). In contrast, the preprocessing of the 2D-sEPSM included a filterbank of fourth-order gammatone filters without any subsequent sharpening. Although sharper auditory filters have been suggested in connection with certain phenomena, such as peripheral (two-tone) suppression (e.g., Robles and Ruggero, 2001), the use of the wider gammatone filters has been successful in various modeling studies on signal-in-noise detection (e.g., Jepsen et al., 2008), comodulation masking release (CMR; Moore et al., 1990; Piechowiak et al., 2007) and speech intelligibility (e.g., Beutelmann et al., 2010; Rennies et al., 2011). Hence, the prediction results suggest that fourth-order gammatone filters seem adequate to account for the speech intelligibility data at moderate stimulus levels as considered in the present and previous studies.

2.5.4 The role of the frequency weighting for predicting speech intelligibility

The 2D-sEPSM and sEPSM^X do not include any explicit audio or modulation frequency weighting, consistent with the STMI and the original sEPSM. Frequency weighting is only reflected by limiting the processing to "audible" audio and modulation frequencies. This is different from the empirical weighting of individual frequency bands that has otherwise been a common feature of speech intelligibility prediction metrics such as the AI, SII, and STI. The general rationale for the weighting in these metrics has been that certain frequency regions appear to be perceptually more relevant for speech intelligibility than other frequency regions (French and Steinberg, 1947; Kryter, 1962; Houtgast and Steeneken, 1985; Warren et al., 2005). For example, in the case of the STI, the weighting has been separated into two types: (1) audio-frequency weighting of individual octave bands within the frequency range 0.125 and 8 kHz, and (2) modulation-frequency weighting in the form of a truncation of the maximum modulation frequency included in the MTF, typically at 16 Hz (Houtgast et al., 1980; Houtgast and Steeneken, 1985). The reasonable agreement between the predicted and measured intelligibility obtained with the sEPSM approaches

suggests that an explicit frequency weighting might not be necessary to account for the data, if the metric that is assumed to be related to speech intelligibility is appropriate.

2.5.5 Relation to speech intelligibility prediction using short-term correlation

An alternative approach to predicting speech intelligibility is the short-time objective intelligibility (STOI) model (Taal et al., 2011), where the decision metric is a short-term correlation coefficient between the original clean speech envelope and the processed (noisy) speech envelope at the output of a number of $^1/_3$-octave bandpass filters. A key step in the STOI model is the normalization and clipping of the processed envelope, such that effects of level differences between the two signals are removed from the correlation coefficient. As a result, STOI effectively measures the similarity of the modulation content from the envelope waveforms of the two signals, whereby any reduction of the correlation may be assumed to result from noise modulations or other non-speech modulations. One benefit of the correlation-based metric is that it includes information about the envelope phase within each audio channel, which is not captured by the power metric used in the SNR_{env}. This implies that the STOI model might be sensitive to changes in the envelope phase caused by phase jitter distortion within individual channels, such that this model might not require an explicit across-channel mechanism to account for phase jitter. The within-channel change in envelope phase, as measured by the cross-correlation, and the across-channel change in envelope power, as measured by the across channel variance, may be two ways of capturing the same loss of speech information. In the STOI approach, the cause for a reduced correlation between the clean and processed envelopes is difficult to directly relate to a physical source, since it can result from any difference between the clean and the processed noisy signal. In contrast, with an envelope-power based metric, the physical source that causes a reduction in the decision metric must be a change in the envelope power of the speech, the noise or both. Moreover, the sEPSM includes additional aspects of the human auditory processing, in the form of the perceptually (Ewert and Dau, 2000) and physiologically (Langner and Schreiner, 1988; Schreiner and Urbas, 1988; Xiang et al., 2013) motivated modulation filterbank. The modulation filterbank concept might be crucial in other conditions, such

as reverberation, where the STOI metric has limitations (Taal et al., 2011).

2.5.6 Perspectives

The current implementations of both the 2D-sEPSM and the sEPSM^{X} are based on the long-term envelope power of the signals. This approach was sufficient to provide accurate predictions in the conditions considered in the present studied. However, the models are likely to fail in conditions with fluctuating interferers, such as amplitude modulated noise. One approach to overcome this limitation could be to combine the current across-channel mechanisms with the multi-resolution version of the sEPSM proposed by Jørgensen et al. (2013), in which the envelope power is calculated in short-term segments with a modulation filter dependent duration. Such an approach, combined with an across-channel process, might allow to account for conditions with stationary and fluctuating maskers, in addition to conditions with phase jitter or noise reduction via spectral subtraction.

Additionally, the $\mathrm{SNR}_{\mathrm{env}}$ concept could be included as the decision metric in a binaural hearing model that uses "better-ear" listening, such as in Lavandier and Culling (2010), Beutelmann et al. (2010), and Rennies et al. (2011) to account for spatial unmasking. Computing the $\mathrm{SNR}_{\mathrm{env}}$ at the output of a purely binaural process, such as the equalization-cancellation process used by Beutelmann et al. (2010) and Rennies et al. (2011) would require a careful analysis of the effects of the process on the envelope of each ear signal.

The framework could be extended towards a more realistic peripheral processing model (e.g., Jepsen et al., 2008). For example, the model does not include non-linear effects of cochlear processing, such as compression and suppression, which are affected in the case of a sensorineural hearing loss. Such an extension would thus allow investigations of the consequences of hearing impairment on speech intelligibility in the framework of the model.

3

Predicting speech release from masking through spatial separation in distance [a]

Abstract

Speech intelligibility models typically consist of a preprocessing part that transforms stimuli into some internal (auditory) representation and a decision metric that relates the internal representation to speech intelligibility. This study investigated speech intelligibility in conditions of spatial release from masking (SRM) where the masker is moved, on-axis, away from the target. Two binaural models, which use the conventional audio signal-to-noise ratio (SNR) in the decision metric, and two monaural models, using a decision metric based on the SNR in the envelope domain (SNR_{env}), were considered. The predictions were compared to data from Westermann and Buchholz (2015a) in conditions where the target was located 0.5 m in front of the listener and the masker was presented at a distance of 0.5, 2, 5 or 10 m in front of the listener. The data showed an SRM of 10 dB when moving the masker from a distance of 0.5 m to a distance of 10 m. The long-term monaural model based on the SNR_{env} metric was able to account for most of the SRM data, whereas the models that used the audio SNR did not predict any SRM, even when they included an equalization–cancellation-like process. The short-term monaural model based on the SNR_{env} metric predicted a small SRM only in the noise-masker condition. The results suggest that "true" binaural processing is not always crucial to account for speech intelligibility in spatial conditions and that an SNR metric in the envelope domain appears to be more appropriate in conditions of on-axis spatial speech segregation than the conventional SNR. Additionally, none of the models consid-

[a] This chapter is based on Chabot-Leclerc and Dau (2014).

ered grouping cues, which seem to play an important role in the conditions studied.

3.1 Introduction

Understanding of a speech target presented simultaneously with a masker can be improved if the target and maskers are not co-located but separated in space. Typically, this improvement in intelligibility has been studied as a variation of the angle separation between the target and the masker and is known as spatial release from masking (SRM; Kidd et al., 1998; Freyman et al., 2001; Hawley et al., 2004). SRM is expressed as the change in speech reception threshold (SRT) when the masker is moved from the co-located position to another position. SRM is usually explained as consisting of two components: a monaural advantage resulting from an increase in the signal-to-noise ratio (SNR) at one ear due to the head shadow effect, yielding a "better ear" (BE) effect, and a binaural advantage due to binaural unmasking (BU) caused by interaural timing differences (ITDs) between the two ear signals (Bronkhorst and Plomp, 1988).

Models that predict intelligibility in spatial conditions tend to consist of a combination of two components that realize a BE and a BU process. For example, the model of Lavandier and Culling (2010) consists of two such pathways (Lavandier and Culling, 2010). The first evaluates the BE contribution by selecting the best target-to-interferer ratio (TIR) for each peripheral channel and combining them using weights from the speech intelligibility index (SII; ANSI, 1997). The second path evaluates the BU by calculating the binaural masking level difference (BMLD) for each frequency channel, using the interaural coherence of the target, as well as the interaural phase of the target and the interferer. The BMLDs are also combined using the SII weights. The binaural advantage represents the sum of the BE and the BU. The model was shown to account for SRTs (with correlations of 0.95 to 0.97) for conditions with an anechoic target and a noise masker placed at various azimuthal angles and distances in three different rooms. However, because the model considers anechoic targets only, it cannot capture the effects of reverberation on intelligibility. Furthermore, it is limited to continuous noises and can therefore not account for the changes in intelligibility due to modulations in the maskers.

In the binaural speech intelligibility model (BSIM; Beutelmann and Brand, 2006), the BU is implemented as an equalization–cancellation process between

the signals received at each ear, maximizing the effective TIR. The SII is then used to evaluate the intelligibility based on the effective TIR. The BSIM could predict intelligibility with correlations of 0.80 to 0.93 in conditions of a co-located target and masker, and a spatially separated target and masker, in three different rooms. A revised version of the BSIM was introduced to account for fluctuating maskers, which could account for SRTs with a multi-talker babble masker and a single-talker masker (Beutelmann et al., 2010). Similarly to the model of Lavandier and Culling, the BSIM cannot account for the effect of reverberation on the speech itself because it cannot separate the useful from the detrimental parts of the speech.

Recently, Westermann *et al.* demonstrated an SRM of 10 dB in a condition where the distance between the target was increased *on-axis* (Westermann and Buchholz, 2015a). A speech masker was moved from a position co-located with the target at 0.5 m in front of the listener to a position 10 m away from the listener. In this particular condition, BE and BU cues were limited. Therefore, the above-mentioned binaural models based on the audio SNR are unlikely to be able to predict the large observed SRM in the study of Westermann and Buchholz (2015a).

The (monaural) speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011) uses the SNR in the envelope domain (SNR_{env}) as its decision metric, rather than the audio SNR. In the sEPSM framework, the random envelope fluctuations produced by the background noise are taken into to account when predicting speech intelligibility (Jørgensen and Dau, 2011). The sEPSM was shown to account for changes to intelligibility due to additive noise, reverberation, and noise reduction via spectral subtraction. An extension of the sEPSM was proposed to account for the speech masking release due to fluctuating maskers. The extension has been named the multi-resolution sEPSM (mr-sEPSM) because the SNR_{env} is calculated in time windows of different duration, depending on the center frequency of the modulation filters in the processing. It is hypothesized that the mr-sEPSM might account for SRM obtained in the on-axis distance condition because it uses the modulation-domain SNR_{env}, rather than the audio-domain SNR, in the decision stage.

The present study investigates whether masking release due to the on-axis spatial separation of a target and masker can be explained by modulation masking rather than by energetic masking. Predictions by the mr-sEPSM, the long-term sEPSM, the model of Lavandier and Culling, as implemented by Jelfs et al.

(2011), and the BSIM were compared to data from literature (Westermann and Buchholz, 2015a).

3.2 mr-sEPSM description

The mr-sEPSM takes as its inputs the mixture of the speech and the noise, as well as the noise alone. Each signal is processed independently. The signals are filtered using a bandpass filterbank consisting of 22 gammatone filters with 1/3-octave spacing of the center frequencies, covering the range from 63 Hz to 8 kHz. An absolute sensitivity threshold is included such that channels are processed further only if their level at the output of the filtering is above the hearing threshold of a normal-hearing listener. The temporal envelope in each gammatone channel is extracted using the Hilbert transform and low-pass filtered with a cutoff frequency of 150 Hz using a first-order Butterworth filter. Each envelope is subsequently processed by a (modulation) filterbank consisting of second-order band-pass filters, in parallel with a third-order low-pass filter. The cutoff frequency of the low-pass filter is 1 Hz and the band-pass filters have center frequencies from 2 to 256 Hz with octave spacing and a Q-factor of 1. The temporal output of each modulation filter is segmented using rectangular windows without overlap. The duration of the window corresponds to the inverse of the center frequency of the modulation filter, e.g. the windows at the output of the 4 Hz filter are 250 ms long. The envelope power is calculated for each of the segments, i, peripheral channel, p, and modulation filter, n, yielding $P_{\text{env},i(p,n)}$. For each temporal segment, the SNR_{env} is calculated from the envelope power of the noisy speech and of the noise alone as

$$\text{SNR}_{\text{env,i}}(p,n) = \frac{P_{\text{env},S+N,i(p,n)} - P_{\text{env},N,i(p,n)}}{P_{\text{env},N,i(p,n)}}, \tag{3.1}$$

where $P_{\text{env},S+N}$ and $P_{\text{env},N}$ represent the normalized envelope power of the noisy speech and the noise alone. For each modulation channel, the SNR_{env} values are averaged across-time. The time-averaged SNR_{env} are combined across modulation filters and then across gammatone filters. The combined SNR_{env} is converted to the probability of correctly recognizing the speech items using in "ideal observer" (Jørgensen and Dau, 2011). More details about the mr-sEPSM can be found in Jørgensen et al. (2013).

3.3 Method

Model predictions were compared to the data from Westermann and Buchholz (2015a) obtained with the coordinate response measure (CRM; Bolia et al., 2000) material. The CRM consists of sentences of the form "Ready [call sign] go to [color] [number] now", with eight call-signs, four colors (red, green, blue and white) and eight numbers (1 through 8). Only male talkers were used, and only the "Baron" call-sign was considered for the target. The speech maskers were sentences of the same form, but with different male talkers. For the speech maskers, two randomly selected sentences were used against each target. In addition to the speech masker, a speech-modulated noise was created by applying the Hilbert envelope of the speech maskers to noise with the same long-term spectra as the speech target.

Two conditions were studied. In experiment 1, the target was placed 0.5 m in front of the listener and the speech maskers were placed either 0.5, 2, 5 or 10 m in front of the listener, on-axis with the target. In experiment 2, the target was also placed 0.5 m in front of the listener, but the masker was a two-talker speech-modulated speech-shaped noise presented from a distance of 0.5 or 10 m from the listener, on axis. The long-term spectra of the maskers were equalized to match the long-term spectrum of the target for each of the measured spatial configurations, to minimize effects that may result from differences in the signals' long-term spectra. Equalized anechoic maskers (either speech, or speech-modulated speech-shaped noise) were convolved with binaural room impulses responses (BRIRs) recorded using a B&K Head and Torso Simulator (Westermann and Buchholz, 2015a). The room had a reverberation time of $T_{60} = 1.9$ s at 2 kHz and a volume of approximately 1150 m^3. The delay introduced by the sound propagating further was removed by time-aligning the direct sound components. The masker was kept at a root mean square (RMS) level of 55 dB SPL and the speech was adjusted to produce the required TIR at the listener.

3.3.1 Models setup

For the predictions, the sEPSM and mr-sEPSM parameters were calibrated to the intelligibility data for unprocessed speech-shaped noise (SSN) with the same long-term spectrum as the target. Data obtained with the CRM speech material presented in SSN were taken from Brungart et al. (2001). The simulation with

the Lavandier and Culling model were obtained using the implementation by Jelfs et al. (2011), available in the AMToolbox (Søndergaard and Majdak, 2013). This implementation differs from the original in that it uses the BRIRs directly to calculate the BE and BU, rather than convolving the BRIRs with noises. The BSIM simulations were obtained using the "BSIM demonstrator" available at HearCom (2014).

3.4 Results

The left panel of Figure 3.1 shows the measured SRM and simulated SRM for experiment 1, where the maskers were speech and placed at distances of 0.5, 2, 5, or 10 m from the listener. Measured SRM (open symbols) ranged from 2.4 dB, when the masker was 2 m from the listener, to 9.8 dB, when the masker was 10 m away from the listener. The mr-sEPSM (filled upward triangles) predicted an SRM of −1.9 dB when the maskers were at 2 and 5 m. No improvement in intelligibility was predicted when the maskers were at 10 m. The long-term sEPSM (filled downward triangles) predicted an SRM ranging from 1.1 dB when the maskers were 2 m away to 7.1 dB when the maskers were at 10 m. The Jelfs model (filled squared) predicted no SRM, independently of masker distance. The BSIM (filled circles) predicted no SRM for the nearer maskers and an SRM of 1.2 dB when the maskers were 10 m away.

The left panel of Figure 3.1 shows SRM when the masker was speech-modulated SSN and only for the positions where the masker was either 0.5 or 10 m from the listener. The data (open symbols) showed no SRM when moving the masker away. The mr-sEPSM (filled upward triangles) predicted an SRM of 3.2 dB with the masker 10 m away. The sEPSM (filled downward triangles) showed an SRM of 4.5 dB when the masker was at the 10 m position. For the Jelfs *et al.* model and the BSIM, the predictions were the same as for the experiment with the speech maskers, since the models consider only the BRIRs and the long-term spectrum of the masker, respectively.

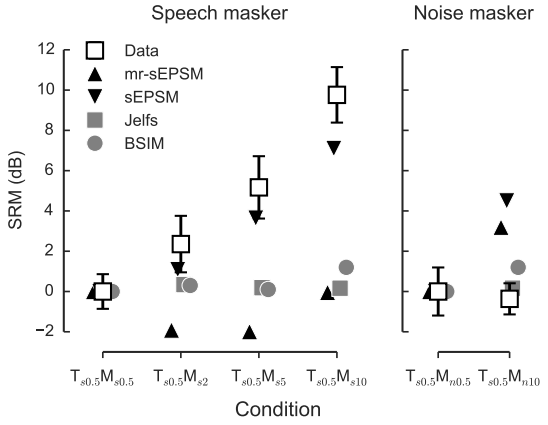


Figure 3.1: Measured SRMs and model SRMs as a function of the masker distance for speech maskers (left) and SMSSN (right). Vertical bars represent one standard deviation.

3.5 Discussion

3.5.1 The role of the decision metric and the binaural processing in the long-term models

The present study investigated whether SRM due to on-axis distance separation in speech-on-speech masking condition could be accounted for solely by energetic masking or modulation masking (Westermann and Buchholz, 2015a). Predictions obtained with the long-term binaural models that use the audio SNR as the decision metric (Lavandier and Culling, 2010; Beutelmann et al., 2010) predicted 0 to 1.2 dB of SRM, compared to the 9.8 dB observed in the data. The two models made the exact same predictions for the noise masker conditions as for the speech masker conditions because they are not sensitive to the change in the nature of the masker. The data, however, did not show any SRM when moving the noise masker. The models based on the long-term SNR could not account for the release from masking because no increase in SNR was created by moving the masker away.

Predictions by the sEPSM, which uses the SNR_{env} as its decision metric, could account for 7.1 dB of the 9.8 dB of the spatial release in the speech masker condition. However, the sEPSM predicted a 4.5 dB SRM in the noise condition, in contrast to the data where no SRM was observed. The sEPSM predicted an increase in intelligibility because the increased distance in the room has a low-pass modulation filter effect on the maskers (Houtgast and Steeneken,

1985). The low-pass filtering in the modulation domain of the masker effectively increases the SNR_{env} for modulation frequencies above 4 Hz and for audio frequencies above 1 kHz. However, the sEPSM has been shown to fail to account for intelligibility data obtained with fluctuating maskers (Jørgensen et al., 2013). The maskers studied in Jørgensen et al. (2013) included, among others, a conversation between two people, SSN that was fully amplitude modulated by an 8-Hz sinusoid, and the International Speech Test Signal (Holube et al., 2010). In those conditions, the modulated maskers caused a *decrease* of SNR_{env} compared to the steady-state SSN because they contain more modulation energy then the SSN. Thus, predictions with the sEPSM obtained for condition with fluctuating maskers are not consistent with the data. However, the prediction of an SRM by the sEPSM suggests that a release from modulation masking may be involved when the speech maskers are moved away from the target.

Binaural processing did not contribute to the prediction of intelligibility in the spatial condition considered. In the model of Jelfs, the BE process amounted to −0.34 dB and the BU process 0.5 dB, leading to a combined binaural advantage of 0.16 dB when the maskers were 10 m away. The BE component was limited by the equalization applied to the masker to compensate for the effect of the room on the masker's spectrum. The BU component was limited by the interaural coherence of the interferer, which is negatively affected by small direct-to-reverberant ratios. For example, at the 10 m position, the acoustic clarity (C_{50}), defined as the ratio of the energy of the first 50 ms to the energy of the remaining part, was −7.62 dB.

Long-term audio SNR and binaural processing of the BRIRs could thus not account for the SRM observed in the speech-on-speech masking condition when the maskers were moved away from the target. The long-term SNR_{env}, as implemented in the sEPSM, could account for most of the SRM, although this model has been shown to fail in monaural conditions with fluctuating maskers.

3.5.2 The role of the decision metric in the short-term models

The mr-sEPSM, which considers the SNR_{env} in short time-windows, did not predict any SRM in the speech masker condition when the masker was at 10 m, but did predict a *negative* release from masking when the masker was at 2 or 5 m. In the noise condition, the mr-sEPSM incorrectly predicted a 3.2 dB SRM when the masker was at 10 m. A possible reason for why the mr-sEPSM did not predict a SRM in the speech-on-speech condition, is that the time

windows used were too short. They can be as short as 3.9 ms. Similarly to what is observed in the audio domain, reverberation fills the "gaps" present in the masker, effectively increasing the SNR_{env} in some time segments (when the target and the maskers were time-aligned in the co-located condition), but also decreasing it in other segments (when the masker was weak or absent). The result is an overall decrease in SNR_{env}. It is possible that limiting the window length in the multi-resolution process to a longer duration would allow the mr-sEPSM to account for part of the release from modulation masking. Other short-term models, such as the ESII and BSIM, used longer window lengths of 12 ms (Beutelmann et al., 2010; Rhebergen and Versfeld, 2005). However, even if the modifications to the mr-sEPSM may improve its performance in the speech masker conditions, it is unlikely that it would account for the entire SRM observed in the data.

In order to demonstrate how a short-term model that uses the audio SNR would perform in the conditions studied, simulations were performed with the extended SII (ESII; Rhebergen and Versfeld, 2005). The ESII calculates the audio SNR in short time-windows of different durations depending on the center frequency of the peripheral filter. Only the speech masker condition was used. Figure 3.2 shows that the ESII predicts a decrease in intelligibility when the masker is moved away. The reason for this is that the masker becomes *less* fluctuating when moved further away, i.e. the gaps are "filled" by the reverberation. Therefore, there are fewer "glimpses" where the SNR is positive when the masker is far away than when it is co-located with the target. The short-term extension of the BSIM would probably not predict an increase in intelligibility either, because it uses the same metric as the SII (Beutelmann et al., 2010). Furthermore, the equalization–cancellation process would not increase the internal SNR, relative to the co-located condition, because the interaural decorrelation caused by the room limits the efficiency of the EC process.

3.5.3 Energetic, modulation, and informational masking

It has been argued that most of what is considered energetic masking (EM) may, in fact, be a form of modulation masking (MM) (Stone et al., 2012; Stone and Moore, 2014). Following this argument and the fact that both models based on the audio SNR failed to predict any SRM, it is difficult to argue that the increase in intelligibility resulting from moving the masker away in the speech-masker condition is due to a release from energetic masking. The prediction by the

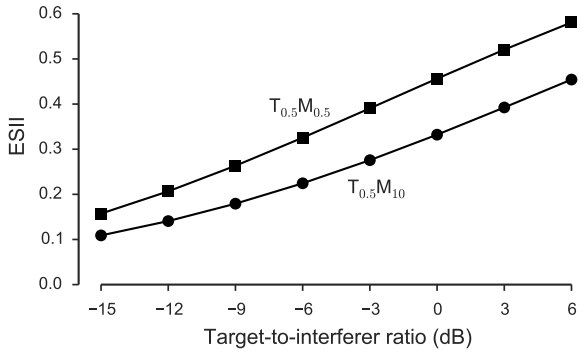


Figure 3.2: ESII prediction for two speech maskers when they are either co-located with the target, at 0.5 m from the listener, or placed 10 m away, on axis.

sEPSM of an SRM of 7.1 dB supports that moving the masker away provides a release from modulation masking. However, the prediction obtained with the mr-sEPSM, which uses the same SNR_{env} metric, did not show any SRM.

It has been argued that when target and speech maskers are co-located, higher threshold are caused by a combination of EM (and likely MM) and informational masking (IM) (Freyman et al., 2001). The two main factors contributing to IM have been suggested to be: (1) a difficulty segregating the target from the masker(s) and/or (2) a difficulty selecting which sound features belong to the target in the presence of similar maskers (Ihlefeld and Shinn-Cunningham, 2008). Westermann and Buchholz argued that a release from IM when moving the masker away was likely to be due to a reduced confusion between the target and the maskers (Westermann and Buchholz, 2015a). The distance-related cues that aid the segregation for the listener were not considered in the modeling frameworks. It would be necessary to extend the models with a process that can benefit from those distance cues by, for instance, identifying the number of streams (Christiansen et al., 2014; Elhilali et al., 2009b). In the context of the present study, a two-stream percept would contribute positively to the intelligibility prediction because it would suppose that the model could distinguish between the target and the masker, such as in the far-masker condition. The co-located condition would possibly produce a one-stream percept, which would negatively impact intelligibility. The intelligibility modeling framework could also be augmented by a more complex front-end that can localize and separate a target from an acoustic scene (May et al., 2012).

3.6 Conclusions

Predictions by four intelligibility models, the mr-sEPSM, the sEPSM, the model of Lavandier and Culling (2010) and the BSIM were compared to literature data that showed an SRM effect when speech maskers were moved from co-located position to a position further away. Such an SRM was not observed when the maskers were noise. The predictions showed that the SRM in the speech-on-speech masking condition could not be explained by changes in the audio SNR nor by binaural processes. Part of the SRM could be explained by a release from modulation masking but additional cues, such as grouping, seem to be required to account for the entire effect of the SRM.

4

Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain[a]

Abstract

This study proposes a binaural extension to the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen et al., 2013). It consists of a combination of better-ear (BE) and binaural unmasking (BU) processes, implemented as two monaural realizations of the mr-sEPSM combined with a short-term equalization–cancellation process, and uses the signal-to-noise ratio in the envelope domain (SNR_{env}) as the decision metric. The model requires only two parameters to be fitted per speech material and does not require an explicit frequency weighting. The model was validated against three data sets from the literature, which covered the following effects: the number of maskers, the masker types (speech-shaped noise (SSN), speech-modulated SSN, babble, and reversed speech), the masker(s) azimuths, reverberation on the target and masker, and the interaural time difference (ITD) of the target and masker. The Pearson correlation coefficient between the simulated speech reception thresholds (SRTs) and the data across all experiments was 0.91. A model version that considered only BE processing performed similarly (correlation coefficient of 0.86) to the complete model, suggesting that BE processing could be considered sufficient to predict intelligibility in most realistic conditions.

[a] This chapter is based on Chabot-Leclerc et al. (2016).

4.1 Introduction

Cherry (1953) coined the term "cocktail party problem" to describe the ability of listeners to "recognize what one person is saying when others are speaking at the same time". It is known that this ability is typically improved if the listeners can use both of their ears, relative to either ear alone, and if the target and maskers are spatially separated. Various models have been designed to disentangle which part of this binaural advantage can be attributed to a selection process between left versus right ear (i.e., a "better-ear" process), a "purely" binaural process where the signals from both ears interact, or a combination of both. The models typically focused on a few aspects affecting speech intelligibility at a time, such as the spatial separation of the target and the maskers, the effects of reverberation on the target or on the maskers, the role of temporal fluctuations in the masker, and the effects of multiple interferers. None of the models can account for all of these aspects at once. In the current study, recent advances in monaural intelligibility predictions are combined with binaural modeling approaches in an attempt to provide a model that can account for all of the aforementioned aspects.

Binaural intelligibility models usually combine a monaural model with some form of binaural processing to capture binaural cues attributed to head shadows and binaural interactions (Bronkhorst and Plomp, 1988). When a masker is placed elsewhere than in front of the listener, the head casts an acoustical shadow on the side opposite to the source. If the target is placed towards the ear that is in the shadow, the target-to-interferer ratio (TIR) is improved, yielding a better ear (BE), which helps the listener understand the target. These interaural level differences are denoted as ILDs. Correspondingly, different source azimuths produce different interaural time differences (ITDs). Binaural interactions rely on the ITD between target and maskers to facilitate their segregation, denoted as "binaural unmasking" (BU). The equalization–cancellation (EC) theory (Durlach, 1963) suggests that binaural unmasking can be explained by the ability of the central auditory system to "cancel" the interferers, effectively maximizing the target-to-interferer ratio.

4.1.1 Models with independent processing of ILDs and ITDs

Binaural models that predict intelligibility in spatial conditions tend to consist of a combination of two components that realize the BE and the BU processes.

For example, the model of Lavandier and Culling (2010) first evaluates the BE contributions by selecting the best long-term target-to-interferer ratio for each peripheral channel, using stationary speech-shaped noise (SSN) convolved with the binaural impulse responses between the listener and the sources as the target and masker "probe signals", and combining them using the speech intelligibility index (SII) weights (ANSI, 1997). The BU path evaluates the binaural masking level difference (BMLD) in each channel using an equation based on the EC concept, which incorporates the ITDs of the target and masker, as well as the interaural coherence of the masker (Culling et al., 2005). The BMLD values are also combined using the SII weights, and then summed with the BE to yield the overall binaural advantage (BA), relative to the colocated condition. Their model could account for conditions with an anechoic target and a noise masker placed at different distances and azimuths in three different rooms. However, because the model considers anechoic targets only, it cannot capture the effects of reverberation on the target itself because reverberation does not strongly affect the envelopes of the convolved SSN probe signals. Furthermore, it is limited to stationary maskers and thus cannot account for intelligibility changes due to amplitude modulations in the maskers because the model only considers the long-term properties of the signals.

This model of Lavandier and Culling (2010) was expanded to include head shadow and multiple stationary maskers in anechoic (Jelfs et al., 2011) and reverberant conditions (Lavandier et al., 2012). Those two model versions used binaural room impulse responses (BRIRs) directly to calculate the TIRs and BMLD values. Although those extended model implementations are computationally more efficient and have more predictive power than the previous one, they still have the same inherent limitations, i.e., they cannot account for any release from masking due to modulations in the maskers and cannot describe effects of temporal smearing of the target at low direct-to-reverberant ratios. Those are similar to the limitations of the SII, on which those models are based; the models would predict good intelligibility at infinite SNRs but low direct-to-reverberant ratios, which is in contrast to the reduced intelligibility observed in such conditions.

Collin and Lavandier (2013) proposed another extension of the original work of Lavandier and Culling (2010) to account for the effects of modulated interferers, whereby the BE and BU calculations are performed in short-time frames of 12 ms on the filtered signals, rather than directly on the BRIRs. The

short-time frames are averaged over the duration of the signals, similar to the processing in the extended speech intelligibility index (ESII; Rhebergen and Versfeld, 2005). Collin and Lavandier used an SSN target, rather than speech, because it was assumed that gaps in the speech would produce negative TIRs even though they carry information that should contribute positively to the intelligibility. Collin and Lavandier (2013) varied the masker distance as well as its modulation depth using either stationary SSN, 1-, 2-, or 4-talker modulated SSN. The model was demonstrated to qualitatively account for the correct trends in the data for different masker distances and modulation depths, although measured and simulated effects were small (from less than 1 dB to about 2 dB). However, while the short-time approach seemed successful when predicting intelligibility in modulated maskers, it would fail to account for the effects of reverberation on the speech because SSN is used instead of speech for the target signal.

4.1.2 Models combining the SII and the EC concept

In the binaural speech intelligibility model (BSIM; Beutelmann et al., 2010), which is a revision and simplification of the original implementation (EC/SII; Beutelmann and Brand, 2006), the BU process is implemented as a frequency-independent equalization and cancellation (Durlach, 1963) of the long-term signals received at each ear. The SII is then used to evaluate the intelligibility based on the effective TIR in each frequency band. The reference SII value corresponding to the speech reception threshold (SRT) is selected only once for all conditions and is defined as the SII predicting 50% intelligibility for the monaural presentation of the Oldenburg Sentence Test in noise (Wagener et al., 1999). The BSIM could predict SRTs of normal-hearing (NH) listeners in conditions with colocated target and stationary speech-shaped noise maskers (SSN), as well as with spatially separated target and maskers, in anechoic conditions and in three different rooms (a listening room, a classroom, and a church). Beutelmann et al. (2010) also extended the BSIM to account for fluctuating maskers by computing the SII after EC processing in short time windows with an effective length of 12 ms. The extension was named short-term BSIM (stBSIM). The stBSIM could account for the release from masking due to modulated maskers in anechoic conditions, but was less accurate when reverberation was introduced; the mean differences between predicted and observed SRTs varied between −4.1 and −2.7 dB. Furthermore, similar to the model of Lavandier and

Culling (2010) and its extensions, the stBSIM cannot account for the effects of reverberation on the speech itself because it cannot separate the useful from the detrimental part of the speech.

Rennies et al. (2011) proposed several modifications of the long-term BSIM to better account for the deleterious effect of reverberation: (1) an extension based on the modulation transfer function, (2) a compensation factor based on the room "definition" (a room acoustical property), and (3) a separation of the speech signal based on the useful and detrimental parts. Extension (2) provided the best fit of the three models in anechoic and reverberant conditions with stationary maskers. Although the proposed modifications increased the predictive power of the model, they also reduced its generality because the model now required access to the room impulse response in addition to the speech and noise signals.

Wan et al. (2010) introduced an application of the EC model of Durlach (1963), which they later denoted as the steady-state EC model (SSEC). Their approach is similar to that of the BSIM (Beutelmann et al., 2010) but differs in a few important ways: the decision device based on the SII selects the best SNR from the left ear, the right ear, or from the cancelled pathway for each frequency channel, rather than from the cancelled pathway only; the EC process resolution is limited by applying frequency-independent and time-varying jitters in both amplitude and time to the output of each peripheral filter, instead of adding uncorrelated noise to each ear signal; a different SII criterion is selected for each combination of number of maskers, and type of maskers, rather than using a single SII criterion. The model was evaluated for different maskers types, 1 to 3 simultaneous maskers, and different masker azimuth angles. Wan et al. (2010) showed that the model could predict SRTs correctly when the maskers were SSN or speech-modulated SSN, but failed when the maskers were speech or reversed speech.

Wan et al. (2014) proposed the short-time EC model (STEC) to extend the SSEC. In contrast to the SSEC, the equalization parameters of the EC process are calculated in overlapping 20 ms windows and can vary as a function of time, which improves cancellation of the dominant masker across time. The cancelled signal is then resynthesized from the short-time windows and the SNR is calculated from the long-term spectrum. This means that only the BU process is applied in a short-time fashion and *not* the BE process. The STEC predictions were more accurate in conditions with speech-modulated SSN; however, the

agreement with the data was worse than with the SSEC for reversed-speech maskers. The STEC described the spatial release from masking occurring with speech maskers slightly better than the SSEC did, but it still failed to account for the large 9 dB release from masking observed in Marrone et al. (2008) when two speech maskers are moved from being colocated with the target to being placed at ±15° azimuth angles. This may be due to differences in informational masking (IM) across the conditions. The STEC still has the same inherent limitation as the SSEC in that the model fitting has to be done for each combination of masker type and number of maskers. Further, it has never been tested in reverberant conditions.

4.1.3 Modulation-domain models

In contrast, Van Wijngaarden and Drullman (2008) extended the speech transmission index (STI; Houtgast and Steeneken, 1973b; IEC, 2003) to consider binaural hearing. The STI considers the integrity of the modulations of a reference signal (or speech) after processing as the decision metric, assessed by the modulation transfer function (MTF). The MTF can capture the effects of reverberation on speech because of the reduction in modulation in the reference signal. The binaural interaction of the binaural STI is based on interaural cross-correlograms. Van Wijngaarden and Drullman (2008) showed that the binaural STI extension could account for consonant-vowel-consonant (CVC) word scores for stationary maskers presented in multiple rooms (anechoic, a listening room, a classroom, and a large church). However, this approach is limited because it cannot be extended to more realistic conditions where the maskers are also modulated, since modulations are then coming from both the target and maskers and they can no longer be distinguished.

In order to account for different amounts of target and masker modulations, Jørgensen and Dau (2011) proposed the monaural speech-based envelope power spectrum model (sEPSM), which considers the signal-to-noise envelope power ratio (SNR_{env}) at the output of a modulation filterbank (Ewert and Dau, 2000) as the decision metric. In addition to conditions with additive maskers, the sEPSM can also account for the effects of reverberation, as well as noise reduction via spectral subtraction because it captures the increase in the masker's modulation power *after* processing. The sEPSM was extended to account for conditions with fluctuating maskers by using a "multi-resolution" process (Jørgensen et al., 2013). In the corresponding multi-resolution model,

the mr-sEPSM, the SNR_{env} is calculated in windows of different length (akin to the ESII of Rhebergen and Versfeld (2005)) according to the center frequency of the modulation filters. The mr-sEPSM was validated using various fluctuating noises, including cafe noise, two-band speech modulated noise, the international speech test signal (ISTS; Holube et al., 2010), and a reversed talker. In contrast to the SII and STI metrics, the SNR_{env} metric can account for both the effects of reverberation on the target and the masker as well as for the release from masking due to fluctuations in the maskers. However, the model has not yet been applied to spatial conditions using two-ear processing. Therefore, using the mr-sEPSM framework in a binaural model could yield a model that can account for all the aforementioned aspects of binaural speech intelligibility: the spatial separation of the target and the maskers, the effects of reverberation on the target and on the maskers, the role of temporal fluctuations in the masker, and the effects of multiple interferers.

None of the models previously mentioned can account for the deleterious effects of colocated concurrent speakers on speech intelligibility. The difference between the measured intelligibility and intelligibility predicted using energy-based model is often labeled as "informational masking".

4.1.4 Proposed modeling framework

Here, a model is proposed that combines concepts from different modeling approaches. Specifically, it integrates a short-time equalization-cancellation process (Wan et al., 2014), a temporal modulation filterbank (Dau et al., 1997a; Ewert and Dau, 2000), the SNR_{env} metric (Jørgensen and Dau, 2011), and a better-ear process in the envelope power domain. The model was evaluated using a set of critical experimental conditions from the literature to tease apart the contributions of the decision metric, the short-time processing, the better-ear process, and the binaural unmasking for predicting intelligibility in spatial conditions. Experiment 1 focused on conditions with multiple maskers in anechoic conditions, Experiment 2 considered conditions with only a single masker, but in a reverberant environment, and Experiment 3 investigated a single-masker condition where only ITD but no ILD information was provided.

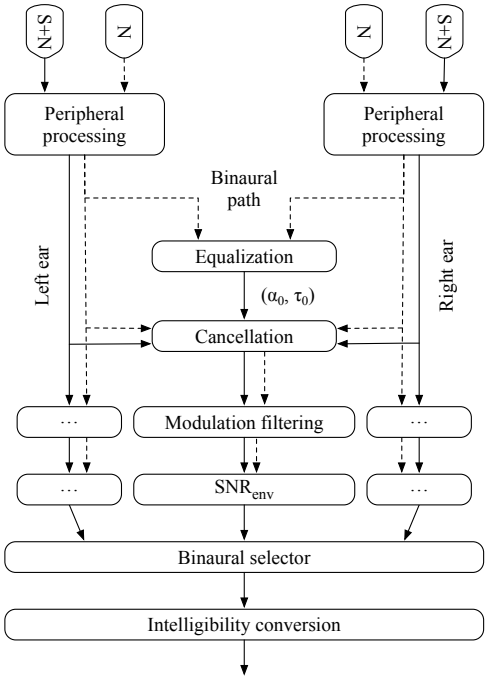


Figure 4.1: Diagram of the model structure. Solid lines denote the path of the speech-plus-noise ($S+N$) mixture and the dash lines show the path of the noise alone (N). The values α_0 and τ_0 represent the optimal parameters selected by the equalization process.

4.2 Model description

4.2.1 Overall model structure

Figure 4.1 shows a sketch of the model proposed in the present study, which is an extension of the monaural mr-sEPSM (Jørgensen et al., 2013). The model consists of realizations of the monaural mr-sEPSM for the left and right ear, and a "central" pathway where binaural unmasking takes place using an EC process (Wan et al., 2014). In contrast to the original mr-sEPSM, the model employs a binaural processing stage. Binaural processing is limited by peripheral transduction, which does not preserve fine-structure information at high frequencies (Bernstein and Trahiotis, 1996). Peripheral transduction is therefore modeled using half-wave rectification and low-pass filtering. A binaural selection stage combines the outputs of the left, right and central pathways. The subsequent output is then converted to intelligibility using an ideal observer concept. The extension to the mr-sEPSM is presented below; further details and justifications about the mr-sEPSM approach can be found in Jørgensen et al. (2013).

4.2.2 Monaural processing stage

The inputs of the model are the noisy speech and the noise alone for each ear. The first stage of each monaural model consists of 22 gammatone filters covering the frequency range from 63 Hz to 8 kHz with a third-octave spacing. The channels are processed further only if the level of the noisy speech for that channel is above the diffuse-field threshold in quiet (ISO, 2005). The envelope of each channel output is extracted using half-wave rectification and low-pass filtered using a fifth-order Butterworth filter with a cut-off frequency of 770 Hz (Breebaart et al., 2001). Jitter in amplitude and time is applied to each envelope independently to limit the efficacy of the EC process; all jitters are zero-mean Gaussian processes with standard deviations of $\sigma_\delta = 105$ μs for the time jitter and of $\sigma_\epsilon = 0.25$ (dimensionless) for the amplitude jitter (Durlach, 1963). The resulting envelopes are further processed by a modulation filterbank consisting of eight second-order Butterworth band-pass filters. A third-order low-pass filter with a 1 Hz cut-off frequency is applied in parallel, which completes the filterbank. Conceptually, this filter can be considered as the lowest frequency band in the filterbank. Only modulation filters with center frequencies below one-fourth of their respective peripheral-filter center frequency are used (Verhey et al., 1999).

The output of each modulation filter, n, is segmented in non-overlapping rectangular windows of durations inversely proportional to the center-frequency of the respective modulation filter, e.g., the windows at the output of the 8 Hz modulation filter are 125 ms long. The power, $P_{\text{env},i}(p,n)$, of each segment, i, is defined as the variance of the segment:

$$P_{\text{env},i}(p,n) = \frac{1}{[\overline{E}(p,t)]^2/2}\overline{[e_i(p,n,t)-\overline{e_i}(p,n)]^2}, \tag{4.1}$$

where p is the corresponding peripheral filter, $E(p,t)$ is the envelope at the output of the peripheral filter, $e_i(p,n,t)$ is the envelope at the output of the modulation filter for the segment i, t is time, and the bar indicates the average over time. $\overline{e_i}$ is the average over a time segment, i, of varying duration according to the center frequency of the modulation filter. $\overline{E}$ is averaged over the whole sentence duration. The lower limit of the envelope power is set to −30 dB relative to 100 percent amplitude modulation.

The $\text{SNR}_{\text{env},i}$ for each segment is computed from the envelope power of the

noisy speech and the noise alone:

$$\mathrm{SNR}_{\mathrm{env},i}(p,n)=\frac{P_{\mathrm{env},S+N,i}(p,n)-P_{\mathrm{env},N,i}(p,n)}{P_{\mathrm{env},N,i}(p,n)}, \tag{4.2}$$

where $S+N$ denotes the noisy speech and N denotes the noise alone.

4.2.3 Binaural processing stage

The binaural unmasking stage is implemented as described in Wan et al. (2014). The jittered peripheral envelopes from the monaural stages are used as inputs to the EC process. The EC processing is assumed to be independent in each channel, and performed in short overlapping time frames. A time–frequency unit is denoted as $U(p,k)$, where p again denotes the peripheral filter, and k is the kth frame, which differs from the ith segment of the modulation-domain multi-resolution process. Each frame, k, is 20 ms, whereas the multi-resolution segments, i, can vary in duration. The overlap between frames is 50% (10 ms). The equalization process in each unit selects the optimal interaural time difference (ITD), τ_0, and the optimal interaural level difference (ILD), α_0, using the following equations:

$$\begin{aligned}\tau_0(p,k)&=\arg\max_{\tau}\left\{\rho_{p,k}\right\},\ |\tau|<\frac{\pi}{\omega_p}, \text{ and}\\ \alpha_0(p,k)&=\sqrt{\frac{F_{N,L}(p,k)}{F_{N,R}(p,k)}},\end{aligned} \tag{4.3}$$

where $\rho_{p,k}$ is the normalized cross-correlation function of the left and right ears within the unit, $F_{N,L}(p,k)$ and $F_{N,R}(p,k)$ are the masker energy for the left and right ear, respectively, and ω is the center frequency of channel p. The unmasked output, $Y_{p,k}(t)$, for the unit $U(p,k)$ after cancellation is calculated as:

$$\begin{aligned}Y_{p,k}(t)=W_k(t)\Bigg\{&\frac{1}{\sqrt{\alpha_0(p,k)}}E_L(p,t)\left(t+\frac{\tau_0(p,k)}{2}\right)\\ &-\sqrt{\alpha_0(p,k)}E_R(p,t)\left(t-\frac{\tau_0(p,k)}{2}\right)\Bigg\},\end{aligned} \tag{4.4}$$

where the subscripts L and R denote the left and right ear, respectively, and $W_k(t)$ is a rectangular window function for the frame k, which can be expressed

as:

$$W_k(t) = \begin{cases} 1, & (k*10)\,\text{ms} \le t \le (k*10)+20\,\text{ms} \\ 0, & \text{otherwise.} \end{cases} \tag{4.5}$$

Subsequently, the binaural signal, B_p, is reconstructed for each channel by summing over all overlapping frames:

$$B_p(t) = \frac{1}{2}\sum_k Y_{p,k}(t). \tag{4.6}$$

The unmasked outputs for the noisy speech, $B_{S+N,p}$, and the noise alone, $B_{N,p}$, are then used as inputs to the modulation filtering stage of the mr-sEPSM, and, subsequently, to the SNR_{env} calculation. This yields BU-$\text{SNR}_{\text{env}}(p,n,t)$, a binaurally unmasked SNR_{env}, for each peripheral channel, modulation channel, and multi-resolution frame.

4.2.4 Binaural selection stage

The binaural selection device selects the best SNR_{env}, denoted as the "binaural SNR_{env}" (B-SNR_{env}), between the better-ear SNR_{env}, (BE-$\text{SNR}_{\text{env},i}$) and the binaurally unmasked $\text{SNR}_{\text{env},i}$ (BU-$\text{SNR}_{\text{env},i}$) for each multi-resolution segment (note that the (p,n,t) indexing has been omitted for the sake of brevity):

$$\text{B-SNR}_{\text{env},i} = \max(\text{BE-SNR}_{\text{env},i}, \text{BU-SNR}_{\text{env},i}), \tag{4.7}$$

where BE-$\text{SNR}_{\text{env},i}$ is the maximum between the left and right $\text{SNR}_{\text{env},i}$ for each segment:

$$\text{BE-SNR}_{\text{env},i} = \max(\text{SNR}_{\text{env},L,i}, \text{SNR}_{\text{env},R,i}). \tag{4.8}$$

The B-SNR_{env} is then averaged over all segments, I_n, of each modulation channel:

$$\text{B-SNR}_{\text{env}}(p,n) = \frac{1}{I_n}\sum_{i=1}^{I_n}\text{B-SNR}_{\text{env},i}(p,n), \tag{4.9}$$

yielding a 9×22 array of values. The time-averaged B-SNR_{env} is first combined across modulation filters:

$$\text{B-SNR}_{\text{env}}(p) = \left[\sum_{n=1}^{9}\text{B-SNR}^2_{\text{env}}(p,n)\right]^{1/2} \tag{4.10}$$

and then across peripheral filters:

$$\text{B-SNR}_{\text{env}} = \left[\sum_{p=1}^{22} \text{B-SNR}_{\text{env}}^2(p) \right]^{1/2}. \tag{4.11}$$

4.2.5 Decision device

The overall $\text{B-SNR}_{\text{env}}$ is converted to a sensitivity index, d', of an "ideal observer" (Jørgensen and Dau, 2011), using the relation:

$$d' = k(\text{B-SNR}_{\text{env}})^q, \tag{4.12}$$

where k and q are parameters independent of the experimental conditions. d' is converted to intelligibility using an m-alternative forced choice decision model, combined with an unequal variance Gaussian model expressed as:

$$P_{\text{correct}}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}} \right), \tag{4.13}$$

where Φ denotes the cumulative normal distribution. The values of σ_N and μ_N are determined by the number of response alternative, m (see the Appendix of Jørgensen and Dau (2011) for details). For open-set paradigms, m is set to 8000, which reflects the number of words in a normal listener's vocabulary. The value of σ_S is a free parameter fixed by fitting model predictions to speech intelligibility data in a condition with a SSN masker. The percentage correct at the output of the model is denoted as B-sEPSM.

Replacing the $\text{B-SNR}_{\text{env}}$ by either $\text{BE-SNR}_{\text{env}}$ or $\text{BU-SNR}_{\text{env}}$ in Eqs. (4.9) to (4.13) provides two alternative model outputs, BE-sEPSM and BU-sEPSM, where only the better-ear or only binaural-unmasking stages are used, respectively.

4.3 Methods

4.3.1 Experiment 1: Multiple maskers in an anechoic condition

Rationale

This experiment investigated the effects of multiple spatially distributed maskers in an anechoic condition on spatial release from masking (SRM) using the data

of Hawley et al. (2004). They systematically measured SRTs as a function of masker azimuth, masker type, and number of maskers using the Harvard IEEE corpus (Rothauser et al., 1969). The interferers were either speech (not considered in the current study because of potential differences in informational masking compared to the other conditions), reversed speech (from the same corpus), speech-shaped noise (SSN), or speech-modulated speech-shaped noise (SMSSN). All maskers were matched to the spectrum of the target sentences and either one, two, or three maskers were presented at once. Hawley *et al.* showed that SRM was larger when multiple voiced interferers were located at different locations from the target, compared to conditions when a single voiced masker was presented. This suggested that a short-term binaural process is critical. Wan et al. (2010) and Wan et al. (2014) used these same data to validate their long-term and short-term EC models.

Experimental conditions

The Loizou (2007) recording of the Harvard IEEE corpus, sampled at 25 kHz, was used for the target material. The SSN was also taken from Loizou (2007) and was created by filtering stationary noise to have the same long-term spectrum as the speech material. The SMSSN was created by applying the broadband envelope of a sentence from the speech corpus to the SSN. The envelope was extracted by low-pass filtering the half-wave rectified speech signal with a first-order Butterworth filter with a 40-Hz cutoff frequency (Hawley et al., 2004). The stimuli were spatialized using the head-related transfer functions (HRTFs) of the HMS II artificial head (HEAD acoustics GmbH, Germany) from the AUDIS database (Blauert et al., 1998), at angles of 0, 30, 60 or 90 degrees. One to three maskers were located in the front (0°, 0°, 0°), to the side (90°, 90°, 90°), distributed on the right (30°, 60°, 90°), or distributed to the left and the right (−30°, 60°, 90°) (see Table I in Hawley et al. (2004) for the full layout). The speech level was fixed at 65 dB SPL and the masker levels were independently set to desired SNRs, before HRTF filtering; adding maskers increased the total interferer level.

Simulations

Simulations obtained with the proposed model (B-sEPSM) were carried out for SNRs ranging from −24 to 12 dB in 3 dB increments. The simulated SRTs corresponded to the SNR at which the simulated intelligibility was 50 percent,

Table 4.1: Calibrated values of the parameters k and q, and of the constants σ_s and m of the ideal observer for the different experiments.

Condition	k	q	σ_s	m
Exp 1: Hawley et al. (2004)	0.82	0.31	0.6	8000
Exp 2: Beutelmann et al. (2010)	0.04	1.42	0.9	50
Exp 3: Lőcsei et al. (2015)	1.14	0.235	0.6	8000

using linear interpolation where necessary. The final SRT represented the average SRT for 30 randomly selected sets of target and maskers. The condition with a single SSN masker, colocated with the target, and spatialized using the AUDIS HRTFs was considered as the reference condition. Because word score data were unavailable in this condition, a Gaussian psychometric function, $c(x)$, with an anechoic SRT, μ_a, and a standard deviation, σ, was first fitted based on the anechoic word score data of Bernstein and Grant (2009) using:

$$c(x) = \mathrm{erfc}\left(\frac{-(x-\mu_a)}{\sqrt{2}\sigma}\right)/2, \tag{4.14}$$

where x represents the SNRs, c is the proportion correct, and erfc is the complementary Gauss error function. Then, μ_a was replaced by the SRT measured by Hawley et al. (2004) in the colocated condition with a single SSN masker (−3.40 dB SNR), yielding a modified psychometric function, $c'(x)$. The parameters of the model's ideal observer, k and q, were adjusted to minimize the root-mean-square error (RMSE) between the simulations obtained with the "left ear" model and the psychometric function $c'(x)$. The constants σ_s and m of the observer were fixed to 0.6 and 8000, respectively. The observer parameters were kept constant throughout Experiment 1. Table 4.1 shows the summary of the ideal observer parameters and constants for all three experiments.

4.3.2 Experiment 2: Single masker in reverberant conditions

Rationale

In contrast to Exp. 1, Exp. 2 considered the effects of a single masker of different types, but in reverberant conditions, using the data of Beutelmann et al. (2010). They measured binaural SRTs in a combination of four different rooms, three target–masker azimuth separations, and three masker types. The speech ma-

terial was the Oldenburg Sentence Test in noise (Wagener et al., 1999), which consists of a closed set of meaningful five-word sentences. The rooms included an anechoic room, a standard IEC listening room (not used in the current study), a typical classroom ($7 \times 6.9 \times 3.2$ m^3, approximately 210 m^3) and a church (outer dimensions: $63 \times 32 \times 22$ m, approximately 22 000 m^3). The interferers were either stationary SSN (stationary), 20-talker babble (babble), or single-talker modulated noise (single-talker). Beutelmann et al. (2010) found an effect of azimuth on the SRM (a 105° separation yielded a larger SRM than a 45° separation) and this effect was largest in the anechoic condition. They also showed that the masker type had a significant effect on the SRM and that speech intelligibility was positively correlated with the modulation depth of the masker. SRM was larger in the anechoic conditions, than in the reverberant conditions. The masker types used by Beutelmann et al. (2010) were similar to the ones used by Hawley et al. (2004) but the different room types are critical to validate that the proposed model can capture the effects of reverberation on SRM.

Conditions

The SSN was the one provided with the Oldenburg Sentence Test, which was created by a random superposition of the material's sentences. The multi-talker babble was the "CD101RW2" noise from the Auditec CD, created as the mixture of 20 talkers reading different texts (Auditec, 2006). The single-talker modulated noise was the "ICRA5" noise (Dreschler et al., 2001). All stimuli were sampled at 44.1 kHz. The noise level was fixed at 65 dB SPL and the target level was adjusted to the desired SNRs. Both the target and masker levels were adjusted *after* HRTF filtering. The stimuli were spatialized using virtual impulse responses created with the ODEON software version 8.0 (Kongens Lyngby, Denmark; Christensen, 2005). The anechoic, the classroom and the church conditions were used. Three spatial setups were used: 1) the target and the masker were colocated 3 m in front of the listener, 2) the target was 3 m in front of the listener and the masker was 2 m away, at 105° azimuth, and 3) the target was 6 m in front of the listener and the masker was 4 m away, at −45° azimuth. In the third condition, the listener was placed close to a wall on the right side. See Beutelmann et al. (2010) for complete details about the listening test setup.

Simulations

Simulations were obtained for SNRs ranging from −36 to 6 dB in 3 dB steps. The final simulated SRT was the average SRT for 30 randomly selected target and masker pairs. The reference psychometric function, p, was created following Wagener and Brand (2005):

$$p(L, \mathrm{SRT}, s) = 100 * \frac{1}{1 + e^{4s(\mathrm{SRT}-L)}}, \tag{4.15}$$

where L represents the given SNRs, s is the slope around the 50 percent point, and SRT is the SNR at the 50 percent points. s was set to 0.18/dB according to Wagener and Brand (2005, their Table 4) for the OLSA material with an SSN masker, and SRT was the median SRT in the spatialized condition measured by Beutelmann et al. (2010) (−7.23 dB, their Fig. 6) for the same material. The ideal observer parameters in the proposed model were fit such as to minimize the RMSE between the "left-ear" of the model and that psychometric function. The parameters were kept constant for all other conditions in this experiment. The observer's constants, σ_s and m, were fixed to 0.9 and 50, respectively, to account for the increased redundancy in the speech material.

4.3.3 Experiment 3: ITD-only condition

Rationale

Experiment 3 investigated the role of the EC process using a condition where the target and masker were lateralized to the left or to the right, using a fixed interaural delay (ITD) of 687.5 μs (Lőcsei et al., 2015). The speech was played in the presence of SSN that was either lateralized to the same side as the speech, denoted as condition S11, or to the opposite side, denoted as condition S01. Lőcsei et al. (2015) found a masking release of about 4 dB when the masker was lateralized to the opposite side. In this condition, no better-ear benefit can be expected because the signal at both ears is the same, except for a short delay used for the lateralization. Therefore, the only cues available should be interaural differences, which should be captured by the EC process.

This experiment was akin to the $S_\pi N_0$ condition often used as an example of pure-tone binaural masking level difference (BMLD; see Levitt and Rabiner (1967b), and Culling et al. (2004)). In such a condition, listeners showed a masking release as large as 12 dB when the target tone was presented out of

phase (π), compared to the in-phase presentation of the target (0; Levitt and Rabiner, 1967b). Release from masking due to ITD or out-of-phase presentation has successfully been modeled for pure-tone signals (Levitt and Rabiner, 1967b) and for speech signals using an EC-like process (Culling et al., 2004).

Conditions

The speech material was the DAT corpus (Nielsen et al., 2014), sampled at 48 kHz and recorded with female speakers. The DAT corpus consists of unique meaningful Danish sentences constructed as a fixed carrier sentence with two interchangeable target words. The masker was stationary noise shaped to have the same long-term spectrum as the speech material. The target level was fixed at 65 dB SPL and the masker level was adjusted to the desired SNR.

Simulations

Simulations were obtained for 30 randomly selected sentences and SSN maskers, and for SNRs from −12 to 9 dB in 3 dB steps. The signals were lateralized to the left or right using a fixed 33 sample delay (687.5 μs). The final simulated SRT was the average across target sentences. The ideal observer's parameters were fit to minimize the RMSE between the "left-ear" of the model and the word-scores as a function of SNR in the colocated, S11, condition, as measured by Lőcsei et al. (2015). The ideal observer's σ_s and m were set to 0.6 and 8000, respectively (Jørgensen et al., 2013).

4.4 Results

4.4.1 Experiment 1: Multiple maskers in an anechoic condition

Figure 4.2 shows the simulated SRTs obtained with the proposed model (B-sEPSM; black squares), those obtained with the better-ear only version of the model (BE-sEPSM; dotted line) as well as the binaural-unmasking version (BU-sEPSM; dashed line) as a function of the masker(s) angle(s). Furthermore, the STEC predictions from Wan et al. (2014) (grey triangles) and the measured data from Hawley et al. (2004) (open squares) are shown. The three columns correspond to one (left), two (middle), or three maskers (right), respectively. The upper panels show data and simulations for the stationary speech-shaped noise maskers (SSN), the middle panels for speech-modulated SSN maskers

(SMSSN), and the bottom panels for reversed speech. Figure 4.3 is a replot of the data and predictions of Fig. 4.2 where the thresholds are represented in terms of a spatial release from masking (SRM) relative to the condition where the target and the maskers were colocated.

Overall, there was a good agreement between the B-sEPSM simulations and the data. The Pearson correlation coefficient across all conditions was 0.91 and the prediction RMSE was 3.0 dB. For the STEC, the correlation coefficient was 0.97 the RMSE was 1.3 dB SNR. Thus, the RMSE was larger for the B-sEPSM than for the STEC but, unlike the STEC, the B-sEPSM was fit only once for all conditions. In contrast, the STEC was fit to the 90° condition for each combination of n maskers and masker type, i.e., for each sub-figure of Fig. 4.2 (Wan et al., 2014).

In the SSN condition (upper panels), the B-sEPSM simulations were slightly lower than in the data but the amount of SRM was well described for all numbers of maskers. In the SMSSN masker condition (middle panels), the B-sEPSM correctly accounted for the masker-type dependency of the SRTs in the case of the single masker. The B-sEPSM predicted an increase in SRTs with increasing number of maskers, consistent with the measured data; however, the SRTs were on average 4.76 dB larger than in the data in the condition with three SMSSN maskers. The simulated SRM was found to be the same as in the data with two SMSSN maskers, but was larger by about 4 dB with three maskers when all maskers were at different locations. The B-sEPSM predicted SRTs up to 8 dB higher in the three SMSSN maskers condition compared to the three SSN maskers condition. This is in contrast to the data, where the SRTs differed, on average, by only 1 dB between the SSN and SMSSN conditions when there were multiple maskers. Thus, the addition of a second or third SMSSN masker decreased the SNR_{env} more than the addition of SSN maskers. In the reversed-speech masker condition (lower panels), the B-sEPSM simulated SRTs were below the measure ones. However, as in the data, the simulated SRTs increased with the number of maskers, suggesting that the B-sEPSM could correctly account for intelligibility as a function of the number of reversed-speech maskers.

4.4.2 Experiment 2: Single masker in reverberant conditions

Figure 4.4 shows the measured SRTs from Beutelmann et al. (2010) (open squares), together with the B-sEPSM predictions (black squares), the simulations ob-

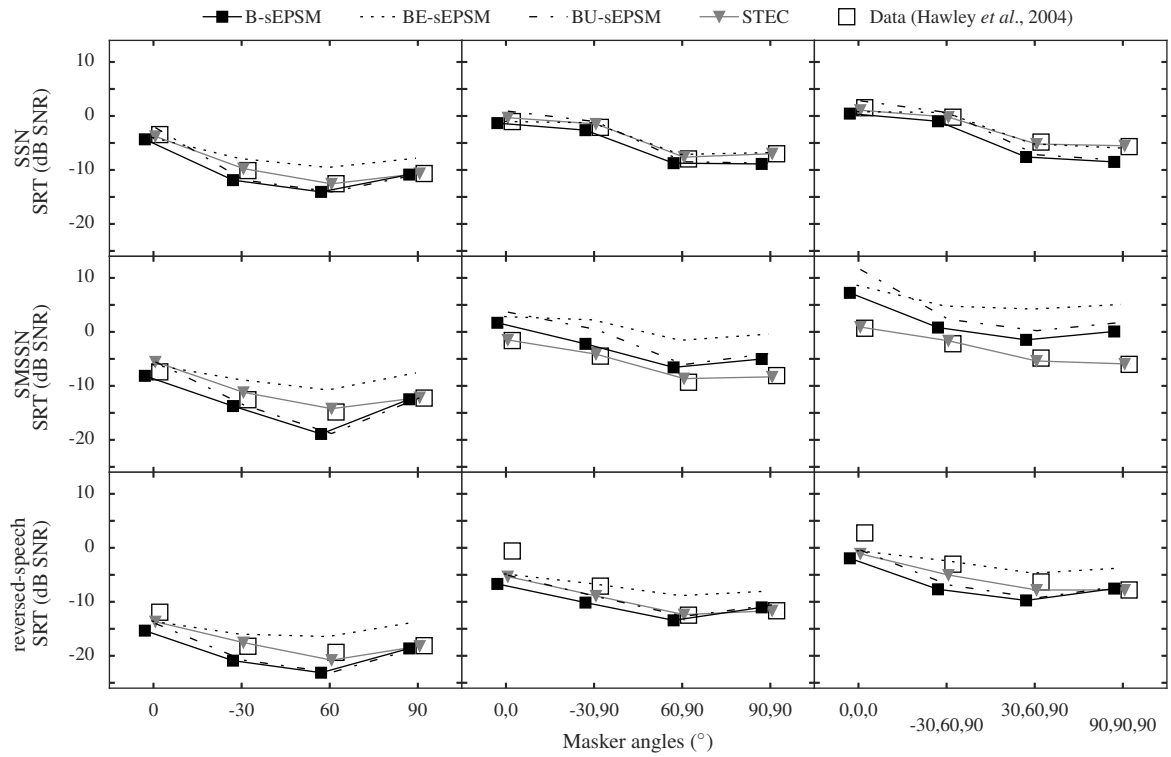


Figure 4.2: Mean speech reception threshold data (open squares; Hawley et al., 2004) and predictions obtained with the proposed model (black squares; B-sEPSM) and its alternate outputs, BE-sEPSM (dotted lines) and BU-sEPSM (dashed lines) as a function of masker(s) angle(s) for speech-shaped noise (SSN) masker(s) (upper panels), speech-modulated SSN (SMSSN; middle panels), and reversed speech (bottom panels). For comparison, STEC model predictions are shown as grey triangles for reference (Wan et al., 2014). The left panels show the condition with one masker only; the middle panels show the conditions with two maskers; and the right panels show the conditions with three maskers.

tained with the better-ear (BE-sEPSM; dotted lines) and binaural-unmasking (BU-sEPSM; dashed lines) versions of the model as a function of the masker azimuth. Furthermore, the stBSIM predictions (grey bullets; replotted from Beutelmann et al., 2010) are shown for comparison. The three columns correspond to the anechoic, classroom, and church conditions, respectively. The upper panels show data and predictions for the stationary masker, the middle panels show the corresponding results for the babble masker, and the bottom panels show the results obtained for the single-talker modulated noise masker.

Overall, there was a good agreement between the predictions and the data. The B-sEPSM Pearson correlation coefficient across all conditions was 0.91 and the average prediction RMSE for the B-sEPSM was 6.5 dB. In contrast, the

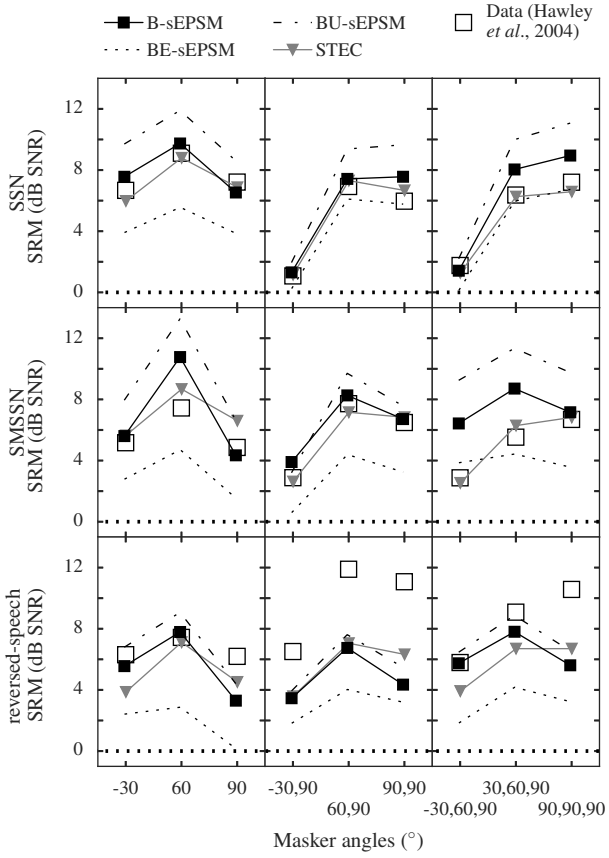


Figure 4.3: Replot of the data and predictions of Fig. 4.2 as spatial release from masking (SRM) relative to the colocated condition.

Pearson coefficient for the stBSIM was 0.89 and the RMSE was 3.65 dB.

In the anechoic condition (left panels), the B-sEPSM produced a larger SRM than that found in the data when the masker was stationary noise or single-talker noise. A similar SRM as in the data was found when the maskers were babble noise. In the classroom condition (middle column), the B-sEPSM accurately accounted for the SRM but there was a negative offset for all masker types. In the church condition (right column), the SRM was also correctly accounted for by the B-sEPSM, except for a negative offset which was largest for the single-talker babble noise. Overall, the B-sEPSM offset seemed to be partly due to the BU-sEPSM contributions, which were consistently lower than the BE-sEPSM contributions to the overall SNR_{env}. Nonetheless, the large offset observed in the reverberant conditions for all maskers was due to the particulars

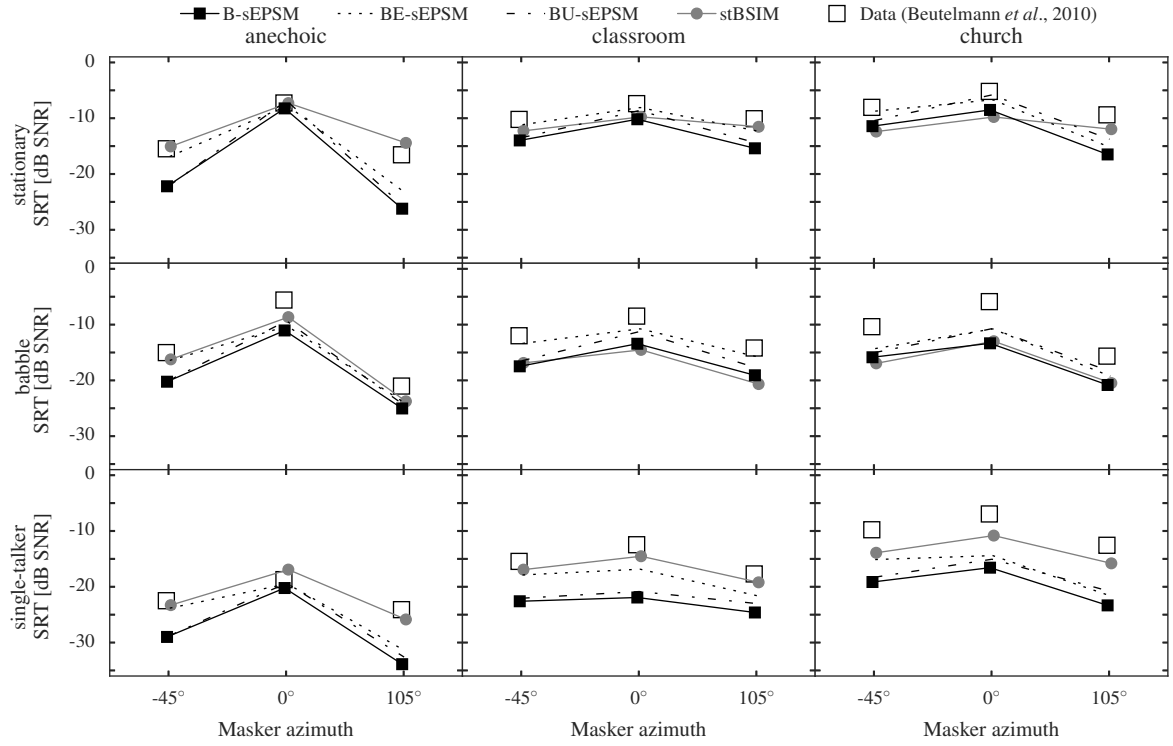


Figure 4.4: Median speech reception thresholds data measured by Beutelmann et al. (2010), (open squares), B-sEPSM predictions (black squares), BE-sEPSM predictions (dotted lines), BU-sEPSM predictions (dashed lines) and stBSIM predictions (grey bullets; Beutelmann et al., 2010) as a function of the azimuth of stationary speech-shaped noise (SSN; upper panels), babble noise (middle panels), or a single-talker modulated noise (bottom panels).

of the intelligibility transformation for the B-sEPSM. The sEPSM framework is sensitive to the type of SSN used in the reference condition; white-noise filtered to have the same long-term average spectrum as speech, and SSN created by the random superposition of speech signals yield different amounts of masking release. A smaller offset could be obtained if the ideal observer was fit to the B-sEPSM, rather than to the "left-ear" SNR_{env}, but the resulting binaural model could not be analyzed in terms of the benefit compared to one ear alone in the colocated condition.

Figure 4.5 is a replot of the data and predictions from Figure 4.4 as spatial release from masking relative to the colocated condition. The data from Beutelmann et al. (2010) showed that SRM decreased with increasing amounts of reverberation, probably as the result of reduced head shadow effect which decreases the BE benefit (Lavandier and Culling, 2010; Plomp, 1976). Reverber-

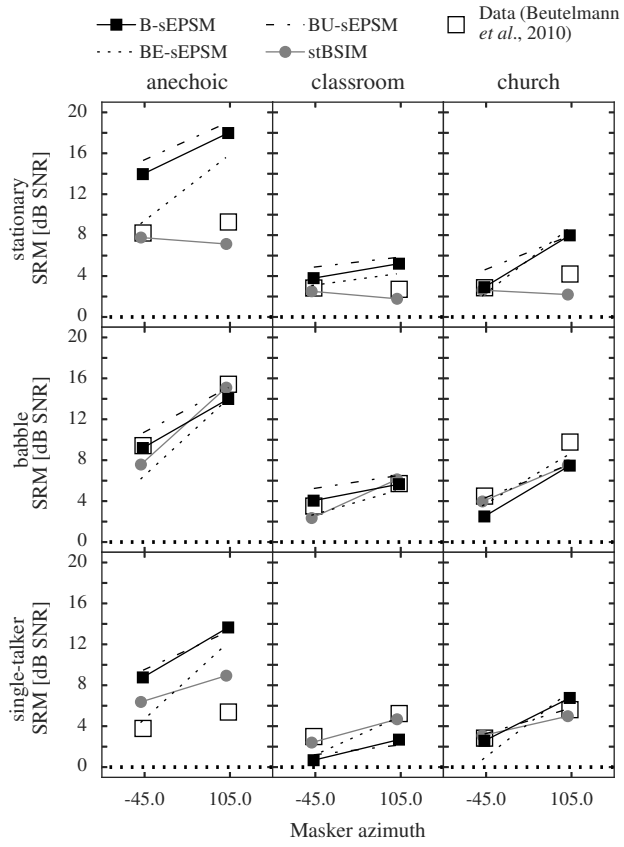


Figure 4.5: Replot of the data and predictions of Fig. 4.4 as spatial release from masking relative to the colocated condition.

ation also decorrelates the signals that reaches both ears, which reduces the efficacy of the EC process (Lavandier and Culling, 2007). These effects were captured by the BE-sEPSM and the BU-sEPSM outputs, respectively, and therefore, by the B-sEPSM, for all masker types, as shown by the correctly predicted SRM (c.f., Fig. 4.5, second and third columns).

Release from masking with a fluctuating masker, relative to a stationary masker, was also reduced in the presence of reverberation; the SRT in the colocated single-talker masker church condition was about 12 dB higher (−18.7 to −6.95 dB SNR) than in the anechoic condition. None of the models accurately predicted this large SRT increase; the B-sEPSM predicted an increase of 3.65 dB and the stBSIM an increase of 6.10 dB.

4.4.3 Experiment 3: ITD-only condition

The left panel of Fig. 4.6 shows the measured SRTs (open squares) from Lőcsei et al. (2015), the B-sEPSM predictions (black squares), as well as the predictions from the better-ear-only version of the model (BE-sEPSM; dotted line and diamonds) and the binaural-unmasking version (BU-sEPSM; dashed line and circles). Target and masker were colocated to the left in the S11 condition. In the S01 condition, the target was lateralized to the left and the masker was lateralized to the right. The right panel shows the same data and predictions replotted as SRM relative to the S11 condition. The B-sEPSM predicted SRTs lower then the measured ones in the S11 condition even though the model was fitted to that condition. This can be explained by the fact that the B-sEPSM was fit to the "left-ear" output only, rather than to the complete model output. Consequently, there seems to be a small advantage inherent to the binaural presentation in this condition, compared to the monaural presentation. The B-sEPSM produced an SRM of about 2 dB, compared to the 4 dB observed in the data. The BE-sEPSM output predicted no release from masking because there was no BE benefit possible; the masker was identical in both ears except for the fixed delay which is about an order of magnitude shorter than all processing windows in the model. In contrast, the BU-sEPSM output could account for all the SRM observed in the data.

4.5 Discussion

This study described a binaural extension of the mr-sEPSM model framework, which combined monaural implementations of the mr-sEPSM with the EC model implementation of Wan et al. (2014). The regular mr-sEPSM process was applied to the envelopes at the output of the EC process, and a selection stage selected the best SNR_{env} from the left-ear, the right-ear—equivalent to better-ear processing—or the binaural unmasking pathway. The model was validated against the data of Hawley et al. (2004), Beutelmann et al. (2010), and of Lőcsei et al. (2015). Overall, the correlation coefficients between simulated and measured SRTs were equal to 0.91. See Table 4.2 for a summary of all model performances.

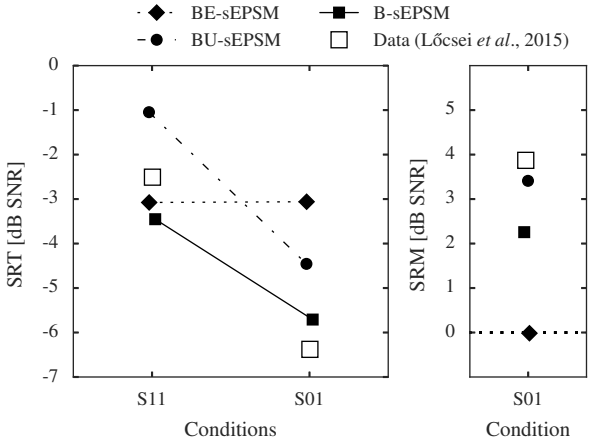


Figure 4.6: The left panel shows speech reception threshold data (open squares) from Lőcsei et al. (2015), B-sEPSM predictions (black squares), BE-sEPSM predictions (solid lines), and BU-sEPSM predictions (dashed lines) as a function of condition. In condition S11, both target and maskers are lateralized to the left and in S01, the target is lateralized to the left and the masker to the right. The right panel shows the same data and predicted, replotted as spatial release from masking relative to the S11 condition.

Table 4.2: Summary of correlation coefficients, r^2, and biases (in dB) for each model and in Exps. 1 and 2. The proposed model is the B-sEPSM and the BE-sEPSM and the BU-sEPSM are alternate outputs which consider only the better-ear, or only the binaural unmasking, respectively. The STEC model is from Wan et al. (2014) and the stBSIM is from Beutelmann et al. (2010). There are not performance values for Exp. 3 because it consisted of only two data points.

Model	Exp. 1 r^2/Bias	Exp. 2 r^2/Bias
B-sEPSM	0.91/−0.7	0.91/−6.0
BE-sEPSM	0.83/2.8	0.91/−3.3
BU-sEPSM	0.90/0.4	0.92/−4.7
STEC	0.97/−0.3	−
stBSIM	−	0.89/−2.8

4.5.1 Comparison to other modeling approaches

Both the proposed model and the STEC (Wan et al., 2014) had correlation coefficients above 0.9 in Exp. 1 (Hawley et al., 2004). However, the two models differ in a few important ways. Unlike the STEC, the B-sEPSM required a single parameter fit for the intelligibility transform for the whole experiment, rather than once per sub-condition. In fact, the B-sEPSM, and sEPSM framework in

general, requires a single parameter fit per speech material. In contrast, the STEC approach was validated using a different SII criterion (SII corresponding to 50 % intelligibility) for each type and number of maskers. The generality of the sEPSM approach to model fitting was validated in the current study, as well as in Jørgensen and Dau (2011) and Jørgensen et al. (2013). Given the appropriate reference condition, which is typically in the presence of an SSN masker, the mr-sEPSM and its variants could account for a large range of processing or masker types, which means that the model requires less *a priori* knowledge about each condition. Another difference is that in the B-sEPSM, the BE and BU pathways are processed using similar time-frames, i.e. all pathways use the same multi-resolution approach to slice the time signals into segments. This means that the B-sEPSM can account for the monaural presentation of speech against a modulated masker because all pathways include short-term processing of the signals, and therefore the B-sEPSM would be compatible with the monaural mr-sEPSM. In contrast, only the BU pathway of the STEC considers a short-term process; the left- and right-ear pathways are applied to the long-term signals only. This is an important limitation of the STEC approach, considering the ability of the auditory system to extract information from BE glimpses, even if they shift across ears (Brungart and Iyer, 2012).

In Experiment 2, the proposed model had a similar correlation coefficient as the stBSIM, but a slightly larger RMSE. Both the B-sEPSM and the stBSIM required a single parameter fit to convert the output of their decision metric to intelligibility. Unlike the B-sEPSM and the STEC, which explicitly separate the BE from the BU processes, the stBSIM implicitly includes the BE process in its closed-form calculation of the effective SNR (Beutelmann et al., 2010, their Eq. 12). It would be possible, however, to create a BE-only version of the stBSIM by removing the ITD-related parameters from that equation, or conversely, to create a BU-only version of the model by removing the ILD-related parameters. However, this binding of the two binaural processes limits the feasibility of modifying the processes independently, e.g., to use different time scales for the frame processing, or to introduce different amounts of sluggishness in each pathway (Culling and Summerfield, 1998; Culling and Mansell, 2013).

Neither the model of Lavandier and Culling (2010) nor any of its extensions was considered in the present study (Lavandier et al., 2012; Collin and Lavandier, 2013). Of the extensions, only the one of Collin and Lavandier (2013) could possibly account for the masking release due to fluctuating maskers used in the

majority of conditions considered in this study, because it is the only version that includes a short-term process. Those models are fundamentally limited because they cannot account for the effect of reverberation on the speech itself because they are not "signal-based", i.e., they do not use speech signals as targets, but rather rely on SSN as the target or on binaural room impulses. These simplifications make those model faster to compute than the proposed model as well as the STEC and the stBSIM, which makes them better tools for, e.g., acoustical room design but limits their applicability in certain scenarios.

Compared to the other models (STEC, stBSIM, Lavandier and Culling (2010), and even the binaural STI (Van Wijngaarden and Drullman, 2008)), the B-sEPSM avoids the need for the explicit frequency weighting from the SII. Instead, the frequency and modulation frequency weightings are limiting the processing to "audible" audio and modulation frequencies (Chabot-Leclerc et al., 2014). Therefore, although the B-sEPSM includes the additional modulation-frequency dimension to the model framework, it reduces the number of fitted parameters required.

Overall, the modeling approach taken by the B-sEPSM, the STEC, and the BSIM did not differ largely. All three models combined a short-term EC process with time–frequency-specific cancellation parameters and a (short- or long-term) BE process. The main difference lay in the decision metric used by the B-sEPSM, namely the SNR_{env} rather than the audio SNR, and the fact that the B-sEPSM included an envelope-domain audio-frequency-selective process.

4.5.2 Role of the decision metric

The SII-based models would fail in conditions with nonlinear processing, such as noise reduction (Rhebergen et al., 2009). The stBSIM as well as the model Collin and Lavandier (2013) are also fundamentally limited in that they cannot account for the effects of reverberation on the speech itself, because they do not use speech as target signal. Only the binaural STI model (Van Wijngaarden and Drullman, 2008), which uses the modulation power reduction after processing as the decision metric, could account for effects of modulation processing, but this approach is also limited because it cannot account for the intelligibility with modulated maskers. The B-sEPSM is the only binaural modeling framework that could account for multiple modulated maskers, reverberation on the target and maskers, as well as non-linear processing. Although these types of processing were not considered in the current study, the mr-sEPSM has been

validated in such conditions (Jørgensen and Dau, 2011; Jørgensen et al., 2013; Chabot-Leclerc et al., 2014). No audibility-based model has been demonstrated to account for the change of intelligibility due to amplitude compression (Rhebergen et al., 2009). Although the mr-sEPSM was not shown to account for the deleterious effect of amplitude compression on speech intelligibility, it could account for spectral subtraction, and, also to phase jitter, given the addition of an across-channel process (Chabot-Leclerc et al., 2014).

4.5.3 Contributions of better-ear and binaural unmasking processes

The explicit separation of the BE and BU pathways in the B-sEPSM makes it possible to analyze their contributions separately. Moreover, the performance of those alternate models can be an indicator of the respective importance of the processes involved in binaural hearing. Overall, the BE- and BU-only simulations, denoted as BE-sEPSM and BU-sEPSM, respectively, showed good agreements between data and simulations. They are depicted as dotted and dashed lines, respectively, in Figs. 4.2–4.6. In Exp. 1, the BE-sEPSM had an overall correlation coefficient of 0.83 and the BU-sEPSM a correlation coefficient of 0.90, which both compare favorably with the complete model's correlation of 0.91 (see Table 4.2 for overview). The performances were similar in Exp. 2, with a correlation of 0.91 for the BE-sEPSM, 0.92 for the BU-sEPSM, and 0.91 for the complete B-sEPSM. In Exp. 3, the BE-sEPSM model failed completely to account for the masking release due to ITDs, as expected, whereas the BU-sEPSM predicted the masking release. The performance of the BE-only model supported the idea that better-ear glimpsing, both in time and in frequency, can account for large parts of spatial release from masking (Brungart and Iyer, 2012; Culling and Mansell, 2013) in realistic conditions. Glyde et al. (2013) suggested this statement to be valid only if the maskers produced mostly energetic masking, i.e., did not cause any confusion between the target and the maskers. This is in contrast to conditions where informational masking may be dominant, such as with certain speech maskers. Therefore, the good performance of the BE-sEPSM can be attributed to the fact that the maskers considered in the present study may have provided a similar degree of informational masking (SSN, SMSSN, multi-talker babble, and reversed-speech).

The BU-sEPSM model performed equally well as the complete model (B-sEPSM) overall, and could account for the entire SRM in Exp. 3. The difference in simulated SRT between the BE-sEPSM and the BU-sEPSM can be attributed

to the fact that both models used the same "left-ear" reference for the fitting of the ideal observer. This discrepancy suggests that either they should be fitted separately, or that the processes should be modified as to produce the same SNR_{env} values in the same colocated condition. In Exp. 1 and 2, the BU output "dominated" the B-sEPSM output, because its SNR_{env} values were larger than that of the BE-sEPSM (which leads to lower SRTs), as it is especially clear in Fig. 4.2. Also, the BU-sEPSM tended to predict a larger masking release than the BE-sEPSM (c.f., Figs. 4.3 and 4.5). It is unclear if this dominance of the BU pathway is an artifact of the modeling or if it is a property of the human binaural system. If the lower SRT predicted by the BU-sEPSM compared to the BE-sEPSM are modeling artifacts, then they could possibly be mitigated by the inclusion of sluggishness to the EC process Culling and Summerfield (1998) and Culling and Colburn (2000) or by an increase of the EC jitters, which would limit its efficacy. Additionally, it may be that the constant short 20 ms windows of the EC process give the BU an advantage over the monaural pathways, where the multi-resolution approach is used. The EC window lengths could be adjusted or limited to restrict this advantage.

Some binaural models of speech intelligibility consider binaural unmasking as an additive process, while others do not. According to Culling and Mansell (2013), intelligibility benefits due to ILD and ITD seem to be additive. The modeling approach of Lavandier and Culling (2010) works under the same assumption that the total binaural advantage is the sum of the BE advantage and the advantage due to ITD processing (BMLD). In the model, only ILDs are considered in the BE pathway and only ITDs are considered in the BMLD pathway. The BSIM approach also indirectly uses this approach, where the ITD contributions can improve the SNR beyond the "better-ear" SNR (Beutelmann et al., 2010). In contrast, the B-sEPSM and the STEC use a selection between the BU and BE, as if they are two separate processes and one of them can outperform the other in a given situation. In these two models, both ILDs and ITDs are considered in the BU pathway. Culling et al. (2004) studied the role of ILDs and ITDs using a subset of the conditions presented by Hawley et al. (2004). They considered the conditions with three speech or three SSN maskers, but presented binaural signals that had only ILDs, only ITDs, or were unmodified. They found the SRT patterns of the ITD-only and unmodified conditions to be similar, although the ITD-only condition had smaller differences between the spatial configurations. The ILD-only condition showed an SRM only when

all maskers were on the right, otherwise the SRTs were the same as when all maskers were colocated with the target. For both masker types, considering the overall binaural advantage as the sum of the BE SRM and of the ITD SRM would lead to a large overestimation of the SRM in the unmodified condition. Therefore, in this condition, an "additive" binaural process is not appropriate and a selection process, such as in the B-sEPSM and STEC, seems more suitable.

4.5.4 Informational masking

The B-sEPSM predicted the correct spatial release from masking (SRM) in Exp. 1 with reversed-speech maskers (c.f., Fig. 4.3 although simulated SRTs were lower than the data (c.f., Fig. 4.2)). A similar difference was observed with the SSEC and the STEC (Wan et al., 2010; Wan et al., 2014) in the same condition. However, the models could not account for the increased thresholds observed when target and speech, or reversed-speech maskers, were colocated (Westermann and Buchholz, 2015b; Carlile and Corkhill, 2015). This limitation was even more clearly illustrated by Wan et al. (2014) in the conditions of Marrone et al. (2008), where the target was placed at 0° azimuth and speech or reversed-speech maskers were either colocated with the target or symmetrically placed around it. The models predicted SRTs lower than the data in the colocated condition because they could not account for the increased IM. In this case, IM is attributed to a failure in bottom-up grouping and streaming caused by target–masker similarities (Shinn-Cunningham, 2008). This is in contrast to the other portion of IM which can be attributed to top-down processes that cannot select the proper stream due to object similarity and target uncertainty (Shinn-Cunningham, 2008).

Being a purely bottom-up model, the B-sEPSM could only be sensitive to the similarity-based IM. However, the B-sEPSM has "perfect" segregation because of its access to the noisy mixture and to the maskers-alone signals and therefore cannot account for any IM. This means that the B-sEPSM requires fitting to a condition without IM, otherwise other simulated thresholds, where IM is not dominant, will be systematically elevated (e.g., in spatially separated conditions). On the converse, simulated SRTs in IM-dominated conditions will be too low if the B-sEPSM is fitted to an IM-free condition, which is the "default" approach for the mr-sEPSM framework. To account for the discrepancy between predicted and measured SRT in IM-dominated conditions, the B-sEPSM would require an estimate of the bottom-up confusion. Chabot-Leclerc and Dau (2014)

showed that it was possible to capture 7 of the 10 dB of SRM observed when a speech maskers was moved, on-axis, from 0.5 m to 10 m away from the target in a reverberant environment (Westermann and Buchholz, 2015a) using the long-term sEPSM. Models based on the audio SNR (e.g., SII, BSIM) did not predict any SRM. Therefore, it seems that it is possible to capture some of the similarity/dissimilarity in the envelope-power representation which is not available in the audio domain. Consequently, it should be possible to evaluate the similarity between the speech and maskers using an estimate of the clean speech representation ($\hat{S} = (S+N)-N$) and the maskers-alone representation in the envelope power domain. A simple "distance" or "contrast" estimate between the clean speech estimate and the maskers could be a promising measure of confusions. A more complex approach for estimating confusions would be to pair the B-sEPSM with a streaming model (e.g., Elhilali and Shamma, 2008; Christiansen et al., 2014) and combine their outputs considering that there are more confusions in a one-stream percept than in a two-stream percept. It would be particularly interesting to apply this approach to the output of the binaural unmasking pathway considering that BE seems to be sufficient to account for SRM when there is no IM (Glyde et al., 2013; Brungart and Iyer, 2012; Carlile and Corkhill, 2015).

4.6 Conclusions

The B-sEPSM is a general model framework for predicting spatial release from masking in realistic and artificial conditions. It combines an explicit combination of better-ear and binaural unmasking processes using monaural implementations of the mr-sEPSM (Jørgensen et al., 2013) and an EC process (Wan et al., 2014). The B-sEPSM uses the SNR_{env} as the decision metric and was shown to predict the SRT dependence on: the number of maskers, different masker types (SSN, SMSSN, babble, and reversed speech), the masker(s) azimuths, reverberation on the target and masker, and the ITD of the target and masker.

Acknowledgements

Thank you to Rainer Beutelmann for providing us with the room impulse responses used in Experiment 2, and to Gustav Lőcsei for the material used in Experiment 3. This research was supported in part by the National Science and

Engineering Research Council of Canada (NSERC), Phonak, and the Technical University of Denmark.

5

Predicting masking release of lateralized speech[a]

Abstract

Lőcsei et al. (2015) measured speech reception thresholds (SRTs) in anechoic conditions where the target speech and the maskers were lateralized using interaural time delays. The maskers were speech-shaped noise (SSN) and reversed babble with 2, 4, or 8 talkers. For a given interferer type, the number of maskers presented on the target's side was varied, such that none, some, or all maskers were presented on the same side as the target. In general, SRTs did not vary significantly when at least one masker was presented on the same side as the target. The largest masking release (MR) was observed when all maskers were on the opposite side of the target. The data in the conditions containing only energetic masking and modulation masking could be accounted for using a binaural extension of the speech-based envelope power spectrum model (sEPSM; Jørgensen et al., 2013), which uses a short-term equalization–cancellation process to model binaural unmasking. In the conditions where informational masking (IM) was involved, the predicted SRTs were lower than the measured values because the model is blind to confusions experienced by the listeners. Additional simulations suggest that, in these conditions, it would be possible to estimate the confusions, and thus the amount of IM, based on the similarity of the target and masker representations in the envelope power domain.

[a] This chapter is based on Chabot-Leclerc et al. (2015).

5.1 Introduction

Listeners benefit from listening with two ears compared to a single ear in complex listening situations. This binaural benefit is usually explained in terms of "better-ear" (BE) and binaural unmasking (BU) concepts. The former relies on interaural level differences (ILDs) caused by the acoustical "shadow" cast by the head, which creates an advantageous signal-to-noise ratio (SNR) at the ear contra-lateral to the masker. In the latter, the interaural time differences (ITDs) give the hearing system the ability to increase the effective SNR by "cancelling" some of the masker signals (equalization–cancellation (EC) theory; Durlach, 1963).

The BE benefits are typically modeled in terms of audibility (Beutelmann et al., 2010; Lavandier and Culling, 2010; Wan et al., 2014), with a decision metric such as the speech intelligibility index (SII; ANSI, 1997). In other words, those models consider only energetic masking (EM), where EM is defined as masking of the peripheral representation of the signal. However, Stone et al. (2012) showed that noises that are typically considered "steady", such as speech-shaped noise (SSN), actually behave more as modulation maskers than as energetic maskers, i.e., they provide "modulation masking" (MM). Yet, EM and MM may not be sufficient to account for speech intelligibility data for some masker types, such as speech, in which case the unaccounted-for masking is labeled as "informational masking" (IM). According to Watson (2005), IM can be divided into two categories, uncertainty and similarity. Uncertainty is explained as a listener's inability to identify the target, whereas similarity prevents a listener from segregating the target and the masker. Multiple factors can reduce the similarity between target and masker, such as spatial separation and fundamental frequency (F_0) information, and thus reduce IM (Bronkhorst, 2000).

The present study investigated the contributions of MM and IM and their interactions in an ITD-only binaural condition with a variable number of maskers (Lőcsei et al., 2015) using a binaural extension of the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen et al., 2013; Chabot-Leclerc et al., 2016). The mr-sEPSM framework considers MM using the SNR in the envelope domain (SNR_{env}) as the decision metric and was shown to account well for intelligibility where IM was not the dominating factor, such as with SSN maskers, sinusoidally modulated maskers, or multi-talker babble. Here, the maskers under consideration were SSN and time-reversed speech

maskers, the latter known to produce informational masking, although not as much as regular speech (Rhebergen et al., 2005). In particular, the focus was to analyze how well the SNR_{env} metric could capture the intelligibility change as a function of the total number of maskers and the masker configuration and what could be attributed to IM.

5.2 Model description

The structure of the proposed model is presented in Fig. 5.1.[a] It consists of two monaural realizations of the mr-sEPSM (Jørgensen et al., 2013) and a binaural unmasking pathway implemented as an EC process (Wan et al., 2014).

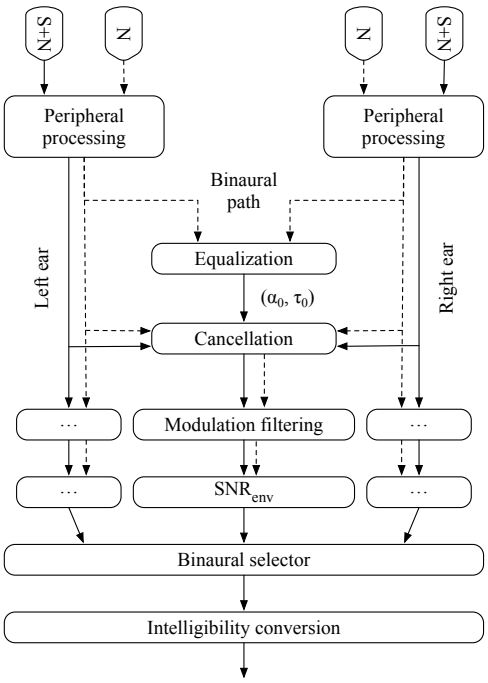


Figure 5.1: Structure of the proposed model.

The model takes as input the noisy speech and the noise-alone signals for each ear. Each signal is processed through a filterbank of 22 gammatone filters covering the frequency range from 63 Hz to 8 kHz with a third-octave spacing. The sub-band envelopes are then extracted using half-wave rectification

[a] The model described here is described in more details in Sec. 4.2. The only difference here is that the ideal observer has been replaced by a simple Gaussian psychometric function.

followed by a fifth-order Butterworth low-pass filter with a cutoff frequency of 770 Hz (Breebaart et al., 2001). Jitter in the time and amplitude domain is applied independently to each sub-band envelope to limit the efficacy of the EC process; all jitters are zero-mean Gaussian processes with standard deviations of $\sigma_\delta = 105$ μs for the temporal jitter and of $\sigma_\epsilon = 0.25$ for the amplitude jitter (Durlach, 1963). In the monaural pathways, the envelopes are further processed by a modulation filterbank consisting of eight second-order bandpass filters with octave spacing between 2 and 256 Hz. A third-order low-pass filter with a 1-Hz cutoff frequency is applied in parallel to the filterbank. Only modulation filters with center frequencies below one-fourth of their respective peripheral-filter center frequency are used (Verhey et al., 1999). The output of each modulation filter is then divided into non-overlapping segments of duration inversely proportional to the modulation filter's characteristic frequency, i.e., the output of the 4 Hz filter is divided into 250 ms segments. The power of each segment is calculated as its variance and the lower limit of the envelope power is set to −30 dB relative to 100 % modulation. The $\mathrm{SNR}_{\mathrm{env}}$ for each segment, i, peripheral channel, p, and modulation channel, n, is calculated as:

$$\mathrm{SNR}_{\mathrm{env},i}(p,n) = \frac{P_{\mathrm{env},S+N,i}(p,n) - P_{\mathrm{env},N,i}(p,n)}{P_{\mathrm{env},N,i}(p,n)}, \tag{5.1}$$

where $P_{\mathrm{env},S+N}$ is the power of the noisy speech mixture and $P_{\mathrm{env},N}$ is the power of the noise alone.

The binaural unmasking stage is implemented as described in Wan et al. (2014). The jittered envelopes at the output of the peripheral filterbank are the inputs to the EC process, which is applied independently in each channel as well as in short 20 ms time frames, k. For each time-frequency frame, the equalization stage selects the optimal ITD, τ_0, and ILD, α_0, using the following equations:

$$\tau_0 = \arg\max_{\tau} \{\rho\}, \ |\tau| < \frac{\pi}{\omega}, \text{ and} \tag{5.2}$$

$$\alpha_0 = \sqrt{\frac{E_{N,L}}{E_{N,R}}}, \tag{5.3}$$

where ρ is the normalized cross-correlation function of the left and right ears within the frame, $E_{N,L}$ and $E_{N,R}$ are the masker energies for the left and right ear, respectively, and ω is the center frequency of the channel of interest. Subse-

quently, the sub-band signal, B_p, is reconstructed for each channel by summing over all frames.

The unmasked outputs for the noisy speech and the noise alone are then used as inputs to the modulation filtering stage of the mr-sEPSM and processed similarly to the monaural pathways, yielding a binaurally unmasked $\mathrm{SNR}_{\mathrm{env}}$, $\mathrm{BU–SNR}_{\mathrm{env}}$.

A selection stage then selects the best $\mathrm{SNR}_{\mathrm{env}}$ of the left, right and binaural pathways, yielding the complete model's output, the $\mathrm{B\text{-}SNR}_{\mathrm{env}}$. The $\mathrm{B\text{-}SNR}_{\mathrm{env}}$ is then averaged across time, and combined optimally across modulation and peripheral filters:

$$\mathrm{B\text{-}SNR}_{\mathrm{env}} = \left[\sum_{p=1}^{22} \sum_{n=1}^{9} \mathrm{B\text{-}SNR}_{\mathrm{env}}^2(p,n) \right]^{1/2}. \tag{5.4}$$

The final $\mathrm{B\text{-}SNR}_{\mathrm{env}}$ is then converted to intelligibility using a Gaussian psychometric function. The left- and right-ear pathways are combined and converted similarly, yielding alternate model outputs for each ear.

More details about the mr-sEPSM framework and the EC process implementation can be found in Jørgensen et al. (2013) and Wan et al. (2014), respectively.

5.3 Methods

In this experiment, the speech and masker signals were lateralized individually to the left or right using fixed 33-sample delays (687.5 μs) and the spatial distribution of maskers was systematically varied. The speech material was the DAT corpus (Nielsen et al., 2014), sampled at 48 kHz and recorded by female speakers. The DAT corpus consists of unique meaningful Danish sentences built as a fixed carrier sentence with two interchangeable target words. The maskers were either of one stationary SSN, denoted as $\mathbf{S}_{xy}$ conditions, or 2, 4, or 8 time-reversed sentences from the GRID corpus (Cooke et al., 2006), denoted as $\mathbf{C}_{xy}$ conditions, where y is the total number of maskers and x is the number of maskers on the same side as the target. Both the SSN and the GRID material were shaped to have the same long-term spectrum as the target speech material. The maskers were either all on the same side as the target (e.g., $\mathbf{C}_{44}$), half on the same side (e.g., $\mathbf{C}_{24}$), or all on the opposite side (e.g., $\mathbf{S}_{04}$). The target level was fixed at 65 dB SPL and the maskers were summed before their levels were

adjusted to the desired SNR. Model predictions were calculated for 30 randomly selected sentences and for SNRs ranging from −12 to 9 dB in 3 dB steps. The predicted SRT was the average across target sentences. The mean and standard deviation of the psychometric function were fitted to minimized the square error between the "left-ear" of the model and the word-scores as a function of SNR in the collocated condition ($\mathbf{S}_{11}$), as measured by Lőcsei et al. (2015).

5.4 Results

Figure 5.2 shows the speech reception thresholds (SRT) measured by Lőcsei et al. (2015) (open squares), the predictions by the proposed model (B-sEPSM; filled squares), as well as the predictions by the left- and right-ear outputs of the B-sEPSM (left- and right-pointing triangles, respectively) for each masker type and configuration. In the $\mathbf{S}_{x1}$ conditions with SSN maskers, the B-sEPSM predicted SRTs lower than the data by 0.5 to about 3 dB, but captured the MR when the maskers were moved to the opposite side. In the $\mathbf{C}_{x8}$ condition, the B-sEPSM accurately captured the MR when 4 and then all 8 reversed-speech maskers were lateralized to the other side. In the $\mathbf{C}_{x4}$ condition, the B-sEPSM predicted a similarly progressive MR as in the $\mathbf{C}_{x8}$ condition, as 2 or all 4 maskers were lateralized to the other side. This is in contrast to the data, where the SRT was constant at about −2.5 dB when 4 or 2 of the maskers were on the same side as the target and then there was about 5 dB of MR once all maskers were on the other side. In the $\mathbf{C}_{x2}$ condition, the B-sEPSM predicted constant SRTs of about −10 dB, irrespective of the masker arrangement. In contrast, the data SRTs were about the same when 2 or 1 masker(s) were collocated with the target at about −4 dB (not significant differences ($p < 0.05$; Lőcsei et al., 2015)) and then decreased by 4 dB once all maskers were on the other side, similar to the $\mathbf{C}_{x4}$ condition. The SRTs predicted by the left- and right-ear models (left- and right-pointing triangles) depended only on the total number of masker and masker type, irrespective of their configuration. The SRTs were highest in the $\mathbf{C}_{x8}$ and lowest in the $\mathbf{C}_{x2}$ condition, consistent with the increased number of dips in the two-masker condition. Overall, the Pearson correlation coefficient between the B-sEPSM predictions and the data was 0.78 and the mean absolute error was 2.24 dB.

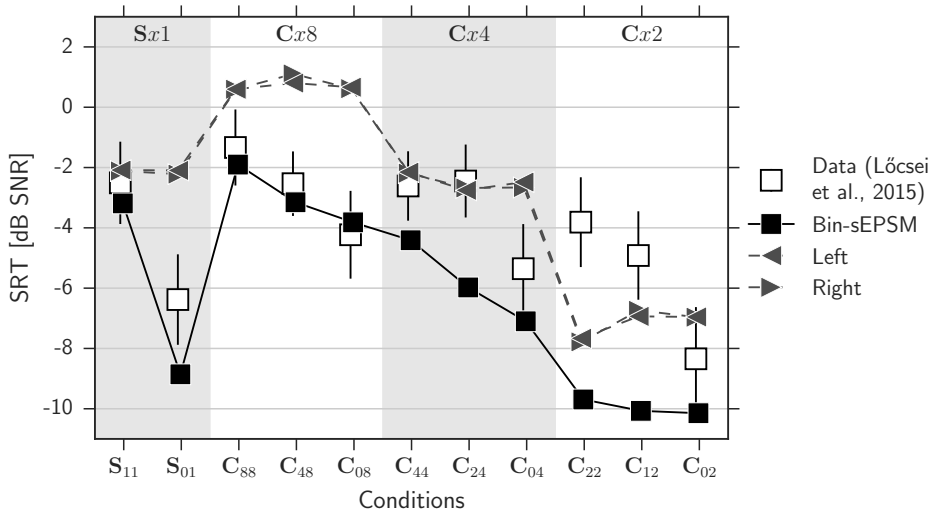


Figure 5.2: Measured SRTs (open squares; Lőcsei et al., 2015) and predictions by the B-sEPSM (filled squares) and the 'left-' and 'right-ear' models (triangles) for each condition. **S** conditions are with SSN maskers and **C** conditions are with reverse-babble maskers.

5.5 Discussion

The B-sEPSM could account well for the MR due to lateralization in the conditions with the SSN masker ($\mathbf{S}_{x1}$ conditions) and also accurately predicted the SRTs and MR in the $\mathbf{C}_{x8}$ conditions. However, the model was "too good" once the number of maskers was small enough such that IM became the dominating factor, i.e., in the conditions $\mathbf{C}_{x4}$ and $\mathbf{C}_{x2}$. A possible explanation framework has been put forward by Best et al. (2013), where it was suggested that intelligibility has a "lower limit" (of SRT) corresponding to the EM/MM present in the condition. In this case, the model's failure can be explained by the fact that it is blind to IM, and thus predicts the lower limit of intelligibility, given EM and MM only.

It is assumed that the mr-sEPSM framework has "perfect segregation" due to its access to the noisy-speech mixture and the noise-alone signals. Therefore, if most of the IM is due to confusion caused by the similarity between the target and maskers, and not to uncertainty about the target, then the B-sEPSM is blind to those confusions (Watson, 2005). An estimate of those confusions in the model would allow it to account for some of the IM in the listener. A possible approach would be to use a model of streaming, such as Elhilali and Shamma (2008) or Christiansen et al. (2014), and to combine its output with

the intelligibility model's output; a single-stream percept would lead to worse intelligibility than a multi-stream percept. Although that approach might prove powerful and possibly more realistic, it would greatly increase the complexity of the models, to the extent that two internal representations would be required. Figure 5.3 shows a potential similarity measure, calculated as a "modulation distance" between the speech estimate (i.e., $(S+N)-N$) and the noise-alone representations, N, as a function of the SNR and for different masker configurations. Given the three-dimensional representation of the envelope power as a function of sub-band frequency, modulation frequency, and time frames, i, the "modulation distance", d_{mod}, is calculated as the Euclidean distance between the sub-band and modulation frequency representation (i.e., a 2D matrix) of the speech estimate and the noise for each time frame:

$$d_{\text{mod},i} = \sqrt{\left(P_{\text{env},S+N,i} - P_{\text{env},N,i}\right)^2 - P_{\text{env},N,i}^2} \tag{5.5}$$

where $P_{\text{env},S+N,i}$ is the envelope power of the mixture, $S+N$, for each segment i, and similarly for the noise, $P_{\text{env},N,i}$. The "distance" is then averaged across all time frames:

$$d_{\text{mod}} = \frac{1}{I}\sum_{i}^{I} d_{\text{mod},i}. \tag{5.6}$$

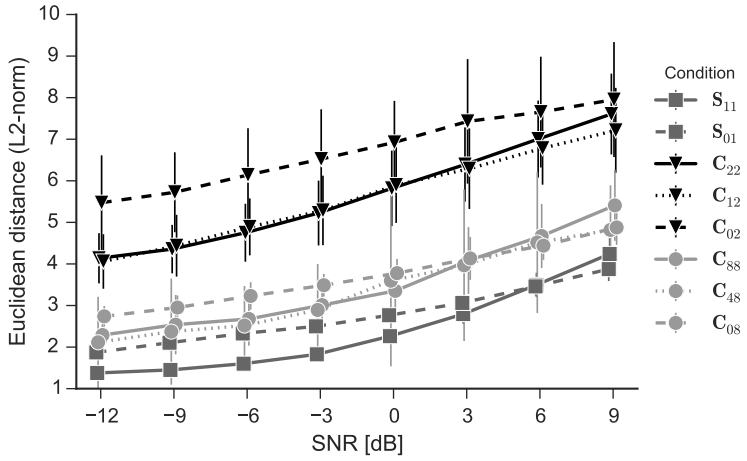


Figure 5.3: Euclidean distance between the speech estimate and the noise in the envelope power domain, as a function of SNR. Each line represents a different condition.

In Fig. 5.3, the black lines show the distance for the $\mathbf{C}_{x2}$ condition, where most IM was observed. The distance was largest in $\mathbf{C}_{02}$ condition (dashed line),

whereas the distances for conditions $\mathbf{C}_{22}$ and $\mathbf{C}_{12}$ (solid and dotted lines) were almost the same. This mirrors the data, where an MR was observed once all maskers were not collocated with the target, i.e., confusions were resolved once spatial cues were available. In contrast, the distance varied much less as a function of masker location when MM was the dominating factor, such as in the SSN maskers conditions (dark gray lines, $\mathbf{S}_{x1}$) and in the eight-reversed speech masker conditions (light gray lines, $\mathbf{C}_{x8}$).

5.5.1 Application of the "modulation distance" to a monaural experiment

Chapter 3 considered data from a study by Westermann and Buchholz (2015a), where they measured SRTs as function of the distance of the maskers. The maskers were either speech modulated SSN (SMSSN) or two talkers of the same gender as the target. They were colocated with the target, 0.5 m away in front of the listener, or they were moved to 2, 5 or 10 m away. The listeners experienced a release from masking of about 10 dB when the speech maskers went from being 0.5 m away to 10 m away. The SRTs measured with the SMSSN masker were about the same as when the speech maskers were at the 10 m position. In these conditions, the mr-sEPSM did not predict any release from masking; neither with speech maskers nor with the SMSSN maskers, presumably because it could not capture the change in amount of confusion experienced when the distance of speech maskers was changed. The "modulation distance" suggested above was introduced in the mr-sEPSM, denoted as mr-sEPSM-D, in order to validate the hypothesis that it could account for the change in the amount of confusions. The only difference with the regular mr-sEPSM is that the time-averaged $\mathrm{SNR}_{\mathrm{env}}(p,n)$ values as a function of audio frequency, p, and modulation frequency, n, are weighted by the inverse of the modulation distance, d_{mod}:

$$\mathrm{SNR}_{\mathrm{env}}(p,n) = \frac{\mathrm{SNR}_{\mathrm{env}}(p,n)}{d_{\mathrm{mod}}}. \tag{5.7}$$

Figure 5.4 replots the data shown in Chap. 3 as SRTs, rather than as values of spatial release from masking. The left panel shows SRTs in the speech maskers condition, and the right panel in the SMSSN masker condition. The predictions obtained with the mr-sEPSM-D are shown as filled squares and the predictions obtained with the regular mr-sEPSM are shown as grey triangles. The predictions were fitted to the condition where the speech maskers were at the 10 m position,

where IM was assumed to be at a minimum. The mr-sEPSM predicted almost constant SRTs, independently of the masker type and position, as in Chap. 3. In contrast, the mr-sEPSM-D could account for the large SRT difference between the speech maskers placed at 0.5 or 10 m. The mr-sEPSM-D predicted SRTs higher than the measured ones when the SMSSN masker was nearby and about the same as the measured ones when it was far.

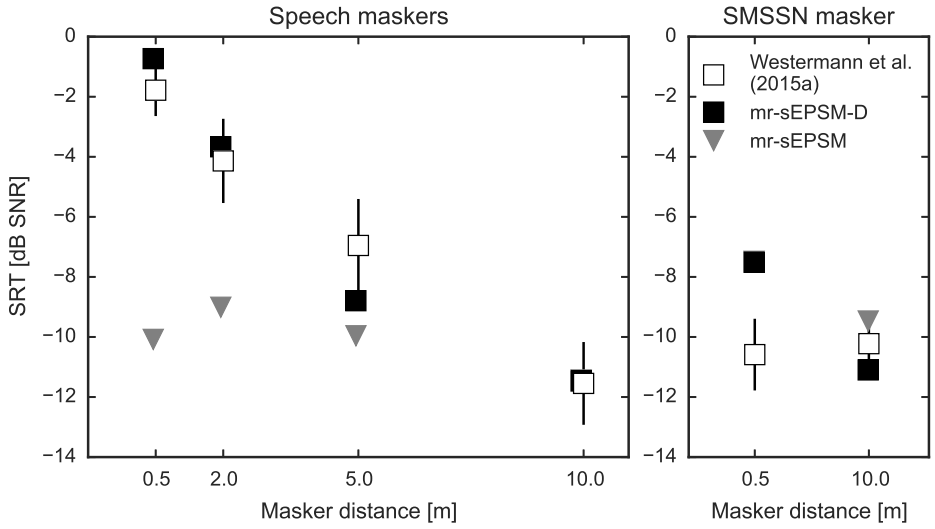


Figure 5.4: Measured SRTs and model predictions as a function of masker distance for speech maskers (left panel) and SMSSN maskers (right panel). The open square symbols denote the data measured by Westermann and Buchholz (2015a), the filled squares represent the predictions obtained with the mr-sEPSM-D, which includes the "modulation distance" metric, and the grey circles denote predictions obtained with the mr-sEPSM. See Section 3.3 for more details about the experimental setup.

The simulation results obtained with the mr-sEPSM-D suggest that the modulation distance is actually proportional to the inverse of what would be considered as confusions. In other words, an increasing modulation distance correlates with decreasing intelligibility, at least in the monaural mr-sEPSM framework. This is in contrast to the direction in which the modulation distance varied at the output of the BU process of the B-sEPSM (c.f., Fig. 5.3), where an increase in the modulation distance correlated with an increase in intelligibility. For example, the distance was larger when the target and the maskers were localized in different ears than when they were localized in the same one. Therefore, it it still unclear how the modulation distance should be considered in the B-sEPSM framwork. However, the modulation distance seems to be a reliable

measure in that it depends on the properties of the target and the maskers, and not only on the audio SNR or SNR_{env}.

5.6 Summary and conclusion

The B-sEPSM could accurately predict SRTs when the dominating factor was modulation masking, but failed when IM became more prevalent. It seems that similarity information between the target estimate and the maskers is available in the multi-resolution envelope power representation and that it could be used to account for some of the IM. However, more work is required in order to combine this information with the binaural model predictions.

Acknowledgments

This research was supported in part by the National Science and Engineering Research Council of Canada (NSERC), Phonak, and the Technical University of Denmark.

6 Overall discussion

6.1 Summary of main results

Chapter 2 presented an analysis of the role of spectro-temporal modulation processing for speech perception. Predictions obtained with the (long-term) speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011) were compared to predictions made by two model variants. The first, named the 2D-sEPSM, used a spectro-temporal modulation front-end inspired by the work of Elhilali et al. (2003). The second, named the sEPSM^{X}, kept the same one-dimensional pre-processing as the regular sEPSM, but included a measure of the variability of the modulation power across audio-frequency at the output of the modulation filterbank. This across-channel process was inspired by models of comodulation masking release (CMR; Par and Kohlrausch, 1998; Piechowiak et al., 2007; Dau et al., 2013). All three sEPSM versions use the SNR_{env} metric. The role of the decision metric was evaluated by comparing the predictions obtained with the three sEPSM versions to the predictions obtained with the spectro-temporal modulation index (STMI; Elhilali et al., 2003), which uses a modulation transfer function (MTF) metric. A diagram of the pre-processing and decision metric combinations and how they relate to the different intelligibility models is shown in Fig. 6.1. Only the models that considered an across-channel process could account for the intelligibility in the phase jitter condition where the spectral structure of the speech was destroyed. The sEPSM^{X} performed as well as the 2D-sEPSM and the STMI, although it used a simpler long-term measure of the across-channel variability. Therefore, a complex spectro-temporal modulation filter front-end does not seem to be necessary in the conditions considered. It was also shown that only the models that use the SNR_{env} metric could account for the intelligibility of noisy speech processed by noise reduction via spectral subtraction. The STMI, like the speech transmission index (STI), failed because it used a MTF-based metric (Ludvigsen et al., 1993; Dubbelboer and Houtgast, 2007). Both the STMI

and the STI predicted an increase in intelligibility, instead of a decrease, when spectral subtraction was applied. The MTF concept fails because it considers the modulation energy introduced by the noise reduction scheme to be contributing positively to intelligibility. In contrast, the SNR_{env} correctly captures the increase in the intrinsic noise produced by the noise reduction, predicting a decrease in intelligibility. The decision metric did not play a role in the phase jitter condition. Overall, the $sEPSM^X$ was the most accurate model, with both the largest correlation coefficient across conditions (0.99) and the smallest root-mean-square error.

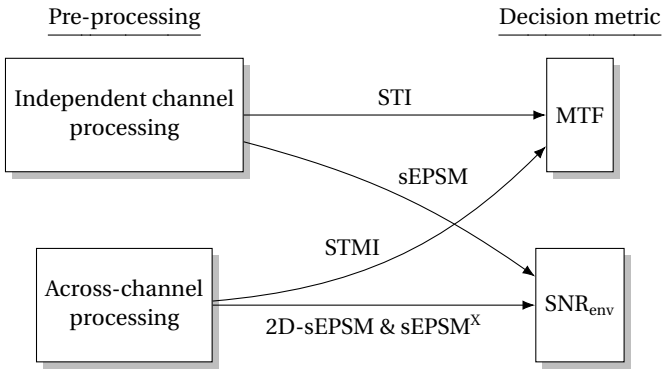


Figure 6.1: Relationships between different types of pre-processing and decision metrics and how they relate to different speech intelligibility models.

Chapter 3 investigated the role of the decision metric and of binaural processing in a reverberant condition where a target was placed in front the listener and a pair of speech maskers were either colocated with the target or moved away, on axis (measured by Westermann and Buchholz, 2015a). The long-term spectral coloration introduced by the increased distance was compensated for and the SNR was set at the ears of the listener, rather than at the sources, such that the distance perception was not conveyed through level cues. Predictions by the long-term (monaural) sEPSM and multi-resolution sEPSM (mr-sEPSM; Jørgensen et al., 2013), which both consider modulation masking (MM) using the SNR_{env} metric, were compared to predictions obtained with the (long-term) binaural speech intelligibility model (BSIM; Beutelmann et al., 2010) and the binaural model of Jelfs et al. (2011) (an implementation of the model of Lavandier and Culling (2010) that requires only binaural room impulse responses), which both consider energetic masking (EM) using the audio SNR for their deci-

sion metric. None of the SNR-based models predicted a release from masking, even though Westermann and Buchholz (2015a) measured a masking release of 10 dB in the binaural condition and 7 dB in the diotic condition (considered to be the monaural release from masking). The long-term (monaural) sEPSM predicted all 7 dB of monaural release from masking. The sEPSM was successful because it captured the modulation low-pass filtering effect of the room on the masker, which increased the SNR_{env}. However, the mr-sEPSM, which considers the SNR_{env} in short time frames, did not predict any release from masking. It was hypothesized that the short-term model predicted a decrease in intelligibility when the maskers were moved away because the maskers effectively became "more stationary" due to temporal smearing. Additional simulations using the extended speech intelligibility index (ESII; Rhebergen and Versfeld, 2005), which considers the SNR in short time windows, supported this hypothesis. The ESII predicted a decrease in intelligibility when increasing the maskers' distance, in contrast to the data. These simulation results suggested that the increase in EM due to temporal smearing is counteracted by a decrease in MM, which is captured by the mr-sEPSM. However, the decrease in MM alone was not sufficient in the mr-sEPSM to account for the release from masking observed in the data. Therefore, it seems that additional cues that were not captured by any of the models are required to account for the complete release from masking.

Chapter 4 introduced a binaural extension of the mr-sEPSM, denoted as B-sEPSM. The B-sEPSM combines a monaural realization of the mr-sEPSM for each ear, as well as a third realization following an equalization–cancellation (EC; Wan et al., 2014) process modeling the binaural unmasking (BU) operation. A mechanism selects the largest SNR_{env} from the left, right, or the BU pathways for each audio and modulation frequency and for each time frame. Alternatively, it is also possible to select only the SNR_{env} from the left and right ear pathways, yielding a "better-ear" model (BE-sEPSM), or to only select the BU pathway (BU-sEPSM). The B-sEPSM was validated against a wide range of conditions from the literature. The proposed model could predict the SRT dependency on: the number of maskers, different maskers types (speech-shaped noise (SSN), speech-modulated SSN, multi-talker babble, and reversed speech), the masker(s) azimuths, reverberation on the target and the masker, and the ITD of the target and the masker. In the realistic conditions that contained both interaural level differences (ILDs) and interaural timing differences (ITDs), the model based only on the BE performed almost as well as the complete model,

suggesting that BE processing might be sufficient to account for intelligibility in conditions dominated by EM and MM (Glyde et al., 2013). The BE-sEPSM supports the observation of the hearing system's ability to build a "better-ear" signal using glimpses from the left and right ears (Brungart and Iyer, 2012; Glyde et al., 2013). The BU-only model also performed as well as the full model. This is in disagreement with the observations that benefits from BE and BU processing seems to be additive (Culling and Mansell, 2013), which is also the assumption made in the model of Lavandier and Culling (2010). In contrast, the model of Wan et al. (2014) relies mainly on the BU processing in the conditions with fluctuating maskers, suggesting that BE and BU process are not additive but are rather separate processes that are selected at a later stage. Therefore, the modeling results obtained in this chapter further supported the assumptions of Wan et al. (2014) regarding the selection of the BE and the BU components. Finally, the proposed model did not require explicit frequency weightings and its mechanism for conversion to intelligibility required only one calibration per speech material, in contrast to models based on the SII or STI, such as the short-term binaural speech intelligibility model (BSIM; Beutelmann et al., 2010) and the short-term EC model (STEC; Wan et al., 2014).

Chapter 5 further tested the binaural model presented in Chap. 4 in an ITD-only condition (Lőcsei et al., 2015) chosen to investigate the contributions of BU processing in the presence of EM, MM and so-called informational masking (IM). The maskers considered were either SSN or two, four, and eight reversed speech maskers from another talker. The speech target was lateralized to one side using a fixed 680 μs ITD and either i) all maskers were colocated with target, ii) some maskers were colocated with the target and some were located on the opposite side, or iii) all maskers were on the opposite side. The B-sEPSM predicted the correct SRTs only in the conditions with an SSN masker and in the condition with eight reversed talkers, i.e., in the conditions where MM was dominant. In the conditions with two or four reversed talked maskers, the predicted SRTs were consistently lower than the measured ones because the model was insensitive to the confusion experienced by the listeners, i.e., when IM was present. It was proposed to capture some of those confusions by looking at the "distance" between the estimate of the speech representation in the modulation domain ($(S+N)-N$) and the noise-alone representation. The magnitude of the "modulation distance" correlated with the release from IM in conditions with a given number of maskers; the distance increased with larger

release from masking, even though the SNR_{env} did not change.

6.2 Capturing aspects of informational masking in the modulation domain

Informational masking (IM) is the general term used here to define threshold elevations that cannot be explained in terms of energetic masking (EM) or modulation masking (MM). According to Watson (2005), IM can be considered to represent two components. The first is a bottom-up process caused by uncertainty at the signal level caused by a trial-to-trial uncertainty, such as a possible change of target, masker, or dimension of either signal, or the presence or absence of certain stimulus features. The second is a top-down process where the listener experiences confusions as to what the target should be because of similarities between the target and the masker. The latter is typically associated with speech recognition tasks. Shinn-Cunningham (2008) proposed a slightly different framework to explain IM, where bottom-up object-formation is contrasted with top-down attention. The object-formation can be hindered by similarities between the target and the masker, which prevents primitive grouping and streaming. The attention directs the object selection and can be impeded by similarities between the target and the maskers and/or by the uncertainty of the target. In both approaches, target–masker similarities are considered to be an important component of IM. The mr-sEPSM is not expected to account for IM, because it has no mechanism to estimate neither the target–masker similarities nor "failures" of attention. When the mr-sEPSM was evaluated in conditions that contain IM, it showed a consistent behavior of predicting SRTs that were lower than the measured ones. For example, Jorgensen and Dau (2013) tested the mr-sEPSM in conditions where speech was presented against SSN, speech-modulated SSN (SMSSN), a different-talker (DT), a time-reversed different talker (DT-R), and a time-reversed same-talker masker (ST-R). The mr-sEPSM accounted for the correct SRTs in the SSN and SMSSN conditions, but underestimated the measured SRTs by 2–3 dB in the DT conditions, and by 12 dB in the ST talker condition, where the target–masker similarities were large. Similar offsets were described in Chap. 4 and 5, where the mr-sEPSM predicted lower SRTs than observed in the data, presumably due to confusions experienced by the listener that are not considered by the model. This offset can be explained by the fact that the ideal observer of the model is fit only once for

a given speech material. Therefore, the model predicts only an offset between the condition under consideration and the reference condition This approach works well as long as the amount of IM does not change between conditions, for example if the model is fit to a condition with SSN and used in other conditions with SMSSN, car-cabin noise, or other maskers that do not contain IM. However, if the model is fit to a condition without IM, it will predict lower SRTs than those observed in the data in conditions that contain IM, because it assumes perfect segregation and predicts the (lowest) SRT that is only limited by MM and EM.

However, the simulations obtained in Chapters 3 and 5 suggested that it would be possible to estimate target–masker confusions in the envelope power domain representation. Figure 6.2 shows a conceptual representation of the results obtained in Chap. 3, where two speech maskers were moved away from the listener, on axis, in a reverberant environment. The ESII, which considers EM in short time frames, predicted a decrease in intelligibility due to an increased amount of masking (since the maskers became more stationary due to reverberation and temporal smearing). The sEPSM, which considers long-term MM, predicted an increase in intelligibility, as in the data, because it captured the low-pass filtering effect of the room in the modulation domain. The mr-sEPSM, which considers MM in short time frames, predicted no change in intelligibility; its use of the SNR_{env} metric counteracted the increased masking caused by the maskers becoming more stationary. Both SNR_{env}-based models predicted changes in intelligibility that were closer to the data, unlike the SNR-based model, suggesting that they captured signal aspects that were not considered in models that take only the audio domain into account. Therefore, it indicates that the listeners might have experienced improved segregation between the target and the masker because of changes in the modulation spectrum of the masker, not because of changes in EM.

Chap. 5 attempted to quantify this improved segregation, in a condition with multiple reversed-talker maskers. It was assumed that, although the sEPSM framework does not have access to the clean speech alone, it is possible to calculate an estimate of it by subtracting the noise representation from the noisy speech representation. This operation is already done in order to calculate the SNR_{env} (c.f., Eq. 2.5). Confusions were then assumed to be correlated to the Euclidean distance between the two-dimensional envelope power (audio-frequency–modulation-frequency), averaged over time, in a manner analogous to the calculation of the overall SNR_{env} in the mr-sEPSM. It is yet unclear how

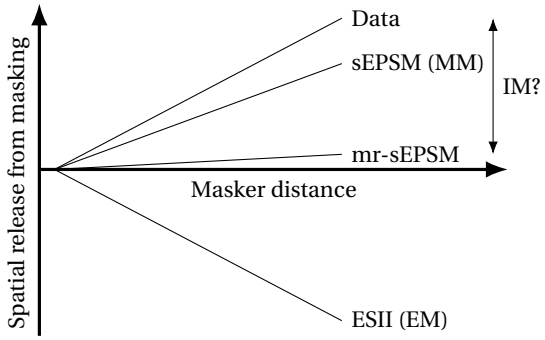


Figure 6.2: Conceptual representation of the data and simulations from in Fig. 3.1. Each line represents the spatial release from masking as a function of the masker distance for the data and different models. EM stands for 'energetic masking', and MM standands for 'modulation masking'. The arrow labeled 'IM?' informational masking' represents the portion of the release from masking that could be attributed to informational masking.

this additional metric should be integrated in the modeling framework but the fact the modulation distance covaries with the confusions seems promising for future modeling work.

6.3 Possible simplification to the sEPSM framework

The ideal observer concept for intelligibility conversion was proposed by Jørgensen and Dau (2011) and further defined by Jørgensen et al. (2013). It is an elegant construct that first converts the $\mathrm{SNR}_{\mathrm{env}}$ value to a d', which is then transformed to a proportion of correct response using an m-alternative forced-choice model, combined with an unequal-variance Gaussian model. This conversion requires the selection of four parameters: k and q, which are considered to be variables set experimentally, and σ_s and m, which are assumed to be constants linked to the response set size and redundancy. Although these parameters were shown to vary in the correct directions considering their definition, e.g., σ_s is inversely proportional to the slope of the psychometric function and is larger for speech materials with little redundancy than for highly redundant materials, four variables are still necessary in order to perform a transformation which could be made with a basic two-parameter psychometric function, such as a Gaussian cumulative distribution function or a logistic function. A two-parameter transform would have the advantage of directly providing the $\mathrm{SNR}_{\mathrm{env}}$ value which corresponds to the SRT for a given speech material, which would

simplify the use of the model in an m-alternative forced-choice experiment framework. Therefore, replacing the ideal observer with a simpler transform would not go against the original sEPSM concept but would simplify it and it would make the relationship between $\mathrm{SNR}_{\mathrm{env}}$ and intelligibility more transparent. There would be a single transform from $\mathrm{SNR}_{\mathrm{env}}$ to percentage of correct responses, rather than two transformations, first to d' and then to percentage of correct responses.

6.4 Binaural processing using the sEPSM framework

The binaural extension of the mr-sEPSM, the B-sEPSM, proposed in Chap. 4, consists of a combination of the monaural mr-sEPSM and the EC process as implemented by Wan et al. (2014). A selection device selects the best $\mathrm{SNR}_{\mathrm{env}}$ value from the left or right ear, modeling a better-ear (BE) process, or from the binaural unmasking (BU) pathway. The rate of selection is not limited in any way other than by the window durations used in the multi-resolution process, and by the fixed 20 ms windows used in the EC-process stage (Wan et al., 2014). Culling and Mansell (2013) measured spatial release from masking (SRM) in an experiment where square-wave modulated noise maskers were placed symmetrically on either side of a listener. The maskers were modulated out of phase at different rates in order to test the speed at which the system could switch back and forth between the two ears. They demonstrated that the binaural system was very sluggish and that modulations of 5 Hz or more reduced the BE benefit from about 6 dB to 2–3 dB. The BU process was less affected by switching and the benefit varied between 2 and 3 dB for unmodulated signals to modulations up to 20 Hz. Therefore, the B-sEPSM likely allows for a switching mechanism that is too fast compared to the real system. The temporal resolution of the binaural process could be reduced at different stages of the model using the concept of a binaural temporal window (Kollmeier and Gilkey, 1990; Culling and Summerfield, 1998). The rate limiting could be placed before the selection of the cancellation parameters in the EC process (Culling and Colburn, 2000) or after the parameter selection, suggesting a slow update of the cancellation parameters (Yost, 1985). Moreover, it would be possible to consider different amounts of sluggishness in the BE and the BU processes due to the structure of the B-sEPSM. A faster process for the BE pathway than for the BU pathway would be in line with measurements by Brungart and Iyer (2012) and by Glyde

et al. (2013) who created BE stimuli based on glimpses in 20 ms windows and showed that the system could take full advantage of those glimpses. In contrast, the BU processing, based on ITD processing, seems to require much longer windows of the order of 100 ms (Culling and Mansell, 2013).

Another area worth exploring is the selection process between the BE and the BU contributions, which would replace the proposed "maximum" operation. For example, there is evidence that binaural unmasking is only realized at low instantaneous SNRs (George et al., 2012); therefore, the BU should probably contribute to intelligibility only in those cases. The direct integration of this process in the mr-sEPSM framework would require the calculation of an estimate of the audio SNR, but could help limit the dominance of the BU pathway in the current implementation.

Another interesting addition would be to use the cue selection mechanism proposed by Faller and Merimaa (2004) within the B-sEPSM framework. Their model was developed for source localization and is based on the coherence of the left- and right-ear inputs. ITD and ILD cues are only considered if the interaural coherence is above a certain threshold. Their model was shown to be able to localize concurrent speakers in anechoic and reverberant conditions. The inclusion of such a cue selection process in the B-sEPSM could contribute to accounting for observations that the BU pathway contributes to informational masking (IM) unmasking (Glyde et al., 2013). Interaural coherence information has already been shown to a relevant feature for intelligibility modeling (Cosentino et al., 2014).

Finally, the B-sEPSM performs the BE selection based on the SNR_{env} values, rather than the audio SNR, even though the BE concept is usually defined in term of the audio SNR, which is easier and simpler to manipulated experimentally. No studies were found to support the assumption that considering the BE SNR_{env} is equivalent to considering the BE audio SNR. Although it is possible that the BE SNR_{env} follows the BE audio SNR, it would be interesting to investigate this relationship. The tools develop by Decorsière et al. (2015) to reconstruct signals based on their envelope representation could be used to replicate the experiments of Glyde et al. (2013) or Brungart and Iyer (2012), whereas signals would be created based on the BE SNR_{env} rather than the BE SNR.

6.5 Perspectives

It is well known that hearing-impaired (HI) listeners suffer from a reduced benefit from azimuth separation between target and maskers, compared to normal-hearing (NH) listeners (Bronkhorst, 1989; Bronkhorst and Plomp, 1992; Peissig and Kollmeier, 1997). This impairment is partially attributed to elevated hearing thresholds at high audio frequencies, such that the HI listeners may sometimes not have access to the signal at the ear with the most favorable SNR (Bronkhorst, 1989). In contrast, intelligibility gains due to ITDs tend to be close to normal in listeners with symmetric hearing losses, but are smaller if the hearing loss is asymmetric (Bronkhorst, 1989). Several intelligibility models (Beutelmann and Brand, 2006; Beutelmann et al., 2010) have already attempted to account for effects of hearing impairment by considering elevated hearing thresholds and other components. The B-sESPM, as well as the sEPSM and the mr-sEPSM in their current forms cannot account for the higher SRTs measured with HI listener, but can account for the reduced masking release, compared to NH listener, in conditions with fluctuating maskers (Scheidiger et al., 2014). The modeling of the effects of hearing impairment on speech intelligibility in the sEPSM framework is currently an important area of research.

The B-sEPSM could be used to study the effects of aided hearing on speech intelligibility. Because the mr-sEPSM has been shown to be able to account for the effects of some types of non-linear processing, it could be used to investigate the impact of different binaural noise reduction techniques, such as beam-forming. Additionally, it could be used to study the effects of synchronized, or non-synchronized, wideband or multi-channel compression on speech intelligibility (Kollmeier et al., 1993; Moore et al., 1992; Ricketts et al., 2001).

7

Python auditory modeling toolbox

The following section describes the Python Auditory Modeling Toolbox (an unfortunate name; Toolbox being a MATLAB *term).*

The Python Auditory Modeling Toolbox (PAMBOX) is a Python package for auditory modeling. It is inspired by both the Auditory Modeling Toolbox (AMToolbox), a MATLAB collection of tools and models for auditory modeling, and scikit-learn, a Python package for machine learning, know for its great programming interface (Buitinck et al., 2013). PAMBOX contains model stages, such as gammatone filterbanks, envelope extraction models, and modulation filterbanks; complete speech intelligibility models, such as the sEPSM, the mr-sEPSM and the SII, as well as some helper functions to, for example, play audio, fit psychometric functions, and set signal levels. Stages and models are implemented as Python objects, which store the parameters, or attributes of a given stage or model. With this approach (see listing 7.1), it is not necessary to pass all parameters for each function call (see listing 7.2).

Listing 7.1: Filtering with PAMBOX. Defining the filterbank and using it are two separate step. `x` is the signal to filter, `order` is the filter order, and `center_frequencies` is a sequence of center frequencies.

```
fb = GammatoneFilterbank(center_frequencies, order)
y = fb.filter(x)
```

Listing 7.2: Filtering in MATLAB requires passing all parameters for call to the `gammatonefilterbank` function. `x` is the signal to filter, `order` is the filter order, and `center_frequencies` is a sequence of center frequencies.

```
y = gammatonefilterbank(x, center_frequencies, order)
```

Each model stage and intelligibility model conforms to a consistent interface, or "application programming interface" (API), such that stages and models can easily we swapped and compared. For example, each intelligibility model has a `predict` method that accepts the clean speech, noisy speech, and noise signals, whereas filterbanks have a `filter` method.

In addition to models and stages, PAMBOX contains a framework to create speech intelligibility experiments. An experiment consists of:

- One or multiple speech intelligibility models;
- A speech corpus (a collection of sentences);
- A list of SNRs;
- None, or one distortion, e.g., reverberation, noise reduction, phase jitter, etc.;
- A list of distortion parameters, e.g., reverberation times, over-subtractions factors, etc.

The experiment framework takes care of the permutations of models, sentences, SNRs, and distortion parameters. It also supports being run across multiple computing cores as well as over a cluster to speed up computations.

Models that are not original to PAMBOX are validated against their original implementations, where available. PAMBOX is distributed under the Modified BSD License.

PAMBOX is currently (2016-09-23) hosted on GitHub, at http://github.com/achabotl/pambox/ and is also accessible at http://pambox.org. The documentation can be read at http://pambox.readthedocs.io or direcly on the PAMBOX repository (http://github.com/achabotl/pambox/docs).

Bibliography

ANSI (1997). *American National Standard methods for calculation of the Speech Intelligibility Index*. ANSI S3.5. New York: American National Standards Institute.

Auditec (2006). *CD101RW2, Audio CD*. http://www.auditec.com. (date last viewed 28/09/15).

Bellamy, J. C. (2000). *Digital Telephony*. 3rd ed. New York: John Wiley & Sons.

Bernstein, J. G. W. and K. W. Grant (2009). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners". In: *J. Acoust. Soc. Am.* 125.5, pp. 3358–3372.

Bernstein, L. R. and C. Trahiotis (1996). "The normalized correlation: Accounting for binaural detection across center frequency". In: *J. Acoust. Soc. Am.* 100.6, pp. 3774–3784.

Berouti, M., R. Schwartz, and J. Makhoul (1979). "Enhancement of speech corrupted by acoustic noise". In: *ICASSP* 4.1, pp. 208–211.

Best, V., E. R. Thompson, C. R. Mason, and G. Kidd (2013). "An Energetic Limit on Spatial Release from Masking". In: *JARO* 14.4, pp. 603–610.

Beutelmann, R. and T. Brand (2006). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners". In: *J. Acoust. Soc. Am.* 120.1, pp. 331–342.

Beutelmann, R., T. Brand, and B. Kollmeier (2010). "Revision, extension, and evaluation of a binaural speech intelligibility model". In: *J. Acoust. Soc. Am.* 127.4, pp. 2479–2497.

Blauert, J. et al. (1998). "The AUDIS catalog of human HRTFs". In: *J. Acoust. Soc. Am.* 103.5, pp. 3082–3082.

Bolia, R. S., W. T. Nelson, M. A. Ericson, and B. D. Simpson (2000). "A speech corpus for multitalker communications research." eng. In: *J. Acoust. Soc. Am.* 107.2, pp. 1065–1066.

Breebaart, J., S. van de Par, and A. Kohlrausch (2001). "Binaural processing model based on contralateral inhibition. I. Model structure." eng. In: *J. Acoust. Soc. Am.* 110.2, pp. 1074–1088.

Bronkhorst, A. W. (1989). "Binaural speech intelligibility in noise for hearing-impaired listeners". In: *J. Acoust. Soc. Am.* 86.4, p. 1374.

Bronkhorst, A. W. and R. Plomp (1992). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing". In: *J. Acoust. Soc. Am.* 92.6, p. 3132.

Bronkhorst, A. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". In: *Acta Acustica united with Acustica* 86.1, pp. 117–128.

Bronkhorst, A. and R Plomp (1988). "The effect of head-induced interaural time and level differences on speech intelligibility in noise". In: *J. Acoust. Soc. Am.* 83, pp. 1508–1516.

Brungart, D. S. and N. Iyer (2012). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers". In: *J. Acoust. Soc. Am.* 132.4, p. 2545.

Brungart, D. S., B. D. Simpson, M. A. Ericson, and K. R. Scott (2001). "Informational and energetic masking effects in the perception of multiple simultaneous talkers". In: *J. Acoust. Soc. Am.* 110.5, pp. 2527–2538.

Buitinck, L. et al. (2013). "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Carlile, S. and C. Corkhill (2015). "Selective spatial attention modulates bottom-up informational masking of speech". In: *Sci. Rep.* 5, 8662 EP.

Chabot-Leclerc, A. and T. Dau (2014). "Predicting speech release from masking through spatial separation in distance". In: *Proc. 7th Forum Acust.* SS16_17.

Chabot-Leclerc, A., S. Jørgensen, and T. Dau (2014). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction". In: *J. Acoust. Soc. Am.* 135.6, pp. 3502–3512.

Chabot-Leclerc, A., E. N. MacDonald, and T. Dau (2015). "Predicting masking release of lateralized speech". *Proc. of the 5th International Symposium on Auditory and Audiological Research, ISAAR*, Submitted.

Chabot-Leclerc, A., E. N. MacDonald, and T. Dau (2016). "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain". *J. Acoust. Soc. Am.*, Accepted.

Cherry, E. C. (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *J. Acoust. Soc. Am.* 25.5, pp. 975–979.

Chi, T, Y Gao, M. C. Guyton, P Ru, and S Shamma (1999). "Spectro-temporal modulation transfer functions and speech intelligibility." In: *J. Acoust. Soc. Am.* 106.5, pp. 2719–2732.

Chi, T., P. Ru, and S. A. Shamma (2005). "Multiresolution spectrotemporal analysis of complex sounds". In: *J. Acoust. Soc. Am.* 118.2, pp. 887–906.

Christensen, C. L. (2005). *Odeon Room Acoustics Program, Version 8.0*. http://www.odeon.dk. (date last viewed 28/05/15). Kgs. Lyngby.

Christensen, C. L. (2009). *Odeon Room Acoustics Program, Version 10.1, User Manual, Industrial, Auditorium and Combined Editions*. Last viewed 23/07/12. Kgs. Lyngby, Denmark: Odeon A/S, pp. 1–145.

Christiansen, C., M. S. Pedersen, and T. Dau (2010). "Prediction of speech intelligibility based on an auditory preprocessing model". In: *Speech Commun.* 52.7-8, pp. 678–692.

Christiansen, S. K., M. L. Jepsen, and T. Dau (2014). "Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in "primitive" auditory stream segregation". In: *J. Acoust. Soc. Am.* 135.1, pp. 323–333.

Collin, B. and M. Lavandier (2013). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers." eng. In: *J. Acoust. Soc. Am.* 134.2, pp. 1146–1159.

Cooke, M. (2006). "A glimpsing model of speech perception in noise". In: *J. Acoust. Soc. Am.* 119.3, pp. 1562–1573.

Cooke, M., J. Barker, S. Cunningham, and X. Shao (2006). "An audio-visual corpus for speech perception and automatic speech recognition". In: *J. Acoust. Soc. Am.* 120.5, pp. 2421–2424.

Cosentino, S., T. Marquardt, D. McAlpine, J. F. Culling, and T. H. Falk (2014). "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals". In: *The Journal of the Acoustical Society of America* 135.2, pp. 796–807.

Culling, J. F. and H. S. Colburn (2000). "Binaural sluggishness in the perception of tone sequences and speech in noise". In: *J. Acoust. Soc. Am.* 107.1, pp. 517–527.

Culling, J. F. and E. R. Mansell (2013). "Speech intelligibility among modulated and spatially distributed noise sources". In: *J. Acoust. Soc. Am.* 133.4, pp. 2254–2261.

Culling, J. F. and Q. Summerfield (1998). “Measurements of the binaural temporal window using a detection task”. In: *J. Acoust. Soc. Am.* 103.6, pp. 3540–3553.

Culling, J. F., M. L. Hawley, and R. Y. Litovsky (2004). “The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources”. In: *J. Acoust. Soc. Am.* 116, pp. 1057–1065.

Culling, J. F., M. L. Hawley, and R. Y. Litovsky (2005). “Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]”. In: *J. Acoust. Soc. Am.* 118.1, pp. 552–552.

Dau, T., T. Piechowiak, and S. Ewert (2013). “Modeling within-and across-channel processes in comodulation masking release”. In: *J. Acoust. Soc. Am.* 133, pp. 350–364.

Dau, T. (2008). “Auditory Processing Models”. In: *Handbook of Signal Processing in Acoustics*. Ed. by D. Havelock, S. Kuwano, and M. Vorländer. Springer New York, pp. 175–196.

Dau, T., B Kollmeier, and A Kohlrausch (1997a). “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers”. In: *J. Acoust. Soc. Am.* 102, pp. 2892–2905.

Dau, T., B Kollmeier, and A Kohlrausch (1997b). “Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration”. In: *J. Acoust. Soc. Am.* 102, p. 2906.

Decorsière, R., P. Søndergaard, E. N. MacDonald, and T. Dau (2015). “Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations”. In: *Audio, Speech, and Language Processing, IEEE Transactions on*, pp. 46–56.

Depireux, D, J. Z. Simon, D. J. Klein, and S Shamma (2001). “Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex.” In: *J. Neurophysiol.* 85.3, pp. 1220–1234.

Derleth, R. P. and T. Dau (2000). “On the role of envelope fluctuation processing in spectral masking.” eng. In: *J. Acoust. Soc. Am.* 108.1, pp. 285–296.

Dreschler, W. A., H. Verschuure, C. Ludvigsen, and S. Westermann (2001). “ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment”. In: *Audiology* 40.3, pp. 148–157.

Dubbelboer, F. and T. Houtgast (2007). “A detailed study on the effects of noise on speech intelligibility.” In: *J. Acoust. Soc. Am.* 122.5, pp. 2865–2871.

Dubbelboer, F. and T. Houtgast (2008). “The concept of signal-to-noise ratio in the modulation domain and speech intelligibility.” In: *J. Acoust. Soc. Am.* 124.6, pp. 3937–3946.

Duquesnoy, A. J. and R. Plomp (1980). “Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis”. In: *J. Acoust. Soc. Am.* 68.2, pp. 537–544.

Durlach, N. (1963). “Equalization and Cancellation Theory of Binaural Masking-Level Differences”. In: *J. Acoust. Soc. Am.* 35, pp. 1206–1218.

Elhilali, M. and S. A. Shamma (2008). “A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation”. In: *J. Acoust. Soc. Am.* 124.6, pp. 3751–3771.

Elhilali, M., T. Chi, and S. A. Shamma (2003). “A spectro-temporal modulation index (STMI) for assessment of speech intelligibility”. In: *Speech Commun.* 41.2-3, pp. 331–348.

Elhilali, M., J. Xiang, S. A. Shamma, and J. Z. Simon (2009a). “Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene.” eng. In: *PLoS Biol.* 7.6, e1000129.

Elhilali, M., L. Ma, C. Micheyl, A. J. Oxenham, and S. A. Shamma (2009b). “Temporal coherence in the perceptual organization and cortical representation of auditory scenes”. In: *Neuron* 61.2, pp. 317–329.

Ewert, S. D. and T. Dau (2000). “Characterizing frequency selectivity for envelope fluctuations.” In: *J. Acoust. Soc. Am.* 108.3 Pt 1, pp. 1181–1196.

Falk, T., C. Zheng, and W.-Y. Chan (2010). “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 18.7, pp. 1766–1774.

Faller, C. and J. Merimaa (2004). “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence”. In: *J. Acoust. Soc. Am.* 116, p. 3075.

Fletcher, H. (1923). “Physical Measurements of Audition and Their Bearing on the Theory of Hearing*”. In: *Bell System Technical Journal* 2.4, pp. 145–180.

French, N. R. and J. C. Steinberg (1947). “Factors governing the intelligibility of speech sounds”. In: *J. Acoust. Soc. Am.* 19.1, pp. 90–119.

Freyman, R. L., U Balakrishnan, and K. S. Helfer (2001). “Spatial release from informational masking in speech recognition”. In: *J. Acoust. Soc. Am.* 109.5 Pt 1, pp. 2112–22.

George, E. L. J., J. M. Festen, and S. T. Goverts (2012). “Effects of reverberation and masker fluctuations on binaural unmasking of speech”. In: *J. Acoust. Soc. Am.* 132.3, p. 1581.

Glyde, H., J. Buchholz, H. Dillon, V. Best, L. Hickson, and S. Cameron (2013). “The effect of better-ear glimpsing on spatial release from masking.” eng. In: *J Acoust Soc Am* 134.4, pp. 2937–2945.

Goldsworthy, R. L. and J. E. Greenberg (2004). “Analysis of speech-based speech transmission index methods with implications for nonlinear operations”. In: *J. Acoust. Soc. Am.* 116.6, pp. 3679–3689.

Green, D. M. and J. A. Swets (1988). “Signal detection theory and psychophysics”. In: Los Altos California: Peninsula Publishing, pp. 238–239.

Hawley, M., R. Litovsky, and J. Culling (2004). “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer”. In: *J. Acoust. Soc. Am.* 115, pp. 833–843.

HearCom (2014). *HearCom: Predicting Binaural Speech Intelligibility*. URL: http://hearcom.eu/prof/RoomAcoustics/BinSI-Intro/BSIMDescription.html (visited on 05/27/2014).

Hilkhuysen, G., N. Gaubitch, M. Brookes, and M. Huckvale (2014). “Effects of noise suppression on intelligibility. II: An attempt to validate physical metricsa)”. In: *The Journal of the Acoustical Society of America* 135.1, pp. 439–450.

Holube, I., S. Fredelake, M. Vlaming, and B. Kollmeier (2010). “Development and analysis of an international speech test signal (ISTS)”. In: *Int. J. Audiol.* 49.12, pp. 891–903.

Houtgast, T. and H. Steeneken (1973a). “The modulation transfer function in room acoustics as a predictor of speech intelligibility”. In: *J. Acoust. Soc. Am.* 54, pp. 557–557.

Houtgast, T. and H. Steeneken (1985). “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria”. In: *J. Acoust. Soc. Am.* 77.March 1985, pp. 1069–1077.

Houtgast, T, H. J. M. Steeneken, and R Plomp (1980). “Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics”. In: *Acustica* 46.1, pp. 60–72.

Houtgast, T. and H. J. Steeneken (1973b). “The modulation transfer function in room acoustics as a predictor of speech intelligibility”. In: *Acta Acustica united with Acustica* 28.1, pp. 66–73.

IEC (2003). *IEC60268-16 Sound system equipment—Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index*. IEC 60268–16-2003. Geneva, Switzerland: International Electrotechnical Commission, pp. 1–28.

ISO (2005). *Reference zero for the calibration of audiometric equipment – Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions*. Tech. rep. 389-7. International Organization for Standardization, Geneva, Switzerland, pp. 1–10.

Ihlefeld, A. and B. Shinn-Cunningham (2008). “Spatial release from energetic and informational masking in a selective speech identification taska)”. In: *J. Acoust. Soc. Am.* 123.6, pp. 4369–4379.

Jeffress, L. A. (1948). “A place theory of sound localization.” In: *Journal of comparative and physiological psychology* 41.1, pp. 35–39.

Jelfs, S., J. F. Culling, and M. Lavandier (2011). “Revision and validation of a binaural model for speech intelligibility in noise”. In: *Hear. Res.* 275.127, pp. 96–104.

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). “A computational model of human auditory signal processing and perception.” In: *J. Acoust. Soc. Am.* 124.1, pp. 422–438.

Johansson, M. S. K. and S. D. Arlinger (2002). “Binaural masking level difference for speech signals in noise”. In: *Int. J. Audiol.* 41.5, pp. 279–284.

Jørgensen, S. and T. Dau (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing”. In: *J. Acoust. Soc. Am.* 130.3, pp. 1475–1487.

Jorgensen, S. and T. Dau (2013). “The role of high-frequency envelope fluctuations for speech masking release”. In: *Proc. Meet. Acoust.* Vol. 19. 1, p. 060126.

Jørgensen, S., S. D. Ewert, and T. Dau (2013). “A multi-resolution envelope-power based model for speech intelligibility”. In: *J. Acoust. Soc. Am.* 134.1, pp. 436–446.

Joris, P. X., P. H. Smith, and T. C. Yin (1998). “Coincidence Detection in the Auditory System”. In: *Neuron* 21.6, pp. 1235–1238.

Kates, J. M. and K. H. Arehart (2005). “Coherence and the speech intelligibility index”. In: *J. Acoust. Soc. Am.* 117, pp. 2224–2237.

Kidd Jr, G, C. R. Mason, T. L. Rohtla, and P. S. Deliwala (1998). "Release from masking due to spatial separation of sources in the identification of non-speech auditory patterns". In: *J. Acoust. Soc. Am.* 104.1, pp. 422–31.

Kleinschmidt, M. (2002). "Methods for Capturing Spectro-Temporal Modulations in Automatic Speech Recognition". In: *Acta Acustica united with Acustica* 88.3, pp. 416–422.

Kohlrausch, A, R Fassel, and T Dau (2000). "The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers". In: *J. Acoust. Soc. Am.* 108.2, 723–734.

Kollmeier, B. and R. H. Gilkey (1990). "Binaural forward and backward masking: evidence for sluggishness in binaural detection". In: *J. Acoust. Soc. Am.* 87.4, pp. 1709–1719.

Kollmeier, B., J. Peissig, and V. Hohmann (1993). "Real-time multiband dynamic compression and noise reduction for binaural hearing aids". In: *J. Rehabil. Res. Dev.* 30, pp. 82–82.

Kowalski, N., D. A. Depireux, and S. Shamma (1996). "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses tomoving ripple spectra". In: *J. Neurophysiol.* 76.5, pp. 3503–3523.

Kryter, K. D. (1962). "Methods for the Calculation and Use of the Articulation Index". In: *J. Acoust. Soc. Am.* 34.11, pp. 1689–1697.

Lőcsei, G., J. Hefting Pedersen, S. Laugesen, S. Santurette, T. Dau, and E. N. MacDonald (2015). *Lateralized speech perception, temporal processing and cognitive function in NH and HI listeners.* Poster presented at Speech in Noise workshop (Copenhagen, Denmark). Copenhagen, Denmark.

Langner, G. and C. E. Schreiner (1988). "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms". In: *J. Neurophysiol.* 60.6, pp. 1799–1822.

Lavandier, M. and J. F. Culling (2007). "Speech segregation in rooms: Effects of reverberation on both target and interferer". In: *J. Acoust. Soc. Am.* 122, pp. 1713–1723.

Lavandier, M. and J. F. Culling (2010). "Prediction of binaural speech intelligibility against noise in rooms". In: *J. Acoust. Soc. Am.* 127.1, pp. 387–99.

Lavandier, M., S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin (2012). "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources." eng. In: *J. Acoust. Soc. Am.* 131.1, pp. 218–231.

Lee, E. A. and D. G. Messerschmitt (1994). *Digital Communication*. 2nd. Boston: Kluwer Academic Publishing, p. 893.

Levitt, H and L. Rabiner (1967a). "Binaural release from masking for speech and gain in intelligibility". In: *J. Acoust. Soc. Am.* 42.3, pp. 601–608.

Levitt, H and L. Rabiner (1967b). "Predicting binaural gain in intelligibility and release from masking for speech". In: *J. Acoust. Soc. Am.* 42.4, pp. 820–829.

Licklider, J. C. R. (1948). "The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise". In: *J. Acoust. Soc. Am.* 20.2, p. 150.

Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice*. 1st. 608 p. CRC press.

Ludvigsen, C, C Elberling, and G Keidser (1993). "Evaluation of a noise reduction method–comparison between observed scores and scores predicted from STI." In: *Scand. Audiol. Suppl.* 38, pp. 50–55.

Lyon, R. and S. Shamma (1996). "Auditory representations of timbre and pitch". In: *Auditory Computation, Volume 6 of Springer Handbook of Auditory Research*. New York: Springer-Verlag, pp. 221–270.

Marrone, N., C. R. Mason, and G. Kidd (2008). "Tuning in the spatial dimension: Evidence from a masked speech identification task". In: *J. Acoust. Soc. Am.* 124.2, pp. 1146–1158.

May, T., S. van de Par, and A. Kohlrausch (2012). "A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation". In: *IEEE Audio, Speech, Language Process.* 20.7, pp. 2016–2030.

McDermott, J. H. and E. P. Simoncelli (2011). "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis". In: *Neuron* 71.5, pp. 926–940.

Mesgarani, N., M. Slaney, and S. Shamma (2006). "Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations". In: *IEEE Audio, Speech, Language Process.* 14.3, pp. 920–930.

Moore, B. C., B. R. Glasberg, and G. P. Schooneveldt (1990). "Across-channel masking and comodulation masking release." eng. In: *J Acoust Soc Am* 87.4, pp. 1683–1694.

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing*. 6th editio. Emerald, Bingley, UK, pp. 1–441.

Moore, B. C., J. S. Johnson, T. M. Clark, and V. Pluvinage (1992). "Evaluation of a dual-channel full dynamic range compression system for people with sensorineural hearing loss." In: *Ear and Hearing* 13.5, pp. 349–370.

Müsch, H. and S. Buus (2001). "Using statistical decision theory to predict speech intelligibility. I. Model structure". In: *J. Acoust. Soc. Am.* 109.6, pp. 2896–2909.

Nemala, S., K. Patil, and M. Elhilali (2013). "A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition". In: *IEEE Audio, Speech, Language Process.* 21.2, pp. 416–426.

Nielsen, J. B. and T. Dau (2009). "Development of a Danish speech intelligibility test." In: *Int. J. Audiol.* 48.10, pp. 729–741.

Nielsen, J. B., T. Dau, and T. Neher (2014). "A Danish open-set speech corpus for competing-speech studies". In: *J. Acoust. Soc. Am.* 135.1, pp. 407–420.

Nilsson, M., S. D. Soli, and J. A. Sullivan (1994). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise". In: *J. Acoust. Soc. Am.* 95.2, pp. 1085–1099.

Par, S. van de and A. Kohlrausch (1998). "Analytical expressions for the envelope correlation of narrow-band stimuli used in CMR and BMLD research". In: *J. Acoust. Soc. Am.* 103.6, pp. 3605–3620.

Pavlovic, C. (1987). "Derivation of primary parameters and procedures for use in speech intelligibility predictions". In: *J. Acoust. Soc. Am.* 82, pp. 413–422.

Payton, K. L. and L. D. Braida (1999). "A method to determine the speech transmission index from speech waveforms." In: *J. Acoust. Soc. Am.* 106.6, pp. 3637–3648.

Peissig, J. and B. Kollmeier (1997). "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners". In: *J. Acoust. Soc. Am.* 101.3, pp. 1660–1670.

Piechowiak, T., S. D. Ewert, and T. Dau (2007). "Modeling comodulation masking release using an equalization-cancellation mechanism". In: *J. Acoust. Soc. Am.* 121.4, pp. 2111–2126.

Plomp, R (1976). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)". In: *Acustica* 34, pp. 200–211.

Rennies, J., T. Brand, and B. Kollmeier (2011). "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet". In: *J. Acoust. Soc. Am.* 130.5, pp. 2999–3012.

Rhebergen, K. S. and N. J. Versfeld (2005). “A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners”. In: *J. Acoust. Soc. Am.* 117.4, pp. 2181–2192.

Rhebergen, K. S., N. J. Versfeld, and W. A. Dreschler (2005). “Release from informational masking by time reversal of native and non-native interfering speech”. In: *J. Acoust. Soc. Am.* 118.3, pp. 1274–1277.

Rhebergen, K. S., N. J. Versfeld, and W. A. Dreschler (2009). “The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise.” In: *J. Acoust. Soc. Am.* 126.6, pp. 3236–45.

Richards, V. M. (1987). “Monaural envelope correlation perception”. In: *J. Acoust. Soc. Am.* 82.5, pp. 1621–1630.

Ricketts, T., G. Lindley, and P. Henry (2001). “Impact of compression and hearing aid style on directional hearing aid benefit and performance”. In: *Ear and Hearing* 22.4, pp. 348–361.

Robles, L. and M. A. Ruggero (2001). “Mechanics of the mammalian cochlea”. In: *Physiol. Rev.* 81.3, pp. 1305–1352.

Rothauser, E. et al. (1969). “IEEE recommended practice for speech quality measurements”. In: *IEEE Trans. Audio Electroacoust.* 17.3, pp. 225–246.

Scheidiger, C., S. Jørgensen, and T. Dau (2014). “Modeling Speech Intelligibility in Hearing Impaired Listeners”. In: *Proc. 7th Forum Acust.*

Schreiner, C. E. and J. V. Urbas (1988). “Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields”. In: *Hearing Research* 32.1, pp. 49–63.

Shamma, S., R. Chadwik, W. Wilbur, K. Morrish, and J. Rinzel (1986). “A Biophysical Model of Cochlear Processing: Intensity Dependence of Pure Tone Response”. In: *J. Acoust. Soc. Am.* 80.1, pp. 133–145.

Shinn-Cunningham, B. G. (2008). “Object-based auditory and visual attention”. In: *Trends in Cognitive Sciences* 12.5, pp. 182–186.

Søndergaard, P. and P. Majdak (2013). “The Auditory Modeling Toolbox”. In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer, pp. 33–56.

Steeneken, H. J. and T Houtgast (1980). “A physical method for measuring speech-transmission quality.” In: *J. Acoust. Soc. Am.* 67.1, pp. 318–326.

Stone, M. A. and B. C. Moore (2014). “On the near non-existence of “pure” energetic masking release for speech”. In: *J. Acoust. Soc. Am.* 135.4, pp. 1967–1977.

Stone, M. A., C. Füllgrabe, R. C. Mackinnon, and B. C. J. Moore (2011). “The importance for speech intelligibility of random fluctuations in "steady" background noise.” eng. In: *J. Acoust. Soc. Am.* 130.5, pp. 2874–2881.

Stone, M. A., C. Füllgrabe, and B. C. J. Moore (2012). “Notionally steady background noise acts primarily as a modulation masker of speech.” eng. In: *J. Acoust. Soc. Am.* 132.1, pp. 317–326.

Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen (2011). “An algorithm for intelligibility prediction of time–frequency weighted noisy speech”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 19.7, pp. 2125–2136.

Van Wijngaarden, S. and R. Drullman (2008). “Binaural intelligibility prediction based on the speech transmission index”. In: *J. Acoust. Soc. Am.* 123, pp. 4514–4523.

Verhey, J. L., T. Dau, and B. Kollmeier (1999). “Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model”. In: *J. Acoust. Soc. Am.* 106.5, pp. 2733–2745.

Wagener, K, V Kühnel, and B Kollmeier (1999). “Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test”. In: *Z. Fur Audiologie, Audiological Acoustics* 38, pp. 4–15.

Wagener, K. C. and T. Brand (2005). “Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters”. In: *Int. J. Audiol.* 44.3, pp. 144–156.

Wan, R., N. I. Durlach, and H. S. Colburn (2010). “Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers”. In: *J. Acoust. Soc. Am.* 128.6, pp. 3678–3690.

Wan, R., N. I. Durlach, and H. S. Colburn (2014). “Application of a short-time version of the Equalization-Cancellation model to speech intelligibility experiments with speech maskers”. In: *J. Acoust. Soc. Am.* 136.2, pp. 768–776.

Wang, K. and S. Shamma (1994). “Self-normalization and noise-robustness in early auditory representations”. In: *Speech and Audio Processing, IEEE Transactions on* 2.3, pp. 421–435.

Warren, R. M., J. A. Bashford Jr, and P. W. Lenz (2005). “Intelligibilities of 1-octave rectangular bands spanning the speech spectrum when heard separately and paired.” eng. In: *J. Acoust. Soc. Am.* 118.5, pp. 3261–3266.

Watson, C. S. (2005). “Some comments on informational masking”. In: *Acta Acustica united with Acustica* 91.3, pp. 502–512.

Westermann, A. and J. M. Buchholz (2015a). “The effect of spatial separation in distance on the intelligibility of speech in rooms”. In: *J. Acoust. Soc. Am.* 137.2, pp. 757–767.

Westermann, A. and J. M. Buchholz (2015b). “The influence of informational masking in reverberant, multi-talker environmentsa)”. In: *J. Acoust. Soc. Am.* 138.2, pp. 584–593.

Woolley, S. M. N., T. E. Fremouw, A. Hsu, and F. E. Theunissen (2005). “Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds.” In: *Nat. Neurosci.* 8.10, pp. 1371–1379.

Xiang, J., D. Poeppel, and J. Z. Simon (2013). “Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations”. In: *J. Acoust. Soc. Am.* 133.1, EL7–EL12.

Yost, W. A. (1985). “Prior stimulation and the masking-level difference”. In: *J. Acoust. Soc. Am.* 78.3, p. 901.

Zaar, J. and T. Dau (2015). “Sources of variability in consonant perception and their auditory correlates”. In: *Journal of the Acoustical Society of America* 137.4, pp. 2306–2306.

van de Par, S. and A. Kohlrausch (1998). “Comparison of monaural (CMR) and binaural (BMLD) masking release.” In: *J Acoust Soc Am* 103.3, pp. 1573–1579.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.

The end.

To be continued...

Developing models to predict intelligibility can help develop a better understanding of the essential “features” of speech, how those features are extracted by the auditory system, and how they are combined and used to create understanding. This dissertation expands on a model named the speech-based envelope power spectrum model (sEPSM), which uses the signal-to-noise ratio in the envelope power domain (SNR_{env}) as the decision metric. The sEPSM was analyzed and compared to several other models that either use different front-ends or different decision metrics, such as the audio SNR. The goal was to tease apart the essential components of intelligibility models in a range of conditions known to be challenging. Furthermore, a quantitative model was developed in an attempt to predict the speech intelligibility measured in conditions where listeners are known to benefit from using both ears, compared to using either ear alone, such as in a noisy “cocktail party”. The model represents a binaural extension of the sEPSM, denoted as B-sEPSM. Overall, the results of this thesis support the hypothesis that the SNR_{env} is a powerful metric for intelligibility prediction. Furthermore, the B-sEPSM could be used to investigate the impact on intelligibility of different binaural noise reduction techniques, such as beam-forming, and of various binaural hearing aid compression strategies.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk