Modeling speech intelligibility based on the signal-to-noise envelope power ratio

Jørgensen, Søren; Dau, Torsten

Publication date:
2014

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

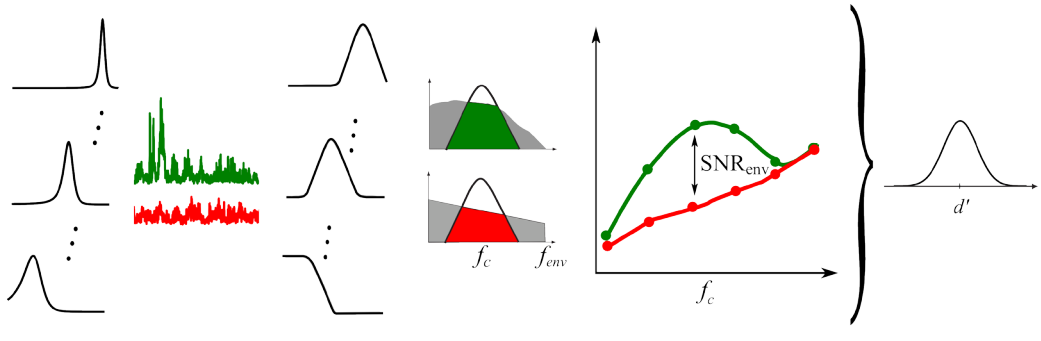
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 15

Søren Jørgensen

Modeling speech intelligibility based on the signal-to-noise envelope power ratio



Centre for Applied Hearing Research
Department of Electrical Engineering

Modeling speech intelligibility based on the signal-to-noise envelope power ratio

PhD thesis by
Søren Jørgensen

Centre for Applied Hearing Research

Technical University of Denmark
2014

Cover illustration: Søren Jørgensen.
The thesis was defended on 16 January 2014.

This PhD-dissertation is the result of a research project at the Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark (Kgs. Lyngby, Denmark).

The project was partly financed by DTU (1/3) and partly by a consortium of the Danish hearing-aid companies Oticon, Widex and GN Resound (2/3).

The assessment committee consisted of Assoc. Prof. Jonas Brunskog, Prof. Martin Cooke, and Prof. John Culling.

Supervisor

Prof. Torsten Dau
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

The intelligibility of speech depends on factors related to the auditory processes involved in sound perception as well as on the acoustic properties of the sound entering the ear. However, a clear understanding of speech perception in complex acoustic conditions and, in particular, a quantitative description of the involved auditory processes provides a major challenge in speech and hearing research. This thesis presents a computational model that attempts to predict the speech intelligibility obtained by normal-hearing listeners in various adverse conditions. The model combines the concept of modulation frequency selectivity in the auditory processing of sound with a decision metric for intelligibility that is based on the signal-to-noise envelope power ratio (SNR_{env}). The proposed speech-based envelope power spectrum model (sEPSM) is demonstrated to account for the effects of stationary background noise, reverberation and noise reduction processing on speech intelligibility, indicating that the model is more general than traditional modeling approaches. Moreover, the model accounts for phase distortions when it includes a mechanism that evaluates the variation of envelope power across (audio) frequency. However, because the SNR_{env} is based on the long-term average envelope power, the model cannot account for the greater intelligibility typically observed in fluctuating noise compared to stationary noise. To overcome this limitation, a multi-resolution version of the sEPSM is presented where the SNR_{env} is estimated in temporal segments with a modulation-filter dependent duration. This multi-resolution approach effectively extends the applicability of the sEPSM to account for conditions with fluctuating interferers, while keeping its predictive power in the conditions with noisy speech distorted by reverberation or spectral subtraction. The relationship between the SNR_{env} based decision-metric and psychoacoustic speech intelligibility is further evaluated by generating stimuli with different SNR_{env} but the same overall power SNR. The results from the corresponding psychoacoustic data generally support the above relationship. However, the model is limited in conditions with manipulated clean speech since it does not account for the accompanied effects of speech distortions on intelligibility.

The value of the sEPSM is further considered in conditions with noisy speech

transmitted through three commercially available mobile phones. The model successfully accounts for the performance across the phones in conditions with a stationary speech-shaped background noise, whereas deviations were observed in conditions with “Traffic” and “Pub” noise.

Overall, the results of this thesis support the hypothesis that the SNR_{env} is a powerful objective metric for speech intelligibility prediction. Moreover, the findings suggest that the concept of modulation-frequency selective processing in the auditory system is crucial for human speech perception.

Resumé

Vores hørelse er vital for talekommunikation, og vellykket kommunikation kræver en tilstrækkelig god taleforståelighed. Det er dog endnu ikke klarlagt præcis hvilke akustiske karakteristika og hvilke underliggende auditive processer der er afgørende for taleforståeligheden i komplekse akustiske miljøer. Denne afhandling præsenterer en signalbehandlingsmodel der kan forudsige taleforståelighed for normalthørende under forskellige vanskelige forhold. Modellen kombinerer begrebet modulations-frekvensselektivitet i hørelsens signalbehandling med en beslutningsparameter for taleforståelighed der er baseret på signal-til-støj modulationseffekt forholdet (SNR_{env}). Den præsenterede tale-baserede modulationseffekt spektrum model (sEPSM) vises i denne afhandling at kunne redegøre for indflydelsen af stationær baggrundsstøj, efterklang og støjreduktions-signalbehandling på taleforståeligheden. Yderligere ses det at modellen kan redegøre for indflydelsen af faseforvrængninger i talesignaler, når der inkluderes en mekanisme der analyserer variationen af modulationseffekten på tværs de auditive frekvenskanaler. Modellen kan dog ikke forklare den større taleforståelighed der typisk observeres i fluktuerende støj i forhold til i stationær støj, da SNR_{env} er baseret på den langsigtede gennemsnitlige modulationseffekt. For at overvinde denne begrænsning, præsenteres en multi-opløsnings version af sEPSM, hvor SNR_{env} estimeres i tidsmæssige segmenter med en varighed der afhænger af modulationsfiltrenes centerfrekvens. Denne fremgangsmåde udvider funktionaliteten af modellen til også at kunne redegøre for situationer med fluktuerende støjkilder, mens den samtidig beholder sine prædiktive egenskaber i betingelser med støjfyldt tale der er forvrænget af efterklang eller støjreduktion. Sammenhængen mellem beslutnings-parameteren (SNR_{env}) og den psykoakustiske taleforståelighed evalueres eksplicit med stimuli der har forskellig SNR_{env}, men det samme overordnede signal-til-støj effekt forhold. Resultaterne fra de tilsvarende psykoakustiske data understøtter overordnet hypotesen om SNR_{env}. Dog er modellen begrænset i tilfælde hvor tale er blevet manipuleret da den ikke kan redegøre for effekten af de medfølgende modulationsforvrængninger på taleforståeligheden. De praktiske anvendelsesmuligheder af sEPSM evalueres desuden i betingelser hvor støjfyldt tale bliver transmitteret gennem tre kommercielt tilgængelige mobiltelefoner. Modellen kan redegøre for forskellene i taleforståeligheden af den transmitterede tale for de

forskellige telefoner i betingelser med en stationær taleformet baggrundsstøj, men afviger i betingelser med to andre typer støj. Samlet set støtter resultaterne i denne afhandling hypotesen om at SNR_{env} er en brugbar objektiv beslutningsparameter til forudsigelse af taleforståelighed. Desuden tyder resultaterne på at konceptet om modulations-frekvensselektivitet i det auditive system er afgørende for den menneskelige taleopfattelse.

Acknowledgments

This thesis is the result of three years of research conducted at the Center for Applied Hearing Research at the Technical University of Denmark to obtain the PhD-degree. It has been an enriching experience, professionally and personally.

I would like to thank my supervisor Torsten Dau for giving me the opportunity to focus on a very specific topic - in incredible detail - and teaching me how to communicate the results to a broader audience. I am thankful for his many critical questions, forcing me to see things from a different perspective, and for his magnificent support throughout the entire project.

I also want to thank my colleagues at CAHR for a fantastic working environment, especially Claus Jespersgaard and Rémi Decorsière for the many discussions on speech and envelope processing, and Caroline van Oosterhout for all the help on the practical aspects of being a PhD-student. Moreover, I thank Stephan Ewert for a fruitful collaboration.

Finally, I want to thank my future wife Bettina for her endless love, and for all her support during this work.

Søren Jørgensen, 22 November 2013.

Related publications

Journal papers

- Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (2014). The importance of auditory spectro-temporal modulation filtering and decision metric for predicting speech intelligibility. *J. Acoust. Soc. Am.*, submitted.
- Jørgensen, S., Decorsiere, R., and Dau, T. (2014) Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility. *J. Acoust. Soc. Am.*, submitted.
- Jørgensen, S., Cubick, J., and Dau, T. (2014) Perceptual and model-based evaluation of speech intelligibility in mobile telecommunication systems. *Speech Commun.*, submitted.
- Jørgensen S., Ewert S. D., and Dau T. (2013). A multi-resolution envelope power based model for speech intelligibility. *J. Acoust. Soc. Am.*, 134, 436-446.
- Jørgensen S. and Dau T. (2013). Modelling speech intelligibility in adverse conditions. *Adv. Exp. Med. Biol.*, 787, 343-351.
- Jørgensen S., Dau T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.*, 130, 1475-148.

Conference papers

- Jørgensen, S., and Dau, T. (2013). The role of high-frequency envelope fluctuations for speech masking release. *Proceedings of Meetings on Acoustics, International Congress on Acoustics, 2013* pp 060126.

- Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (2013). The role of across-frequency envelope processing for speech intelligibility. *Proceedings of Meetings on Acoustics, International Congress on Acoustics, 2013* pp 060128.

- Jørgensen, S., and Dau, T. (2013). Predicting speech intelligibility in conditions with nonlinearly processed noisy speech. *Proceedings of AIA-DAGA 2013, Joint 39th German and Italian Convention on Acoustics*, Merano, Italy, March 2013. pp 220-223.

- Jørgensen, S., and Dau, T. (2012). Prediction of speech masking release for fluctuating interferers based on the envelope power signal-to-noise ratio. *Proceedings of ACOUSTICS 2012*, Hong Kong, May 2012.

- Jørgensen, S., and Dau, T. (2011). Predicting speech intelligibility in adverse conditions: evaluation of the speech-based envelope power spectrum model. *Proceedings of the 3rd International Symposium on Auditory and Audiological Research (ISAAR)* , Helsingør, Denmark, August 2011, 307-314.

- Jørgensen, S., and Dau, T. (2011). Predicting the effect of spectral subtraction on the speech recognition threshold based on the signal-to-noise ratio in the envelope domain. *Proceedings of the Forum Acusticum*, Aalborg, Denmark, June 2011.

- Dau, T. and Jørgensen, S. (2011). Predicting the intelligibility of processed noisy speech based on the signal-to-noise ratio in the modulation domain. *Forschritte der Akustik DAGA'11, 37th German Convention on Acoustics*, Germany, March 2011.

Book chapters

- Jørgensen, S., and Dau, T. (2013). "Modelling Speech Intelligibility in Adverse Conditions" in *Basic Aspects of Hearing*, edited by Moore, B.C.J. et al. (Springer Science+Business Media, New York), Chapter 38, 343-351.

Published abstracts

- Jørgensen, S., and Dau, T. (2013). The role of high-frequency envelope fluctuations for speech masking release. *J. Acoust. Soc. Am.*, 133, 3391.
- Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (2013). The role of across-frequency envelope processing for speech intelligibility. *J. Acoust. Soc. Am.*, 133, 3391.
- Jørgensen S., Decorsiere, R., MacDonald E. W., and Dau, T. (2013). The relationship between background noise envelope power and speech intelligibility in adverse conditions. *Association for Research in Otolaryngology (ARO), 36th Mid-Winter Meeting*, Baltimore, MA, February 2013.
- Jørgensen, S., and Dau, T. (2012). Prediction of speech masking release for fluctuating interferers based on the envelope power signal-to-noise ratio. *J. Acoust. Soc. Am.*, 131, 3341.
- Dau, T. and Jørgensen, S. (2011). Predicting speech intelligibility based on the envelope power signal-to-noise ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.*, 129, 2384.

Contents

1 General introduction

The auditory system is a remarkable sensory organ that allows us to interact and interpret the environment that surrounds us. The hearing sense is crucial for the development of spoken language and speech perception, and thus, for our speech communication. We rely on our ability to communicate via speech, either face-to-face or by using telecommunication. However, sufficient intelligibility of the speech is crucial for successful communication. The field of psychoacoustics provides researchers with the tools to quantify speech intelligibility and its dependence on environmental and human factors.

The traditional approach to study speech intelligibility has been via perceptual tests where a number of listeners are asked to repeat a well defined speech corpus presented to them acoustically; either in quiet or in combination with one or more interferers. The ratio of the number of correctly repeated speech items, for example words, to the number of presented items has thereby been defined as an indicator of the speech intelligibility. Pioneering work was achieved in connection with the development of the telephone in the early 20'th century regarding the factors that influence speech intelligibility (e.g., Steinberg, 1929; Knudsen, 1929; Miller, 1947; Fletcher and Galt, 1950). For example, the masking of speech by an interfering sound (such that the speech becomes partly unintelligible) was demonstrated to depend greatly on the temporal and spectral properties of the interferer. Generally, when the interferer itself is not speech, a stationary sound with the same long-term spectrum as the speech produces the most masking; in contrast, a pure-tone interferer has only little influence on the intelligibility. In addition to factors related to the acoustic environment, the intelligibility depends on the target speech itself. For example, speech items consisting of unfamiliar nonsense words typically lead to lower intelligibility compared to familiar

words. Moreover, speech intelligibility depends on the statistical properties of the test design. For example, if the listener is given a fixed set of speech items to respond from, the number of items in the response set greatly influences the measured intelligibility because the probability of reporting the correct item is greater for a small response set than for a large set.

However, despite decades of research on the factors that influence speech intelligibility, the auditory processes and acoustic features that govern our speech perception are still not fully understood. Several models of speech intelligibility have been proposed to account for the experimental findings. An early model, know as the Articulation Index (AI; French and Steinberg, 1947), predicts speech intelligibility based on the long-term acoustic power of the speech and the noise interferer in a number of frequency bands covering the speech spectrum. The AI assumes that the masking of speech is caused by the masker providing a greater excitation of the inner ear than the speech. This concept has been referred to as *energetic* masking. The AI can be evaluated from independent physical measurements of the speech and the noise, which makes the model applicable in practical situations. However, the AI-model does not include any aspects related to the temporal structure of the speech or the noise and can therefore not account for effects of temporal distortions on speech intelligibility, such as those caused by reverberation (Knudsen, 1929; Fletcher and Galt, 1950).

The temporal structure of speech can be characterized by its *temporal fine structure*, i.e., the instantaneous variation in the sound pressure, and its *envelope*, referring to the slow variation in the overall amplitude. The envelope has been suggested to be an important carrier of the semantic information in the speech (Smith *et al.*, 2002; Fogerty, 2011), and this aspect was included in an alternative speech intelligibility model denoted as the speech transmission index (STI; Houtgast and Steeneken, 1971). The STI performs a frequency analysis of the temporal envelope within individual peripheral (audio) frequency channels. The integrity of the speech envelope is measured by the ratio of the envelope spectrum of the distorted (transmitted) speech and that of the clean speech, denoted as the modulation transfer function (MTF). This approach was demonstrated to account for effects of noise and reverberation and has been especially useful in applications involving room acoustics where reverberation has

a large influence on the intelligibility. However, with the introduction of digital technology in telecommunication and hearing-aid devices, it became possible to perform nonlinear noise reduction processing on the noisy speech signal. In such conditions, the classical AI and STI models predict that intelligibility should increase after the processing. However, this has typically not been confirmed by psychoacoustic data (Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007), which has been referred to as the "noise reduction paradox". This indicates that the knowledge reflected in the classical models, regarding the critical acoustic features and the auditory processes that govern speech intelligibility is incomplete.

In a related area of auditory research, scientists performed experimental work on the characterization of the selective properties of the auditory system with respect to the processing of amplitude modulated signals. This led to the concept of *modulation* masking, whereby the modulation characteristics of an interfering signal can influence the detection of a target signal modulation (e.g., Bacon and Grantham, 1989; Houtgast, 1989). Frequency selectivity in the envelope frequency domain was demonstrated, analogous to the frequency selectivity in the audio-frequency domain (Dau *et al.*, 1997a). Dau *et al.* (1999) and Ewert and Dau (2000) demonstrated that a decision metric based on the signal-to-noise envelope power ratio, measured at the output of modulation frequency selective bandpass filters, could account for the psychoacoustic modulation masking data. The corresponding model was denoted the envelope power spectrum model (EPSM), in analogy to the classical power spectrum model of masking in the audio-frequency domain (Fletcher, 1940). Recently, it has been suggested that a similar decision metric could account for the apparent "noise reduction paradox" of speech perception (Dubbelboer and Houtgast, 2008). This indicated that speech intelligibility might also be affected by modulation masking, in addition to factors reflecting the integrity of the speech envelope measured by the STI and the effect of energetic masking as measured by the AI, respectively.

The present thesis describes a model of speech intelligibility that differs considerably from the above speech modeling approaches. The proposed model extends the concept of the EPSM towards speech intelligibility and evaluates the usefulness of the signal-to-noise envelope power ratio as a decision metric for speech intelligibility.

Chapter 2 describes the details of the proposed model framework, denoted the speech-based EPSM (sEPSM), where the signal-to-noise envelope power ratio (SNR_{env}) is measured from the envelope power spectra of noisy speech and noise alone at the output of a modulation bandpass filterbank following peripheral auditory filtering. The parameters of the processing stages are determined from psychoacoustic data from the literature. While the original EPSM assumed a simple decision criterion relating SNR_{env} to detectability, the sEPSM applied the concept of an "ideal observer", which includes parameters that are related to the speech material and the test design. This enables the model to account for differences in the response set size and redundancy of the considered speech stimuli. Model predictions are compared to speech intelligibility data in conditions with different speech materials, background noise, reverberation and noise reduction processing.

Chapter 3 investigates the role of spectro-temporal envelope processing in intelligibility modeling by comparing predictions from the sEPSM from Chapter 2 to two modified versions. One version assumes a two-dimensional modulation filtering stage, inspired by earlier work of Elhilali *et al.* (2003). The other model version keeps the one-dimensional (temporal) modulation filtering as in the original sEPSM, but introduces an across (peripheral) audio-frequency mechanism, inspired by models of comodulation masking release (CMR; Buus, 1985; Piechowiak *et al.*, 2007; Dau *et al.*, 2013). The role of the decision metric is studied by comparing the predictions obtained with the SNR_{env} with those obtained with the MTF. The predictions from all models are evaluated with regard to the psychoacoustic data in conditions with reverberation, spectral subtraction processing and phase jitter distortion.

Chapter 4 presents a multi-resolution version of the sEPSM (mr-sEPSM), inspired by a short-term intelligibility model based on the AI (Rhebergen and Versfeld, 2005). Instead of estimating the SNR_{env} from the long-term envelope power spectra, the multi-resolution version estimates the SNR_{env} in short-term segments with durations that are inversely related to the center frequencies of a the modulation filters, i.e., long segments are used for low modulation center frequencies and short windows for high modulation center frequencies. This allows the model to capture information reflected in both slowly varying low-frequency modulations as well as fast fluctuating high-frequency

modulations in the noisy speech. Predictions using the multi-resolution sEPSM are compared to experimental data in conditions with various stationary and fluctuating interferers, as well as in conditions with reverberation and spectral subtraction.

Chapter 5 provides an analysis of the contributions of different audio and envelope frequency regions to speech intelligibility using the multi-resolution sEPSM framework. Predictions are compared to data from Christensen *et al.* (2013), which considered the effect of attenuating the high-rate modulations from the target and the interfering talker on speech intelligibility. Furthermore, model predictions in conditions with speech mixed with different types of fluctuating interferers are compared to corresponding data from Festen and Plomp (1990).

Chapter 6 further challenges the underlying hypotheses of the sEPSM framework for speech intelligibility prediction. The relationship between $\mathrm{SNR}_{\mathrm{env}}$ and speech intelligibility is investigated using stimuli that were manipulated to have different values of $\mathrm{SNR}_{\mathrm{env}}$, while keeping the same long-term energy SNR. This was done by modifying the modulation power of speech before mixing it with unprocessed stationary speech-shaped noise or by modifying the modulation power of the noise before mixing it with the unprocessed speech.

Chapter 7 evaluates the usefulness of the sEPSM to predict speech intelligibility in a practical application. The purpose is to evaluate the intelligibility of speech transmitted through modern telecommunication systems, such as mobile phones, and to assess whether the sEPSM can be used to predict the intelligibility performance of the phones. Predictions are compared to perceptual data obtained with three different mobile phones in conditions with three different types of background noise.

Finally, Chapter 8 summarizes the main findings and discusses the limitations and perspectives of the proposed model.

2

Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing*

A model for predicting the intelligibility of processed noisy speech is proposed. The speech-based envelope power spectrum model (sEPSM) has a similar structure as the model of Ewert and Dau [(2000). J. Acoust. Soc. Am. **108**, 1181–1196], developed to account for modulation detection and masking data. The model estimates the speech-to-noise envelope power ratio, SNR_{env}, at the output of a modulation filterbank and relates this metric to speech intelligibility using the concept of an ideal observer. Predictions were compared to data on the intelligibility of speech presented in stationary speech-shaped noise. The model was further tested in conditions with noisy speech subjected to reverberation and spectral subtraction. Good agreement between predictions and data was found in all cases. For spectral subtraction, an analysis of the model's internal representation of the stimuli revealed that the predicted decrease of intelligibility was caused by the estimated noise envelope power exceeding that of the speech. The classical concept of the speech transmission index (STI) fails in this condition. The results strongly suggest that the signal-to-noise ratio at the output of a modulation frequency selective process provides a key measure of speech intelligibility.

* This chapter is based on Jørgensen and Dau (2011).

2.1 Introduction

Speech is fundamental for human communication. People rely on the ability to understand speech, even in adverse acoustic environments, and the effect of noise on speech intelligibility has been studied for many years. The prediction of speech intelligibility has been of interest for various applications, such as for the acoustic design of lecture halls, for assessing the effect of noise reduction in communication channels and for the evaluation of hearing-aid algorithms. A broad range of prediction models has been presented, including the Articulation Index (AI; ANSI S3.5, 1969), the Speech Intelligibility Index (SII; ANSI S3.5, 1997), and the Speech Transmission Index (STI; IEC60268-16, 2003). It has been demonstrated that the slowly varying level fluctuations in the envelope of speech, the so-called speech modulations, are affected by noise. Houtgast *et al.* (1980) presented the STI, which predicts speech intelligibility based on changes in the modulation index of an amplitude modulated probe signal. The STI is based on the assumption that any changes in the modulation of the probe signal, due to noise or other processing, will have the same effect on the modulations in a speech signal. Conceptually, the STI measures the reduction of the modulations of the clean speech signal over the range of audio-frequency bands and modulation-frequency bands thought to contribute most to speech intelligibility. This concept has been very successful in predicting speech intelligibility in noisy and reverberant conditions (e.g., Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985).

Despite this success, the STI concept does not work well when noisy speech is subjected to further nonlinear processing, such as deterministic envelope reduction (Noordhoek and Drullman, 1997), envelope compression (e.g., Drullman, 1995; Hohmann and Kollmeier, 1995; Rhebergen *et al.*, 2009) or spectral subtraction (e.g., Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007). For example, in the case of spectral subtraction, the STI predicts a large improvement of speech intelligibility, while the experimental data show that speech intelligibility typically decreases by 1-5% (e.g., Lim, 1978; Boll, 1979; Ludvigsen *et al.*, 1993). To extend the applicability of the STI to nonlinear processing, Payton and Braida (1999) and Goldsworthy and Greenberg (2004) presented various modifications of the STI, generally known as speech-based STI methods (sSTI). The main difference between the original STI and

the sSTI methods is that speech is used as the probe signal; the concept of considering the reduction of clean speech modulations resulting from the processing remains the same. Goldsworthy and Greenberg (2004) proposed that the sSTI could predict the effect of various kinds of nonlinear processing. However, this was never evaluated explicitly using a comparison between predicted and measured speech intelligibility.

An alternative approach to predicting the effect of nonlinear processing of noisy speech was proposed by Dubbelboer and Houtgast (2008). Instead of considering the changes only in the modulations of clean speech, as in the STI, Dubbelboer and Houtgast also considered the changes of a modulation *noise floor* arising from the nonlinear interaction between the clean speech and the noise waveforms. The noise floor was assumed to consist of so-called spurious modulations, i.e. noise modulations and interaction modulations. It was hypothesized that the nonlinear processing of noisy speech affected both the speech part and the spurious noise part of the modulations in the signal, and that the ratio between these two components was critical for speech intelligibility. The ratio of the speech modulations to the spurious modulations was defined as the signal-to-noise ratio in the modulation domain, $(S/N)_{mod}$, and the concept was shown to account qualitatively for the effects of different signal processing schemes applied to noisy speech. However, a central aspect of their method was that the interaction modulations, which do not exist until the speech and the noise are actually mixed, play an essential role for the noise floor and thus for speech intelligibility. $(S/N)_{mod}$ could therefore not be estimated from the clean signal and the noise waveforms separately, or from the statistics of the stimuli. Difficulties with estimating the interaction modulations from the broadband noisy speech stimulus led to the use of a narrowband probe signal, a 1-kHz pure tone with an imposed 4-Hz sinusoidal amplitude modulation. This makes $(S/N)_{mod}$ difficult to generalize to speech signals so that effects of speaking style or speech material (words versus sentences) could be studied.

The concept of considering the signal-to-noise ratio in the modulation domain as an important metric in auditory processing is not new. The envelope power spectrum model (EPSM; Dau *et al.*, 1999; Ewert and Dau, 2000) used a similar measure to predict amplitude modulation detection and masking data, assuming that signal

modulation imposed on a carrier is detectable if the envelope power signal-to-noise ratio at the output of the modulation filter tuned to the signal frequency exceeds a critical value. In contrast to Dubbelboer and Houtgast (2008), the signal modulation in the EPSM was assumed to be masked by the inherent modulations in the noise, not by interaction modulations. In the present study, the EPSM was extended to predict speech intelligibility. The calculation of the envelope power signal-to-noise ratio, denoted here as $\mathrm{SNR}_{\mathrm{env}}$ to distinguish this from $(\mathrm{S/N})_{\mathrm{mod}}$, is considered to be the key element of the approach. In contrast to Dubbelboer and Houtgast (2008), the proposed model (i) includes peripheral and modulation filtering inspired by the human auditory system, (ii) uses a decision device that takes the response-set size and the redundancy of a given speech material into account and, (iii) bases predictions on the hypothesis that the intelligibility of processed noisy speech is influenced mainly by the intrinsic fluctuations in the envelope of the noise waveform, not by interaction modulations. The model was evaluated for conditions of speech in the presence of stationary noise, using data from the literature. Furthermore, the model was evaluated for conditions where noisy speech was distorted by linear and nonlinear processing.

2.2 Description of the model

2.2.1 Overall structure

The structure of the speech-based envelope power spectrum model (sEPSM) is shown in Fig. 2.1. The first stage represents the filtering of the basilar membrane. This is followed by envelope extraction and modulation processing. A major change to the original EPSM is the inclusion of the processing of very low modulation frequencies, between 0.5 and 4 Hz, which have been shown to be essential for speech intelligibility (Elliott and Theunissen, 2009; Füllgrabe *et al.*, 2009). Another major change was in the decision device. The original EPSM related the envelope signal-to-noise ratio to modulation detection by assuming a certain signal-to-noise ratio at threshold. In contrast, the sEPSM applies an *m*-alternative forced choice (*m*AFC) decision model (an

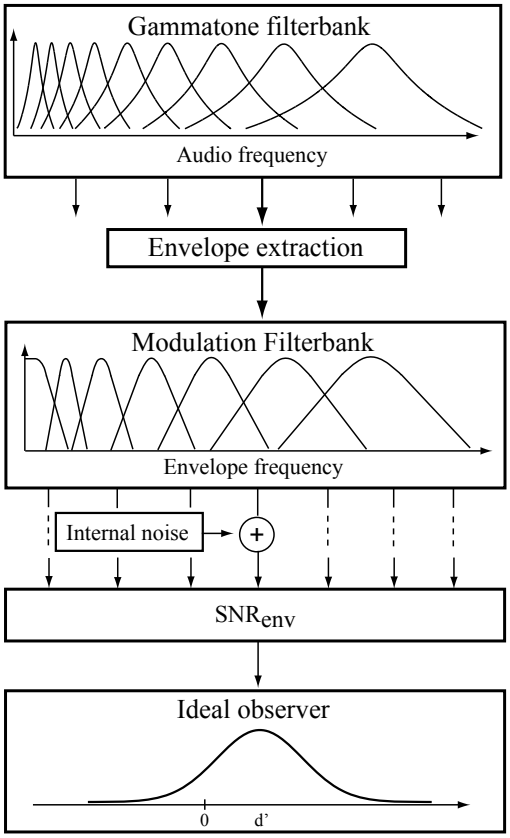


Figure 2.1: Block diagram of the structure of the speech-based envelope power spectrum model (sEPSM). The model consists of a gammatone bandpass filterbank followed by envelope extraction via Hilbert transformation and a modulation bandpass filterbank. The envelope signal-to-noise ratio, SNR_{env}, is calculated from the long-term integrated envelope power at the output of each modulation filter and 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 the concept of an "ideal observer".

ideal observer), which defines the relationship between SNR_{env} and percent correctly recognized items in a speech intelligibility test.

2.2.2 Processing stages in the model

Peripheral filtering and modulation processing

The first stage of the model is a bandpass filterbank, consisting of 22 fourth-order gammatone filters with 1/3-octave spacing of the center frequencies, which covers the range from 63 Hz to 8 kHz. This spacing was chosen to minimize the correlation between adjacent filters. An absolute sensitivity threshold is included such that only filters with output energy above the hearing threshold are considered for further processing[1]. The envelope of the output of each gammatone filter is extracted using the Hilbert transform. The resulting envelope function provides the input to a modulation filterbank consisting of a third-order low-pass filter in parallel with 6 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. The ac-coupled power of the filtered envelope is calculated by integrating the power density over the transfer-range of each modulation filter:

$$P_{env} = \frac{1}{S(0)} \int_{f_{env}>0}^{\infty} S(f_{env}) W_{fc}(f_{env}) df \tag{2.1}$$

where $S(f_{env})$ denotes the power density of the input envelope as a function of the envelope frequency, f_{env}. $S(0)$ represents the dc component of the envelope power spectrum and $W_{fc}(f_{env})$ denotes the squared transfer function of the modulation filter centered at the envelope frequency, f_c. In order to limit the model's sensitivity to the

[1] The spectrum level of the input stimulus was calculated in 1/3-octave bands as described in ANSI S3.5 (1997), using the same center frequencies as the audio filters of the sEPSM. The spectrum level of the input is compared to the diffuse field threshold in quiet (ISO389-7, 2005).

smaller modulations in a speech signal, envelope power below -20 dB is set to a level of -20 dB[2].

Calculation of the envelope power signal-to-noise ratio

Assuming that noise (N) and noisy speech ($S+N$) stimuli are available separately at the input of the model, the envelope power of the noisy speech ($P_{env,S+N}$) and that of the noise alone ($P_{env,N}$) are available at the output of the modulation processing stage. The effect of adding the noise to the speech is assumed to be three-fold. The noise (i) reduces the envelope power of the mixture compared to the clean speech by "filling up" the low-level parts of the speech, (ii) introduces a noise floor due to the intrinsic fluctuations in the noise itself, and (iii) creates new modulations from the nonlinear interaction between the speech and noise. Here, the envelope power of noise alone ($P_{env,N}$) represents an estimate of the noise floor. The interaction modulations are not included in this estimate, as they are assumed to have a negligible influence on intelligibility compared to the other two effects. In order to estimate the envelope power of the speech in the mixture ($\hat{P}_{env,S}$), the noise floor envelope power is subtracted from the noisy speech envelope power:

$$\hat{P}_{env,S} = P_{env,S+N} - P_{env,N} \tag{2.2}$$

Note that the value of $\hat{P}_{env,S}$ is different from the envelope power of the original clean speech. The $\mathrm{SNR_{env}}$ is calculated by taking the ratio of the estimated speech envelope

[2] The internal noise threshold is motivated by Noordhoek and Drullman (1997), who showed that modulations in the envelope of speech in quiet could be reduced to -20 dB before reaching 50 % intelligibility. Thus, the threshold reflects the minimum envelope power needed to recognize 50 % of the presented speech. The threshold deviates from the -30 dB used in the original EPSM, which reflected the empirically found AM detection threshold (Viemeister, 1979; Ewert and Dau, 2000). However, the sEPSM is concerned with the prediction of speech intelligibility, and the minimum envelope power related to the threshold for recognizing speech appears to be different from the minimum envelope power needed for detecting sinusoidal AM. The deviation might not have been apparent if the task had been speech detection, rather than speech recognition.

power and the noise envelope power:

$$\mathrm{SNR_{env}} = \frac{P_{env,S+N} - P_{env,N}}{P_{env,N}} \tag{2.3}$$

$$= \frac{\hat{P}_{env,S}}{P_{env,N}} \tag{2.4}$$

It is assumed that the envelope power of the noise alone does not exceed the envelope power of the noisy speech, such that the numerator of Eq. 2.3 cannot be negative:

$$P_{env,S+N} = \max\{P_{env,S+N}, P_N\} + \varepsilon. \tag{2.5}$$

where ε is a small positive constant which prevents the numerator from being zero in the case of $P_{env,S+N} = P_{env,N}$. Since the model contains 7 modulation filters for each of the 22 audio filters, the output of the $\mathrm{SNR_{env}}$ calculation stage is a multi-channel representation containing 7 x 22 $\mathrm{SNR_{env}}$ values.

Information integration across channels

Within the model, the information is combined across the audio and modulation frequency channels. The observations from n channels are combined using:

$$\mathrm{SNR_{env}} = \left[\sum_{i=1}^{n} (\mathrm{SNR_{env,i}})^2\right]^{1/2} \tag{2.6}$$

The use of overlapping filters in both the audio and the modulation domain results in partly correlated outputs, implying that the assumption of statistically independent observations is not fulfilled. The amount of correlation could be decreased by using fewer filters and a wider spacing between them. However, the information loss resulting from a reduction of the number of observations was earlier shown to affect the output more than the gain from statistical independence (Ewert and Dau, 2000). Using Eq. 2.6, the seven $\mathrm{SNR_{env}}$ values from the modulation filters are combined yielding a single value for each audio filter. The values for all audio filters are then combined using the same concept, resulting in the overall $\mathrm{SNR_{env}}$.

Transformation from $\mathrm{SNR_{env}}$ to percent correct

The overall $\mathrm{SNR_{env}}$ is converted to a sensitivity index, d', of an "ideal observer" using the relation:

$$d' = k \cdot (\mathrm{SNR_{env}})^q \tag{2.7}$$

where k and q are constants. This relation has been used previously to describe the relationship between d' and SNR in the audio domain (Egan, 1965; Egan *et al.*, 1969). Here, it is assumed that the relation can be applied to the SNR in the envelope domain. The constants k and q are assumed to be independent of speech material and experimental condition, so only one value is used for each constant. The value of d' is converted to the percentage of correct responses using an mAFC model (Green and Birdsall, 1964) in combination with an unequal-variance Gaussian model (e.g., Mickes *et al.*, 2007). 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. It can be shown that the value of x_S is a random variable with mean 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 correct as often as the value of x_S is larger than $x_{N,max}$. The corresponding probability of having selected the correct item is estimated from the difference distribution of x_S and $x_{N,max}$:

$$P_{correct}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right) \tag{2.8}$$

where Φ denotes the cumulative normal distribution. The values of σ_N and μ_N are determined by the response-set size, m, of the speech material. σ_S is a free parameter, assumed here to be related to the redundancy of the speech material. For example, speech material consisting of meaningful sentences with high redundancy would have a low value of σ_S and a material consisting of single-syllable words with low redundancy would have a higher value of σ_S. The expressions for estimating σ_N and μ_N can be found in the Appendix.

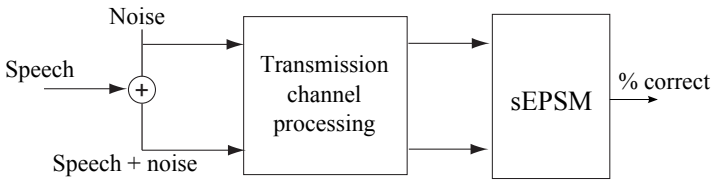


Figure 2.2: Block diagram showing the structure for estimating speech intelligibility using the sEPSM. See the main text for a description of the individual stages.

2.2.3 Prediction of speech intelligibility data

The assumptions used to predict speech intelligibility data are as follows: (i) noise and speech are available separately prior to any processing, (ii) the noise alone provides a representative estimate of the noise within the noisy speech, (iii) the information in the different processing channels is uncorrelated and combined optimally across channels, and (iv) only the amplitude of the temporal envelope of the speech influences intelligibility, i.e. temporal fine structure (TFS) and envelope phase information is neglected. The general structure of the model framework for predicting intelligibility of noisy speech using the sEPSM is shown in Fig. 2.2. The structure consists of two parallel processing paths, each divided into two stages. The input signals are the noise alone (N) and the noisy speech ($S+N$), respectively. The first stage represents transmission-channel processing applied to the noise and the noisy speech separately. It is noted that the same processing is applied to the two paths. Here, the noise alone provides an estimate of the effect of the transmission-channel processing on the intrinsic noise modulations within the noisy speech. The processed stimuli are then used as input to the sEPSM described above and the output is a percent correct prediction.

Figure 2.3 illustrates the concept of predicting the intelligibility of processed noisy speech. The top panel represents schematically the $\mathrm{SNR_{env}}$, i.e. the input to the ideal observer, as a function of the input SNR for unprocessed (black) and processed (gray) noisy speech. The bottom panel shows the predicted percent correct based on the ideal observer. Comparison of the two panels reveals that the ideal observer essentially converts the model's representation of the $\mathrm{SNR_{env}}$ to the probability of

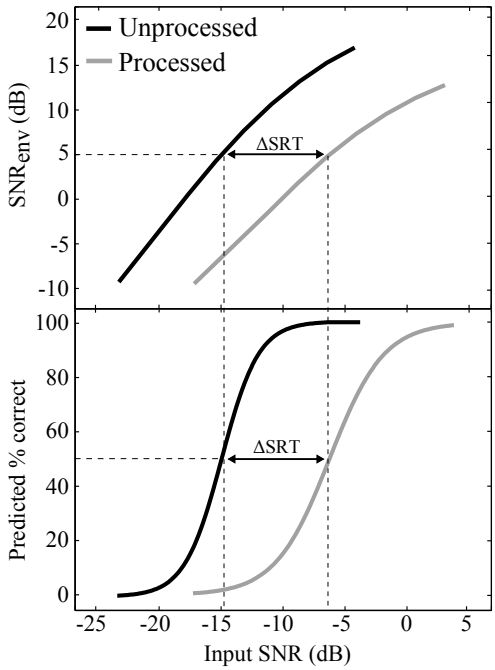


Figure 2.3: Illustration of the effects of noise and transmission-channel processing on predictions with the sEPSM framework. Top panel: SNR_{env} as a function of the input SNR for unprocessed (black) and processed (gray) noisy speech. Bottom panel: corresponding predicted percentage of correct responses as a function of the input SNR. The change in speech recognition threshold (SRT) for a given type of transmission-channel processing is estimated by measuring the horizontal shift of the predicted psychometric function.

correctly recognizing the noisy speech. Initially, the parameters of the ideal observer are adjusted so that it predicts the shape of the psychometric function for a given unprocessed speech material (black curve in the bottom panel). With these parameters fixed, the noisy speech is subjected to the transmission-channel processing, such as reverberation or spectral subtraction, and the simulated result is the predicted percent correct after the processing (gray curve in the bottom panel). Since the parameters of the ideal observer remain the same as for the unprocessed condition, it is only the *transmission-channel processing* of the stimuli that causes the shift of the predicted psychometric functions. By measuring the horizontal shift in terms of the input SNR, the change in any point on the functions can be estimated. For example, the change

in the speech recognition threshold, ΔSRT, can be predicted by measuring the shift at the 50 % point. From the top panel of Fig. 2.3 it can be seen that ΔSRT may also be estimated by considering the $\mathrm{SNR_{env}}$ value corresponding to 50 % correct in the unprocessed condition (5 dB in this example) and measuring the input SNR needed to provide the same $\mathrm{SNR_{env}}$ after processing through the transmission channel. The central hypothesis of the present model framework is that the predicted change in intelligibility arises because the transmission-channel processing changes the input SNR needed to obtain the $\mathrm{SNR_{env}}$ corresponding to a given percent correct.

2.3 Method

2.3.1 Speech material

Model predictions were compared to data from the literature on the intelligibility of noisy speech. Three Danish speech materials were considered: (i) the Conversational Language Understanding Evaluation (CLUE) test consisting of unique meaningful sentences (Nielsen and Dau, 2009), (ii) the DANTALE II sentence test consisting of grammatically correct but meaningless sentences constructed from a set of 50 words (Wagener *et al.*, 2003), and (iii) the DANTALE word test consisting of meaningful monosyllabic words (Keidser, 1993). For all test materials, the speech items were mixed with a speech-shaped stationary noise constructed specifically to match the long-term spectra of the respective materials. The three types of materials lead to different *normative* psychometric functions, i.e. they differ in the measured intelligibility as a function of the SNR when averaged over a large number of normal-hearing listeners. Table 2.1 shows the normative SRT and the slope, s_{50}, at the 50 % point of the psychometric function for each type of material. The SRTs are different, with $\mathrm{SRT_{CLUE}}$ having the largest value, indicating that a higher SNR is needed to recognize speech from the CLUE material. Furthermore, $s_{50,\mathrm{CLUE}} > s_{50,\mathrm{DANTALE\ II}} > s_{50,\mathrm{DANTALE}}$, indicating that the transition from poor to good understanding with increasing SNR occurs more quickly for the CLUE material than for the other two materials. The differences in slope and SRT can be explained by differences in the redundancy and

Table 2.1: Speech recognition thresholds (SRT) and slopes (s_{50}) of the psychometric functions for three Danish speech materials: CLUE, DANTALE II and DANTALE. Data from Nielsen and Dau (2009), Wagener *et al.* (2003) and Keidser (1993), respectively.

Speech material	SRT [dB]	s_{50} [% per dB]
CLUE	-3.2	18.7
DANTALE II	-8.4	12.6
DANTALE	-8.7	6.0

response-set size of the speech material (Miller *et al.*, 1951; Hirsh *et al.*, 1954; Egan *et al.*, 1956; Kruger and Kruger, 1997). A small number of response alternatives and a low redundancy typically leads to a more shallow slope whereas a larger number and a higher redundancy leads to a steeper slope.

2.3.2 Stimuli and experimental conditions

Speech in noise and reverberation

New data on the intelligibility of noisy speech with reverberation were obtained. The noisy sentences were convolved with an impulse response reflecting a specific reverberation time. The impulse response was generated using ODEON room acoustics software version 10 (Christensen, 2009). The simulated room was rectangular in shape with a volume of 3200 m^3 and the absorption was distributed such that the room had equal reverberation time (T_{30}) in the frequency range from 63 to 8000 Hz. Five different values of T_{30} were used: 0, 0.4, 0.7, 1.3, and 2.3 s.

Speech in noise subjected to spectral subtraction

Spectral subtraction is the common name for a group of signal processing schemes intended to increase the signal-to-noise ratio of a noisy signal by modifying the spectral content (see e.g., Lim, 1978; Berouti *et al.*, 1979; Boll, 1979; Ludvigsen *et al.*, 1993; Tsoukalas *et al.*, 1997). Here, one type of spectral subtraction is applied to noisy speech in order to investigate its effect on speech intelligibility. Sentences were mixed with

speech-shaped noise and subjected to spectral subtraction as defined by Berouti *et al.* (1979):

$$\hat{S}(f) = \left[P_{S+N}(f) - \alpha \hat{P}_N(f)\right]^{1/2} \tag{2.9}$$

where $\hat{S}(f)$ denotes an estimate of the clean speech power spectrum, $\hat{P}_N(f)$ represents an estimate of the noise power spectrum, $P_{S+N}(f)$ is the power spectrum of the noisy speech and α denotes an over-subtraction factor. The over-subtraction factor is included to reduce spectral artifacts (Berouti *et al.*, 1979). The spectral subtraction algorithm was implemented using a 2048-point Short Time Fourier Transform (STFT) with a window (Hanning) length of 24 ms and an overlap of 50%. The STFT was calculated separately for the noise alone and the noisy speech signals. For each frame, the spectral estimate of the noise, $\hat{P}_N(f)$, was calculated as the mean value of the noise power spectral density. This value was then multiplied by the over-subtraction factor and subtracted from each spectral bin of the noisy speech power spectrum. 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. It is noted that this implementation represents an "ideal" form of the spectral subtraction algorithm, since the noise signal is available separately. In practical applications, $\hat{P}_N(f)$ would be estimated from the noisy signal during non-speech periods.

Six different processing conditions were considered with the over-subtraction factor, α, as the parameter. The α values were 0, 0.5, 1, 2, 4 and 8, where $\alpha = 0$ represented the reference condition with no spectral subtraction. An α value from 1 to 4 is expected to provide the optimal noise reduction while $\alpha = 8$ is expected to distort the speech in addition to reducing the amount of noise (Berouti *et al.*, 1979).

2.3.3 Apparatus and procedure

The speech intelligibility data were collected by measuring the SRT using the CLUE speech material. All stimuli were stored digitally at a sampling frequency of 44.1 kHz and presented diotically using a calibrated pair of Sennheiser HD 580 headphones and

a high-quality sound card in a double-walled sound-attenuating booth. The speech had a constant level of 65 dB SPL and noise was added to achieve the appropriate SNR before further processing. Each sentence was presented once with the noise starting 1 second before the sentence began and ending 600 ms after it ended. The noise was ramped on and off using 400-ms squared-cosine ramps. A simple 1 up-1 down adaptive procedure, yielding an estimate of the 50 % correct threshold, was used, with the SNR being adjusted in steps of 2 dB. If all words in a sentence was correctly repeated, the SNR was lowered and if not, the SNR was increased. The threshold therefore represents sentence intelligibility rather than word intelligibility. Ten sentences were used to measure one SRT, which was calculated as the average SNR at the last eight presentations. Listeners were instructed to repeat as much of the presented sentence as possible, and were allowed to guess; no feedback was provided. Each listener was tested three times in the same condition using different lists each time, and five different lists were used for training before the measurement.

2.3.4 Listeners

Five male and three female normal-hearing listeners between 24 and 33 years of age participated in the experiment. They had pure-tone thresholds of 20 dB hearing level or better in the frequency range from 0.25 to 8 kHz, except for one subject, who had a threshold of 30 dB at 8 kHz. All listeners had experience with psychoacoustic measurements.

2.3.5 Model setup and parameters

Prediction of results for speech in noise

To generate model predictions, 150 sentences from the CLUE material, 144 sentences from the DANTALE II material and 100 single words from the DANTALE material were used as samples. The durations of the noise samples were matched to those of the speech and all stimuli were down-sampled to 22.05 kHz to reduce computation time. The percentage of correct responses was calculated for each speech sample, for a set of

Table 2.2: Calibrated values of σ_S and the constants (k,q) of the ideal observer, for three different speech materials: CLUE, DANTALEII and DANTALE. The estimated response-set sizes, m, are also shown.

Speech material	k	q	σ_S	m
CLUE	$\sqrt{1.2}$	0.5	0.6	8000
DANTALE II	$\sqrt{1.2}$	0.5	0.9	50
DANTALE	$\sqrt{1.2}$	0.5	0.9	8000

input SNRs. The final prediction at a given SNR and for a given material was computed as the average across all simulations at that SNR. The model was calibrated by adjusting the constants (k,q) and the parameter, σ_S, of the ideal observer such that the predicted percentage of correct responses matched the relevant data. The calibrated values are given in Table 2.2. A single set of constants (k,q) was found that fitted all three speech materials. The values of σ_S have the relation $\sigma_{S,\text{CLUE}} < \sigma_{S,\text{DANTALEII}} = \sigma_{S,\text{DANTALE}}$, which is consistent with the apparent redundancy of the materials: the meaningful sentences of the CLUE material have a higher redundancy than the mono-syllabic words and nonsense sentences of the DANTALE and DANTALE II materials. The values of m were estimated from the designs of the speech materials. For example, each test-list from the DANTALE II material is constructed from the same set of 50 words, so $m_{\text{DANTALE II}} = 50$. The speech materials CLUE and DANTALE are open-set, for which there are no explicit boundaries of the response-set sizes. Here, m is limited by the size of the listeners' active vocabulary, which was assumed here to comprise 8000 items both for words and sentences (Müsch and Buus, 2001).

Prediction of results for reverberant noisy speech and spectral subtraction

Predictions were generated using 150 different CLUE sentences mixed with noise at seven SNRs, ranging from -9 to 9 dB with 3-dB intervals. For each SNR, the predicted percent correct was obtained as the average across all 150 sentences. A continuous psychometric function was obtained by joining the predicted responses at the seven SNRs with straight lines. The transmission-channel processing, here reverberation and spectral subtraction, was applied to both the noisy speech and the noise alone stimuli (Fig. 2.2). The reverberant stimuli were truncated when the level of their envelopes

reached 5 % of their maximum values. In the case of spectral subtraction, the same noise estimate $\hat{P}_N(f)$ was subtracted from both noisy speech and noise alone. This provided an estimate of the effect of the spectral subtraction processing on the intrinsic noise modulations within the noisy speech. Note that, for the measured data, spectral subtraction was applied to the mixture only. The predicted SRT for a specific processing condition was obtained as the 50% point on the predicted psychometric function. The ΔSRT for a given condition, both for the measured and the predicted data, was obtained as the difference between the SRT for the reference condition and for the processed condition. Positive ΔSRTs indicate an increase of SRT.

2.4 Results

2.4.1 Speech in stationary noise

In this section, predicted and obtained speech intelligibility in stationary noise are compared. Figure 2.4 shows the data (open symbols) and corresponding predictions (filled symbols) for the CLUE material (top), the DANTALE II material (middle) and the DANTALE material (bottom). For all three speech materials, the predicted percent correct values are in very good agreement with the data. This results from the calibration of the ideal observer, i.e. the adjustment of (k,q) and σ_S to the individual materials. The adjustment was performed until the best possible agreement was achieved between the measured and predicted data points, determined by visual inspection. Note that (k,q) is fixed for all materials. Thus, the sEPSM can account for the different psychometric functions of the materials solely by setting the response-set sizes to the appropriate values for the speech materials and fitting the corresponding values of σ_S. To better visualize the trends in the data, a two-parametric function (solid line) (Wagener *et al.*, 2003) fitted to the data is shown with goodness of fit statistics provided in each panel. This function was not used for model predictions.

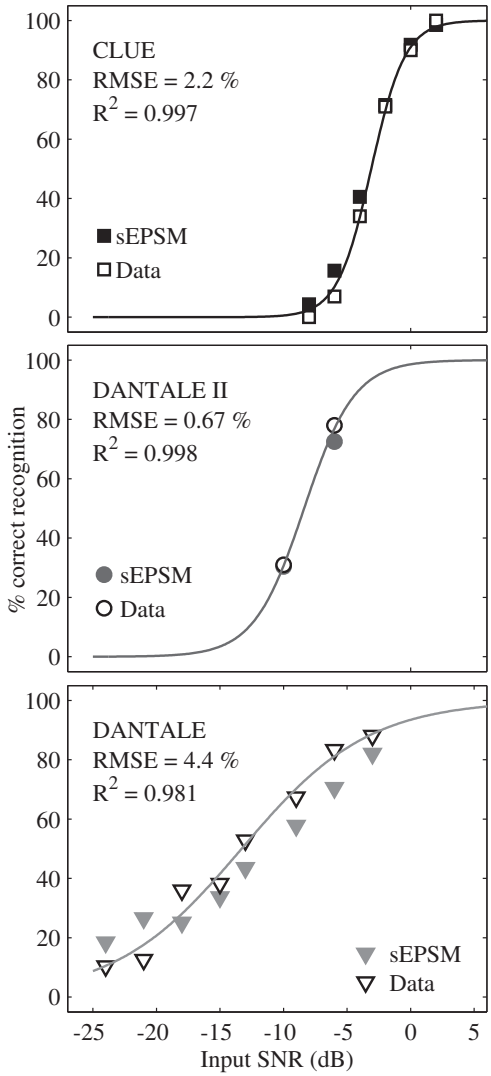


Figure 2.4: Measured data (open symbols) and corresponding simulations (filled symbols) for the CLUE material (top panel), the DANTALE II material (middle-panel) and the DANTALE material (bottom panel). The data are reprinted from Nielsen and Dau (2009), Wagener *et al.* (2003) and Keidser (1993), respectively. To aid visual inspection, a two-parametric function (solid line) (Wagener *et al.*, 2003) is fitted to the measured data using a least-squares procedure. The coefficient of determination (R^2) and the root-mean-squared-error (RMSE) of the fit is indicated on each panel. Note that the prediction at -10 dB SNR for the DANTALE II material is hidden behind the measured data point.

2.4.2 Reverberant speech

Here, predictions and data for reverberant noisy speech are compared. Figure 2.5 shows ΔSRT as a function of the reverberation time. The open squares show data averaged across 6 listeners, and the filled squares show predictions. The mean measured SRT in the reference condition was -3 dB, which is consistent with the normative SRT for the CLUE test (see Table 2.1 and the top panel of Fig. 2.4). The vertical bars indicate the standard deviation of the listeners' mean SRT, which was 1 dB on average. A two-way analysis of variance (ANOVA) performed on the SRT data showed no significant effect of subjects ($F_{5,20} = 3.74$, $p = 0.015$). A significant effect of T_{30} was found ($F_{4,20} = 73.1$, $p < 0.001$). SRT increased with 4 to 8 dB with increasing T_{30}. This is consistent with the data of Duquesnoy and Plomp (1980). A multiple comparison procedure using Tukey's honestly significant difference criterion revealed significant differences ($p < 0.05$) between SRT in the reference condition ($T_{30} = 0$) and all values of T_{30} greater than zero. Significant differences were also found between $T_{30} = [0.4, 0.7]$ and $T_{30} = [1.3, 2.3]$. The predictions also show an increase of SRT with increasing reverberation time. The Pearson correlation coefficient between predicted and measured ΔSRT is 0.98 and the RMSE is 0.71 dB, showing that the predictions follow the data very well.

2.4.3 Spectral subtraction

Figure 2.6 shows ΔSRT (left ordinate) as a function of the over-subtraction factor α. The open squares represent measured ΔSRTs, averaged across 4 subjects. Here, the SRT in the reference condition was obtained at an SNR of -3.3 dB, consistent with the value in Table 2.1. The vertical bars indicate the standard deviation of the mean SRT, which averages about 0.5 dB. A two-way ANOVA performed on the mean SRT data across subjects and conditions showed no significant effect of subjects ($F_{3,15} = 0.39$, $p = 0.18$) but a significant effect of α ($F_{5,15} = 15.08$, $p < 0.001$). For all cases of spectral subtraction ($\alpha > 0$), SRT increased with 1.5 to 2.5 dB, reflecting reduced speech intelligibility compared to the condition without spectral subtraction ($\alpha = 0$). Such decreases in intelligibility are consistent with the data of Boll (1979), Ludvigsen

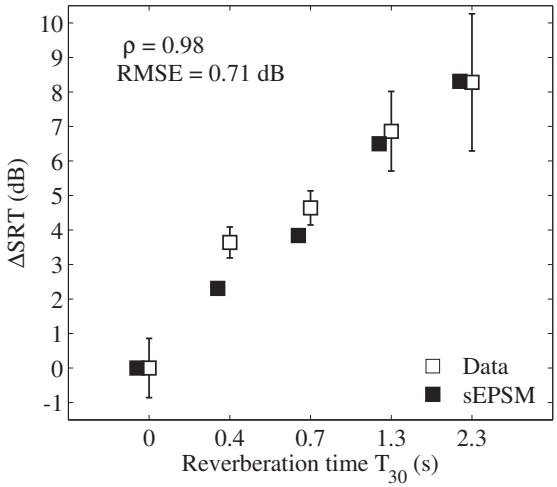


Figure 2.5: The measured change in SRT (open squares), averaged across 6 normal-hearing listeners, as a function of the reverberation time, T_{30}. The mean SRT in the reference condition was -3 dB. Model predictions are indicated by the filled squares. The linear correlation coefficient (ρ) and RMSE is indicated in the upper left corner.

et al. (1993), and Sarampalis *et al.* (2009). Multiple comparison tests showed a significant difference ($p < 0.05$) between the reference condition and all conditions of spectral subtraction. There were also significant differences ($p < 0.05$) between SRTs for $\alpha = [0.5, 1, 2]$ and 8. The filled squares in Fig. 2.6 show the predicted results, which agree well with the measured data but are consistently a bit higher for $\alpha > 0$. The Pearson correlation coefficient between the predictions and the data is 0.99 and the RMSE is 0.48 dB, indicating that the agreement between predictions and measurements is slightly better for spectral subtraction than for reverberation. The filled gray circles in Fig. 2.6 show sSTI values (Houtgast and Steeneken, 1985) (indicated on the right ordinate; note the reverse scaling). The sSTI increased in all conditions of spectral subtraction compared to the reference condition ($\alpha = 0$), thus predicting an increase in speech intelligibility, opposite to the trend in the data. This inconsistency was also found by Ludvigsen *et al.* (1993) and Dubbelboer and Houtgast (2007) and has occasionally been referred to as the “noise reduction paradox”.

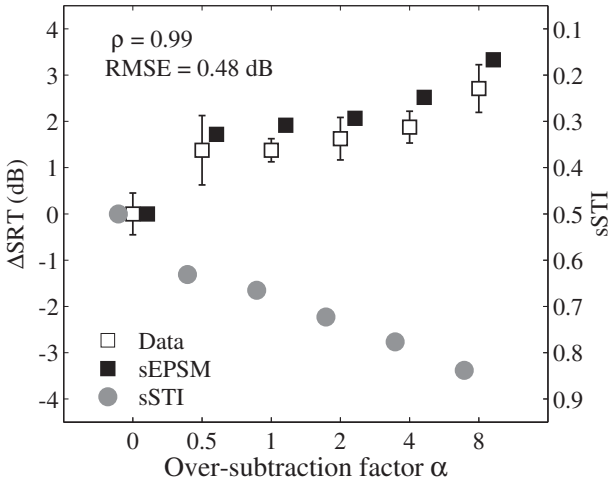


Figure 2.6: ΔSRT (left ordinate) as a function of the over-subtraction factor α for 4 normal-hearing listeners (open squares) and sEPSM predictions (filled squares). The right ordinate (with a reversed scale) shows the corresponding sSTI values as filled gray circles. These values are, however, not converted to ΔSRT values since these would be outside the left ordinate scale.

2.5 Model analysis

2.5.1 Modulation excitation patterns of the stimuli

In this section, we consider the contributions of the different model processing stages to the predicted results. The four panels in Fig. 2.7 show the output of the modulation filterbank, for the four gammatone filters centered at 250 Hz (top left), 1250 Hz (top right), 2500 Hz (bottom left), and 5000 Hz (bottom right). Each point represents the integrated envelope power at the output of a single modulation filter centered at the modulation frequencies labeled on the abscissa (the 1-Hz label refers to the cutoff frequency of the modulation lowpass filter). The corresponding patterns are denoted as “modulation excitation patterns”, inspired by the term “excitation pattern”used in the audio-frequency domain (Moore and Glasberg, 1983). Three patterns are shown in each plot, resulting from the excitation with clean speech (squares), noisy speech at an SNR of 0 dB (filled circles), and noise alone (triangles). Note that clean speech

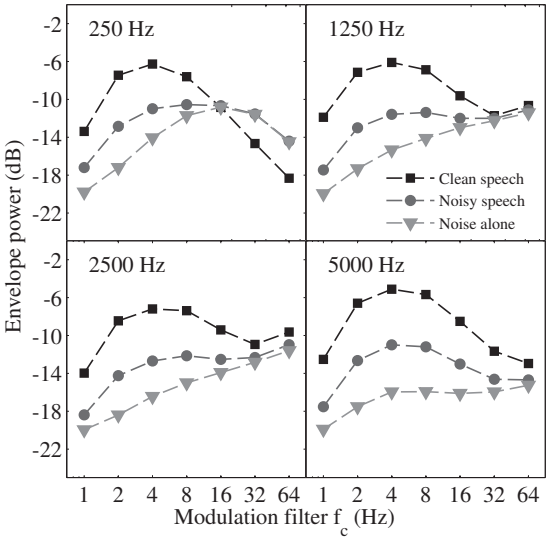


Figure 2.7: Modulation excitation patterns for the outputs of four gammatone filters with center frequencies of 250 Hz (top left), 1250 Hz (top right), 2500 Hz (bottom left) and 5000 Hz (bottom right). Each point represents the long-term integrated envelope power at the output of a modulation filter, plotted as a function of its center frequency (f_c), for three different input stimuli: clean speech (squares), noisy speech at an SNR of 0 dB (filled circles), and noise alone (triangles).

was not considered in Sec. 2.4 but its pattern is shown here for comparison purposes. In all panels, in the range from 1 to 16 Hz, the pattern for clean speech resembles the classical modulation spectra of speech as presented in Houtgast and Steeneken (1985) and Payton and Braida (1999). A peak at 4 Hz is clearly visible, corresponding roughly to the rate of syllables in normal speech. The filters centered at 1250 and 2500 Hz show an increase of the envelope power of the clean speech above about 32 Hz. This increase was not observed in the study of Payton and Braida (1999) since they lowpass filtered all speech envelopes with a cutoff frequency of 31 Hz before calculation of the modulation spectra.

2.5.2 The effects of audio-frequency and envelope-frequency selectivity on $\mathrm{SNR_{env}}$

The difference between the modulation excitation patterns for noisy speech and noise alone determines the value of $\mathrm{SNR_{env}}$; a small difference corresponds to a low $\mathrm{SNR_{env}}$ and a large difference corresponds to a large value of $\mathrm{SNR_{env}}$. Figure 2.7 shows that $\mathrm{SNR_{env}}$ depends on the modulation (center) frequency and has small values above about 16 Hz. Furthermore, $\mathrm{SNR_{env}}$ varies to some extent with audio frequency. This illustrates that the individual audio- and modulation-frequency channels contribute differently to the overall value of $\mathrm{SNR_{env}}$ at the output of the model. This was further investigated by repeating the spectral subtraction predictions using modified versions of the model. Figure 2.8 shows ΔSRT as a function of the over-subtraction factor α. The open squares replot the data from Fig. 2.6. The filled gray circles show predictions when the gammatone filterbank was replaced by a single bandpass filter with cutoff frequencies at 63 Hz and 8 kHz, thus assuming no selectivity in the audio-frequency domain. The filled triangles in Fig. 2.8 shows predictions when the gammatone filterbank was left unchanged but the modulation filterbank was replaced by a single modulation low-pass filter with a cutoff frequency of 150 Hz, thus assuming no selectivity in the modulation-frequency domain. The predicted ΔSRT is substantially decreased for all conditions with $\alpha > 0$, for both modified models, which is inconsistent with the data. This suggests that the frequency selectivity assumed in the two domains is important for accurate prediction of the intelligibility of noisy speech subjected to spectral subtraction.

2.5.3 The effect of spectral subtraction in the envelope-frequency domain

Figure 2.9(A) shows modulation excitation patterns at the output of the 1-kHz gammatone filter for four values of the over-subtraction factor α. The envelope power of the noisy speech (filled circles) increases with increasing α. However, the envelope power of the noise alone (triangles) also increases with increasing α. Thus, the predicted decrease of speech intelligibility with increasing α is caused by the

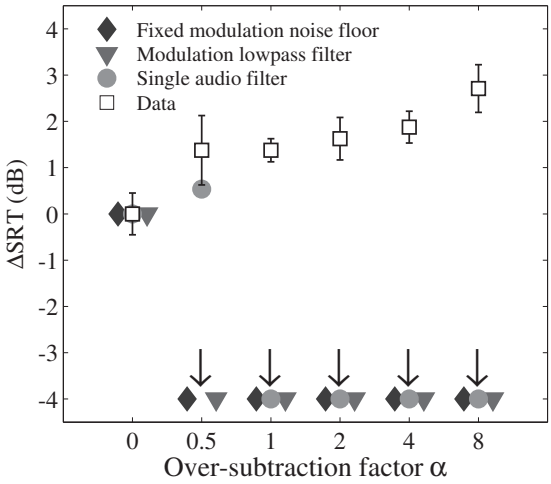


Figure 2.8: ΔSRTs as a function of the over-subtraction factor α. The open squares show the data from Fig. 2.6. The filled circles represent predictions when the gammatone filterbank was replaced by a single bandpass filter with cutoff frequencies at 63 Hz and 8 kHz. The filled triangles show predictions when the modulation filterbank was replaced by a single lowpass filter with a cutoff frequency of 150 Hz, while keeping the original gammatone filterbank. The filled diamonds show predictions with a fixed modulation noise floor, explained in Sec. 2.5.3. The arrows indicate that the corresponding values are outside the ordinate scale.

envelope power of the noise, which is increased by a greater amount than that of the noisy speech. The detrimental effect of spectral subtraction on speech intelligibility may therefore be explained by a decreasing $\mathrm{SNR_{env}}$ with increasing α. Figure 2.9(B) shows $\mathrm{SNR_{env}}$ as a function of the input SNR for the different values of α used in the experiment. Each point represents the overall $\mathrm{SNR_{env}}$ for a given value of α and input SNR. $\mathrm{SNR_{env}}$ generally decreases with increasing α, except for the input SNR of -9 dB. To further illustrate this, Fig. 2.9(C) shows the change in envelope SNR, $\Delta\mathrm{SNR_{env}}$ (left ordinate), obtained at the fixed input SNR of 0 dB as a function of α. $\Delta\mathrm{SNR_{env}}$ decreases from 0 to -2.5 dB with increasing α. For comparison, the open squares replot the data from Fig. 2.6 (right ordinate, reversed scale). The predicted decrease of $\mathrm{SNR_{env}}$ with α, for this particular input SNR, is in quantitative agreement with the increase of SRT observed in the data. It should be noted that the units of

SNR_{env} and SRT are not the same. Therefore, this comparison should only illustrate that the measured change in SRT is paralleled in the SNR_{env} metric.

Since SNR_{env} is calculated from the envelope power of the noisy speech and noise alone, the value of the noise envelope power alone is crucial. To further evaluate this, predictions were made assuming a *fixed* value of the noise-alone envelope power as a function of the processing condition. Specifically, the transmission-channel processing (here spectral subtraction) was not applied to the noise-alone path of the model (see Fig. 2.2). In this case, the increase of the envelope power of the noisy speech, as result of spectral subtraction, would lead to an increase of SNR_{env}. As a consequence, the corresponding predicted SRT decreases. The filled diamonds in Fig. 2.8 show the corresponding prediction, showing a decrease of SRT for all values of $\alpha > 0$. A decrease of SRT corresponds to an increase of intelligibility, which is inconsistent with the data. This shows that, in the model, the envelope power of the noise alone is crucial for prediction of the intelligibility of noisy speech subjected to spectral subtraction.

2.5.4 The effect of reverberation in the envelope-frequency domain

The four panels in Fig. 2.10(A) show modulation excitation patterns at the output of the 1-kHz gammatone filter for conditions with different amounts of reverberation. The top left panel represents the reference condition without reverberation ($T_{30} = 0$). The envelope power of noisy speech decreases above 2 Hz as the reverberation time increases, indicating that reverberation acts as a low-pass filter in the modulation domain (Houtgast and Steeneken, 1985). The envelope power of the noise alone increases at low modulation filter (center) frequencies when the reverberation time increases, which is mainly caused by the increasing duration of the reverberant tail of the noise stimulus. As a result, the overall SNR_{env} decreases with increasing reverberation time, consistent with the decreasing intelligibility seen in Fig. 2.5 (open squares) .

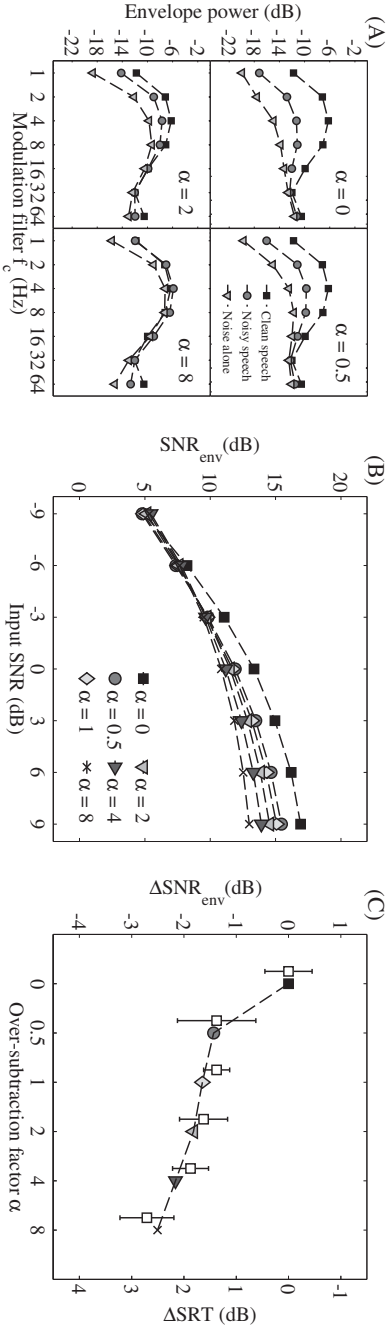


Figure 2.9: (A) Modulation excitation patterns computed at the output of the 1-kHz gammatone filter for four values of the over-subtraction factor: $\alpha = 0$ (top left), $\alpha = 0.5$ (top right), $\alpha = 2$ (bottom left) and $\alpha = 8$ (bottom right). (B) $\mathrm{SNR}_{\mathrm{env}}$ as a function of the input SNR, for all values of α used. (C) $\Delta\mathrm{SNR}_{\mathrm{env}}$ as a function of α for a constant input SNR of 0 dB. The open squares replot the data from Fig. 2.6 with a reversed scale on the right ordinate.

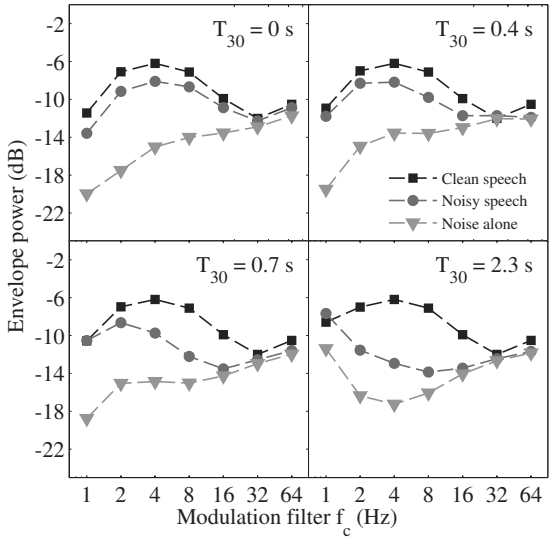


Figure 2.10: Modulation excitation patterns for the output of the 1-kHz gammatone filter for four reverberation times: $T_{30} = 0$ (top left), $T_{30} = 0.4$ (top right), $T_{30} = 0.7$ (bottom left) and $T_{30} = 2.3$ (bottom right). The input SNR was 9 dB.

2.6 Discussion

2.6.1 Relation between sEPSM and STI

The present study shows that the sEPSM framework accounts well for the measured ΔSRT values observed for conditions of noisy speech and reverberation. The STI has earlier been demonstrated to account for similar data (e.g., Duquesnoy and Plomp, 1980; Houtgast and Steeneken, 1973) as well as for conditions in which speech is distorted by reverberation only (Houtgast *et al.*, 1980; Houtgast and Steeneken, 1985). Data for the purely reverberant condition (without noise distortion) can also be accounted for by the sEPSM. In this case, the $\mathrm{SNR_{env}}$ is limited by internal noise (not shown here explicitly). In addition to the conditions of noise and reverberation, the sEPSM was shown to account for the intelligibility of noisy speech subjected to spectral subtraction, where the STI approach clearly fails. The sSTI approach of Goldsworthy and Greenberg

(2004) predicted a slight increase of intelligibility up to $\alpha = 2$, followed by a decrease towards $\alpha = 8$. Although their implementation of the spectral subtraction algorithm differed slightly from the one used in the present study, their predictions contradict the data presented here. The critical distinction between all STI-based models and the sEPSM is that the STI considers the difference between clean speech and noisy speech in the envelope domain, quantified by the modulation transfer function (MTF). This is equivalent to assuming a *fixed* noise floor within the noisy speech. In contrast, the sEPSM considers the difference between the noisy speech and the noise alone, and the noise alone provides an explicit estimate of the intrinsic noise floor. In the case of spectral subtraction, the modulations in the noisy speech envelope are enhanced by the processing, and this leads to an increased envelope power, which causes the STI to predict an increase of intelligibility. However, the modulations inherent in the noise-alone envelope are enhanced as well, in some conditions by a larger amount than the noisy speech (Fig. 2.9(A)). The increased intrinsic noise envelope power is not taken into account by the STI which explains why the approach must fail in these conditions. If a fixed noise floor was assumed in the sEPSM, this would also fail (Fig. 2.8). Thus, the key to the sEPSM approach is the explicit estimation of the intrinsic noise envelope power within the noisy speech.

2.6.2 Audio and modulation frequency weighting

Weighting of individual frequency bands is a common feature of speech intelligibility prediction metrics such as the AI, SII and STI. The general rationale for the weighting is that certain frequency regions appear to be perceptually more relevant for speech intelligibility than other frequency regions (French and Steinberg, 1947; Kryter, 1960; Houtgast and Steeneken, 1985; Warren *et al.*, 2005). In the case of the STI, the weighting can be separated into two types: (i) audio-frequency weighting of individual octave bands within the frequency range 0.125 and 8 kHz, and (ii) 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 truncation of modulation frequencies was motivated by the finding that speech modulation spectra have low amplitudes above 16 Hz (Drullman *et al.*, 1994b,a;

Xu *et al.*, 2005). In contrast to the standardized metrics, the sEPSM does not apply any explicit weighting functions, apart from limitations in terms of absolute sensitivity (Sec. 2.2.2). The agreement between the predicted and measured intelligibility (Fig. 2.5 and Fig. 2.6) 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.

2.6.3 The role of interaction modulations

The measure signal-to-noise ratio in the modulation domain, $(S/N)_{mod}$, proposed by Dubbelboer and Houtgast (2008), quantified the strength of speech modulations relative to a noise floor of spurious modulations. The spurious modulations consist of noise modulations and interaction modulations arising from the nonlinear interaction between the noise and the speech waveforms. The novelty compared to the STI was their hypothesis that speech intelligibility may be reduced not only due to the decrease of the speech modulations, but also due to the increase of the level of the spurious noise modulation, for example in the case of the processing through spectral subtraction. The present study followed the idea that noise modulations can influence speech intelligibility. However, the hypothesis of Dubbelboer and Houtgast (2008) that the interaction modulations represent an essential effect causing the decreased intelligibility of noisy speech processed by spectral subtraction, was not supported here. Instead, the main effect causing decreased intelligibility, according to the sEPSM, is the increase of the intrinsic fluctuations in the noise itself.

Another main assumption of the $(S/N)_{mod}$ concept was that the modulation noise floor, and thus $(S/N)_{mod}$, could only be estimated from the mixed waveform of noise and speech. Because of the difficulties related to envelope filtering of noisy speech, Dubbelboer and Houtgast (2008) used a special narrow-band probe signal, containing a single modulation frequency, for estimating $(S/N)_{mod}$. In contrast, SNR_{env} is calculated here from the speech and the noise separately, so problems with separating the noise and the speech after envelope filtering are avoided. The drawback is that the interaction between the speech and the noise waveforms is neglected. However, given the accuracy

of the present model predictions, it appears that the intrinsic fluctuations in the noise envelope play a larger role for speech intelligibility than interaction modulations.

The assumption that $(S/N)_{mod}$ must be estimated from the mixture of speech and noise makes it difficult to extend from a concept to a model. In contrast, SNR_{env} in the present study was calculated using a model framework, the sEPSM, that includes elements of established hypotheses about human auditory processing. An earlier version of this framework was demonstrated to predict modulation detection and masking data with different types of carriers and modulation maskers (Dau *et al.*, 1999; Ewert and Dau, 2000; Ewert *et al.*, 2002; Ewert and Dau, 2004).

2.6.4 Limitations of the approach

Various studies have demonstrated that speech intelligibility is higher in conditions with non-stationary interferers, such as amplitude modulated noise or competing talkers, than with stationary interferers (e.g., Miller, 1947; Miller and Licklider, 1950; Duquesnoy, 1983; Festen and Plomp, 1990; Füllgrabe *et al.*, 2006). The difference in intelligibility has been called speech masking release. The sEPSM in its present form cannot account for this phenomenon, since predictions are based on the long-term integrated SNR_{env}. For example, if the noise was amplitude modulated at a rate of 4 Hz, this would result in greater noise envelope power at 4 Hz which, in turn, would lead to a very low SNR_{env} at this frequency. The contribution from the 4 Hz modulation filter to the overall SNR_{env} would then be reduced compared to the stationary case, and the model would predict decreased speech intelligibility, in contrast to the experimental data. Thus, benefits from listening in the dips of a modulated noise or a competing talker cannot be accounted for by the model in its present form. Moreover, the model does not include nonlinear effects of cochlear processing, such as compression and suppression, nor does it reflect any influence of the statistics of the integrated envelope power which is a random variable over short times. In this respect, the sEPSM differs substantially from, for example, the more complex processing models described by Dau *et al.* (1997a,b) or Jepsen *et al.* (2008), which were applied to speech intelligibility in the studies of Jürgens and Brand (2009) and Christiansen *et al.* (2010). In particular, the models of

Dau *et al.* (1997a,b) and Jepsen *et al.* (2008) contain an adaptation stage that allows the description of simultaneous and non-simultaneous masking as well as modulation masking effects. Such a stage is not contained in the current model. The sEPSM is a model that is restricted to amplitude modulation processing and makes only a few assumptions about the processing of the stimuli. Even though this model oversimplifies the "real" processing in the auditory system (more than do the complex models), this analytical approach might be helpful for understanding which processing stages are essential for successfully describing key aspects of speech recognition.

2.6.5 Perspectives

The sEPSM can be applied to non-linear conditions that have not been considered in the present study, such as amplitude compression and peak-clipping. In particular, the intelligibility changes due to phase-jitter and phase-shift distortions, as studied by Elhilali *et al.* (2003), pose interesting challenges. Elhilali and colleagues showed that their concept of the spectro-temporal modulation index, STMI, can account for the changes in intelligibility caused by these distortions. The STMI assumes a two-dimensional modulation filtering process, motivated by spectro-temporal response fields in the cortex of animals (Chi *et al.*, 1999), and the degradation of the spectral periodicity in the "auditory spectrogram" due to the non-linear processing is reflected in the STMI, but not in the STI. However, it would be interesting to test to what extent the noise distortions produced by this non-linear process can be accounted for by the sEPSM, which only considers a one-dimensional (temporal) modulation analysis.

In its current form, the sEPSM utilizes an observer that functions as a signal detector, assuming that a certain long-term overall $\mathrm{SNR}_{\mathrm{env}}$ corresponds to a given percent correct. This detector does not perform any direct recognition of words nor consider the difference between individual words from a single-syllable word material. In this way, the sEPSM (and the standardized STI and SII) differs from models that predict smaller details of the speech signal, such as consonant or phoneme confusions (Gallun and Souza, 2008; Jürgens and Brand, 2009). The spectral correlation index (SCI; Gallun and Souza, 2008) measures the modulation energy in a number of modulation bandpass

filters at the output of a number of audio bandpass filters. Instead of considering the total modulation energy of a given stimulus, as performed in the STI and the present model, the SCI considers the specific *pattern* of modulation energy across the different modulation and audio filters. It assumes that a given consonant will have a specific modulation energy pattern, and that a confusion between two consonants can be predicted from the correlation between their corresponding modulation patterns. The SCI showed a correlation with measured consonant confusion error rates of up to 0.82. The ideal observer in the sEPSM could be modified to perform a similar operation as the SCI. However, in contrast to considering a pattern of speech modulation energy, the sEPSM would consider a specific pattern of $\mathrm{SNR_{env}}$-values associated with each consonant. By comparing patterns of $\mathrm{SNR_{env}}$ instead of speech modulation energy, effects of modulation masking (e.g. caused by intrinsic fluctuations of external noise) would be included in the estimates of confusions.

The $\mathrm{SNR_{env}}$ concept could also be incorporated into more advanced models of auditory signal processing (e.g., Chi *et al.*, 2005; Jepsen *et al.*, 2008). Such more detailed models might facilitate investigations of the effect of hearing-impairment on speech intelligibility.

2.7 Summary and conclusions

The speech-based envelope power spectrum model (sEPSM) was proposed for the prediction of the intelligibility of processed noisy speech. The model estimates the envelope signal-to-noise ratio, $\mathrm{SNR_{env}}$, from the long-term integrated envelope power falling in the transfer range of a number of modulation bandpass filters, and combines the $\mathrm{SNR_{env}}$ values optimally across modulation filters and auditory filters.

Model predictions were compared with measured speech recognition thresholds (SRT) for two conditions: (i) reverberant noisy speech and (ii) noisy speech subjected to spectral subtraction. The predictions showed an increase of the SRT both with increasing reverberation time and as a consequence of spectral subtraction, in good agreement with the data. The STI predicts similar results for conditions with

reverberation, but clearly fails for spectral subtraction. The analysis of the model predictions revealed that the spectral subtraction process increases the envelope power of the estimated noise component of the noisy speech by a greater amount than that of the speech itself, which leads to a decreased $\mathrm{SNR}_{\mathrm{env}}$ and thus a decrease in predicted intelligibility.

The sEPSM concept represents a more general concept than the STI and might be powerful when applied to other conditions of processed noisy speech.

Appendix

A. Details of the ideal observer

The conversion from the sensitivity index, d', to percent correct is described by equation 2.8. The values of σ_N and μ_N are determined by the number of alternatives, m, in the mAFC model of Green and Birdsall (1964):

$$\sigma_N = \frac{1.28255}{U_n} \quad \text{and} \quad \mu_N = U_n + \frac{0.577}{U_n} \tag{2.10}$$

where

$$U_n = \Phi\left(1 - \frac{1}{m}\right)^{-1} \tag{2.11}$$

and Φ denotes the cumulative normal distribution. As m increases, μ_N also increases while σ_N decreases. The value of U_n represents the value that would be drawn from a normal distribution with probability $p = 1 - 1/m$. 1/m corresponds to the minimum percent correct of the model's psychometric function, which for three alternatives is 33% correct and for 100 alternatives is 1% correct.

Acknowledgments

We thank two anonymous reviewers for their very helpful comments and suggestions. We also want to thank Brian Moore and Morten Jepsen for valuable suggestions for improvement of an earlier version of the manuscript.

3

The importance of auditory spectro-temporal modulation filtering and decision metric for predicting speech intelligibility†

Speech intelligibility models typically consist of a preprocessing part that transforms the 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 that were (i) subjected to reverberation, (ii) distorted by phase jitter or (iii) 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 not be required for speech intelligibility prediction.

3.1 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 (SNR) 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; IEC60268-16, 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, 1982; ANSI S3.5, 1997). The MTF calculates the reduction of the envelope fluctuations 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. Because the MTF captures the effects of the distortions on the envelope of the reference signal, the STI accounts 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

† This chapter is based on Chabot-Leclerc *et al.* (2014).

(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 STI methods (sSTI). The main difference from the original STI method is that speech is used as the reference signal rather than a modulated wide-band noise, and that the integrity of the temporal envelope is quantified by other metrics than the MTF. Although the sSTI methods seemed promising, they were never 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.

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 modulations 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 is in contrast to the STI, which only considers the integrity of the modulations in the temporal domain. The SMTI considers a two-dimensional (spectro-temporal) MTF as the decision metric, effectively assuming a spectro-temporal modulation band-pass 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 $\mathrm{STMI}^{\mathrm{R}}$ and $\mathrm{STMI}^{\mathrm{T}}$, 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 principal limitations as the STI when noisy speech is processed by spectral subtraction.

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). 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 account for the changes of intelligibility observed in conditions with additive noise, reverberation, and 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 fluctuations 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 essential 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 essential for the prediction of effects of spectral subtraction, but kept the analysis of only *temporal* modulations,

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 2-D spectro-temporal modulation filterbank, as in the STMI, denoted in the following as "2D-sEPSM". The other realization kept the purely temporal (1-D) modulation filterbank and introduced a mechanism that analyzed the variance of the output 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 comodulation masking release (CMR; e.g., van de Par and Kohlrausch, 1998; Piechowiak *et al.*, 2007). CMR refers to the greater detectability of a tone centered in a narrow-band noise, surrounded by one or more flanking noise bands with comodulated waveforms, compared to the same situation with uncorrelated flanking noise bands. The addition of a tone in the comodulated 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 loss of speech information.

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).

3.2 Model descriptions

Figure 3.1 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 power above the normal hearing threshold are considered. The two final stages indicated in Fig. 1 represent the decision module, consisting of the SNR_{env} calculation and an "ideal observer", as defined in Jørgensen and Dau (2011).

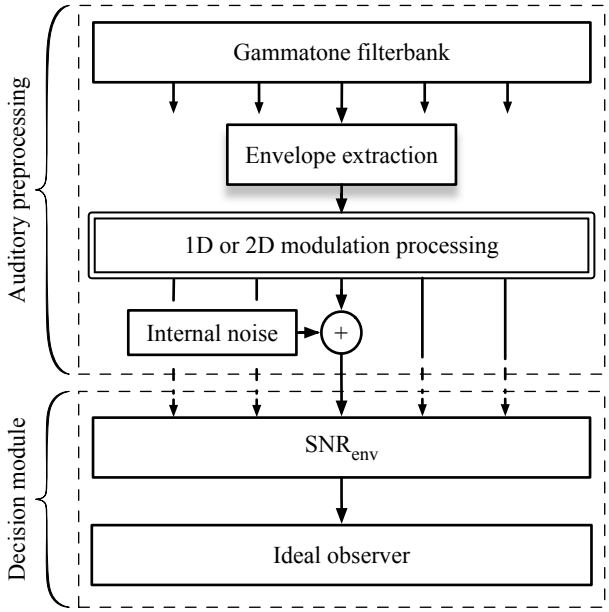


Figure 3.1: Block diagram of the overall structure of the modeling framework. The model consists of a gammatone bandpass filterbank followed by envelope extraction via the Hilbert transform, producing the "auditory spectrogram" of the incoming signal. The modulation filtering differs in the two considered model versions, the 2D-sEPSM and the sEPSM^{X}. The envelope signal-to-noise ratio, SNR$_{env}$, is calculated from the time-averaged "cortical representation" at the output of either the one-dimension modulation filtering (sEPSM^{X}) or the two-dimensional modulation filtering (2D-sEPSM). 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.

3.2.1 Model 1: Two-dimensional 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 90 Hz and 3.5 kHz (24 filters/octave over a 5.3 octave range). The envelope of the output of each gammatone filter is extracted using the Hilbert transform, lowpass filtered using a first-order Butterworth filter with a cut-off frequency of 150 Hz (Ewert

and Dau, 2000; Kohlrausch *et al.*, 2000) and short-term averaged in blocks of 4 ms to form an "auditory spectrogram"-like representation. Next, the joint spectral and temporal modulation content is extracted from the auditory spectrogram using a bank of spectrally and temporally selective modulation-filters (Chi *et al.*, 1999). The output of this processing is referred to as the "cortical representation" of the incoming signal and has the four dimensions time, cochlear frequency, temporal modulation frequency and spectral modulation frequency. 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 cortical 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 cortical representation is calculated as the variance across time, normalized with the squared time-average of the unfiltered temporal envelope, leaving a three dimensional internal representation of the noisy speech mixture, $T_{mix}(f,\omega,\Omega)$, and of the noise alone, $N(f,\omega,\Omega)$. These time-collapsed cortical representations are considered equivalent to the envelope power spectrum of the noisy speech, $P_{env,S+N}$, and of the noise alone, $P_{env,N}$, as defined in Jørgensen and Dau (2011). The $\mathrm{SNR_{env}}$ can therefore be expressed as:

$$SNR_{env} = \frac{T_{mix} - N}{N}. \tag{3.1}$$

It is assumed that the values of the mixture do not exceed the value of the clean speech:

$$T_{mix} = \min(T_{mix}, T_{clean}), \tag{3.2}$$

and that the cortical representation of the noise does not exceed that of the mixture:

$$N = \min(T_{mix}, N) \tag{3.3}$$

The lower limits of T_{mix} and N are represented by a small positive value ε, reflecting an

internal noise threshold:

$$T_{mix} = \max(T_{mix}, \varepsilon) \tag{3.4}$$

$$N = \max(N, \varepsilon) \tag{3.5}$$

In order to limit the model's sensitivity to very small modulations in the stimuli, ε is set to -40 dB. [1].

3.2.2 Model 2: One-dimensional envelope power spectrum model with variance weigthing across frequency (sEPSM^{X})

The sEPSM^{X} assumes 22 gammatone filters with 1/3-octave spacing of the center frequencies, covering the range from 63 Hz to 8 kHz. The envelope of the output of each gammatone filter is extracted via the Hilbert transform and low-pass filtered using a first-order Butterworth filter with a cut-off frequency of 150 Hz (Kohlrausch *et al.*, 2000). The envelope is analyzed by a filterbank consisting of a third-order lowpass 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^{X}, the cortical representation is a three-dimensional function of time, audio-filter center frequency, and modulation-filter center frequency.

At the decision device, the long-term envelope power is calculated from the temporal output of each modulation filter as the variance of the filter output across time, and normalized with the squared time-average (DC) of the unfiltered envelope. The $\mathrm{SNR_{env}}$ is then calculated from the normalized envelope power of the noisy speech, $P_{env,S+N}$, and the noise alone, $P_{env,N}$, at the output of each modulation filter:

$$SNR_{env} = \frac{P_{env,S+N} - P_{env,N}}{P_{env,N}} \tag{3.6}$$

[1] 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.

Similarly to the 2D-sEPSM, the model's' sensitivity is limited and the envelope power below -20 dB is set to -20 dB (Jørgensen and Dau, 2011).

3.2.3 Transformation from $\mathrm{SNR_{env}}$ to probability of being correct

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

$$SNR_{env} = \left[\sum_{g=1}^{G} \sum_{l=1}^{L} \left(SNR_{env,g,l} \right)^2 \right]^{1/2} \tag{3.7}$$

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

$$SNR_{env,g,l} = [\sigma_g^2]^{\beta} \cdot SNR_{env,g,l}, \tag{3.8}$$

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 value of σ_g^2 was determined from the stimuli using several computational steps: first, the long-term envelope power of the noisy speech mixture, $P^*_{env,g,l}$, was computed at the output of modulation-filter g and audio-filter l and normalized with a factor proportional to the bandwidth in Hz 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 $P^*_{env,g,l}$ was computed across the 22 peripheral filters.

Figure 3.2 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 $P^*_{env,g,l}$ represent the center frequencies of the filters. The across-channel variance is calculated across rows in a given column.

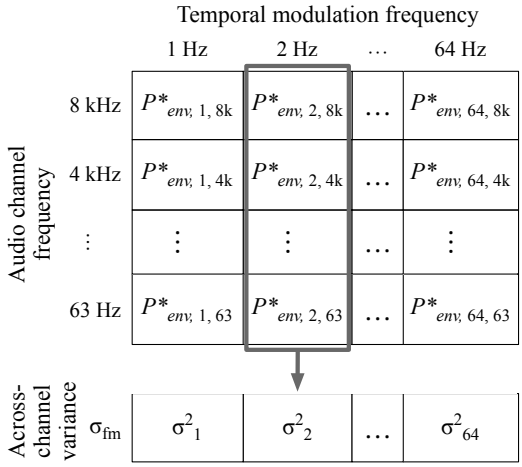


Figure 3.2: Illustration of the calculation of the across-channel variance of the envelope power for a given time frame. Each cell represents the normalized envelope power, $P^*_{env,g,l}$, of the noisy speech at the output of a temporal modulation filter g, and audio filter, l. The across-channel variance, σ_g^2, for a given modulation filter center frequency corresponds to the variance across rows for a given column of the matrix.

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

$$d' = k \cdot (SNR_{env})^q \tag{3.9}$$

where k and q are constants, 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 $x_{N,max}$:

$$P_{correct}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right) \tag{3.10}$$

where Φ designates the cumulative normal distribution. The values of σ_N^2 and μ_N are determined by the number of alternatives, m, which for an open-set paradigm is set to the number of words in a normal listener's active vocabulary (Müsch and Buus, 2001). The value of σ_S is inversely proportional to the slope of the ideal observer's psychometric function, and can be adjusted to account for different degrees of redundancy in the speech materials (see Jørgensen and Dau, 2011).

3.3 Method

3.3.1 Speech 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 conditions, the sentences were mixed with speech-shaped stationary noise having the same long-term average spectrum as the speech material.

3.3.2 Stimuli and experimental conditions

Three conditions of processed noisy speech were considered: reverberant speech, speech processed by spectral subtraction and speech subjected to a phase-jitter distortion. In all conditions, sentences were mixed with the stationary speech noise at a given SNR before processing. The measured data presented here were either collected as part of the present study (phase jitter condition) or were taken from Jørgensen and Dau (2011) (reverberation and spectral subtraction conditions).

Reverberation

The noisy sentences were convolved with an impulse response corresponding to a particular reverberation time. The impulse responses were created using the ODEON room acoustic software version 10 (Christensen, 2009). The simulated room was rectangular in shape with a volume of 3200 m^3 and the absorption was distributed such that the room had very similar reverberation time (T_{30}) measured in octave bands from 63 to 8000 Hz. In the room simulation, the absorption was adjusted such that the room impulse responses corresponded to five different values of T_{30}: 0, 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, were 0.60, -2.9, -6.6, -8.0 dB, respectively.

Spectral subtraction

The spectral subtraction was implemented using the scheme defined by Berouti *et al.* (1979). An estimate of the noise power spectrum was subtracted from the power spectrum of the noisy speech in 20-ms time windows. The amount of noise subtracted was defined by the over-subtraction factor, κ. 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, $s(t)$, with an SNR of 5 dB with a cosine function with a random phase, as described in Elhilali *et al.* (2003):

$$r(t) = \Re\{s(t)e^{j\Theta(t)}\} = s(t)\cos(\Theta(t)), \tag{3.11}$$

where θ(t) is a random process uniformly distributed over $[0, 2\pi]$ ($0 < \alpha < 1$), 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 α = 0.5 and 1, the signal becomes a temporally modulated white noise.

3.3.3 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 driven by a high-quality soundcard in a double-walled sound-attenuating booth. The speech had a constant level of 65 dB sound pressure level and noise was added to achieve the desired SNR before further processing. Each sentence was presented once with the noise starting one second 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 subject: two lists were used for each α value and two sentences per list for each SNR, resulting in four data points for each combination of SNR and distortion parameter per subject. 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.

3.3.4 Listeners

Measurements were obtained with six normal-hearing males, aged from 25 to 27 years. Their pure-tone thresholds were of 20 dB hearing level or better in the frequency range 0.25–8 kHz. All subjects had previous experience with psychoacoustic measurements. All of them were native Danish speakers and students at the Technical University of Denmark. None of them were paid for their participation.

Table 3.1: Calibrated values of the parameters σ_S, m, k, and q of the ideal observer for the CLUE speech material.

Model	k	q	σ_S	m	β
2D-sEPSM	0.7	0.28	0.6	8000	-
sEPSM^X	0.79	0.5	0.6	8000	0.31

3.3.5 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 -7 to 9 dB with 4 dB intervals, except in the phase jitter condition where the SNR was 5 dB only. In all cases, the processing 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 speech reception threshold (SRT) for a specific condition was obtained as the SNR corresponding to 50% intelligibility.

For the EPSM^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 the conditions with phase jitter. 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 to obtain a good agreement with the data in the reference condition. All parameters were then kept fixed in all other experimental conditions and the values are given in Table 3.1.

3.4 Results

3.4.1 Reverberant speech

Figure 3.3 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' average SRT. The measured SRT increased with increasing reverberation time, reflecting a decrease in intelligibility. 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 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 root-mean-square-error (RMSE) was 2.2 dB. In the case of the sEPSM^{X} (black triangles), the Pearson coefficient was 0.99 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. 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}.

3.4.2 Spectral subtraction

Figure 3.4 shows the results for the condition with spectral subtraction. SRTs are shown as a function of the over-subtraction factor κ. The measured SRTs, replotted 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^{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^{X} predictions had a correlation with the data of $\rho = 0.99$ and a RMSE of

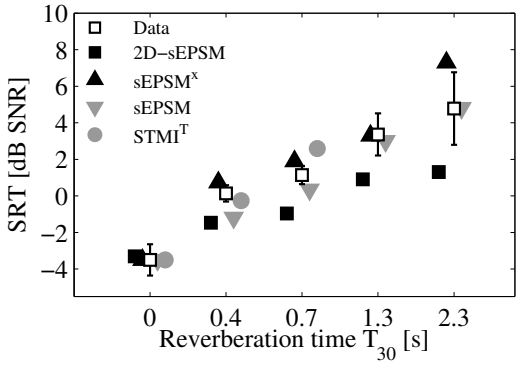


Figure 3.3: SRTs as a function of the reverberation time, T_{30}. The open squares represent a replot of the data in Jørgensen and Dau (2011), with the vertical bars indicating one standard deviation. The filled symbols show predictions obtained with the different models. The black squares and the black 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. 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.

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^T. This model predicted a *decrease* of SRT, i.e. increasing speech intelligibility with increasing κ, in contrast to the measured data.

3.4.3 Phase jitter

The open symbols in Fig. 3.5 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

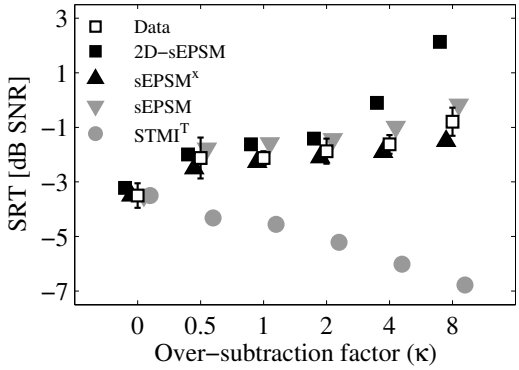


Figure 3.4: SRTs as a function of the over-subtraction factor κ in conditions of spectral subtraction. The open squares represent measured data from Jørgensen and Dau (2011), with the vertical bars indicating one standard deviation. The filled black squares show the predictions obtained with the 2D-sEPSM and the upward triangles represent the results using the sEPSM^{X}. For direct comparison, the filled gray downward triangles show predictions with the original sEPSM (replotted from Jørgensen and Dau, 2011). The filled gray circles show predictions obtained with the STMI^{T} as proposed by Elhilali *et al.* (2003).

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. A two-way analysis 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 above 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 0.93 and the RMSE was 24.5%. The predictions obtained with the sEPSM^{X} followed the data more closely than the 2D-sEPSM, with minima of about 4% correct responses for $\alpha = 0.5$ and 1 and a local maximum of 47% for $\alpha = 0.75$. The correlation between the sEPSM^{X}

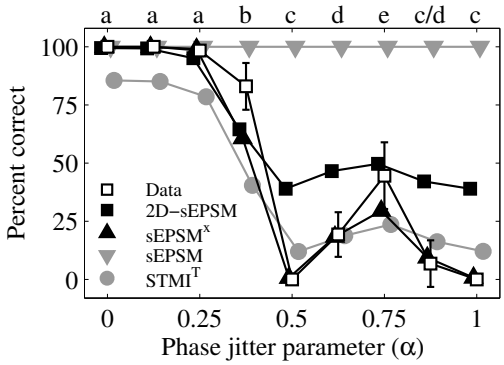


Figure 3.5: Word intelligibility as a function of the phase-jitter parameter α, for a fixed SNR of 5 dB. The open symbols indicate average measured data collected in the present study. Vertical bars show one standard deviation of the average listeners' percentage of word intelligibility. The filled symbols show predictions obtained with the different models. 2D-sEPSM predictions are shown as filled black squares; sEPSM^{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^{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).

predictions and the data was $\rho = 0.98$, with an RMSE value of 9%. 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 obtained with the STMI^{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.

3.5 Discussion

3.5.1 The role of the decision metric

All considered models could, with varying degrees of accuracy, account for the main effect of reverberation on the intelligibility of noisy speech. In contrast, only the models considering the decision metric based on the SNR_{env} could account for the detrimental effect of spectral subtraction. The MTF-like metric of the $STMI^T$ could not account for the spectral subtraction data, possibly because it does not consider the effects of the nonlinear processing on the inherent noise modulations alone. This is consistent with the results from Jørgensen and Dau (2011). The two models presented here, the 2D-sEPSM and the $sEPSM^X$, both 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. This suggests that the SNR_{env} is a powerful metric for speech intelligibility prediction and that it is robust with respect to specific assumptions in the auditory preprocessing.

3.5.2 The role of across-frequency modulation processing

Regarding 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 in this condition. The results obtained with the two 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. 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 cat (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; Verhey *et al.*, 1999; Derleth and Dau, 2000; Jepsen *et al.*, 2008). The approach to measure the amount of coherent modulation activity across frequency after the preprocessing in the model is also consistent with recent concepts in computational auditory scene analysis (Elhilali *et al.*, 2009), 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 ones based on cross-correlation or the temporal coherence of the temporal envelope of neighboring channels, 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 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 were optimized for the reference condition only, while the k and β-parameters in 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 metrics 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.

3.5.3 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 S., 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 (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.

3.5.4 The role of the frequency weighting for predicted 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 explicit 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, 1960; Houtgast and Steeneken, 1985; Warren *et al.*, 2005). For example, in the case of the STI, the weighting has been separated into two types: (i) audio-frequency weighting of individual octave bands within the frequency range 0.125 and 8 kHz, and (ii) 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.

3.5.5 Relation to other speech intelligibility models

An alternative approach for predicting speech intelligibility is the short-time objective intelligibility model (STOI; 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 band pass filters. A key step in the 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. This is conceptually similar to the $\mathrm{SNR}_{\mathrm{env}}$ metric, which effectively measures the ratio of speech and noise modulation power. 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. 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 would 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 caused by the jitter. The main difference is that the sEPSM framework makes specific assumptions about the source of the speech degradation, i.e., modulation masking by noise, while this is not clear in the correlation coefficient. Moreover, the sEPSM includes additional aspects of human auditory processing, in the form of the perceptually and physiologically motivated modulation filterbank, which might be crucial in other conditions, such as reverberation, where the STOI metric have limitations (Taal *et al.*, 2011).

3.5.6 Perspectives

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. Alternatively, since the sEPSM is a model that is restricted to amplitude modulation processing, the $\mathrm{SNR}_{\mathrm{env}}$ concept could also be incorporated into more advanced models of across-channel auditory signal processing (e.g., Chi *et al.*, 2005; Dau *et al.*, 2013). The sEPSM approach simplifies the "real" processing in the auditory system much more than the complex models do. Nevertheless, this simple approach might be helpful for understanding which processing stages are essential for successfully describing key aspects of speech perception.

Acknowledgements

This research was supported in part by the National Science and Engineering Research Council of Canada (NSERC), the Danish Research Foundation, and the Danish hearing aid companies Oticon, Widex, and GN ReSound.

4

A Multi-resolution envelope-power based model for speech intelligibility[‡]

The speech-based envelope power spectrum model (sEPSM) presented by Jørgensen and Dau (2011) estimates the envelope power signal-to- noise ratio (SNR_{env}) after modulation-frequency selective processing. Changes in this metric were shown to account well for changes of speech intelligibility for normal-hearing listeners in conditions with additive stationary noise, reverberation, and nonlinear processing with spectral subtraction. In the latter condition, the standardized speech transmission index (IEC60268-16, 2003) fails. However, the sEPSM is limited to conditions with stationary interferers, due to the long-term integration of the envelope power, and cannot account for increased intelligibility typically obtained with fluctuating maskers. Here, a multi-resolution version of the sEPSM is presented where the SNR_{env} is estimated in temporal segments with a modulation-filter dependent duration. The multi-resolution sEPSM is demonstrated to account for intelligibility obtained in conditions with stationary and fluctuating interferers, and noisy speech distorted by reverberation or spectral subtraction. The results support the hypothesis that the SNR_{env} is a powerful objective metric for speech intelligibility prediction.

[‡] This chapter is based on Jørgensen *et al.* (2013).

4.1 Introduction

Speech intelligibility in noisy environments is largely affected by the spectral and temporal characteristics of the background noise. In an early review, Miller (1947) described the interfering effects of tones, white noise, and speech on the intelligibility of target speech. A stationary noise with a long-term spectrum similar to that of the target speech was indicated to be the most effective masker. In a later study, Miller and Licklider (1950) demonstrated that the amount of speech masking by noise decreased when the noise was periodically interrupted, while keeping the long-term spectrum and the signal-to-noise ratio (SNR) the same as in the stationary case. The corresponding reduction of the speech reception threshold (SRT) for the speech presented in the interrupted noise relative to that obtained in the steady noise was denoted as speech masking release, with SRT representing the SNR at which 50% of the target speech was correctly understood. The observed masking release was attributed to the listeners' ability to take advantage of the portions of the noisy speech with a favorable short-term SNR, i.e., to utilize speech information in the dips of the interferer. Various other studies have explored the masking release effect for speech using different types of interferers (e.g., Dirks *et al.*, 1969; Festen and Plomp, 1990; Howard-Jones and Rosen, 1993; George *et al.*, 2006; Rhebergen *et al.*, 2006; Buss *et al.*, 2009). The size of the masking release effect was found to depend greatly on the type of fluctuations in the masker, e.g., sinusoidal, speech-like, or gated, as well as on the fluctuation rate, modulation depth, and duty cycle.

Classical approaches to predict speech intelligibility, such as the articulation index (AI; ANSI S3.5, 1969) and its successor, the speech intelligibility index (SII; ANSI S3.5, 1997), have failed to account for the masking release effect. Since these models are based on long-term frequency information only, they cannot distinguish between a stationary and a fluctuating masker if the long-term frequency contents of the maskers are the same. In an alternative approach, Festen and Plomp (1990) considered a metric based on the modulation-transfer function (MTF; Houtgast *et al.*, 1980), which reflects the reduction of the intensity modulations of a probe signal caused by the transmission through a processing channel. However, the speech transmission index (STI; IEC60268-16, 2003), which is based on the MTF-concept, makes no distinction

between the interferer and the target fluctuations and thus cannot account for the masking release effect. In order to separate the fluctuations of the target and the interferer, Festen and Plomp (1990) considered a modified version of the STI, proposed by Ludvigsen *et al.* (1990), which is based on a regression analysis of the clean and the noisy modulation spectra. While the corresponding predictions showed the overall trends in the data, the amount of masking release was generally underestimated by this approach.

A common feature of the above models is that their predictions are based on the long-term estimates of speech and noise, i.e., no short-term stimulus information is considered. To overcome this limitation, Rhebergen and Versfeld (2005) and Rhebergen *et al.* (2006) proposed a short-time version of the SII, denoted as the extended SII (ESII). The main property of this metric is the determination of a frequency-band specific SNR in a number of short temporal windows covering the time-course of a given speech token. The short-term SNR values are weighted with empirically derived band-importance functions, and the resulting short-term SII-values are averaged to provide an estimate of the intelligibility of the token. This approach was shown to account for the masking release effect obtained with square-wave interrupted noise, sinusoidally intensity modulation noise, saw-tooth noise, as well as multi-talker babble. However, also substantial discrepancies were demonstrated, e.g., in conditions with speech mixed with interfering speech from a single talker. Furthermore, as the original SII, the ESII fails in conditions where noisy speech is nonlinearly processed, e.g., via amplitude compression (Rhebergen *et al.*, 2009) or noise-reduction processing (Smeds *et al.*, 2011). Thus, the short-term SNR, as calculated in the ESII-metric, does not seem to capture the information underlying speech intelligibility in these conditions. Cooke (2006) presented an alternative short-term model based on “glimpses” of the target speech in the valleys of the fluctuating interferer. A glimpse was hereby defined as a time-frequency (T-F) unit in an auditory spectrogram-like representation with a local SNR above a given threshold. Within this framework, the number of glimpses and the extent of the glimpse-regions in the auditory spectrogram representation were considered to be related to speech intelligibility. However, since also the glimpse-model is based on the short-term SNR of the stimuli, it can be expected to fail when noisy speech is subjected to nonlinear processing.

Inspired by a study of Dubbelboer and Houtgast (2008), Jørgensen and Dau (2011) proposed that the change in SRT caused by nonlinear processing of noisy speech can be accounted for by a metric based on the signal-to-noise ratio in the envelope power domain (SNR_{env}). The key step in this approach is the determination of the SNR_{env} from the envelope power of noisy speech and noise alone at the output of a bank of modulation-frequency selective filters following peripheral filtering. It was earlier demonstrated that such a modulation-frequency selective filterbank is essential to account for perceptual amplitude modulation detection and masking data (e.g., Dau *et al.*, 1997a,b, 1999; Ewert and Dau, 2000). Ewert and Dau (2000) showed that modulation masking patterns can be accounted for by using the envelope power spectrum model (EPSM), where the envelope of a signal is extracted at the output of an auditory filterbank and further analyzed by a bank of overlapping second-order band-pass modulation filters. Jørgensen and Dau (2011) extended this framework to predict speech intelligibility and denoted it the speech-based EPSM (sEPSM). The sEPSM was shown to account for the intelligibility of different speech materials in stationary noise, as well as for the effect of additional reverberation or spectral subtraction processing. In this framework, a shift of the SRT is associated with a shift of the SNR_{env}. However, the sEPSM presented in Jørgensen and Dau (2011) computes the SNR_{env} from the long-term integration of the stimuli's envelope power. Thus, as with the earlier stationary models, this model can fail in conditions with fluctuating interferers.

In the present study, it is suggested that a frame-based estimation of the SNR_{env} computed using a "short-term" estimation of the envelope power will generalize the sEPSM framework to also account for conditions with fluctuating interferers. To be consistent with the envelope power definition of the earlier model, the short-term envelope power, averaged over several frames of an ideal sinusoidal modulation, and the corresponding long-term envelope power should be identical. This implies that the frame-duration should be equal to or exceed the period of the modulation at the output of the modulation filter centered at the modulation frequency. The model presented here, therefore, includes a temporal "multi-resolution" estimation of the SNR_{env}, with modulation-filter dependent frame duration. It is hypothesized, that the SNR_{env} is increased in the dips of a fluctuating interferer relative to a stationary interferer and

that the multi-resolution estimation of the SNR_{env} will account for the masking release effect.

The proposed multi-resolution sEPSM (mr-sEPSM) is evaluated and compared to earlier models using three categories of interferers and nonlinear processing that represent different challenges for the model: (i) speech mixed with three types of stationary noise with widely different spectral characteristic, evaluating the model's ability to account for differences in spectral masking; (ii) speech mixed with five types of fluctuating noise with very different temporal structure, evaluating the model's predictions of speech masking release; and (iii) two conditions with noisy speech subjected to reverberation and spectral subtraction processing, testing the model with respect to convolution and nonlinear distortions in the processed speech.

4.2 Model description

4.2.1 Overall structure

The processing structure of the mr-sEPSM is illustrated in Fig. 4.1. The key improvements over the model previously proposed by Jørgensen and Dau (2011) are found in the modulation-processing stage as well as in the calculation of the envelope power. The upper limit of the model's modulation-frequency selectivity was increased from 65 Hz to 265 Hz, which is more consistent with earlier studies on perceptual modulation-frequency selectivity (Ewert and Dau, 2000; Ewert *et al.*, 2002). Furthermore, the envelope power at the output of each modulation filter was determined in temporal windows with a modulation-filter dependent duration.

4.2.2 Processing stages of the model

The first stage of the model is a "peripheral" bandpass filterbank consisting of 22 gammatone filters with equivalent rectangular bandwidths (Glasberg and Moore, 1990) and third-octave spacing, covering the frequency range from 63 Hz to 8 kHz. A threshold representing the limit of hearing sensitivity is included such that individual

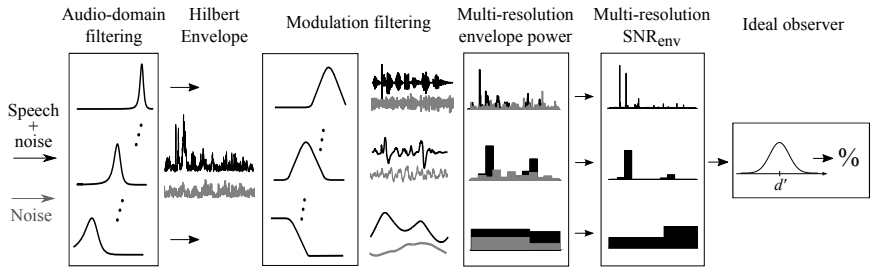


Figure 4.1: Sketch of the multi-resolution sEPSM. Noisy speech (black) and noise alone (gray) are processed separately through a peripheral bandpass filterbank followed by envelope extraction via the Hilbert transform. Each sub-band envelope is processed by a modulation bandpass filterbank. The envelope power is computed at the output of each modulation filter for the noisy speech ($P_{env,S+N}$, black) and the noise alone ($P_{env,N}$, gray) with a segment duration inversely related to the modulation-filter center frequency. The corresponding multi-resolution SNR$_{\text{env}}$ is computed from $P_{env,S+N}$ and $P_{env,N}$ averaged across time and combined across modulation filters and peripheral filters. Finally, the overall SNR$_{\text{env}}$ is converted to a percent-correct prediction assuming an ideal observer as in Jørgensen and Dau (2011).

peripheral filters are considered in the subsequent calculations only if the level of the stimulus at the output is above the diffuse-field hearing threshold in quiet (ISO389-7, 2005). The temporal envelope of each filter output is extracted via the Hilbert-transform and low-pass filtered with a cut-off frequency of 150 Hz, using a first-order Butterworth filter (Ewert and Dau, 2000; Kohlrausch *et al.*, 2000). The resulting envelope is analyzed by a modulation filterbank, consisting of eight second-order bandpass filters with octave spacing and center frequencies covering a range from 2 - 256 Hz. The bandpass filters are in parallel with a third-order low-pass filter with a cutoff frequency of 1 Hz. All modulation filters up to modulation center frequencies below one fourth of the respective peripheral-filter center frequency are used, as earlier proposed in Verhey *et al.* (1999). The temporal output of each modulation filter is segmented using rectangular windows without overlap. The durations of these windows are the inverse of the center frequency of the corresponding modulation filter. For example, the window duration in the 4-Hz modulation channel is 250 ms. In the case of the lowpass modulation filter, the window duration, 1 s, represents the inverse of the cut-off frequency of the filter. For $E(p,t)$ representing the Hilbert envelope as a function of time, t, at the output of the peripheral filter, p, and $e(p,n,t)$ representing the corresponding output of modulation filter n, the envelope power, $P_{env,i}$ of a temporal

segment, i, is a function of the peripheral filter and the modulation filter. $P_{env,i}$ is defined here as the variance of $e(p,n,t)$, calculated over the duration of the segment and normalized with the squared time-average of $E(p,t)$

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

where the bar indicates the average over time. The normalization ensures that the envelope power defined in this way is independent of the overall level of the stimulus. Furthermore, the envelope power of a 100% sinusoidally amplitude modulated pure tone will equal 0 dB (one) at the output of the modulation filter centered at the modulation frequency, for the peripheral filter centered at the carrier frequency. A lower limit of the envelope power at -30 dB (relative to 100% amplitude modulation) is included in each temporal segment. This limit reflects the minimum human sensitivity to amplitude modulation (Ewert and Dau, 2000; Kohlrausch *et al.*, 2000).

For each temporal segment, $\mathrm{SNR}_{\mathrm{env,i}}$ is computed from the envelope power of the noisy speech and the noise alone for that segment as

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

where $P_{env,S+N,i}$ and $P_{env,N,i}$ represent the normalized envelope power of the noisy speech and the noise alone. The middle part of Fig. 4.1 illustrates $e(p,n,t)$ for a given peripheral filter in response to a sentence mixed with noise (black) and in response to the noise alone (gray), and the right part illustrates the corresponding segmental $\mathrm{SNR}_{\mathrm{env,i}}(p,n)$; higher bars indicate larger $\mathrm{SNR}_{\mathrm{env}}$. For each modulation filter, the $\mathrm{SNR}_{\mathrm{env,i}}$-values are averaged across the number of temporal segments, $I(n)$, covering the duration of a given speech token

$$\mathrm{SNR}_{\mathrm{env}}(p,n) = \frac{1}{I(n)} \sum_{i=1}^{I(n)} \mathrm{SNR}_{\mathrm{env},i}(p,n). \tag{4.3}$$

The result is a single $\mathrm{SNR}_{\mathrm{env}}$-value for each of the nine 9 modulation filters, in each of the 22 peripheral filters, representing 9×22 values overall. The values are combined across all modulation filters according to the following equation

$$\mathrm{SNR}_{\mathrm{env}}(p) = \left[\sum_{n=1}^{9} \mathrm{SNR}_{\mathrm{env}}^2(p,n) \right]^{1/2}. \tag{4.4}$$

Similarly, the $\mathrm{SNR}_{\mathrm{env}}$-values are then combined across peripheral filters according to

$$\mathrm{SNR}_{\mathrm{env}} = \left[\sum_{p=1}^{22} \mathrm{SNR}_{\mathrm{env}}^2(p) \right]^{1/2}. \tag{4.5}$$

The combined overall $\mathrm{SNR}_{\mathrm{env}}$ is converted to a sensitivity index, d', of a statistically "ideal observer" using the relation

$$d' = k \cdot (\mathrm{SNR}_{\mathrm{env}})^q, \tag{4.6}$$

where k and q are empirically determined constants. The value of q is set to 0.5 as suggested in Jørgensen and Dau (2011). However, in contrast to Jørgensen and Dau (2011), k is assumed to depend on the speech material used. Finally, d' is converted to the probability of recognizing the speech for the ideal observer using an m-alternative forced choice model (Green and Birdsall, 1964) in combination with an unequal-variance Gaussian model. Conceptually, the ideal observer is assumed to compare the input speech item with m stored alternatives and to 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. In this model, the value of x_S is a random variable with mean d' and variance σ_S. Similarly, the value of $x_{N,max}$ is a random variable with mean μ_N and variance σ_N. The selected item is considered to be correct if the value of x_S is larger than $x_{N,max}$. The corresponding probability of being correct is estimated from the difference distribution of x_S and $x_{N,max}$

$$P_{correct}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \tag{4.7}$$

where Φ denotes the cumulative normal distribution. The values of μ_N and σ_N are determined by the response-set size, m, of the speech material (detailed expressions are given in Jørgensen and Dau, 2011). The value of m can be determined exactly if the material is based on a closed-set paradigm with a fixed number of response alternatives, such as the matrix-test paradigm (e.g., Hagerman and Olofsson, 1982). If the material is based on an open-set paradigm, such as the CLUE-material (Nielsen and Dau, 2009), m is set to the number of words in a normal listener's active vocabulary (Müsch and Buus, 2001). σ_S is related to the inverse of the slope of the ideal observer's psychometric function, assumed here to be mainly determined by the redundancy of a given speech material. For example, speech material consisting of meaningful sentences with high redundancy would have a steep psychometric function and a low value of σ_S, whereas speech material consisting of single-syllable words with low redundancy would have a shallow psychometric function and higher value of σ_S. As the exact relation between speech redundancy and σ_S is difficult to quantify, the value of σ_S was estimated empirically in this study.

4.3 Method

4.3.1 Speech material

The target speech was either Danish sentences from the DANTALE II speech material (Wagener *et al.*, 2003), Danish sentences from the CLUE speech material (Nielsen and Dau, 2009), or Dutch sentences from the speech material developed by Plomp and Mimpen (1979), denoted here as the PM-material. The DANTALE II material consists of grammatically correct but meaningless sentences, based on the paradigm by Hagerman and Olofsson (1982), spoken by a female speaker, while the CLUE and the PM-materials consist of meaningful everyday sentences, similar to the hearing in noise test (Nilsson *et al.*, 1994), spoken by male speakers.

The data presented here partly consist of data collected in this study (Sections 4.3.3 and 4.3.4) and other data collected by Kjems *et al.* (2009), Festen and Plomp (1990), and Jørgensen and Dau (2011).

4.3.2 Experimental conditions

Three categories of conditions were considered: (i) speech mixed with five stationary interferers, (ii) speech mixed with five non-stationary interferers, and (iii) two conditions with linearly and nonlinearly processed noisy speech.

Table 4.1 summarizes the experimental conditions. For each of the three different speech materials, a condition with a matching stationary speech-shaped noise (SSN) was considered. These conditions were used as references for assessing the masking release effect in conditions with fluctuating interferers. Two additional stationary interferers were considered: car-cabin noise (Car) and the sound of bottles on a conveyor belt (Bottle). The fluctuating interferers were as follows: (i) a conversation between two people sitting in a cafe (Cafe), (ii) SSN that was fully amplitude modulated by an 8-Hz sinusoid (SAM), (iii) the speech-like, but non- semantic, International Speech Test Signal (ISTS; Holube *et al.*, 2010), (iv) two-band speech modulated noise (SMN; Festen and Plomp, 1990), and (v) time-reversed speech from a female talker (reversed

Table 4.1: Summary of experimental conditions. "+" indicates that the attribute "steady" or "fluctuating" holds for the noise used in a given condition.

Condition	Material	Steady	Fluctuating
SSN	CLUE	+	
SSN	DANTALE II	+	
SSN	PM	+	
Car	DANTALE II	+	
Bottle	DANTALE II	+	
Cafe	DANTALE II		+
SAM	CLUE		+
ISTS	CLUE		+
SMN	PM		+
RT	PM	+	
SSN + Reverberation	CLUE	+	
SSN + Spectral subtraction	CLUE	+	

talker, RT). Only the data with the SSN_{CLUE}, SAM, and ISTS interferers were collected in the present study.

The conditions with linearly processed noisy speech consisted of sentences from the CLUE- material mixed with SSN and convolved with reverberant impulse responses obtained using the room acoustic simulation software ODEON (Christensen, 2009). The simulated impulse responses corresponded to rooms having reverberation times (T_{30}) of 0.4, 0.7, 1.3, and 2.3 s. The conditions with nonlinearly processed noisy speech consisted of CLUE-sentences in the presence of SSN, processed by an ideal spectral subtraction algorithm using six different values of the spectral subtraction factor α (see Jørgensen and Dau, 2011, for details about the algorithm).

4.3.3 Apparatus and procedure

For the new speech intelligibility data collected here, all stimuli were stored digitally with a sampling frequency of 44.1 kHz and presented diotically using a calibrated pair of Sennheiser HD 580 headphones and a high-quality sound card in a double-walled, sound-attenuating booth. The digital files for the SSN and SAM- noise had duration

of 22 s while the ISTS was 52 s long. The speech level was fixed to a long-term, root-mean-square (RMS), level of 60 dB sound pressure level (SPL). Each sentence was presented once and the added noise was a random token from the original noise files such that it started 1 s before the sentence and stopped 600 ms after it. The noise was ramped on and off using 400-ms squared-cosine ramps and the RMS level of the noise sample was set such that the target SNR was achieved for each individual combination of speech and noise tokens.

A simple 1-up, 1-down adaptive procedure was used, yielding an estimate of the 50% point on the psychometric function. The SNR was adjusted in steps of 2 dB. If every word of a sentence was correctly repeated, the SNR was decreased, otherwise the SNR was increased. The obtained threshold therefore represented sentence intelligibility rather than word intelligibility. Ten sentences were used to measure one SRT, which was calculated as the average SNR after the last eight responses. Listeners were instructed to repeat as much of the presented sentence as possible, and were allowed to guess; no feedback was provided. For each listener, one SRT was measured with each interferer and one list was used for training before the measurement for each interferer.

4.3.4 Listeners

For the new data presented here, five male normal-hearing listeners between 24 and 33 years of age participated in the experiment. They had pure-tone thresholds of 20 dB hearing level or better in the frequency range from 0.25 to 8 kHz. All listeners had experience with psychoacoustic measurements.

4.3.5 Model setup and parameters

To generate the model predictions, 100 sentences from each of the speech materials were used. Each sentence was mixed with a noise token (randomly selected from the full-length noise files) over a discrete range of SNRs. For a given SNR-value, the final percent correct prediction was computed as the average predicted score across all sentences of a given speech material. The prediction at each SNR was then connected

Table 4.2: Values of model parameters k, q, σ_S, and m, for the three different speech materials: CLUE, DANTALE II, and PM.

Speech material	k	q	σ_S	m
CLUE	0.61	0.5	0.6	8000
DANTALE II	0.41	0.5	0.9	50
Plomp and Mimpen (1979)	0.36	0.5	0.6	8000

by straight lines, resulting in a continuous psychometric function, from which the SRT was estimated as the SNR corresponding to 50% correct. For a given run of the model, the inputs were a sentence mixed with a noise token and the noise alone. For the conditions with spectral subtraction, the noise alone was estimated using the method described by Hagerman and Olofsson (2004). For the conditions with reverberation, the noise alone was convolved with the impulse response corresponding to a given reverberation time.

The model was calibrated by adjusting the parameters of the ideal-observer stage such that the predicted SRT matched the data in the SSN-condition for a given speech material (represented by the first three rows of Table 4.1). These parameters were kept fixed for all other conditions for that material. The parameter k (Eq. 4.6) was adjusted to obtain a match of the measured and predicted SRT, while the values for m, q, and σ_S, were taken from Jørgensen and Dau (2011). The values for the PM-material were assumed to be the same as for the CLUE- sentences since both materials consist of meaningful sentences of similar structure and size. The values of all parameters are given in Table 4.2.

4.4 Results

Figure 4.2 shows the results for all considered conditions. Open symbols indicate experimental data and closed symbols show model predictions. The data from Kjems *et al.* (2009) are indicated by diamonds, the data from Festen and Plomp (1990) are shown as right-pointing triangles, and the data from Jørgensen and Dau (2011) are represented as downward pointing triangles. The data collected in the present study

are indicated as squares. Error-bars (if reported) indicate plus/minus one standard deviation across the listeners' SRTs. Corresponding predictions obtained with the mr-sEPSM are represented as black bullets and predictions from the long-term sEPSM are shown by the gray bullets. The root-mean-square error (RMSE) between the data and the mr-sEPSM predictions are indicated in each panel, representing the conditions with stationary interferers (top left), fluctuating interferers (top right), reverberation (bottom left), and spectral subtraction (bottom right). A two-way analysis of variance, performed on the new data (open squares) across five subjects and three conditions, revealed no significant effect of subject ($F_{4,14} = 1.25$, $p = 0.37$), but a significant effect of condition ($F_{2,14} = 185$, $p = 0.37$, $p < 0.0001$).

4.4.1 Stationary interferers

The measured SRTs for the three SSN-conditions (left part of top-left panel) were obtained at SNRs of -3.5, -4.3, and -7.3 dB for the CLUE, PM, and DANTALE II-materials, respectively. The multi-resolution and the long-term sEPSM predictions matched the data since the models were calibrated in these conditions. The SSN-conditions were therefore not included in the RMSE-calculation. The SRTs for the Car and Bottle noises were obtained at SNRs of -20 and -12 dB. The very low SRT obtained with the Car noise is mainly a consequence of the dominant low-frequency content of this noise, such that the masking effect on speech is weaker than in the case of the SSN and the Bottle noise. The mr-sEPSM accounted very well for the SRTs obtained with the Car and Bottle noises, as indicated by the RMSE of 0.48 dB, while the original sEPSM slightly underestimated the SRT for the Bottle noise.

4.4.2 Fluctuating interferers

The results for the conditions with fluctuating interferers are shown in the top-right panel of Fig. 4.2. SRTs were obtained at SNRs of -8.8, -9.5, -18, -7.9, and -10 dB for the Cafe, SAM, ISTS, SMN, and RT interferers, respectively. A multiple comparison analysis using Tukey's honestly significant difference criterion showed significant mutual differences between the SRTs obtained in the $\mathrm{SSN_{CLUE}}$, SAM, and

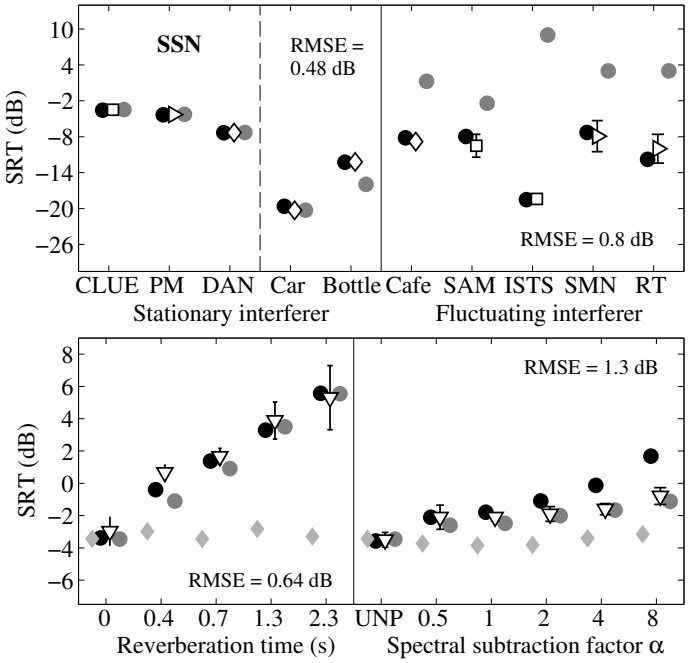


Figure 4.2: Results for the conditions with stationary interferers (top-left panel), fluctuating interferers (top-right panel), reverberant noisy speech (bottom-left panel), and spectral subtraction (bottom-right panel). Measured data are indicated as open symbols. Predictions from the mr-sEPSM and the original "long-term" sEPSM are shown as filled black and gray circles, respectively (all panels). Predictions from the ESII are indicated as filled gray diamonds in the bottom panels. The root-mean-square error (RMSE) between the data and the mr-sEPSM-predictions is indicated in each panel. The perfect match of the predictions to the data in the SSN-conditions (top-left panel) results from the calibration of the model to these conditions, which were therefore not included in the RMSE calculation.

ISTS conditions, demonstrating a clear masking release effect with the fluctuating interferers. Similarly, the SRTs for the Cafe, SMN, and RT-conditions from Kjems *et al.* (2009) and Festen and Plomp (1990) were reported to be significantly lower than the SRTs in the respective SSN-conditions. The mr-sEPSM-predictions are in good agreement with the data for all fluctuating interferers, thus accounting for the masking release effect. However, the match was not as good as that found with the stationary interferers, as reflected in the slightly larger RMSE of 0.84 dB. The ESII has been shown to account for very similar conditions (e.g., Rhebergen and Versfeld, 2005;

Rhebergen *et al.*, 2006), not indicated explicitly in the Fig. 4.2. In contrast, the original sEPSM (gray bullets) clearly failed to account for the fluctuating interferers, predicting higher SRTs for all fluctuating conditions than for the SSN-conditions.

4.4.3 Processed noisy speech

Reverberation

The bottom-left panel of Fig. 4.2 shows the results for the conditions with SSN and reverberation. The measured SRTs from Jørgensen and Dau (2011) increased with increasing reverberation time and both models, the mr-sEPSM (black bullets) with an RMSE of 0.64 dB and the original sEPSM (gray bullets) with an RMSE of 0.71 dB, could account for this effect. Predictions obtained with the ESII, including a forward masking function (Rhebergen *et al.*, 2006), are represented by the filled gray diamonds, where the SRT corresponded to an SII-value of 0.32. The ESII predicted an SRT that was largely independent of the reverberation time, in contrast to the data. When combined with the STI, the ESII has been shown to account for data from similar conditions (George *et al.*, 2008), as further discussed in Section 4.6.3.

Spectral subtraction

The bottom-right panel of Fig. 4.2 shows measured SRTs from Jørgensen and Dau (2011) (open triangles) for six conditions of the spectral subtraction factor α. "UNP" denotes the reference conditions with no spectral subtraction. Jørgensen and Dau (2011) found a significant increase of the SRT with increasing α, demonstrating lower intelligibility with spectral subtraction than without this processing. The mr-sEPSM (black bullets) could predict the trend in the data, although it slightly overestimated the SRTs for $\alpha = 4$ and 8, resulting in an RMSE of 1.3 dB, which was higher than that (0.48 dB) for the original sEPSM (gray bullets). The ESII (filled gray diamonds) predicted an SRT that was largely independent of spectral subtraction, in contrast to the data.

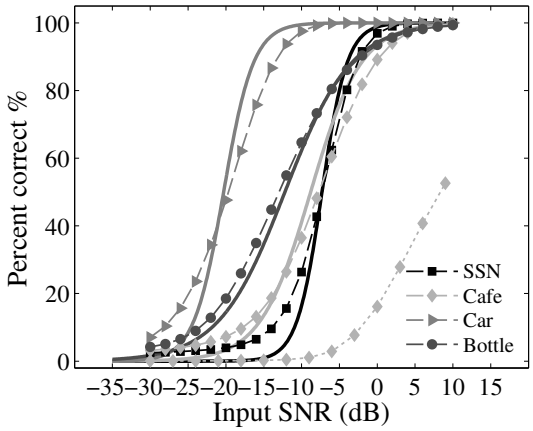


Figure 4.3: Psychometric functions derived from the data obtained by Kjems *et al.* (2009; solid lines) and corresponding predicted functions obtained with the mr-sEPSM (symbols connected by dashed lines), for the SSN, Bottle, Car, and Cafe interferers. Predictions from the "long-term" sEPSM version by Jørgensen and Dau (2011) for the Cafe-interferer are denoted by gray diamonds connected with the dotted lines.

4.4.4 Prediction of psychometric functions

Figure 4.3 shows psychometric functions (solid black and gray curves) derived from the two-parameter model described in Wagener *et al.* (2003). The parameters were the SRT and the slope data obtained by Kjems *et al.* (2009) for the three stationary interferers SSN_{DAN}, Bottle, and Car as well as the fluctuating Cafe-interferer. The mr-sEPSM (symbols connected by dashed lines) accounted well for the shape of the functions, considering that a fixed set of parameters was used for all conditions. In contrast, the long-term sEPSM (diamonds connected by the dotted line) clearly failed with the fluctuating Cafe-interferer, but it produced similar results as the mr-sEPSM in the stationary conditions (not shown here).

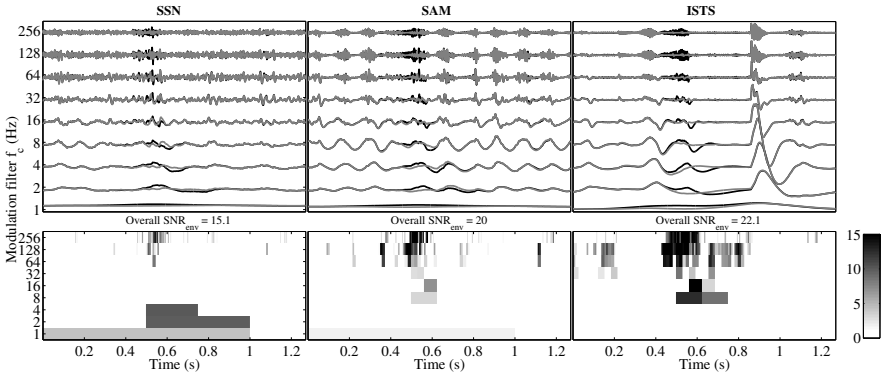


Figure 4.4: (Top) Output from the nine modulation filters (indicated at the left ordinate) as a function of time in response to the interferer alone (gray) and the interferer plus speech (black) for three different interferers: SSN (left), SAM (middle), and ISTS (right). (Bottom) SNR_{env}-values (in dB) per time segment, calculated from the modulation-filter outputs shown in the corresponding upper panels. The overall time-averaged SNR_{env} is indicated directly above each of the bottom panels.

4.5 Model analysis

4.5.1 Prediction of speech masking release

The following analysis shows how the effect of speech masking release is reflected in the internal representation of the stimuli in the framework of the mr-sEPSM. The three panels in the upper row of Fig. 4.4 show the output of the nine modulation filters ranging from 1 to 256 Hz (indicated on the ordinate) in the 1-kHz peripheral filter in response to a 1.2-s noise-alone token (gray traces) and a sentence with the same noise token added (black traces) at an SNR of -5 dB. The three panels show the responses to the SSN (upper left), SAM (upper middle), and ISTS (upper right) interferers, respectively. The most dominant changes due to the presence of the speech occurred at about 0.5 s in all modulation filters. In case of the SAM and ISTS interferers, the changes in the modulation filter outputs were more prominent than in the case of the SSN interferer, particularly in the higher-frequency modulation filters.

The bottom row of Fig. 4.4 shows the corresponding SNR_{env}-values (in dB) per time segment, calculated from the modulation-filter outputs in the upper panels (Eq. 4.2).

Dark elements indicate large SNR_{env}-values, as shown by the shaded bar on the right. As the duration of a segment is inversely related to the modulation-filter center frequency, the segments become shorter with increasing modulation-filter center frequency. The time-averaged SNR_{env} per modulation filter (Eq. 4.3) combined across all modulation filters and peripheral filters (Eq. 4.5), is indicated for each interferer above the respective panel. In the case of the SSN (left panel), the segments at about 0.5 s at very low modulation frequencies (1 - 4 Hz) and at high modulation frequencies ($>$ 64 Hz) provided the main contributions to the overall SNR_{env}. For the SAM interferer (middle panel), the SNR_{env} in the high-frequency modulation filters was considerably increased compared to the SSN. In addition, segments with high SNR_{env} occurred in the dips of the masker (at about 0.3 and 1.1 s). The same trend can be observed for the ISTS interferer (right panel), resulting in an even larger overall SNR_{env}. This larger overall SNR_{env} for the SAM and ISTS interferers, compared to the SSN, led to a larger predicted intelligibility. Thus, in the framework of the mr-sEPSM, the masking release effect is caused by the larger short-term SNR_{env} occurring during the dips of the fluctuating interferers.

To illustrate the relative contribution of individual peripheral and modulation filters to speech intelligibility in the mr-sEPSM, Fig. 4.5 shows the distribution of the time-averaged SNR_{env} across peripheral and modulation filters for the three interferers SSN (left), SAM (middle), and ISTS (right). The SNR_{env}-values represent averages across 150 sentences of the CLUE material at an SNR of -5 dB. Dark elements indicate large SNR_{env}-values (see color bar on the right). The white regions at the bottom and lower right corners of the panels were not considered in the model because they represent modulation frequencies that exceeded a quarter of the center frequency of the corresponding peripheral filters ($f_{cm} < 0.25 f_c$), or because the level of the stimulus in the peripheral channel was below the sensitivity threshold of the model in case of the low-frequency peripheral filters and ISTS (see Section 4.2.2). For the SSN-interferer, the main contribution to the overall SNR_{env} stemmed from the modulation filters tuned to the 1 to 8-Hz region from the peripheral filters in the range from 200 and 3000 Hz. In contrast, for the SAM- noise, the largest contribution originated from modulation filters tuned to frequencies above 16 Hz and peripheral filters above 1000 Hz. It can

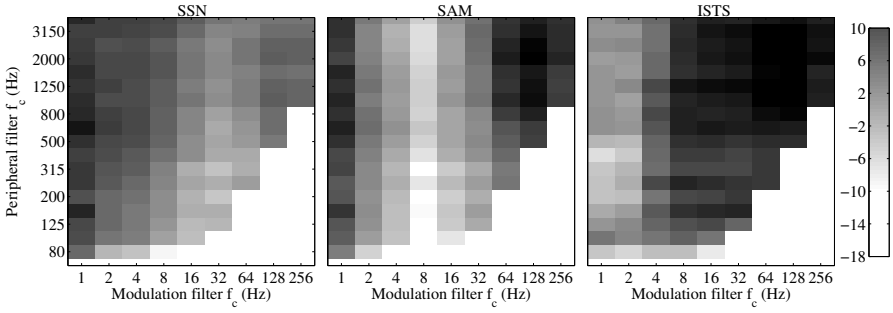


Figure 4.5: Time-averaged SNR_{env} (Eq. 4.3) in dB, averaged across 150 sentences of the CLUE material at an SNR of -5 dB, obtained in the various peripheral filters and modulation filters for the SSN (left), SAM (middle), and ISTS (right) interferers. Dark elements are large SNR_{env}-values as indicated by the scale on the right.

also be observed that the modulation filters centered near the modulation rate of the SAM- noise (8 Hz) had the lowest SNR_{env}- values.

A similar trend is visible in the case of the ISTS interferer. Here, the prominent dips in the interferer were of longer duration and the corresponding region with very low SNR_{env} values ranged from 1 - 4 Hz. The largest SNR_{env}-values for the ISTS were observed in the modulation filters above 4 Hz and in peripheral filters above 800 Hz. Thus, in general terms, the modulation filters that contributed most to the overall SNR_{env} shifted from the modulation-frequency region below 16 Hz in the case of the stationary interferer to the region above 16 Hz in the case of the fluctuating interferers. This resulted from the target-speech envelope fluctuations that were accessible in the valleys of the fluctuating interferers. The modulation- frequency range corresponding to the dominant fluctuation rate of the interferer (such as the 8-Hz rate in the case of the SAM interferer) provided very low SNR_{env}-values, which hardly contributed to the overall SNR_{env}. This observation indicates modulation masking caused by the respective interferer.

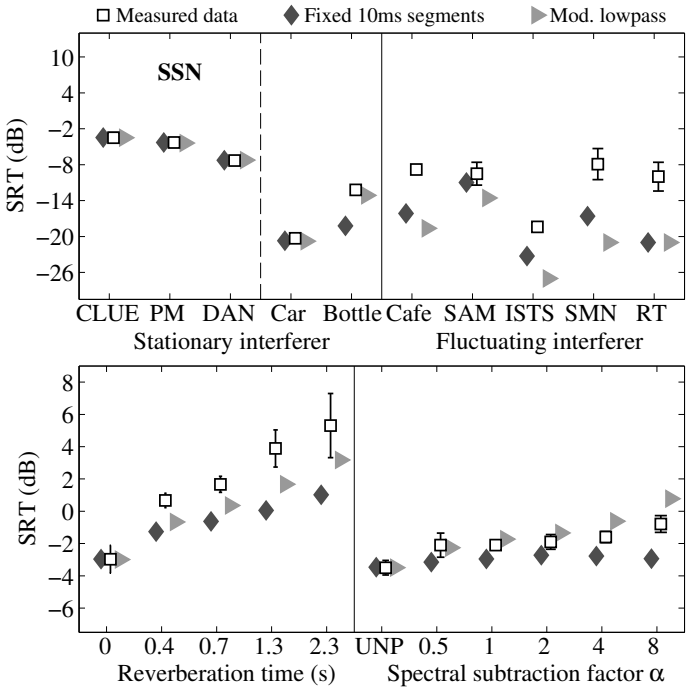


Figure 4.6: Replot of the data (open squares) from Fig. 4.2 together with predictions obtained with two modified versions of the model, including one with a fixed-resolution segmentation of 10 ms (dark gray diamonds) and a version where the modulation filterbank was replaced by a 150-Hz modulation lowpass filter (gray triangles).

4.5.2 The role of modulation filters and multi-resolution SNR_{env}-processing

In comparison to earlier short-term analyses in speech intelligibility prediction (e.g., ESII, Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006), the current approach introduced two major changes: (i) processing by a modulation filterbank, and (ii) multi-resolution temporal segmentation of the filterbank output. To illustrate the relative contribution of the two processes to the predictive power of the mr-sEPSM, Fig. 4.6 shows predictions from two modified versions of the mr-sEPSM. In addition, the data (open squares) from Fig. 4.2 were replotted in the figure for direct comparison.

The first modified model was considered to evaluate the effect of the modulation filterbank processing while maintaining the multi-resolution segmental SNR_{env}-analysis of the mr-sEPSM. In this modification, the filterbank was replaced by a single first-order 150-Hz modulation lowpass filter. The lowpass-filtered envelope was thus effectively analyzed by a parallel set of temporal windows with durations that ranged from 1 to 1/256 s. The corresponding predictions from this "modulation-lowpass" version (filled gray triangles) accounted for the SRTs obtained in all stationary interferers (RMSE = 0.75 dB) and the effect of spectral subtraction was described well (RMSE = 0.88 dB). Moreover, lower SRTs were predicted for the fluctuating interferers than for the SSN-interferers, demonstrating that the modulation filterbank is not essential for predicting masking release.

This suggests that the main feature allowing the mr-sEPSM to predict masking release is the temporal segmentation. However, the masking release effect was highly overestimated (RMSE = 9.8 dB), and the effect of reverberation could only be accounted for qualitatively (RMSE = 2.3 dB), demonstrating that a modulation lowpass filter is not sufficient to account for the data in these conditions.

The second model version maintained the modulation filterbank, but replaced the multi-resolution segmentation by fixed 10-ms segments in all modulation filters. Thus, this simulation evaluated the effect of the multi-resolution SNR_{env}- analysis. This "fixed-resolution"-version (dark gray diamonds in Fig. 4.6) provided predictions with a larger deviation from the data than the unmodified model in the conditions with the stationary interferers (RMSE = 4.3 dB), and the fixed-resolution version also strongly overestimated the amount of masking release for the fluctuating interferers (RMSE = 7.4 dB). Moreover, the effect of reverberation was accounted for only qualitatively (RMSE = 3.7) while little effect of spectral subtraction was predicted (RMSE = 1.3). This demonstrates that even though a fixed 10-ms segmentation may facilitate the prediction of masking release, longer window durations, such as those assumed at low modulation frequencies in the unmodified mr-sEPSM or in the original sEPSM (Jørgensen and Dau, 2011), are required to account for reverberation and spectral subtraction with the present model.

4.6 Discussion

4.6.1 Importance of the multi-resolution analysis and the modulation filterbank

The proposed mr-sEPSM was shown to account for the SRTs obtained in conditions with stationary and fluctuating interferers, as well as in conditions with reverberation and noisy speech processed by spectral subtraction. This demonstrates an improvement over the original sEPSM (Jørgensen and Dau, 2011), which fails in conditions with fluctuating interferers. The model analysis revealed that the key component that allows the mr-sEPSM to account for the fluctuating interferers is the multi-resolution estimation of the SNR_{env}, with the short segments in the modulation filters tuned to modulation frequencies above about 16 Hz contributing most to speech masking release. Moreover, while the mr-sEPSM predicted the data quite well, a modified version where the duration of the segments were fixed to 10 ms failed to account for all aspects of the data (dark gray diamonds in Fig. 4.6). Thus, the relatively long segments (0.5 and 1 s) used at low modulation frequencies by the mr-sEPSM were critical to provide accurate predictions in the conditions with spectral subtraction and reverberation. Simulations with a modified model assuming a modulation-lowpass filtering process highly overestimated the masking release effect and underestimated the effect of reverberation (gray triangles in Fig. 4.6), while it accounted for the data in the conditions with spectral subtraction. In the framework of the present model, this demonstrated that the modulation-frequency selective process was not essential to account for the effects of spectral subtraction, but contributed to the improved accuracy in the conditions with fluctuating interferers and reverberation.

Neither the use of a fixed short-term (10-ms) segmentation nor a single "long-term" analysis, as in the case of the original sEPSM (Jørgensen and Dau, 2011), could account for both a masking release, in conditions with fluctuating interferers, and the effects of reverberation and spectral subtraction. These results suggest that the auditory system is able to access speech information carried in the envelope on different time scales simultaneously, and that the temporal segments with large SNR_{env} contribute the most to overall speech intelligibility. The inverse relation of the modulation bandpass center

frequency and the duration of the segments can be interpreted as a critical sampling of the information conveyed in the individual (band- limited) modulation channels.

4.6.2 Importance of modulation-frequency range for masking release

The model analysis demonstrated that different peripheral and modulation filters contributed maximally to speech intelligibility depending on the type of the interferer. The major contribution to masking release in the framework of the mr-sEPSM was provided at modulation frequencies above 16 Hz. "Local information" in the dips of a fluctuating interferer (increased local SNR_{env}) was mainly observed in the model for time segments with shorter durations than the local dips in the interferer. Conversely, in modulation filters tuned to the average fluctuation rate of the interferer, modulation masking was observed, reflected in minima of the local SNR_{env} values. It is possible that this finding of the model analysis corresponds to an increased perceptual importance of high modulation frequencies of the target speech signal in conditions where a masking release is observed. If so, these high-frequency modulations may convey fundamental frequency (f0) information that could help segregating target and interferer.

4.6.3 Relation to other short-term intelligibility models

The ESII, as described in Rhebergen and Versfeld (2005), considers the conventional energy SNR in short-term frames as the underlying intelligibility metric for predicting speech intelligibility with fluctuating interferers. An updated version that includes properties of forward masking was shown to account for various conditions with fluctuating interferers (Rhebergen *et al.*, 2006), but was largely insensitive to the effects of reverberation and spectral subtraction as considered here (cf. Fig. 4.2). George *et al.* (2008) suggested combining the ESII with the STI in order to account for the effects in conditions with fluctuating interferers in reverberant environments. However, while this approach was proven successful, it applies two different decision metrics in parallel, the conventional energy SNR and the MTF, depending on the experimental condition,

since the individual metrics can only account for part of the speech intelligibility data. Moreover, the ESII model uses a stationary speech-shaped noise as a probe for the actual speech. Thus, the model neglects the effects of the level fluctuations of the target speech itself, which may lead to difficulties when applying the model to conditions with speech interferers.

In contrast to the ESII, the mr-sEPSM framework considers a single decision metric, the $\mathrm{SNR}_{\mathrm{env}}$, estimated from the normalized variance of the envelope fluctuations of noisy speech and noise alone. For some conditions, such as with an SSN interferer, the $\mathrm{SNR}_{\mathrm{env}}$ is related to the conventional energy SNR. Increasing the SNR in a condition with SSN also increases $\mathrm{SNR}_{\mathrm{env}}$. However, the two metrics can also differ considerably, such as in conditions where the modulation depth of the noise has been reduced, but the short-term energy has been kept fixed.

Similar to the ESII, the model by Cooke (2006) defines the useful "glimpse area" based on the short-term energy SNR in a spectrogram-like representation. This metric is therefore most likely not able to account for the results obtained with the processed noisy speech. However, the glimpse model includes a speech recognition backend, which may account for the distortions in a different way, via the mismatch between the learned dictionary and the distorted input. Thus, the inclusion of automatic speech recognition can increase the range of conditions in which intelligibility can be successfully predicted. However, this comes at the cost of greater complexity compared to a purely predictive model.

4.6.4 Limitations of the modeling approach

The current mr-sEPSM and the earlier sEPSM consider the $\mathrm{SNR}_{\mathrm{env}}$ calculated from an analysis of the temporal envelope in a number of parallel peripheral filters, and combine the time-averaged values from the individual filters assuming statistically independent observations. This means that the models are insensitive to the relative timing of segments with high or low $\mathrm{SNR}_{\mathrm{env}}$ in the individual filters, and would therefore be insensitive to phase shifts within a given filter. However, these phase shifts have been shown to severely affect speech intelligibility (Elhilali *et al.*, 2003). It is possible, that

the purely temporal SNR_{env} analysis should be combined with an across peripheral-frequency process, such as an across-channel correlation network, which would be sensitive to across peripheral-channel phase differences.

4.6.5 Perspectives

The presented framework may instruct future work that investigates speech perception in both normal and hearing-impaired listeners. For example, the mr-sEPSM framework may be combined with a model of hearing impairment to facilitate a systematic investigation of how the SNR_{env} representation is affected by different types of hearing loss and to what extent this can account for the reduced speech masking release typically observed for hearing-impaired listeners (e.g., Festen and Plomp, 1990; Strelchyk and Dau, 2009). Moreover, the present approach may be considered as a framework for systematically analyzing the role of different peripheral- and modulation-filters for the speech intelligibility in conditions of speech-on-speech masking, and may be used to separate out the effects of energetic, modulation and informational masking.

4.7 Summary and Conclusions

A multi-resolution short-term SNR_{env} estimation was introduced in the sEPSM (Jørgensen and Dau, 2011). The duration of the segments used for short-term SNR_{env} calculation was assumed to be inversely related to the center frequencies of the modulation bandpass filters in the model. The resulting mr-sEPSM was demonstrated to extend the applicability of the modeling approach to conditions with fluctuating interferers, where the original sEPSM failed as a consequence of the "long-term" SNR_{env} calculation. The mr-sEPSM was shown to successfully account for conditions of stationary and fluctuating interferers as well as the effects of reverberation and spectral subtraction. In addition, the model results suggested that high- frequency modulations ($>$ 16 Hz) contribute to speech intelligibility in the conditions with fluctuating interferers whereas low-frequency modulations ($<$ 16 Hz) are typically dominant in conditions with stationary and/or reverberant interferers.

Acknowledgments

We thank Ewen MacDonald for suggestions for improvement of an earlier version of this paper. This work was supported by the Danish Research Foundation, the Danish hearing aid companies Oticon, Widex, and GN ReSound, and the Deutsche Forschungsgemeinschaft (DFG; FOR 1732 “Individualisierte Hörakustik”, TPE).

5

The role of high-frequency envelope fluctuations for speech masking release[§]

The multi-resolution version of the sEPSM presented in Chapter 4 was shown to successfully predict speech intelligibility in conditions with stationary and fluctuating interferers, reverberation, and spectral subtraction. The key element in the model was the multi-resolution estimation of the signal-to-noise ratio in the envelope domain ($\mathrm{SNR}_{\mathrm{env}}$) at the output of a modulation filterbank. The simulations suggested that mainly modulation filters centered in the range from 1 - 8 Hz contribute to speech intelligibility in the case of stationary maskers whereas modulation filters tuned to frequencies above 16 Hz might be important in the case of fluctuating maskers. In the present study, the role of high-frequency envelope fluctuations for speech masking release was further investigated in conditions of speech-on-speech masking. Simulations were compared to measured data from normal-hearing listeners (Festen and Plomp, 1990; Christensen *et al.*, 2013). The results support the hypothesis that high-frequency envelope fluctuations (> 30 Hz) are essential for speech intelligibility in conditions with speech interferers. While the sEPSM reflects effects of energetic and modulation masking in speech intelligibility, the remaining unexplored effect in some conditions may be attributed to, and defined as, "informational masking".

[§] This chapter is based on Jørgensen and Dau (2013).

5.1 Introduction

Many aspects of speech intelligibility in noise may be accounted for by the availability of the slowly varying fluctuations in the temporal envelope of the speech, also called speech modulations (e.g., Houtgast and Steeneken, 1973; Drullman, 1995). In a series of studies, Houtgast and Steeneken demonstrated that the concept of the modulation transfer function (MTF), which measures the integrity of speech envelope fluctuations in noise, can account for many conditions with varying levels of background noise and reverberation. However, the MTF-concept does not capture effects of nonlinear distortions on the speech envelope, and cannot account for noise reduction processing, such as spectral subtraction (e.g., Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007). Dubbelboer and Houtgast (2008) suggested that the envelope fluctuations of the noise itself, which are not included in the MTF-concept, should also be considered since they may also be affected by the nonlinear processing. They proposed that the strength of the speech envelope fluctuations relative to that of the noise envelope fluctuations could account for some aspects of nonlinear processing of noisy speech. However, this finding was not associated with a functional model for quantitative prediction of speech intelligibility. The concept of considering the relative strength of the target signal and noise envelope fluctuations was earlier used by Dau *et al.* (1999) and Ewert and Dau (2000) in their envelope power spectrum model (EPSM) of modulation masking, which was demonstrated to account for modulation detection and masking data. The EPSM quantified the envelope fluctuations by measuring the envelope power (i.e., the variance of the envelope fluctuations relative to the mean envelope amplitude). The EPSM was extended to speech stimuli by Jørgensen and Dau (2011), arguing that the signal-to-noise envelope power ratio (SNR_{env}) at the output of modulation-frequency selective filtering could be used to predict intelligibility of processed noisy speech. It was demonstrated that the speech-based envelope power spectrum model (sEPSM) and the SNR_{env} were sufficient to account for conditions with stationary noise and reverberation as well as with noisy speech processed by spectral subtraction. However, the sEPSM as presented in Jørgensen and Dau (2011) fails in conditions with fluctuating interferers due to the assumed long-term computation of the envelope power. It cannot account for speech “masking release”, referring to the

higher intelligibility observed in conditions with modulated noise or competing talkers compared to a condition with a stationary noise. In order to overcome this limitation, a multi-resolution version of the sEPSM was proposed (mr-sEPSM; Jørgensen *et al.*, 2013), which estimates the $\mathrm{SNR}_{\mathrm{env}}$ in short temporal segments with a modulation filter dependent duration. The mr-sEPSM was demonstrated to account for speech masking release with fluctuating interferers as well as for conditions with reverberation and spectral subtraction. Moreover, the simulations indicated that the modulation filters centered above 30 Hz were important in the case of fluctuating interferers, suggesting that high-rate envelope fluctuations might be important for speech masking release. In the present study, the role of high-rate envelope fluctuations for speech masking release was further investigated. Model predictions were compared to data from the literature in conditions with stationary noise, amplitude modulated noise and competing talkers as interfering signals. The competing-talker conditions included speech from the same talker (ST) as the target speech and speech from a different talker (DT). Additional conditions included vocoded stimuli where the high-rate envelope fluctuations of the target speech and interferer were attenuated.

5.2 Methods

Predictions from the multi-resolution sEPSM (Jørgensen *et al.*, 2013) were compared to data collected by Festen and Plomp (1990) and Christensen *et al.* (2013). The target speech were either Dutch or Danish meaningful five-to-seven word sentences, spoken by male native speakers (Plomp and Mimpen, 1979; Nielsen and Dau, 2009). Two categories of conditions were investigated. The first category included the conditions considered by Festen and Plomp (1990) where unprocessed sentences were mixed with five different interferers: a steady speech-shaped-noise (SSN); a speech- modulated noise (SMN) consisting of the SSN that was amplitude modulated with the broadband envelope of the target speech; forward and time-reversed speech from a different talker (DT and DT-R); and time-reversed speech from the same talker (ST-R) as the target. The other category included conditions considered by Christensen *et al.* (2013) where processed speech was mixed with either unprocessed SSN or a processed version of the

international speech test signal (ISTS; Holube *et al.*, 2010). The long-term spectrum of the SSN was shaped to the long-term spectrum of the ISTS. The processing consisted of a vocoder with amplitude and frequency modulated pure-tone carriers. The original speech signal was decomposed into 16 spectral bands using gammatone-filters. In each band, the Hilbert envelope and the instantaneous frequency were extracted. A temporally smoothed version of the instantaneous frequency was used to drive a sine generator and the output was amplitude modulated with a low-pass filtered version of the envelope signal. The root-mean-square (RMS) level of each band was equalized to the unprocessed RMS-level of the corresponding band and all bands were recombined by summing the signals with the proper delay to account for time-delays caused by the filters. The resulting processed signal was scaled to have the same long-term RMS-level as the unprocessed signal. More details on the vocoder can be found in Christensen *et al.* (2013). For each of the interferers (SSN or ISTS), two conditions were considered, with envelope lowpass filter cut-off frequencies of 30 Hz or 300 Hz. In addition, the mixed stimulus of speech and noise was either presented broadband (BB) or high-pass (HP) filtered with a cut-off frequency of 500 Hz, resulting in four conditions for each interferer: BB30, BB300, HP30, and HP300. An additional off-frequency masker was added in the HP-conditions, consisting of the SSN with an RMS-level of 12 dB below the speech presentation level.

For the sEPSM predictions, the model parameter from Jørgensen *et al.* (2013) were applied here, except for the conditions with vocoded speech where the k-parameter of the ideal observer was adjusted to provide a good agreement between predictions and data in the SSN-BB300 condition. These parameters were then kept constant for all other experimental conditions. Identical stimuli were used for the simulations and the measurements.

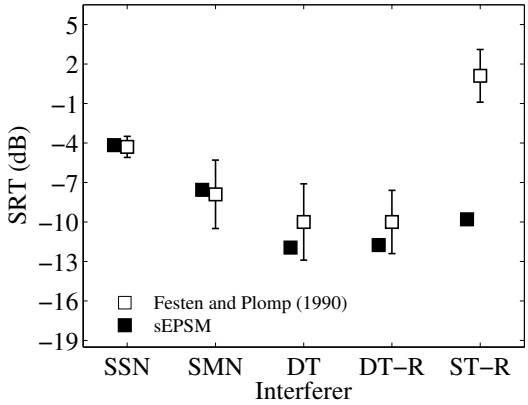


Figure 5.1: SRTs obtained by Festen and Plomp (1990) (open squares) and predicted (filled squares) by the model as function of the interferer type - SSN: stationary speech-shaped noise; SMN: speech-modulated noise; DT: speech from a different talker; DT-R: time-reversed speech from a different talker; ST-R: time-reversed speech from the same talker as the target.

5.3 Results

5.3.1 Fluctuating interferers

Figure 5.1 shows the speech reception thresholds (SRT), corresponding to 50% correct, obtained by Festen and Plomp (1990) for 20 normal-hearing listeners (open squares), for each of the five interferers indicated on the abscissa. The vertical bars denote plus/minus one standard deviation. The SRT in the SSN-condition was obtained at an SNR of about −4 dB, while lower SRTs were obtained for the SMN (−8 dB SNR), and DT interferers (−10 dB SNR), reflecting a clear effect of masking release. There was no effect of the playback mode (forward versus time-reversed) on the SRT for the DT-interferer. In contrast, the SRT for the ST-R interferer was obtained at an SNR of about 1 dB, i.e., 5 dB higher than for the SSN-condition, demonstrating a *negative* masking release. The model predictions (filled squares) were in good agreement with the data for the SSN, SMN, DT, and DT-R interferers, but clearly deviated from the data in the condition of the ST-R interferer, where a much lower SRT was predicted than obtained in the data. Thus, the model accounted for the observed speech masking

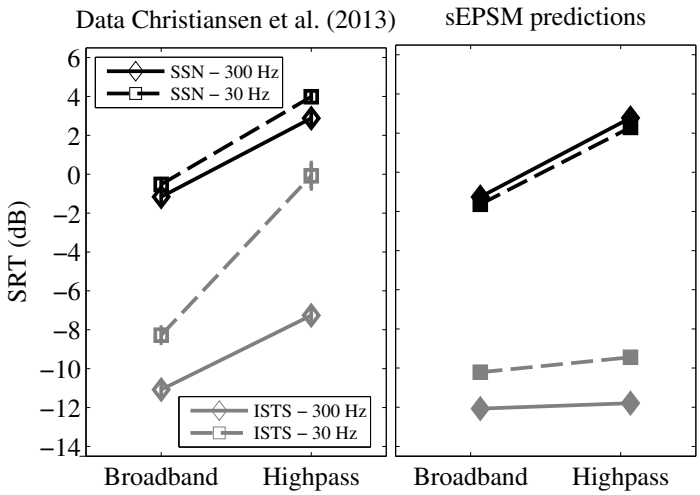


Figure 5.2: Data obtained by Christensen *et al.* (2013) (left panel) and model predictions (right panel) in the conditions with vocoded speech.

release, except for the condition where the competing talker was a time-reversed version of the target speech.

5.3.2 Vocoded stimuli

The left panel of Fig. 5.2 shows the measured SRTs obtained by Christensen *et al.* (2013) for 5 normal-hearing listeners in the conditions with vocoded speech and interferers. For the SSN-interferer (black symbols and lines), the SRT in the BB300-condition (left open diamond) was obtained at an SNR of about -1 dB, while it was slightly higher in the BB30-condition (left open square), demonstrating only a small effect of envelope low-pass filtering. The SRTs were generally obtained at higher SNRs for the HP300 condition (right open diamond) and HP30 condition (right open square) compared to the BB-conditions. For the ISTS interferer (gray symbols and lines), the SRT in the BB300-condition (left gray diamond) was obtained at an SNR of about -11 dB, versus about -8 dB SNR in the BB30 condition (right gray square). This demonstrates a clear effect of envelope low-pass filtering for the ISTS interferer.

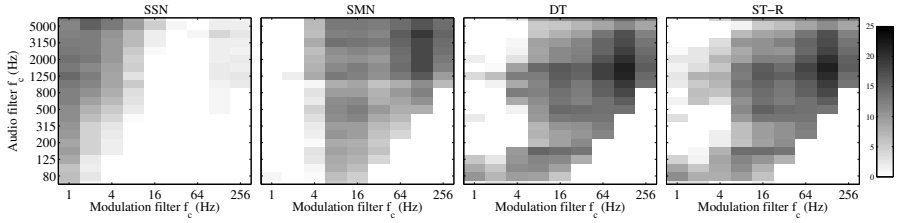


Figure 5.3: SNR_{env} -patterns for the SSN (left), SMN, (middle left), DT (middle right), and ST-R interferers (middle left). The patterns represent the time-averaged SNR_{env} in dB as a function of the audio and modulation filter center frequencies considered in the model, in response to sentences mixed with the corresponding noise at an input SNR of −3 dB. Each pattern represents the average across 120 sentences.

The SRT in the HP300-condition (right part of left panel) was obtained at an SNR of about −7 dB, while it was at about 0 dB SNR in the HP30 condition, which is a much larger difference than for the corresponding BB-conditions. This demonstrates a clear interaction between HP-filtering and envelope low-pass filtering for the ISTS interferer (see Christensen *et al.*, 2013). For the SSN-interferer, the model (black symbols in the right panel of Fig. 5.2 accounted well for the effect of high-pass filtering, while it predicted slightly lower SRTs for the BB30 and HP30 conditions than for the BB300 and HP300 conditions, in contrast to the data. For the ISTS interferer, the model accounted well for the effect of envelope low-pass filtering in the broad-band conditions, but not in the corresponding high-pass-filtered conditions, where only a small effect of high-pass filtering was predicted, in contrast the data.

5.4 Model analysis

The model predictions were in good agreement with the data from Festen and Plomp (1990) in the conditions with unprocessed speech and the SSN, SMN, and DT interferers, but deviated from the data in the case of the ST-R interferer. This result was analyzed further by considering the model's internal representation of the stimuli. Figure 5.3 shows the time-averaged SNR_{env} as function of the audio and modulation-filter center frequencies in the model for the SSN (left), SMN, (middle left), DT (middle right), and ST-R (right) interferers. Each panel represents the contribution

of the individual audio and modulation-filters to intelligibility in the model, where dark areas denote greater contribution. For the SSN interferer, the main contribution stem from modulation-filters centered below about 16 Hz and from a broad range of audio-filter center frequencies. In contrast, for the (fluctuating) SMN, DT and ST-R interferers, the main contribution stem from modulation-filters centered at frequencies above 4 Hz, and mostly from filters centered above 32 Hz. Thus, in the framework of the model, high-rate envelope fluctuations above 30 Hz are important for speech masking release in conditions with fluctuating noise and interfering talkers. However, even though high-rate modulation filters carry the main contribute in the model in the ST-R condition, the measured intelligibility was found to be poor. This demonstrates that the a high SNR_{env} alone, as measured in the model, is not sufficient to account for speech intelligibility in all conditions.

Figure 5.4 shows SNR_{env}-patterns for each condition with vocoded speech mixed with either SNN (top row) or ISTS (bottom row) interferers. For the SSN-interferer in the BB-conditions (left and middle-left panels), the main contributions stem from the modulation-filters centered below 16 Hz and mainly from the audio-filters centered around 125 Hz. This is most likely because the fundamental frequency of the target speech (about 120 Hz) was poorly masked by the SSN-interferer, which was spectrally shaped to the ISTS-interferer that had a higher fundamental frequency of about 200 Hz. The dominating low-frequency audio and modulation-filter contributions in the BB-conditions explains why the model predicts higher SRTs when this information is removed by high-pass filtering (top right panels), and why the effect of envelope low-pass filtering is small. For the ISTS interferer (bottom to row of Fig. 5.4) in the BB-conditions (left panels), the main contributions stem from the modulation filters centered above 32 Hz, and from audio-filters centered above 500 Hz. Thus, the model is less sensitive to high-pass filtering in the case of the ISTS-interferer than for the SSN-interferer. However, the model accounts well for the effect of envelope low-pass filtering with the ISTS-interferer because the main contribution stem from the modulation filters centered above 32 Hz in this condition.

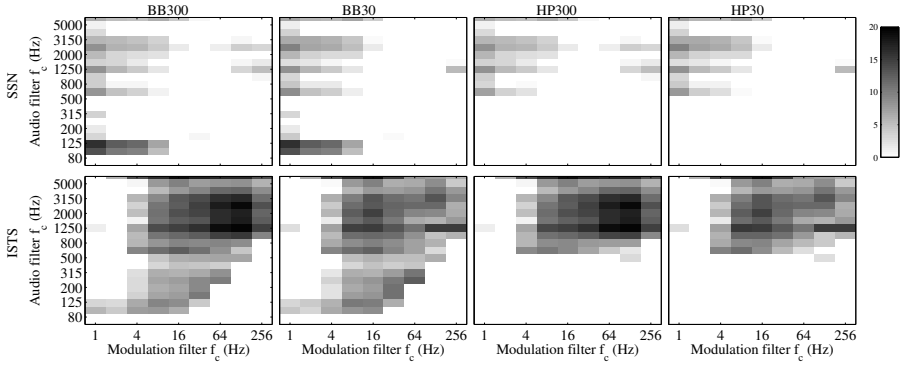


Figure 5.4: SNR_{env} -patterns for each condition with vocoded speech mixed with either SNN (top row) or ISTS (bottom row) interferers. The input SNR was -2 dB and the patterns are based on the average across 150 sentences.

5.5 Discussion

The data by Festen and Plomp (1990) demonstrated better speech intelligibility in the conditions with fluctuating interferers than for steady SSN, with the same long-term spectrum. The predictions from the present study suggested that target speech envelope fluctuations might be available (as quantified by the SNR_{env}-patterns in Fig. 5.3) at high fluctuation rates ($>$ 32 Hz) in conditions with modulated noise and competing talkers but not with SSN, leading to better intelligibility. The model accounted for the data by Festen and Plomp (1990) obtained with a competing talker that was different from the target talker. However, in the condition where the competing talker was a time-reversed version of the target, the model suggested that there was a high SNR_{env}, while the data showed poor human performance. This demonstrates that the SNR_{env} is not always sufficient to account for speech intelligibility. While the SNR_{env} includes aspects of energetic and modulation masking, the remaining unexplored effect may be attributed and defined as "informational masking" (e.g., Brungart, 2001) occurring at levels higher up in the auditory pathway than modulation processing. Informational masking may be caused by the identical fundamental frequencies of the target and interferer, which may confuse the listener and make it difficult to perform a perceptual segregation between the target and the interferer. While the model results suggest that

the high-rate envelope fluctuations may play an important role for speech masking release, the experiments in Festen and Plomp (1990) did not explicitly evaluate this point. However, Christensen *et al.* (2013) considered the effect of removing high-rate envelope fluctuations from the target and interfering talker (represented by the ISTS) on speech intelligibility. Their data showed largely reduced intelligibility when the envelope fluctuations above 30 Hz were attenuated by low-pass filtering. The sEPSM accounted well for this effect, supporting the model-based finding that high-rate envelope fluctuations might be important for speech intelligibility in conditions with speech interferers. The purely envelope power-based model proposed here could account for a large part of the considered data. However, there were specific conditions where this approach was limited. In those cases, additional cues and auditory processes may affect speech intelligibility, such as interference in the processing of carrier or fundamental frequency information, which is not reflected in the model and might be associated with (higher-level) informational masking effects.

5.6 Conclusion

The multi-resolution speech-based envelope power spectrum model accounted well for speech masking release in conditions with modulated noise and all but one of the conditions with competing talkers. The model predictions support the finding of Christensen *et al.* (2013) that high-rate envelope fluctuations ($>$ 30 Hz) play an important role for the speech intelligibility in conditions with speech interferers. However, the model could not account for a condition where the target and the masker were from the same talker, which was attributed to effects of informational masking not reflected in the model. Overall, the results support the hypothesis that the $\mathrm{SNR}_{\mathrm{env}}$ may be a useful metric for speech intelligibility prediction.

6

Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility[¶]

Jørgensen and Dau (2011) suggested a new metric for speech intelligibility prediction based on the signal-to-noise envelope power ratio (SNR_{env}), calculated at the output of a modulation-frequency selective process. In the framework of the speech-based envelope power spectrum model (sEPSM), the SNR_{env} was demonstrated to account for speech intelligibility data in various conditions with linearly and nonlinearly processed noisy speech as well as for conditions with stationary and fluctuating interferers. Here, the relation between the SNR_{env} and speech intelligibility was investigated further by systematically varying the modulation power of either the speech or the noise before mixing the two components, while keeping the overall power ratio of the two components constant. A good correspondence between the data and the corresponding sEPSM predictions was obtained when the noise was manipulated and mixed with the unprocessed speech, consistent with the hypothesis that SNR_{env} is indicative of speech intelligibility. However, discrepancies between data and predictions occurred for conditions where the speech was manipulated and the noise left untouched. In these conditions, distortions introduced by the applied modulation processing were detrimental for speech intelligibility but not reflected in the SNR_{env} metric, thus representing a limitation of the modeling framework.

[¶] This chapter is based on Jørgensen *et al.* (2014b).

6.1 Introduction

Speech intelligibility prediction has been a major research field since the first telecommunication technologies were introduced in the late nineteenth century. One of the first broadly applied methods for predicting speech intelligibility was introduced by French and Steinberg (1947) who presented the concept of the articulation index (AI). Fundamentally, the AI predicts speech intelligibility by calculating the overall power signal-to-noise ratio (SNR) from the speech long-term spectrum and the background noise long-term spectrum in various frequency bands. The AI was later extended to include corrections for hearing sensitivity loss, speech level as well as upward and downward spread of masking. This has led to a revised prediction model denoted the speech intelligibility index (SII; ANSI S3.5, 1997). While the SII model was demonstrated to account for speech intelligibility data in conditions with stationary background noise and low- and high-pass filtering, it has limitations, for example, in reverberant conditions.

Houtgast and Steeneken (1973) demonstrated that reverberation leads to temporal smearing of the speech signal, which is not detected by the conventional power SNR-metric used in the SII. Houtgast and Steeneken (1971) defined the speech transmission index (STI) as a measure of the integrity of the temporal modulations of the speech, and demonstrated that such a metric could account for the detrimental effect of stationary noise and reverberation on speech intelligibility. However, the STI-concept is also limited and fails in conditions where the noisy speech mixture has been processed by noise reduction, such as spectral subtraction (Ludvigsen *et al.*, 1993), possibly because the noise reduction affects the noise modulations as well as the speech modulations (Dubbelboer and Houtgast, 2007; Jørgensen and Dau, 2011). Several extensions of the original STI have been proposed (e.g., Payton and Braida, 1999; Goldsworthy and Greenberg, 2004) all of which, however, were based on a comparison between the clean speech and the noisy transmitted speech. Thus, none of the approaches considered the effect of the noise reduction processing on the amount of "intrinsic" modulations of the noise itself. This was done in an alternative approach by Jørgensen and Dau (2011), which suggested to consider the signal-to-noise envelope power ratio (SNR_{env}) as a measure of the amount of the useful speech modulation content available

to the listener. It was demonstrated that the SNR_{env}-based metric and the STI lead to similar predictions in conditions with reverberation and stationary noise, but only the SNR_{env}-metric can also account for the detrimental effect of spectral subtraction on speech intelligibility (Jørgensen and Dau, 2011).

In the present study, the SNR_{env} was computed using the multi-resolution version of the speech-based envelope power spectrum model (sEPSM) as presented in Jørgensen *et al.* (2013). This model is very similar to the version presented in Jørgensen and Dau (2011) but, instead of measuring the SNR_{env} from the long-term envelope power spectrum, the multi-resolution version of the sEPSM estimates the SNR_{env} in short temporal segments, with segment durations inversely proportional to the center frequencies of the modulation filters considered in the processing. This model was shown to successfully predict the speech reception threshold (SRT) in conditions with speech mixed with various stationary and fluctuating interferers as well as in conditions with noisy speech processed by spectral subtraction and reverberation (Jørgensen *et al.*, 2013).

The main hypothesis of the sEPSM framework is that there is a monotonic relationship between the SNR_{env} of a noisy speech stimulus and the corresponding speech intelligibility. However, this relationship has so far not been evaluated by systematic variation of the amount of SNR_{env} (while keeping the conventional power SNR fixed). Stimuli with different SNR_{env} but the same power SNR can either be obtained by a modification of the modulation content of the speech signal, the noise interferer, or both. The *analysis* of the modulation content of a (speech) signal may be straightforward, whereas it is more challenging to synthesize or process signals such that they possess prescribed modulation properties. For example, Drullman *et al.* (1994b) attempted to filter out certain modulation frequencies from a speech stimulus using an analysis and re-synthesis framework. However, one problem with their approach was that the synthesis step involved the original phase information, which re-introduced the original modulation content after bandpass filtering in the auditory system (Ghitza, 2001). A possible way to avoid the synthesis problem is to iteratively build a signal that has a desired modulation spectrum. The development of spectrogram reconstruction tools (e.g., Griffin and Lim, 1984; Zhu X., 2006; Sun

and Smith III, 2012; Decorsière, 2013) makes this possible by building a real time-domain signal corresponding to a given target spectrogram. Here, a spectrogram is defined as a collection of sub-band envelope signals, resulting from passing the original signal through a bandpass filterbank. Using such an approach, Elliott and Theunissen (2009) analyzed the contribution of independent temporal and spectral modulation frequency bands to the intelligibility of speech. They found that speech intelligibility remained high at about 75 % words correct when restricting the temporal modulations to frequencies below 7 Hz and the spectral modulations to rates below 3.75 cycles/kHz. Restricting this "core" spectro-temporal modulation frequency range further had a large detrimental effect on intelligibility.

In the present study, the spectrogram reconstruction tool described in Decorsière (2013) was used to generate noise and speech stimuli with amplified or attenuated modulation content. Based on the $\mathrm{SNR_{env}}$ concept, the hypothesis was that noise with attenuated modulation content mixed with unprocessed speech as well as speech with amplified modulation content mixed with unprocessed noise should provide better intelligibility than unprocessed speech in noise. The modulation-processed stimuli were used to evaluate the relationship between the $\mathrm{SNR_{env}}$ and speech intelligibility obtained in corresponding psychoacoustic tests. The processing strategy taken here directly manipulated either the clean speech signal or the noise alone before mixing the two components, in an attempt to make the speech more intelligible in a given noisy condition. This approach differs from other modulation-domain speech enhancement strategies (e.g., Paliwal *et al.*, 2010; So and Paliwal, 2011; Wójcicki and Loizou, 2012) that focused on ways to attenuate the noise component of a noisy speech mixture. Here, the focus was to enhance the modulation content of the speech relative to the noise, before mixing the two components. Such an approach could be useful in a situation where there is access to the speech signal before it is transmitted and mixed with environmental noise, such as at a train station.

6.2 Method

6.2.1 Speech material, apparatus, and procedure

Speech reception thresholds (SRT) were measured using the material provided in the Danish Conversational Language Understanding Evaluation (CLUE; Nielsen and Dau, 2009), which is similar to the hearing in noise test (HINT; Nilsson *et al.*, 1994). The speech material in the CLUE test consists of 18 lists of ten unique sentences, recorded in anechoic conditions. Each sentence represents a meaningful everyday sentence containing five words, spoken in a natural manner, by a male speaker. The background noise in the CLUE test is a stationary speech-shaped noise (SSN) constructed by concatenating and superimposing the sentence material to obtain a stimulus with the same long-term spectrum as the average long-term spectrum of the sentences. Five male normal-hearing native Danish speakers, aged between 24 and 38 years, participated in the study. The subjects were sitting in a double-walled insulated booth together with the experimenter who was controlling the procedure via MATLAB. The digital signals, sampled at 44.1 kHz, were converted to analog by a high-end RME DIGI96/8 soundcard. They were presented to the subjects diotically via Sennheiser HD580 headphones. The average sound pressure level (SPL) of the stimuli in the test was 65 dB. After each presentation, the subjects were instructed to repeat the words he/she understood, with the possibility of guessing or passing on misunderstood words. The experimenter recorded the correctly understood words individually.

The SNR was controlled throughout the test by changing the level of the SSN after the listener's response, using an adaptive procedure. If all the words of a sentence were repeated correctly the SNR was lowered by 2 dB, otherwise it was increased by 2 dB. The SRT was determined as the average of the SNRs calculated after the response to the last eight sentences of a list. Further details on the speech material can be found in Nielsen and Dau (2009).

6.2.2 Stimulus conditions

Two stimulus conditions were considered: (i) unprocessed speech mixed with SSN that was processed in the modulation domain as described further below (section 6.2.3) and (ii) speech that was processed in the modulation domain and mixed with unprocessed SSN. The modulation processing either attenuated or amplified the modulation power in a target modulation frequency range between 4 and 16 Hz, while (ideally) providing zero gain outside the target range. Six conditions of the target modulation gain were considered when only the noise was processed: 20, 10, 0 , −5, −10 and −20 dB relative to the unprocessed noise, and seven conditions were considered when only the speech was processed: 20, 10, 5, 0, −6, −10 and −20 dB, whereby 0 dB represented the unprocessed reference condition. Smooth transitions between the amplified/attenuated modulation-frequency band and the zero-gain frequency region were obtained using raised cosine ramps in the transition bands from 1 to 4 Hz and from 16 to 22 Hz. The effectiveness of the modulation processing was analyzed by computing the modulation transfer function (MTF) of the processed signals relative to the unprocessed signal as suggested by Schimmel and Atlas (2005). The MTF of a single channel, m, was defined here as

$$MTF_m = \frac{|\mathscr{F}\{|\hat{p}_m|\}|}{|\mathscr{F}\{|\hat{u}_m|\}|}, \tag{6.1}$$

where p and u represent the processed and unprocessed signals, respectively, $\mathscr{F}$ denotes the Fourier transform, the hat represents the analytical signal, and $|\cdot|$ denotes the modulus. Note that the MTF defined here is different from the MTF-concept used in the STI. The MTF of an entire spectrogram was taken as the average of the subchannel MTFs, where the MTF_m of the individual channels were weighted according to their energy

$$MTF = \frac{1}{\sum_{m=1}^{M} \|u_m\|_2} \sum_{m=1}^{M} \|u_m\|_2 \cdot MTF_m, \tag{6.2}$$

where $\|\cdot\|_2$ is the Euclidean norm.

Figure 6.1 (upper left panel) shows the MTF for the processed noises for each of the five target modulation gains. The dashed curves represent the target MTF and the solid curves show the obtained “actual” MTF at the output of the modulation processing

stage that is described in more detail further below. In a perfect modulation processing system, the dashed and solid lines would coincide. In the processing framework presented here, the actual gain/attenuation is smaller than the target gain/attenuation, particularly for the largest target values of +/-20 dB. The top right panel of Fig. 6.1 shows the corresponding long-term excitation patterns (Glasberg and Moore, 1990) of the processed noises, representing the total energy of the signal at the output of a bank of gammatone filters with one equivalent rectangular bandwidth spacing, plotted as a function of the center frequencies of the filters. The patterns for the three negative gains (-5, -10 and -20 dB) coincide with the one obtained for the unprocessed signal. The pattern for 20 dB gain (light gray) lies above the unprocessed pattern (black) by more than $+5$ dB at very low frequencies (< 50 Hz) and below the unprocessed pattern by more than -5 dB at frequencies above 1200 Hz.

The lower left panel of Fig. 6.1 shows the target MTFs (dashed lines) and the actual MTFs (solid lines) for the six conditions where the speech was processed. Each MTF represents the average across the individual MTFs obtained for 180 sentences without noise. The obtained actual gains/attenuations are below the respective target gains/attenuations, particularly for the target gains of +/- 20 dB. Furthermore, the effective frequency region of the MTF relative to the target range (4 to 16 Hz) is shifted towards higher modulation frequencies in the conditions with modulation enhancements, which is different from the results obtained with the processed noise. The lower right panel of Fig. 6.1 shows the corresponding excitation patterns for the processed speech and the unprocessed speech. The maximal deviation of the patterns for processed speech from the one for unprocessed speech amounts to 5 dB at frequencies above 5 kHz.

6.2.3 Modulation processing framework

The modulation processing consisted of two parts as indicated in Fig. 6.2. In part A, the original unprocessed signal was first passed through a Gammatone filterbank, and the Hilbert envelope was extracted at the output of each filter. The resulting set of envelopes constituted the unprocessed spectrogram. Each envelope of the spectrogram

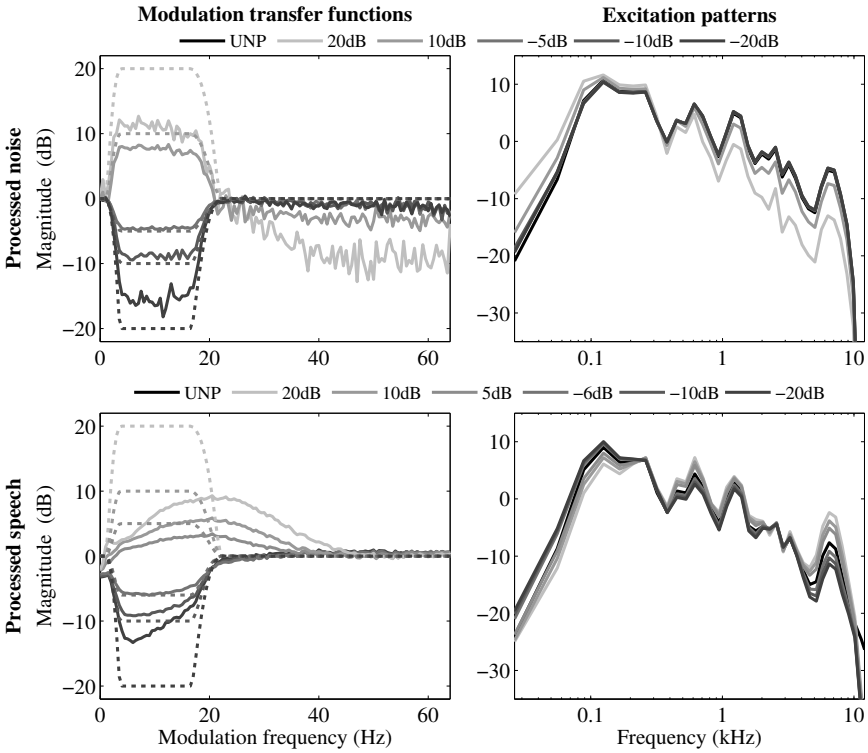


Figure 6.1: Top left: Target modulation transfer functions (MTF; dashed lines) and actual MTFs (solid lines) for the five conditions with processed noise. The grayscale corresponds to different target modulation gains (−20, −10 −5, 10, 20 dB). Each noise signal was 22 seconds long, sampled at 22.05 kHz, and the corresponding MTF was obtained as the average over 50 segments of 2 s each. Top right: Long-term excitation patterns of the five processed noises and the unmodified noise (UNP). Bottom left: Target (dashed lines) and actual (solid lines) MTFs of the processed speech for the six target modulation gains. MTFs were averaged over the 180 sentences of the CLUE material. Bottom right: Long-term excitation patterns of the speech for the six target modulation gains and the unprocessed speech (UNP).

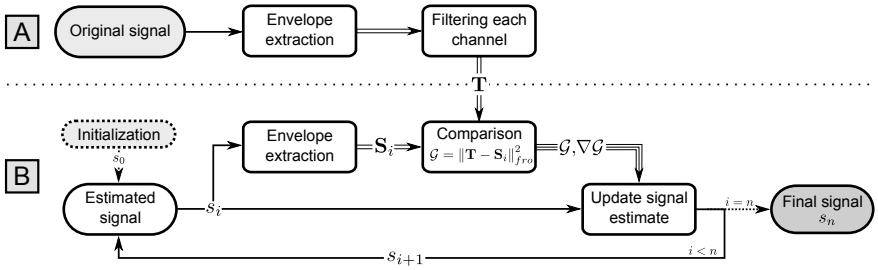


Figure 6.2: Schematic view of the two steps in the modulation processing framework. Single-lined arrows relate to time-domain signals and double-lined arrows to multi-channel envelopes (spectrograms). Step A, upper part, generates a target spectrogram **T** by separately filtering each channel of the original signal's spectrogram. In step B, lower part, a signal is constructed iteratively by comparing the spectrogram $\mathbf{S}_i$ of the current signal estimate s_i to the target **T**. The distance between the two spectrograms $\mathscr{G}$ and its gradient $\nabla\mathscr{G}$ are used to update the current signal estimate until the maximum number of iterations n is reached.

representation was filtered by a zero-phase bandpass filter with a given target MTF. To avoid transients in the filter magnitude response, transitions between the pass-band and the filtered band were smoothed using half raised cosine windows. The filtered envelopes were then individually normalized such that they had the same root mean square (RMS) value as their unfiltered counterpart. This ensured that the total power of the envelopes in each frequency channel was only marginally affected by the envelope filtering such that the processed signal had a similar long-term audio-domain excitation pattern as the original signal (right panels of Fig. 6.1).

To be consistent with the definition of a spectrogram, each processed envelope had to be non-negative. However, filtered envelopes could exhibit significant negative sections, particularly when large positive modulation gains were provided to signals that initially contained large envelope fluctuations, such as speech. To overcome this, the dynamic range of the envelope was limited by raising the envelope of each channel to the power of 1/3 before filtering. After filtering, the original dynamic range was restored by raising the filtered envelope to the power of 3. The resulting filtered spectrogram provided the "target" input, **T**, to the signal reconstruction stage (Part B) of the modulation processing. In the signal reconstruction, indicated as Part B in Fig. 6.2, a time-domain signal, s, was reconstructed iteratively, such that the difference between the spectrogram of the reconstructed signal and the target spectrogram was minimal.

The procedure was initiated by a random noise signal s_0 that, for each iteration i, was updated in the direction that reduced the distance between its spectrogram $\mathbf{S}_i$ and the target spectrogram $\mathbf{T}$. The distance, $\mathscr{G}$, between the spectrograms was given as the square of the Frobenius matrix norm of the difference between the two spectrograms

$$\mathscr{G} = \|\mathbf{T} - \mathbf{S}_i\|_{fro}^2 \tag{6.3}$$

The iterative procedure was terminated after 100 iterations. Details about the signal reconstruction can be found in Decorsière (2013).

6.2.4 Speech intelligibility prediction

The processing structure of the sEPSM is illustrated in Fig. 6.3. The details of the processing can be found in Jørgensen and Dau (2011) and Jørgensen *et al.* (2013). Some of the main stages are described in the following. The first stage is a bandpass filterbank consisting of 22 gammatone filters Glasberg and Moore (1990) with a third-octave spacing between their center frequencies, covering the range from 63 Hz to 8 kHz. The temporal envelope of each output is extracted via the Hilbert-transform and then low-pass filtered with a cut-off frequency of 150 Hz using a first-order Butterworth filter. The resulting envelope is analyzed by a modulation bandpass filterbank. The filterbank consists of eight second-order bandpass filters with octave spacing and center frequencies covering the range from 2 - 256 Hz, in parallel with a third-order lowpass filter with a cut-off frequency of 1 Hz.

The running temporal output of each modulation filter is divided into short segments using rectangular windows with no overlap (Jørgensen *et al.*, 2013) . The duration of the window is specific for each modulation filter and equal to the inverse of the center-frequency of a given modulation filter (or the cut-off frequency in the case of the 1-Hz low-pass filter). For example, the window duration in the 4-Hz modulation filter is 250 ms. For each window, the AC-coupled envelope power (variance) of the noisy speech and the noise alone are calculated separately and normalized with the corresponding long-term DC-power. The $\mathrm{SNR}_{\mathrm{env}}$ of a window is estimated from the envelope power as

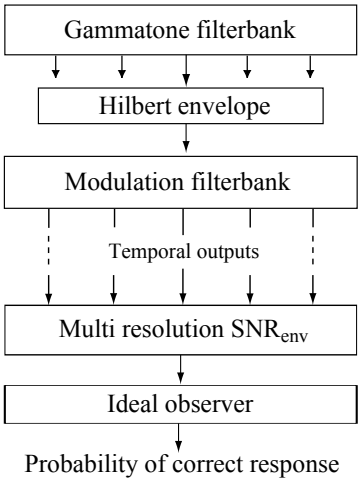


Figure 6.3: Block diagram of the processing structure of the sEPSM (Jørgensen *et al.*, 2013). Noisy speech and noise alone are processed separately through a gammatone bandpass filterbank followed by envelope extraction via the Hilbert transform. Each sub-band envelope is further passed through a modulation bandpass filterbank. The modulation- filtered temporal outputs are segmented with a segment duration inversely related to the modulation-filter center frequency. The envelope power (variance) is computed in each segment for the noisy speech ($P_{env,S+N}$) and the noise alone ($P_{env,N}$), from which the corresponding SNR$_{\text{env}}$ is derived. The segmental SNR$_{\text{env}}$ is then averaged across segments and combined across modulation filters and audio-domain (peripheral) filters. Finally, the overall SNR$_{\text{env}}$ is converted to the probability of correct response assuming an ideal observer as in Jørgensen and Dau (2011) .

$$\text{SNR}_{\text{env}} = \frac{P_{env,S+N} - P_{env,N}}{P_{env,N}} \tag{6.4}$$

where $P_{env,S+N}$ and $P_{env,N}$ represent the normalized envelope power of the noisy speech and the noise alone, respectively. For each modulation filter, the running SNR$_{\text{env}}$-values are averaged across time, assuming that all parts of a sentence contribute equally to intelligibility. The time-averaged SNR$_{\text{env}}$-values from the different modulation-filters are then combined across modulation filters and across Gammatone filters, using the "integration model" from Green and Swets (1988). The combined SNR$_{\text{env}}$ is converted to the probability of correctly recognizing the speech item using the concept of a statistically "ideal observer" (Jørgensen and Dau, 2011) .

For the simulations, 150 sentences from the CLUE material were used. Each sentence

was mixed with a noise token (randomly selected from the full-length noise files) over a discrete range of SNRs. For a given SNR-value, the final percent correct prediction was computed as the average predicted score across all sentences of a given speech material. The prediction at each SNR was then connected by straight lines, resulting in a continuous psychometric function, from which the SRT was estimated as the SNR corresponding to 50% correct. The values of the parameters in the model were kept fixed in all conditions and corresponded to those given in Table II in Jørgensen *et al.* (2013).

6.3 Results

Figure 6.4 shows the results for the conditions where the noise interferer was processed and the speech left untouched. The open symbols show measured speech intelligibility data, represented as the change in SRT (ΔSRT) relative to the unprocessed condition. ΔSRT is shown as a function of the target modulation gain, and a positive ΔSRT reflects worse intelligibility compared to the unprocessed condition. An analysis of variance was conducted to assess the statistical significance of the measured data. The statistical results are presented as asterisks above the data points, with $p < 0.1$ for $(*)$ and $p < 0.01$ for $(**)$ indicating significant differences from the unprocessed condition. A non-monotonic relationship between the obtained SRT and the target modulation gain was observed. In the range of conditions with negative gain, i.e. with attenuated noise modulations, the SRT decreased slightly (up to about 2 dB) with decreasing gain. In the conditions with positive gains, i.e. amplified modulations, a large decrease of the SRT of about 9 dB was observed for the target gain of 20 dB. The filled circles represent the predictions obtained with the sEPSM, which were in good agreement with the data, although the model slightly overestimated the effect at large positive gains. For direct comparison, predictions obtained with the extended SII (ESII; Rhebergen *et al.*, 2006), using the unprocesed SSN as a probe for the speech signal, are also shown in Fig. 6.4 and indicated by the filled diamonds. The ESII predicted the trend in the data for positive modulation gains, but predicted a small positive ΔSRT, i.e. a slight decrease of speech intelligibility, for negative gains, in contrast to the measured data.

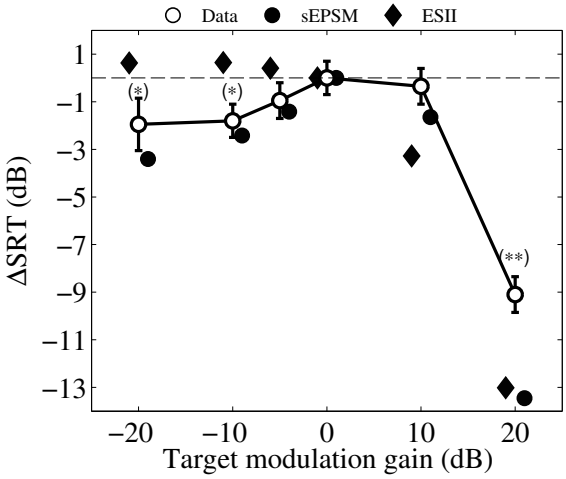


Figure 6.4: Change of the speech reception threshold, ΔSRT, relative to the unprocessed condition (0 dB target gain), as a function of the target modulation gain applied to the noise interferer (but not the speech signal). Open circles represent the measured data, with error bars indicating standard errors. Asterisks indicate a statistically significant difference to the reference condition, with $p < 0.1$ represented as $(*)$ and $p < 0.01$ represented as $(**)$. The filled circles show predictions obtained with the sEPSM, and filled diamonds show predictions using the ESII.

Figure 6.5 shows the results obtained for the conditions with processed speech (in the presence of unprocessed noise). The open symbols show the measured data. The SRT increased by 1.5 dB for a target modulation gain of 10 dB and by 5.5 dB for a gain of 20 dB, i.e. representing a decrease of intelligibility. Similarly, in the conditions with negative gains, the SRT increased by 2.7 dB for a target gain of −10 dB and by 7.3 dB for a target gain of −20 dB. Thus, the intelligibility decreased in the conditions where a large negative or positive modulation gain was applied to the speech (alone).

The corresponding predictions obtained with the sEPSM are shown by the filled circles and were essentially independent of the amount of negative gain, in clear contrast to the data. Moreover, the model predicted a decrease in SRT for the conditions with positive gains. This reflects the underlying assumption in the model linking enlarged modulation power of the speech signal (cf. bottom left panel of Fig. 6.1) to increased speech intelligibility. However, this relation is not supported by the data for the conditions with modulation enhanced speech. For comparison, predictions obtained

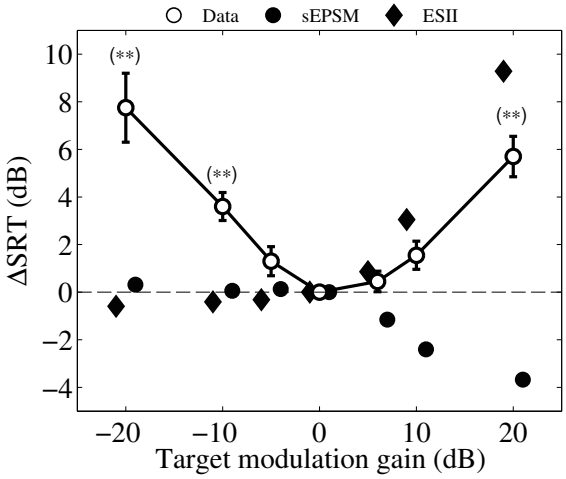


Figure 6.5: ΔSRT relative to the unprocessed condition (0-dB target gain) as a function of the target modulation gain applied to the speech (mixed with unprocessed noise). The open symbols represent the measured data, the filled circles show the sEPSM predictions, and the filled diamonds show predictions using the ESII. The error bars represent standard errors. Asterisks indicate a statistically significant difference to the reference condition, with $p < 0.1$ for $(*)$ and $p < 0.01$ for $(**)$.

with the ESII are indicated by the filled diamonds in the figure where modulation processed SSN was used as a probe for the speech. Thus, the simulations for the processed speech conditions essentially represent a mirrored pattern of the simulations for the processed noise conditions (around the 0-dB ΔSRT axis). The ESII appeared to account at least partly for the reduced intelligibility for positive gains observed in the data. However, because the ESII applied a noise probe for the speech, the inputs to the model in the conditions with processed noise and processed speech were the same: for the simulations in Fig. 6.4 the inputs were unprocessed SSN for the speech and processed SSN for the noise, respectively, and vice versa in the simulations shown in Fig. 6.5. Hence only the labels of the input signal were changed. This makes it difficult to determine whether the trend predicted by the ESII reflects the actual underlying cause for the reduced intelligibility in the data or if it is a consequence of the special noise-probe signal leading to symmetric predictions in the two situations.

6.4 Discussion

6.4.1 Modulation processing of the noise interferer

The perceptual data showed an improvement of speech intelligibility when the noise was processed and the speech left untouched. The predictions obtained with the sEPSM were in good agreement with the measured data in those conditions, demonstrating that the intelligibility improvement could be accounted for by a greater SNR_{env} after the modulation processing, which supports the hypothesis of a relationship between SNR_{env} and speech intelligibility. In the conditions with negative target modulation gain, the decrease of SRT obtained in the data was up to 2 dB. In the framework of the sEPSM, this decrease was due to a reduced amount of modulation masking by the noise, because the modulation power of the noise was effectively reduced in the range from 4-16 Hz (cf. top leftpanel of Fig. 6.1). In contrast, the ESII was insensitive to the effect of the reduced noise modulation power, because the short-term power SNR remained the same as in the reference condition.

In the case of positive target modulation gains, a decrease of SRT of up to 9 dB was observed in the data. Physically, the amplification of the noise modulations led to strong amplitude variations of the noise's temporal waveform, similar to amplitude modulated noise (e.g., Festen and Plomp, 1990). In the framework of the sEPSM, the decreased SRT observed in these conditions could be explained by less modulation masking at relatively high modulation frequencies ($>$ 30 Hz), as demonstrated by Jørgensen *et al.* (2013). The ESII could also account for the results obtained with positive modulation gain, based on a greater short-term power SNR. This is consistent with earlier studies that demonstrated the usefulness of ESII for predicting speech intelligibility in the presence of a modulated noise (Rhebergen *et al.*, 2006). This suggests that, the decreased SRT in the conditions with amplified noise modulation power can be accounted for by less modulation masking as well as by less energetic masking, indicating that those aspects are linked in these conditions. However, the improved SRT in the conditions with attenuated noise modulation power could only be accounted for by less modulation masking.

6.4.2 Modulation processing of clean speech

The perceptual data showed a general reduction of speech intelligibility when the speech was processed and the noise left untouched. The predictions from the sEPSM were not in agreement with this aspect of the data, demonstrating a limitation of the model framework. The sEPSM assumes that any modulation of the speech component of the input signal is contributing to intelligibility. However, this assumption might be violated after the processing of the speech alone, because the modulation processing may have introduced distortions in the form of strong modulation components not related to speech at all. To investigate this, the amount of the speech distortion was assessed using a measure based on the Perceptual Evaluation of Speech Quality (PESQ; ITU-T P.862, 2001). Figure 6.6 shows the amount of speech distortion, defined here as the inverse of the PESQ-rating, scaled to a range between 0 and 1, for the different conditions of modulation processed speech. The distortion is zero for the zero-gain reference condition and increases with increasing or decreasing target modulation gain. This trend corresponds well to the trend observed in the intelligibility data shown in Fig. 6.5. The presence of distortion and the attribute of unnaturalness in the conditions with modulation-processed speech were also reported qualitatively by the listeners, even at high SNRs. Thus, the discrepancy between the data and the sEPSM predicitons might be explained by distortions of the speech signal (resulting from the modulation processing), which are not considered by the sEPSM.

There were at least two sources of the distortion when amplifying the modulation of natural speech. First, the step in the modulation filtering process that generated the target spectrogram represented a “blind” process, i.e. the filtering process had no *a priori* information about the initial spectro-temporal structure of the speech signal. Hence, the amplified modulation was not constrained to follow the natural temporal structure of the speech, which could lead to modulation components not related to speech at all. Moreover, the larger the modulation gain/attenuation was, the further the filtered spectrogram deviated from a faithful representation of speech, i.e. the target spectrogram could represent distorted speech rather than enhanced speech. A second source of distortion was related to the iterative reconstruction process. Since the spectrogram representation is generally of higher dimensionality than a real

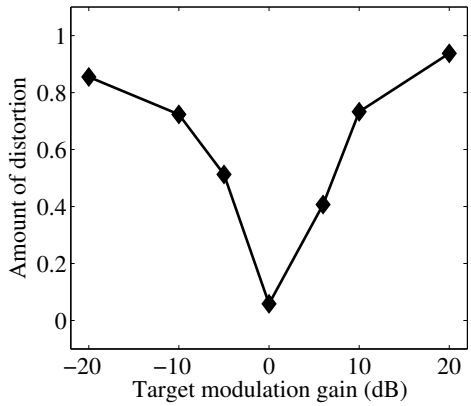


Figure 6.6: Objectively measured distortion of the speech signal for the different conditions of modulation processing. The distortion is defined here as the inverse of the PESQ measure scaled to the range between 0 and 1, and each point represents the average across 50 sentences. A metric such as the $\mathrm{SNR}_{\mathrm{env}}$ assumes that all audible speech modulations contribute to speech intelligibility. The metric cannot account for distortions that are not resulting from the processing applied to the mixture of speech and noise (such as spectral subtraction).

time-domain signal, many theoretical spectrograms could result from the modulation-filtering process, for which no corresponding real time-domain signal would exist (Le Roux *et al.*, 2010). Thus, the target spectrogram obtained in Fig. 6.2 (Part A) did not necessarily correspond to an actual time-domain signal. This implied that the objective function in (Eq. (6.3)) might never reach zero, even if the reconstruction was successful in the sense that it reached the minimum of the objective function. The remaining distance between the target and the obtained spectrograms would translate into an error in the reconstructed signal, in the form of uncontrolled distortions.

6.4.3 Usefulness of modulation processing for speech intelligibility enhancement

Ideally, according to the concept of the $\mathrm{SNR}_{\mathrm{env}}$, amplifying the modulations of a speech signal by a given amount (before mixing it with the unprocessed noise) should lead to a similar positive effect on intelligibility as attenuating the modulations in the noise by the same amount (before mixing the noise with the unprocessed speech). This was

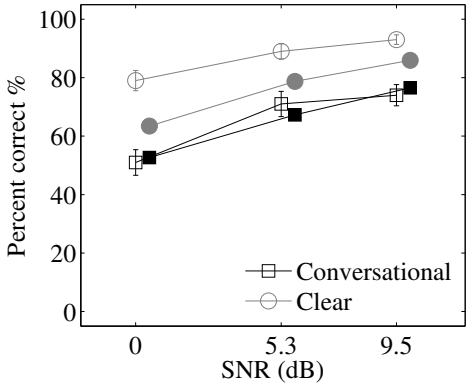


Figure 6.7: Percentage of correct key words for normal hearing listeners reported by Payton *et al.* (1994) in conditions with speech spoken in a conversational (open squares) and in a clear manner (open circles) mixed with noise at 0, 5.3, and 9.5 dB SNR. Corresponding predictions by the sEPSM using the same stimuli are indicated by the filled symbols.

reflected in the predictions by the sEPSM, showing the same predicted SRT when applying a target modulation gain of −20 dB to the noise (left part of Fig. 6.4) as when applying a gain of 20 dB to the speech (right part of Fig. 6.5).

However, the ability of the applied modulation processing scheme to enhance speech intelligibility was limited to the conditions with processing of the noise alone. Attempts to enhance the modulations of the speech did not result in improved intelligibility. The speech considered in the present study was spoken in a conversational manner, so the modulations of that speech might be enhanced in a natural way by articulating the sentences more clearly, creating modulation enhanced speech free from distortions. Such stimuli were considered in the studies of Picheny *et al.* (1985) and Payton *et al.* (1994), which measured better intelligibility for clearly spoken speech compared to conversationally spoken speech in normal-hearing listeners. Figure 6.7 shows the percentage of correct key words reported by Payton *et al.* (1994) in conditions with speech spoken in a conversational manner (open squares) and in a clear manner (open circles) and mixed with noise at 0, 5.3, and 9.5 dB SNR. The word score improved by 28, 18, and 19 % for the three SNRs, respectively, for the clearly spoken speech

compared to the conversational speech. Corresponding predictions[1] obtained with the sEPSM are indicated by the filled symbols in Fig. 6.7, with filled black squares representing the conversational speech conditions and the gray filled circles showing the clear speech conditions. The sEPSM predicted an improved word score for the clearly spoken speech consistent with the data, whereby, however, the improvement was about 10% and thus smaller than in the data. Nevertheless, this supports the hypothesized monotonic relationship between $\mathrm{SNR}_{\mathrm{env}}$ and speech intelligibility, in the case where the modulation enhanced speech was free from distortions. In contrast, the ESII would not account for these conditions, since the differences in speaking style would not be represented by the speech-noise probe used in this model.

The reduced amount of predicted improvement for the clearly spoken vs. conversational speech might result from an inconsistency in the way intelligibility was evaluated by Payton *et al.* (1994) in their measurement procedure and by the model. In the behavioral data, the percentage of correct score represented the number of correctly identified keywords of a given sentence, i.e., not the intelligibility of the sentence as a whole. In contrast, the model predictions reflected the average intelligibility of the whole sentence, including non-keywords, which could bias the model towards smaller amounts of improvement. Nevertheless, the agreement between the predictions and the data support the hypothesis that enhancing the speech modulation should lead to a speech intelligibility improvement. The improvement observed in the data from Payton *et al.* (1994) may well represent an upper limit of the potential benefit provided by an artificial speech modulation enhancement approach.

The benefit of processing either the noise or the speech alone before mixing appeared to be modest compared to other approaches of speech intelligibility enhancement, where the mixture of speech and noise is modified. For example, Wójcicki and Loizou (2012) demonstrated that discarding noise-dominated modulation-domain spectral components based on an $\mathrm{SNR}_{\mathrm{env}}$-like metric led to improvements of intelligibility of up to 13 dB SRT. Their approach was fundamentally different from the one considered in the present

[1] The parameters of the ideal observer in the sEPSM (Jørgensen *et al.*, 2013) were adjusted to minimize the root-mean-square error of the predictions and the data in the conditions with conversational speech. The complete parameter set was $k = 0.42$, $q = 0.5$, $\sigma_S = 5$, and $m = 8000$.

study, since they focused on noise-removal rather than modifying the unmixed signals to optimize intelligibility. The benefit of the present approach is that the enhancement could, in principle, be obtained by a filter-like operation, without having to estimate the noise component. In practice, the enhancement could be performed as some pre-processing of the speech signal that could optimize intelligibility, before it is mixed with noise. However, noise-removal could, in principle, also be achieved with the present modulation processing framework, by modifying the setup of the target spectrogram (Part A of Fig. 6.2) such that it produces a target that is a noise-reduced version of the noisy speech mixture.

6.5 Summary and conclusions

The effect of manipulating the SNR_{env} on speech intelligibility was investigated by systematically varying the modulation power of either the speech or the noise before mixing the two components. Improvements of the SRT for normal-hearing listeners were obtained in conditions where the modulation power of the noise was modified, leaving the speech untouched. Predictions from the sEPSM accounted well for this effect, supporting the hypothesis that the SNR_{env} is indicative of speech intelligibility. However, a large detriment of the SRT was obtained when the speech modulation power was modified and the noise left untouched, which could not be accounted for by the sEPSM. This discrepancy might be explained by distortions introduced by the modulation processing, which were not reflected in the SNR_{env} metric, thus representing a limitation of the modeling framework.

7

Perceptual and model-based evaluation of speech intelligibility in mobile telecommunication systems[‖]

In the development process of modern telecommunication systems, such as mobile phones, it is common practice to use computer models to objectively evaluate the transmission quality of the system, instead of time-consuming perceptual listening tests. Such models have typically focused on the quality of the transmitted speech, while little or no attention has been provided to speech intelligibility. The present study investigated to what extent two state-of-the art speech intelligibility models could predict the intelligibility of noisy speech transmitted through mobile phones. Sentences from the Danish Dantale II speech material (Wagener *et al.*, 2003) were mixed with three different kinds of background noise, transmitted through three different mobile phones, and recorded at the receiver via a local network simulator. The speech intelligibility of the transmitted sentences was assessed by six normal-hearing listeners and model predictions were compared to the perceptual data. A good correspondence between the measured data and the predictions from one of the models was found in the conditions with speech-shaped background noise, whereas deviations were observed in conditions with “Traffic” and “Pub” noise. Overall, the results suggest that speech intelligibility models inspired by auditory signal processing can be useful for the objective evaluation of speech transmission through mobile phones.

[‖] This chapter is based on Jørgensen *et al.* (2014a).

7.1 Introduction

Speech transmission through modern telecommunication devices, such as mobile phones, has traditionally been evaluated mainly in terms of speech quality. One reason for the focus on speech quality, rather than speech intelligibility, might be that mobile phone communication typically occurs at signal-to-noise ratios (SNR) where speech intelligibility is not compromised. In situations with very poor intelligibility, people generally either terminate the conversation or put themselves in a position where the SNR is increased. However, in a mobile telecommunication situation, such as when the talker is situated in a car or a train, the transmitted speech signal can, in fact, be largely affect by the presence of background noise surrounding the talker, which is difficult to move away from. The listener at the receiving end might have difficulty understanding the transmitted speech, even if he or she was situated in a quiet room. In such conditions, estimates of speech intelligibility could provide an additional important performance parameter of the mobile telecommunication system, in addition to a speech quality evaluation.

The perceptual evaluation of speech quality and intelligibility, based on listening tests, has generally been considered to be more reliable than objective prediction models for complex environmental or conversational conditions (Gierlich and Kettler, 2006). However, perceptual evaluations are very time consuming, requiring the use of several listeners and hours of testing per condition. Thus, a reliable objective tool for predicting human speech intelligibility performance in a given acoustic condition would be very valuable in the development process of new telecommunication systems. The objective evaluation method recommended by the International Telecommunication Union is the "perceptual evaluation of speech quality" (PESQ; ITU-T P.862, 2001) model. This model has been shown to correlate well with the perceptual quality ratings quantified by the mean opinion score (MOS; ITU-T P.800, 1996). Several studies have attempted to use PESQ for speech intelligibility prediction in addition to speech quality (e.g., Liu *et al.*, 2006; Beerends *et al.*, 2009; Kondo, 2011). For example, Liu *et al.* (2006) reported a good correlation of the PESQ metric to perceptual scores. However, the listening test applied by Liu *et al.* (2006) was based on a rating of listening effort, rather than a quantitative measure of the number of speech items that had been

understood. It is therefore unclear to what extent the perceptual data reported by Liu *et al.* (2006) reflect aspects of quality rather than intelligibility. In contrast, Beerends *et al.* (2009) found a poor correlation of PESQ predictions to perceptual intelligibility scores of consonant-vowel-consonant stimuli. The PESQ score was generally found to be at floor level for all conditions. Similarly, Kondo (2011) observed a floor effect when comparing PESQ predictions to perceptual speech intelligibility obtained using a dynamic rhyme test. Thus, the PESQ metric as provided in ITU-T P.862 (2001) appears to be inappropriate for speech intelligibility prediction and it might, therefore, be advantageous to consider prediction models designed for speech intelligibility.

Several objective speech intelligibility metrics have been proposed, the first of which and most widely used one is the articulation index (AI; ANSI S3.5, 1969). The AI essentially measures the amount of audible speech, based on information about the SNR and the human hearing threshold in several frequency bands that cover the overall frequency spectrum of speech. The AI model was later modified to include aspects of auditory masking and hearing loss, and was published as a new standard under the name "speech intelligibility index" (SII; ANSI S3.5, 1997). However, the SII is inherently limited to conditions with a stationary background noise, and fails in more realistic conditions with fluctuating noise backgrounds, such as speech babble in a cafeteria. To overcome this limitation, Rhebergen and Versfeld (2005) presented the extended speech intelligibility index (ESII), which could account for the speech intelligibility in conditions with various types of fluctuating interferers. However, the AI, SII, and ESII models were designed to account for the effects of reduced bandwidth of early telecommunication systems, and cannot directly be applied to conditions where the noisy speech mixture has been processed by nonlinear noise reduction and low bit-rate coding used in modern telecommunication systems.

Recently, an alternative metric for predicting speech intelligibility was proposed (Jørgensen and Dau, 2011), which is based on a measure of the SNR in the envelope domain ($\mathrm{SNR_{env}}$). The $\mathrm{SNR_{env}}$ measures the relative strength of the speech and the noise envelope fluctuations, inspired by earlier work indicating that such a metric might be related to speech intelligibility (Dubbelboer and Houtgast, 2008). The metric is computed using the speech-based envelope power spectrum model (sEPSM; Jørgensen

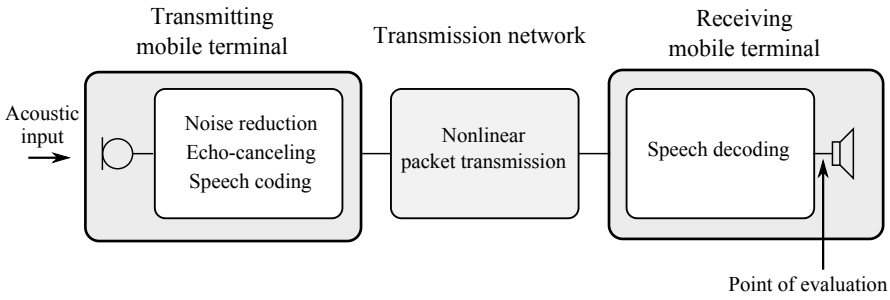


Figure 7.1: Illustration of the basic elements of a transmission chain in a modern telecommunication system, such as a mobile phone. The acoustic input is picked up by the microphone of the transmitting device and typically processed by noise reduction and echo-canceling algorithms. The signal is transformed via a speech coder and the resulting coefficients are transmitted using a digital transmission network. Finally, the signal is picked up by the receiving device, decoded, and played back via the devices' loudspeaker. The vertical arrow indicates the point of evaluation considered in the present study.

and Dau, 2011; Jørgensen *et al.*, 2013), which effectively mimics key aspects of human auditory signal processing. The key difference between the ESII and the sEPSM is that the sEPSM analyses the temporal modulation characteristics of the noisy speech and those of the background noise, using the concept of a modulation filterbank. The model was demonstrated to account for conditions with various stationary and fluctuating interferers as well as the effects of reverberation and nonlinear noise reduction. The sEPSM framework might therefore be applicable for predicting the speech intelligibility performance of mobile phones.

Several aspects of the mobile phone signal transmission chain may influence the quality and intelligibility of transmitted speech. These include the surrounding acoustic environment, the microphone characteristics, the digital signal processing in the phone (typically including noise reduction and echo-canceling) as well as the digital transmission network (Gierlich and Kettler, 2006). Figure 7.1 illustrates the very basic elements of such a transmission chain. Several studies have focused on the evaluation of different aspects of the transmission chain, such as speech coding/decoding algorithms (McLoughlin *et al.*, 2002), echo-canceling (Hänsler, 1994), noise reduction algorithms (Westerlund *et al.*, 2005), and effects of network transmission (ETSI TR 102 251, 2003). However, in order to include the combined effect of the various nonlinear steps

of the transmission chain, an evaluation should consider the transmission chain as a whole, from the acoustic input through the transmitting phone to the signal picked up by the receiver (Gierlich and Kettler, 2006). In the present study, the transmitted signal was evaluated at the point just after it had been decoded by the receiving phone, as indicated by the vertical arrow on Fig. 7.1. Thus, the loudspeaker characteristics or other components of the signal processing that were specific to the receiving phone were not included. The aim was to obtain a realistic simulation of a telecommunication situation, by including the combined effects of the acoustic environment surrounding the transmitting mobile phone, the signal processing specific to the phone, and the digital transmission network.

The present investigation consisted of two parts. First, speech intelligibility was measured perceptually in six normal hearing-listeners considering different noisy conditions and the effects of the transmission through three commercially available mobile phones. The aim was to investigate to what extent speech intelligibility performance varied across the mobile phones in the different noise environments. Second, predictions using the sEPSM and the ESII were obtained and compared to the measured data, in order to investigate to what extent these speech intelligibility models could be used for the objective evaluation of speech transmission through the mobile phones.

7.2 Method

7.2.1 Stimuli

Sentences from the Danish Dantale II speech corpus (Wagener *et al.*, 2003), spoken by a female talker, were used as the target speech stimuli. The speech corpus consists of 160 different five-word sentences with a grammatically correct structure (name + verb + numeral + adjective + object), but with unpredictable meaning, which allows reusing the sentences as many times as required for the same listener. The sentences were acoustically mixed with noise and recorded digitally using a setup according to ETSI EG 202 396-1 (2008), which simulated a one-way mobile telecommunication situation,

as illustrated in Fig. 7.2. The setup consisted of a Brüel & Kjær 4128-D Head and torso simulator (HATS) geometrically centered between four loudspeakers in a standardized listening room (IEC 268-13, 1985). The mobile phone under test was attached to the HATS using the Brüel & Kjær Handset Positioner Type 4606. Binaural noise signals were supplied to the loudspeakers such that the two left loudspeakers played back the left channel and the right speakers played back the right channel of the binaural signal. The four loudspeaker signals were decorrelated by introducing delays of 0, 11, 17, and 29 ms to the individual signals. The mouth speaker of the HATS played back the target speech. The frequency response of the loudspeakers as well as the mouth speaker of the HATS were equalized using the ear-microphones of the HATS as described in ETSI EG 202 396-1 (2008). The mobile phone under test was connected to a Rohde and Schwarz CMD 55 local network simulator (LNS) through a locally established GSM network and the electrical output signal from the LNS was recorded using the Matlab software package and an RME Fireface UCX soundcard.

Individual recordings of all 160 sentences mixed with each noise type at each SNR for each phone were stored digitally with a sampling rate of 44.1 kHz. In addition to the signal from the LNS, a reference signal from a Brüel & Kjær 4938 1/4-inch microphone positioned close to the input microphone of the mobile phone was recorded for all conditions. In this setup, and throughout this paper, the SNR was defined as the relative speech and noise levels measured with the reference microphone.

7.2.2 Perceptual evaluation

Conditions

Three commercially available mobile phones from three different manufactures were considered, denoted here as A, B, and C. The phones were released on the market in the years 2002, 2008, and 2010, respectively. Three types of background noise were considered: Dantale II Speech shaped noise (SSN; Wagener *et al.*, 2003), “Traffic”, and “Pub” noise from the noise database provided in ETSI EG 202 396-1 (2008). Moreover, two reference conditions with SSN background noise were considered: the broadband signal from the reference microphone, denoted as “Ref”, and the signal from

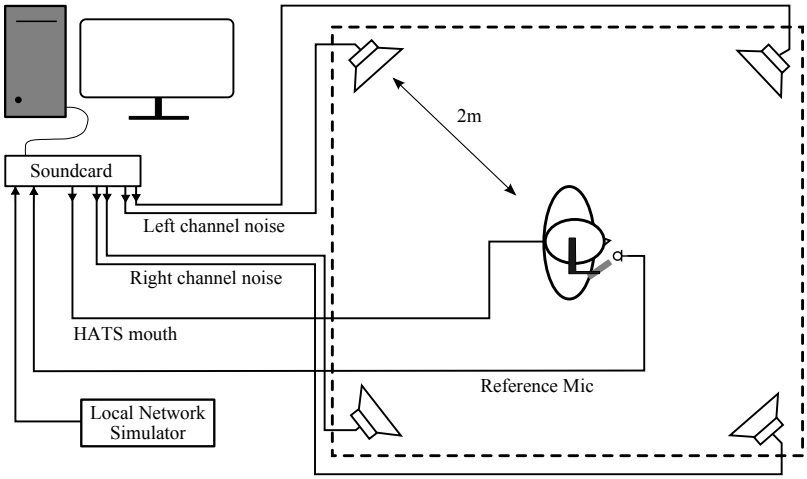


Figure 7.2: Sketch of the setup used to simulate a one-way telecommunication situation in the present study. See main text for further details.

the reference microphone filtered with the modified intermediate reference system (IRS; ITU-T P.830, 1996) transfer function (bandpass filter with −10-dB cut-off frequencies of 260 Hz and 3750 Hz), denoted as "Ref BP". All conditions were evaluated at four SNRs, which were chosen to cover the range from about 40% to 90% intelligibility in a given condition, based on pilot measurements. All conditions and SNRs were tested twice for each listener. In total, 2 x 44 (3 phones x 3 noises x 4 SNRs + 2 Ref conditions x 4 SNRs) conditions were evaluated per listener.

Apparatus and procedure

The perceptual evaluation was conducted in a double-walled sound insulated booth where the listener and the experimenter were seated. The sentences were presented to the listener's right ear via Sennheiser HD 650 headphones, which were equalized to have a flat frequency response at the ear reference point. The stimuli were filtered with the modified IRS receive transfer function to simulate a standard acoustic output of a receiving mobile phone. A noisy speech recording was presented to the listener,

with the noise starting 1.6 seconds before the sentence and ending 1.2 seconds after it. The stimulus was faded in and out using 100 ms Hanning ramps. The gain of the playback system was adjusted independently for the Ref and Ref BP conditions, such that sentences mixed with SSN at 0-dB SNR were presented at 70 dB SPL, respectively. Similarly, the gain was adjusted such that the sentences mixed with SSN at 0-dB SNR transmitted through phone A were presented at 70 dB SPL. This level was then kept fixed and used for all other mobile phone conditions. Therefore, the presentation level depended on the noise type, the phone type and the SNR condition. The task of the listeners was to repeat as many words of the presented sentences as possible. The listeners were allowed to guess. The experimenter noted the number of correct words on a computer screen hidden from the listener. The test was divided into three sessions of approx. two hours per listener. Two lists of 10 sentences were used for each phone and SNR configuration. Additional 80 sentences were used for training before the first and the second session, and 40 sentences before the third session, in order to familiarize the listeners with the different test conditions.

Listeners

Six native Danish listeners with audiometric thresholds below 20 dB HL from 250 Hz to 8 kHz participated in the evaluation and were paid for their participation. None of the listeners had previous experience with psychoacoustic testing.

7.2.3 Simulations

The model predictions were based on a subset of 30 of the 160 sentences for each condition. The simulations for each sentence were performed separately in all conditions, and the results were averaged across the 30 sentences. The stimuli were truncated in time such that the onsets for the speech and the noise were the same.

Separation of speech and noise components

The models considered here require separate access to the speech and the noise at the output of the transmission system. However, the separate transmission of the speech and the noise through the mobile phones would not reflect the situation considered in the perceptual evaluation, since the nonlinear behavior of the transmission system affected the noisy speech mixture (used in the perceptual evaluation) in a different way than the speech and the noise alone. Therefore, estimates of the noise and the speech components were obtained from the noisy speech mixture after mobile phone transmission. This was achieved using the method suggested by Hagerman and Olofsson (2004), briefly described in the following. Two input signals a_{in} and b_{in} were defined from the speech and noise signals as:

$$a_{in}(t) = s(t) + n(t) \text{ and} \tag{7.1}$$

$$b_{in}(t) = s(t) - n(t), \tag{7.2}$$

where s and n denote the speech and the noise components at the input to the mobile phone, respectively and t represents time. The separated speech and noise components, $s'_{out}(t)$ and $n'_{out}(t)$, were then obtained as:

$$s'_{out}(t) + \frac{1}{2}E_1(t) = \frac{1}{2}(a_{out}(t) + b_{out}(t)) \text{ and} \tag{7.3}$$

$$n'_{out}(t) + \frac{1}{2}E_2(t) = \frac{1}{2}(a_{out}(t) - b_{out}(t)), \tag{7.4}$$

where $a_{out}(t)$ and $b_{out}(t)$ denote the recorded signal mixtures. The error terms, $E_1(t)$ and $E_2(t)$ can be estimated using the methods described by Hagerman and Olofsson (2004), but were neglected here for simplicity.

sEPSM-based simulations

The sEPSM framework assumes *a priori* information about the noise component of a noisy speech mixture. Thus, for the sEPSM-predictions, the inputs to the model for a

given condition were the transmitted mixture, a_{out}, and the estimated noise component n'_{out}. From these inputs, the $\mathrm{SNR_{env}}$ was computed and converted to a probability of correct responses using an "ideal observer", as described in detail in Jørgensen *et al.* (2013). The ideal observer included four parameters that were taken from Table II in Jørgensen *et al.* (2013). However, the parameter k of the model's ideal observer was adjusted here such that the model predictions provided a good fit to the perceptual data in the Ref condition. All model parameters were then kept fixed in all other conditions.

ESII-based simulations

The ESII required access to the long-term level of the speech component and the waveform of the noise component at the output of the speech transmission system under test. A SSN signal was used as a probe for the speech component, following Rhebergen and Versfeld (2005) and Rhebergen *et al.* (2006). Specifically, the inputs to the ESII in a given condition were the transmitted SSN-speech probe with a root-mean-squared (rms) level equal to the rms level of $s'_{out}(t)$ and the separated noise component $n'_{out}(t)$. The output of the ESII was a SII-value between 0 and 1, which was transformed to a percentage of correct responses using the transfer function suggested by Amlani *et al.* (2002):

$$P_{correct} = 1 - 10^{(-SII+K)/Q}. \tag{7.5}$$

K and Q were free parameters with values that were obtained using a minimum-mean-square fit of the transfer function to the perceptually obtained percent correct values in the two reference conditions (Ref and Ref BP). Figure 7.3 shows the perceptual percentage of correct words as a function of the corresponding SII (circles), together with the obtained transfer function (solid line). This function was then used for all other conditions.

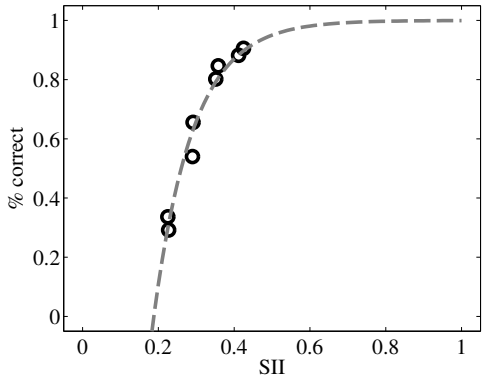


Figure 7.3: Psychoacoustic data plotted as a function of the corresponding SII (black circles), for the two reference conditions, and the corresponding ESII transfer function (dashed line).

7.3 Results

7.3.1 Perceptual data

The percentage of correct responses for the individual listeners (averaged across the two repeated presentations) are shown in Fig. 7.4, for the Ref condition (left panel) and the Ref BP condition (right panel). A psychometric function with two parameters (Wagener *et al.*, 2003), the slope (S_{50}) and the 50-% point (SRT_{50}), was fitted to the mean results of each individual listener in a given condition:

$$P(SNR) = \left[1 + e^{-4S_{50}\cdot(SNR - SRT_{50})}\right]^{-1}. \tag{7.6}$$

The parameters were averaged across the listeners to obtain a group-mean psychometric function. The obtained functions for the Ref and the Ref BP conditions are shown as solid black lines in Fig. 7.4. Using the same approach, Fig. 7.5 shows the psychometric functions for all conditions with SSN (left panel), Pub noise (middle panel), and Traffic noise (right panel). The shaded area surrounding each function represents one standard error of the function's parameters. The psychometric functions for the Ref and Ref BP conditions were clearly above those for the mobile phones for input SNRs above

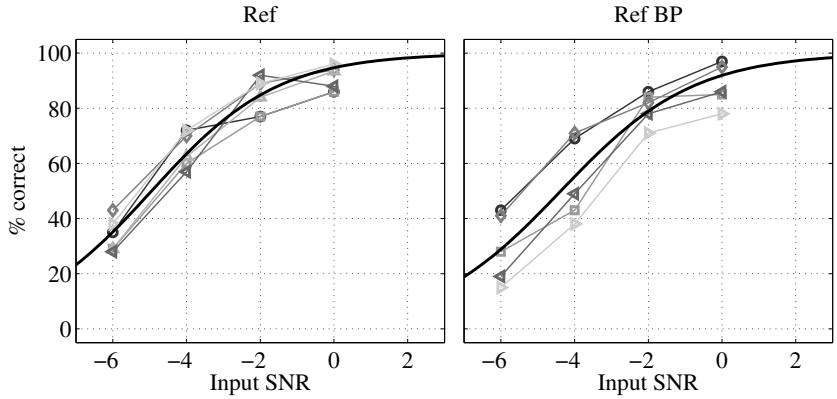


Figure 7.4: Percentage of correct responses for the individual listeners for the Ref (left panel) and the Ref BP conditions (right panel), and the corresponding group-mean psychometric functions (solid black line).

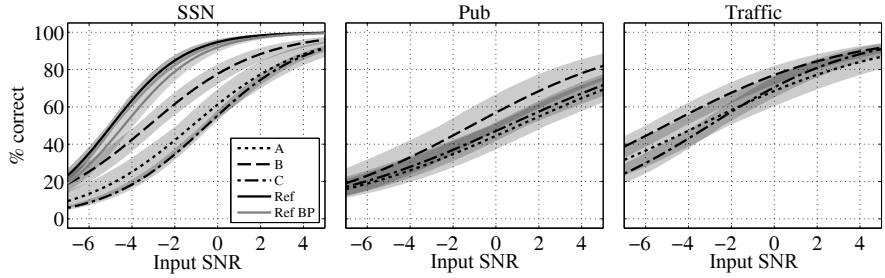


Figure 7.5: Psychometric functions for all the considered conditions with SSN (left panel), with Pub noise (middle panel), and with Traffic noise (right panel). The shaded area surrounding each function represents one standard error of the parameters of the psychometric function.

at and −4 dB. This demonstrates worse intelligibility for the phone conditions than for the Ref condition. Moreover, the slopes of the functions in the Ref and Ref BP conditions were steeper than those of the phone conditions. Specifically, for the SSN conditions (left panel), the psychometric functions for the three phones differed in their horizontal position. In contrast, the functions were much closer to each other in the conditions with Pub (middle panel) and Traffic (right panel) noise. This reflects a greater variability in the perceptual data for the Pub and Traffic-noise conditions than for the SSN conditions.

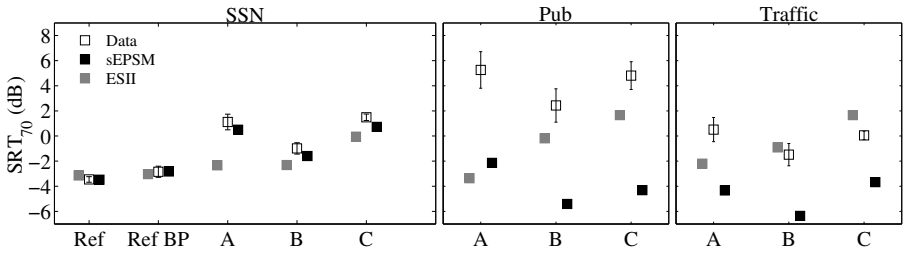


Figure 7.6: SRT_{70} obtained from the perceptual psychometric functions (open squares), for all the conditions with SSN (left panel), with Pub noise (middle panel), and with Traffic noise (right panel). The vertical bars denote one standard error. Predictions from the sEPSM is indicated by the filled black squares, and predictions from the ESII is indicated by the filled gray squares.

In order to quantify the differences between the obtained results, the following analysis was based on the speech reception thresholds determined from the psychometric functions as the SNR corresponding to 70% correct (SRT_{70}). The SRT_{70} was chosen because this point lied roughly in the middle of the measured range of percent correct. Figure 7.6 shows SRT_{70} (open squares) obtained from the perceptual data for the conditions with SSN (left panel), Pub noise (middle panel), and Traffic noise (right panel). The vertical bars denote one standard error.

For the SSN conditions, the SRT_{70} for the Ref condition was obtained at an SNR of -3.5 dB, followed by a slightly higher SRT_{70} for the Ref BP condition at -3.0 dB. The SRT_{70} for the three phones were obtained at higher SNRs than the reference conditions, with the lowest SRT_{70} (best intelligibility) obtained for phone B at an SNR of -0.9 dB, followed by phones A and C, which were both obtained at an SNR of 1.2 dB. Similar patterns of results were found for the Pub noise with a generally higher SRT_{70}, and for the Traffic noise with a generally lower SRT_{70}. For any given phone, the lowest SRT_{70} was obtained for the Traffic noise followed by the SSN and the Pub noise.

A one-way analysis of variance performed on the data for the three phones revealed that the SRT_{70} was significantly different across the phones for the SSN conditions ($p < 0.05$), but not for the Pub and the Traffic noises. A multiple comparison analysis with Bonferroni correction performed on all SSN conditions showed that the SRT_{70} for the Ref condition was significantly lower than the SRT_{70}s for all three phones

($p < 0.05$). Thus, the transmission chain led to significant decreases of the intelligibility for all phones, with respect to the reference condition. Moreover, the SRT_{70} for phone B was significantly lower than for phone C. Finally, the SRT_{70}s for the Ref and the Ref BP conditions were not significantly different, demonstrating that band limiting imposed by the IRS-filter had very little influence on the obtained intelligibility.

A two-way analysis of variance with phone-type and noise-type as factors and individual listener SRT_{70}s as observations revealed a significant effect of phone type ($F_{2,36} = 5.1$, $p = 0.0112$) and noise type ($F_{2,36} = 20.38$, $p < 0.0001$), which means that the listener group average cannot be considered equal across phones, nor across noises for a given phone. There was no significant interaction, which means that the two factors were additive, i.e., the pattern of SRTs for the phones was not significantly different across the three noise types.

7.3.2 Model predictions

Psychometric functions (Eqn. 7.6) were fitted to the predicted results in the corresponding conditions. Figure 7.6 shows predicted SRT_{70} obtained with the sEPSM (black filled squares) and the ESII (gray filled squares) for all considered conditions. The sEPSM accounted well for the data in the SSN conditions, with a root-mean-squared error (RMSE) of 0.51 dB. Moreover, the sEPSM predicted the same pattern of results as seen in the data for the Pub and Traffic noises, but with a vertical offset of about -8 dB for the Pub noise and -5 dB for the Traffic noise. The ESII accounted for the two reference conditions, and predicted the right trend for phones B and C. However, the ESII failed to predict the overall pattern of results across the three mobile phones seen in the data.

7.4 Discussion

7.4.1 Simulation of a realistic one-way communication situation

One aim of the present study was to investigate to what extent speech intelligibility performance varied across three commercially available mobile phones in realistic telecommunication situations. The situations included different (noisy) acoustic environments where the mobile phones transmitted speech via a local network simulator. This setup included occasional signal dropouts which were reflected in the corresponding recordings of the noisy speech as silent periods, sometimes removing entire words from the target sentence and, thus, potentially affecting the intelligibility. The dropouts were randomly occurring, as they would in a real telecommunication system and were, therefore, considered to contribute to the realism of the systems that were evaluated. Another element of the communication system simulation was that the gain of the playback system was fixed for the perceptual and objective evaluation procedures. This implied that the overall level of the presented stimuli varied across the phones, reflecting the effect of the noise reduction and the signal processing specific to a given phone. This level variation was considered as an inherent part of the system, contributing to the realism of the simulation, and was therefore not compensated for in the playback system.

One drawback of simulating the transmission chain as a whole was that it is difficult to disentangle which aspects of the transmission led to the differences in the performance across the phones. However, such a level of analytical detail was sacrificed in order to provide an overall impression of the performance of a given phone in a realistic communication situation.

7.4.2 Perceptual evaluation of speech intelligibility in modern telecommunication

The results from the perceptual evaluation showed that the conditions with SSN provided the lowest variability across the listeners and, thus, the highest sensitivity in

terms of detecting significant performance differences across the phones. The pattern of results for the different phones were very similar across the noise types, suggesting that it may be sufficient to use SSN for assessing the relative performance across the devices. This is in agreement with a recent study by Wong *et al.* (2012), which concluded that the SSN was a good predictor of the intelligibility in more realistic background noises.

The maximal difference in SRT_{70} obtained for the three phones with the SSN amounted to about 2 dB, which was reflected in a difference of the psychometric functions for phone C and B of 25% at an SNR of 0 dB. Such a difference might be crucial in an everyday communication situation, and motivates the use of intelligibility as a critical performance parameter of mobile phones. The measured psychometric functions also revealed a difference of 40% between phone C and the Ref condition at an SNR of 0 dB. This shows that there should be potential for improved speech intelligibility in future mobile phone systems.

7.4.3 Performance of the prediction models

Another aim of the present study was to investigate to what extent an objective speech intelligibility model could replace a panel of human listeners for evaluating the performance of a modern mobile phone. The sEPSM framework accounted qualitatively for the differences in SRT_{70} across the three different mobile phones considered in this study, in all three considered background noise conditions, i.e. the predicted rank order across the phones was in agreement with the data. Moreover, the model accounted quantitatively for the data obtained with SSN showing an RMS error of only 0.51 dB. This suggested that this metric might be useful for performance discrimination in the development process of mobile phones. However, the predictions for the Pub and Traffic noises were offset vertically by -8 and -5 dB, respectively, implying that the model was too sensitive in those conditions. For all mobile phones, the sEPSM predicted a better intelligibility for the Pub noise compared to SSN, which is in contrast to the data. Thus, the predicted rank order across the three noises was *not* in agreement with the data. This discrepancy might be related to possible errors in the separation process of the speech and noise signals after the transmission through the mobile phone.

Further analysis of the model's internal representation of the stimuli may reveal the source of the prediction errors, which was beyond the scope of this study.

The predictions from the ESII were in agreement with the data for the Ref and Ref BP conditions, but failed to accurately account for the different SRT_{70} across the mobile phones. Therefore, this metric did not seem appropriate for performance discrimination of the mobile phones considered in the present study.

7.5 Summary and conclusions

Speech intelligibility performance was assessed for three commercially available mobile phones in three different noise scenarios. The results showed clear differences in speech intelligibility across the three mobile phones of up to 2 dB of the SRT_{70} and across the three noise conditions. Speech transmission via the considered mobile phones led to a decreased speech intelligibility, relative to the reference condition without speech transmission through the mobile phones, demonstrating a substantial room for improvement with respect to optimization of the mobile phone signal processing for speech intelligibility. A good correspondence between the predictions from the sEPSM and the measured perceptual intelligibility was obtained in the conditions with a stationary speech-shaped background noise. This suggests that the sEPSM might be useful for objective speech intelligibility prediction in the development and evaluation of mobile phone algorithms.

Acknowledgments

The authors thank Wookeun Song, Lars Birger Nielsen, and Claus Blaabjerg from Brüel & Kjær for providing measurement equipment and support for recording of the stimuli. This work was supported financially by the Danish Sound Technology Network, and the Danish hearing aid companies, Widex, Oticon and GN Resound.

8

General discussion

8.1 Summary of main results

This thesis described the development and evaluation of a new model framework for speech intelligibility prediction. The key hypothesis of the approach, outlined in Chapter 2, was that the signal-to-noise envelope power ratio (SNR_{env}) may be used as the "decision metric" for predicting speech intelligibility in noise. This was inspired by earlier work by Dubbelboer and Houtgast (2008), where a similar metric was used to predict speech intelligibility, based on a simple modulated pure-tone signal. Moreover, the model included modulation frequency selective processing, motivated by several studies demonstrating that such a concept is crucial in order to account for perceptual data obtained with amplitude modulated stimuli (e.g., Dau *et al.*, 1997a,b; Kohlrausch *et al.*, 2000; Dau *et al.*, 1999; Ewert and Dau, 2000; Ewert *et al.*, 2002).

Chapter 2 described the details of the new framework, where the SNR_{env} was measured from the envelope power spectra of noisy speech and noise alone at the output of a modulation bandpass filterbank following peripheral auditory filtering. The envelope power of the stimuli were normalized with the DC[1] component of the envelope power spectrum, implying that the SNR_{env} effectively reflects the amount of modulation masking by the noise. This approach was earlier considered in the framework of the envelope power spectrum model (EPSM; Ewert and Dau, 2000), where the signal-to-noise metric in the envelope domain was shown to predict amplitude

[1] The DC component of the envelope power spectrum is mathematically equivalent to the squared time-averaged amplitude of the envelope, from which the spectrum is calculated.

modulation detection and masking data. The new model was therefore denoted the speech-based EPSM (sEPSM).

While the original EPSM assumed a simple decision criterion relating SNR_{env} to detectability, the sEPSM introduced the concept of an "ideal observer", which included parameters that were assumed to be related to a specific speech material. This enabled the model to account for differences in the response set size and redundancy of the considered speech stimuli, assuming that the model had access to the noisy speech mixture and the noise alone. The sEPSM was demonstrated to account for effects of background noise, reverberation, and spectral subtraction processing on the speech reception thresholds in normal-hearing listeners. It outperformed the classical speech transmission index (STI) model, which failed to account for the effect of spectral subtraction. It was suggested that the success of the sEPSM was caused by the estimation of the noise envelope power, as a part of the SNR_{env} decision metric. The noise envelope power, and thus the amount of modulation masking, increased after the spectral subtraction processing, which could account for the reduced intelligibility obtained in these conditions. In contrast, modulation masking is not reflected in the STI since it measures the reduction in modulation power of the speech, which is equivalent to assuming a constant modulation noise floor. Overall, the results of Chapter 2 indicated that the modulation filterbank and the decision metric based on the SNR_{env} were crucial components in the model for successful speech intelligibility prediction. However, it was unclear whether the model would account for distortions that affect mainly the spectral content (but not the temporal modulation content) of the speech, such as phase jitter.

Chapter 3 investigated the role of spectro-temporal modulation processing for speech perception by comparing predictions from the sEPSM from Chapter 2 to two modified versions. One version, denoted as 2D-sEPSM, assumed a two-dimensional modulation filtering stage, inspired by earlier work of Elhilali *et al.* (2003) that suggested such a stage might be crucial to account for the effect of phase jitter distortion on speech intelligibility. The other model version kept the one-dimensional (temporal) modulation filtering as in the original sEPSM, but introduced an across audio-frequency mechanism that evaluated the variability of the modulation filter outputs across the (peripheral)

audio-frequency axis, inspired by models of comodulation masking release (CMR; Buus, 1985; Piechowiak *et al.*, 2007; Dau *et al.*, 2013). This version was denoted sEPSM^{X}. The role of the decision metric was studied by comparing predictions from the different sEPSM versions, which all considered the SNR_{env} as the decision metric, to predictions from the spectro-temporal modulation index (STMI; Elhilali *et al.*, 2003) and the STI, which both incorporate a measure of the integrity of the speech modulations (via the the modulation transfer function, MTF) as the decision metric. The predictions from all models were compared to measured data in conditions with reverberation, spectral subtraction processing and phase jitter distortion. The results showed that only the models based on the SNR_{env} decision metric could account for the whole data set, whereas the MTF-based models failed to account for spectral subtraction processing. Moreover, only those models that considered some across-channel processing could account for the effects caused by the phase jitter distortion. The sEPSM^{X} provided the best predictions (in terms of least root-mean-squared errors), suggesting that the 2D modulation filterbank processing as assumed in the STMI (and implemented in the 2D-sEPSM) might not be crucial for speech intelligibility prediction.

The models presented in Chapter 2 and 3 considered the SNR_{env} based on *long-term* estimates of the envelope power. These models would therefore not be able to account for conditions in which speech is mixed with a non-stationary interferer. In such conditions, speech intelligibility has been demonstrated to be better than in conditions with a stationary interferer; an effect referred to as speech masking release. A non-stationary interferer would actually contain a dominant long-term envelope power, which would lead to more modulation masking compared to the situation with a stationary noise, i.e., the opposite to a speech masking release. To overcome this limitation, Chapter 4 presented a multi-resolution version of the sEPSM (mr-sEPSM), inspired by the short-term estimation of the SII introduced with the Extended SII model (ESII; Rhebergen and Versfeld, 2005). Instead of estimating the envelope power from the long-term envelope power spectra, the multi-resolution version estimated the envelope power from the time-domain envelope waveforms at the output of the modulation filterbank. The outputs from the modulation filterbank were segmented with durations that were inversely related to the center frequencies of a the modulation

filters, i.e., long windows were used at low modulation center frequencies and short windows at high modulation center frequencies. This allowed the model to capture information reflected in both slowly varying low-frequency modulations as well as fast fluctuating high-frequency modulations in the noisy speech. This effectively allowed the model to evaluate the SNR_{env} in the temporal dips of the fluctuating interferer, and thus to capture the potential release from modulation masking in those time instances. The multi-resolution sEPSM was demonstrated to account for the effect of various stationary and fluctuating interferers, as well as for conditions with reverberation and spectral subtraction, demonstrating an improvement over the original (long-term) sEPSM and the ESII, which failed for reverberation and spectral subtraction processing. The combination of the multi-resolution model with the across-channel variance mechanism described in Chapter 3 should allow the model to also account for the conditions with phase jitter.

An analysis of the internal representation of the stimuli in the mr-sEPSM revealed that the release from masking predicted by the model resulted from a greater SNR_{env} in the modulation frequency channels centered at modulation frequencies above about 16 Hz. This suggested that relatively high-rate modulations were important for the masking release effect. This was investigated further in Chapter 5 where predictions from the multi-resolution sEPSM were compared to data from Christensen *et al.* (2013), which considered the effect of attenuating high-rate modulations from the target and the interfering talker on speech intelligibility. Their data showed that envelope low-pass filtering with a cutoff frequency of 30 Hz led to lower speech intelligibility than with a cutoff frequency of 300 Hz. The model accounted well for this effect, supporting the model-based finding from Chapter 4 that high-rate modulations can be important for speech intelligibility in conditions with speech interferers. Chapter 5 also compared model predictions in conditions with speech mixed with different types of fluctuating interferers to corresponding data from Festen and Plomp (1990). The sEPSM predictions were in good agreement with the data in the conditions with fluctuating noise and interfering talkers, except for one condition where the interferer consisted of time-reversed speech from the same talker as the target. In this case, the intelligibility data may have been influenced by the listeners' ability to segregate the target and the interfering speaker. Such an effect is not included in the sEPSM

framework, which "only" represents the amount of modulation masking. The additional masking observed in the data may be related to "informational masking", which might reflect an interference in the processing of carrier (or fundamental frequency) information that is not reflected in the model.

Chapter 6 challenged further the general application of the sEPSM framework for speech intelligibility prediction. The hypothesized relationship between SNR_{env} and speech intelligibility was investigated using stimuli that were manipulated to have different value of SNR_{env}, while keeping the same acoustic long-term power SNR. This was done by modifying the modulation power of speech before mixing it with unprocessed stationary speech-shaped noise or by modifying the modulation power of the noise before mixing it with the unprocessed speech. The predictions were in agreement with the corresponding perceptual data in the conditions with processed noise. However, the model failed to capture the effects of manipulating the speech (alone) on the intelligibility. The model predicted that speech intelligibility should increase when the modulation power of the speech was enhanced; in contrast, a reduction of intelligibility was observed in the data. An analysis of the stimuli suggested that the processed speech stimuli were distorted, which might explain the reduced intelligibility. The distortions were introduced by the modulation processing framework, because the enhanced modulation content of the speech was not restricted to follow the natural modulation of the clean speech, but might have contained modulation components not related to speech at all. This was not captured in the SNR_{env}-metric, because the sEPSM considers all modulations in the noisy speech input that are not related to the noise as belonging to the speech - and thus as beneficial. The conditions with processed speech alone thus demonstrated a clear limitation of this model framework. However, it was argued that this failure is not a principal limitation of the SNR_{env} concept, but rather a limitation of the model used to estimate the SNR_{env}. This was supported by simulations of conditions where speech modulations were enhanced in a natural way by instructing the talker to pronounce the words very clearly (Payton *et al.*, 1994). The model predicted a greater intelligibility for the clearly spoken speech compared to conversationally spoken speech, in agreement with the perceptual data. The results of Chapter 6, thus, supported the hypothesis of a relationship between SNR_{env} and speech intelligibility.

Chapter 7 evaluated the usefulness of the sEPSM to predict speech intelligibility in a practical scenario. The aim of this study was to evaluate the intelligibility of speech transmitted through modern telecommunication systems, such as mobile phones, and to assess whether the sEPSM could be used to predict the intelligibility performance of the phones. Predictions were compared to perceptual data obtained with three different mobile phones in conditions with three different types of background noise. The perceptual data showed that the intelligibility became worse when the speech was transmitted through any of the phones, compared to the reference condition without any phones. The intelligibility of the transmitted speech with a given type of background noise varied across the different phones, most likely reflecting differences in the noise reduction processing in the three phones. The variation across the phones was similar for all types of background noise, suggesting that one of the noise types would be sufficient to discriminate the performance across phones. The conditions with stationary speech-shaped noise led to the lowest variability across the listeners, and thus the most reliable data. The sEPSM accounted for the overall pattern of results across the three phones for all noise types, and closely matched the data in the conditions with speech-shaped noise. However, the model generally predicted too good performance for the two other noise types. This discrepancy might be related to the way the noise component of the transmitted noisy speech was estimated, but further analysis is required to clarify this effect in the model. Overall, the results from Chapter 7 demonstrated that the model was applicable as an objective tool for speech intelligibility assessment in the development process of mobile phones.

8.2 The role of modulation frequency selectivity for speech intelligibility

One main hypothesis in the sEPSM framework is that modulation frequency selectivity is crucial for speech intelligibility. However, several speech intelligibility models like the SII and ESII, have been presented in the literature that do not include this concept, and nevertheless, they were shown to perform well in various conditions. This raises the question of whether the modulation frequency selectivity is a crucial component

of a speech intelligibility model. One critical condition that appears to discriminate the models with and without modulation frequency selectivity is reverberation. This condition is interesting because reverberation represents a highly ecologically relevant condition, which is very common in everyday environments. The predictions from the "modulation low-pass" version of the multi-resolution sEPSM without a modulation filterbank (gray triangles in Fig. 4.6) could account for the data in the conditions with reverberation and spectral subtraction, indicating that modulation frequency selectivity might not be crucial for these conditions. This was in contrast to the analysis described in Chapter 2, which suggested that the modulation filterbank actually was important for the model to account for spectral subtraction. However, the model version without the modulation filterbank did, in fact, include modulation frequency selective processing in the form of the multi-resolution segmentation, because the different segment durations preserved modulation frequencies down to $1/t_d$ Hz, where t_d was the duration of the segment. In contrast, as demonstrated in Chapter 4, a model without any modulation frequency selectivity process, such as the ESII, fails to account for the effect of reverberation, even though it represents a fine temporal resolution. Similarly, the short time objective intelligibility (STOI; Taal *et al.*, 2011) model, which considers a modulation-based decision metric but no modulation frequency selectivity, fails to account for the effects of reverberation. Thus, in such a condition, the failure of these models might not be due to the decision metric, but because of the lack of modulation frequency selectivity. Overall, the concept of modulation frequency selectivity, either in the form of a modulation bandpass filterbank or a multi-resolution segmentation, was crucial for the sEPSM to account for reverberation. This is consistent with the STI model, which also included modulation frequency selectivity and can account for effects of reverberation. The results of this thesis therefore further supports the hypothesis that modulation frequency selective processing is crucial for speech intelligibility in some conditions. The concept of modulation frequency selectivity has also been supported by results from physiological studies in animals (Langner and Schreiner, 1988; Schreiner and Urbas, 1988), neural imaging in humans (Xiang *et al.*, 2013), auditory modeling (e.g., Dau *et al.*, 1997a; Verhey *et al.*, 1999; Derleth and Dau, 2000; Jepsen *et al.*, 2008), sound texture synthesis (McDermott and Simoncelli, 2011), and auditory scene analysis (Elhilali *et al.*, 2009).

8.3 Limitations of the modeling framework

One general limitation of the proposed model framework is related to the decision metric: the $\mathrm{SNR}_{\mathrm{env}}$ metric is a concept based on a the envelope power ratio of the speech and the noise components, estimated from a noisy speech mixture. As such, the $\mathrm{SNR}_{\mathrm{env}}$ is, in principle, only valid in conditions where the noise component of the noisy speech mixture can be specified. This is the case when the noise is additive, i.e., when the noise and the speech components are independent signals from which the envelope power can be computed. In contrast, the $\mathrm{SNR}_{\mathrm{env}}$ would be difficult to determine in conditions where, for example, an ideal binary mask (IBM) was applied to a noisy speech mixture with a very low SNR (e.g., Kjems *et al.*, 2009). In such a case, the noise and speech components, and thus their envelope power, are not independent because the noise is modulated with the spectro-temporal structure of the speech. However, even though the $\mathrm{SNR}_{\mathrm{env}}$-metric has limitations, the concept of an intelligibility decision metric defined in the envelope domain might still be valid and could be defined in a different way than using the envelope power. For example, Goldsworthy and Greenberg (2004) suggested to compute an SNR in the envelope domain from the cross-correlation coefficient r of the clean and the processed envelope signals:

$$SNR_{\mathrm{env}}{}^{*} = \frac{r}{1-r}. \tag{8.1}$$

In this case, r (the numerator) would represent the amount of target speech information, and $1-r$ (the denominator) would describe the amount of non-target speech information. Similar approaches have also been considered recently where r was used as the decision metric directly (Christiansen *et al.*, 2010; Taal *et al.*, 2011). The STOI model (Taal *et al.*, 2011) was demonstrated to successfully predict speech intelligibility in conditions with IBM-processed speech and in conditions with single-channel noise reduction. This approach may also account for some of the conditions with processed speech considered in this thesis, such as the conditions with modulation processed speech considered in Chapter 6, where the envelope-power based metric failed. One benefit of the correlation-based metric is that it includes information about the envelope phase, which is not captured by the envelope power metric. A decision metric based on

correlation will therefore be sensitive to within-channel phase differences between the processed and the original envelope, which might be crucial in some cases, such as with the phase jitter distortion considered in Chapter 3. This suggests that an $\mathrm{SNR}_{\mathrm{env}}$-metric based on a cross-correlation between the clean speech and the processed (noisy) speech envelope, via a template matching process, might present a more successful strategy in some cases. One drawback of a cross-correlation based decision metric is, however, that 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. Thus, the envelope power might still be an attractive concept because it allows to characterize intelligibility based on a physically measurable quantity, namely the envelope power spectrum of the noise and the mixture, without assuming a specific template.

8.4 Perspectives

The sEPSM framework is effectively a monaural model of speech intelligibility. However, a general model should be applicable in realistic scenarios with spatially distributed targets and interferers. Typically, speech intelligibility is greater when the target and the interferer are spatially separated, compared to the situation where they are colocated, an effect referred to as “spatial release from masking” (e.g, Plomp and Mimpen, 1981; Bronkhorst and Plomp, 1992; Hawley *et al.*, 2004). Spatial release from masking may be regarded as consisting of two components, one arising from improvements in SNR at the “better ear” closest to the target, mainly caused by headshadow, and another due to “binaural interaction”, possibly facilitated by interaural time differences between the target and the interferer (Durlach, 1963; Culling and Summerfield, 1995; Breebaart *et al.*, 2001). The sEPSM could be extended to separately evaluate the input signals to the left and right ear, whereby the better ear effect could be modeled. In order to account for effects of binaural interaction, the sEPSM should be combined with a model of binaural processing, such as the equalization

cancelation model (EC; Durlach, 1963). Such a framework might allow the evaluation of the relative contribution of the better-ear effect and the binaural interaction on the spatial release from masking in various target and interferer configurations. Several model approaches have already considered speech intelligibility in spatial scenarios (e.g., Beutelmann and Brand, 2006; Beutelmann *et al.*, 2010; Lavandier and Culling, 2010; Lavandier *et al.*, 2012). However, these models are typically based on the SII or STI that have been demonstrated to be limited in certain monaural conditions. Thus, it would be interesting to investigate if a binaural model combined with the SNR_{env} decision metric could provide a more general framework for prediction speech intelligibility in realistic spatial scenarios.

So far, the sEPSM has relied on *a priori* information about the noise component of the noisy speech mixture in order to predict speech intelligibility. However, this might be difficult to achieve outside the laboratory, and a "blind" estimation of the SNR_{env} would therefore be valuable for speech intelligibility evaluation in practical applications. This might be obtained by estimating the noise and speech sources using principles from computational auditory scene analysis (Wang, 2005). For example, binaural cues and pitch information might help to estimate an ideal binary modulation mask, which would directly correspond to a blind estimation of SNR_{env} that could be used to predict speech intelligibility. A corresponding approach was considered by Wójcicki and Loizou (2012) where the aim was to perform noise reduction in the modulation domain, rather than to predict intelligibility. They demonstrated that the blind estimation of the noise led to qualitatively similar results as using *a priori* information. It would be interesting to investigate if a similar approach could be applied in the sEPSM framework for speech intelligibility prediction.

The current version of the sEPSM assumes that one of the sources of reduced speech intelligibility is modulation masking. Moreover, it assumes that the listener can always distinguish between the target and the interferer. However, in some conditions of speech mixed with a speech interferer, the listener may confuse target and interfering talkers. This confusion has often been referred to as "informational masking" (e.g., Brungart, 2001). Informational masking has been suggested to be associated with the listeners' inability to perceptually segregate the target and the interfering sources

(Shinn-Cunningham, 2008), which is related to the concept of auditory streaming, where acoustic events are grouped into distinct perceptual objects depending on their spectro-temporal properties (Bregman, 1990). Recently, models of computational auditory scene analysis have been shown to account for a one-versus-two stream percept of synthetic signals, as well as the transition between the two percepts (Elhilali *et al.*, 2009; Shamma *et al.*, 2011; Christiansen *et al.*, 2014). A similar modeling concept might be applied to determine the perceptual grouping of one or more speech and noise sources, whereby such a model could be combined with the sEPSM to predict aspects of informational masking in relation to auditory stream segregation.

References

Amlani, A. M., Punch, J. L., and Ching, T. Y. C. (**2002**). "Methods and applications of the audibility index in hearing aid selection and fitting", Trends Amplify **6**, 81–129.

ANSI S3.5 (**1969**). "American national standard methods for the calculation of the articulation index", (Acoustical Society of America, New York) .

ANSI S3.5 (**1997**). "Methods for the calculation of the speech intelligibility index", (Acoustical Society of America, New York) .

Bacon, S. P. and Grantham, D. W. (**1989**). "Modulation masking: Effects of modulation frequency, depth, and phase", The Journal of the Acoustical Society of America **85**, 2575–2580.

Beerends, J. G., Van Buuren, R., Van Vugt, J., and Verhave, J. (**2009**). "Objective speech intelligibility measurement on the basis of natural speech in combination with perceptual modeling", J. Audio Eng. Soc. **57**, 299–308.

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). "Enhancement of speech corrupted by acoustic noise", Proc. IEEE Int. Conf. Acoust., Speech, Signal Proces. (ICASSP) **4**, 208–211.

Beutelmann, R. and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners.", J. Acoust. Soc. Am. **120**, 331–342.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model", J. Acoust. Soc. Am. **127**, 2479–2497.

Boll, S. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust., Speech, Signal Process. **27**, 113–120.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**). "Binaural pro- cessing model based on contralateral inhibition. i. model structure", J. Acoust. Soc. Am. **110**, 1074–1089.

Bregman, A. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press).

Bronkhorst, A. W. and Plomp, R. (**1992**). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing", J. Acoust. Soc. Am. **92**, 3132–3139.

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

Buss, E., Whittle, L. N., Grose, J. H., and Hall, J. W. I. (**2009**). "Masking release for words in amplitude-modulated noise as a function of modulation rate and task", J. Acoust. Soc. Am. **126**, 269–280.

Buus, S. (**1985**). "Release from masking caused by envelope fluctuations", J. Acoust. Soc. Am. **78**, 1958–1965.

Chabot-Leclerc, A., Jøergensen, S., and Dau, T. (**2014**). "The importance of auditory spectro-temporal modulation filtering and decision metric for predicting speech intelligibility", J. Acoust. Soc. Am. (Submitted).

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility", J. Acoust. Soc. Am. **106**, 2719–2732.

Chi, T., Ru, P., and Shamma, S. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds", J. Acoust. Soc. Am. **118**, 887–906.

Christensen, C., MacDonald, E., and Dau, T. (**2013**). "Contribution of envelope periodicity to release from speech-on-speech masking", J. Acoust. Soc. Am. **134**, 2197–2204.

Christensen, C. L. (**2009**). *Odeon Room Acoustics Program, Version 10.1, User Manual, Industrial, Auditorium and Combined Editions* (ODEON A/S, Kgs. Lyngby, Denmark), URL http://www.odeon.dk/pdf/OdeonManual.pdf.

Christiansen, C., Pedersen, M. S., and Dau, T. (**2010**). "Prediction of speech intelligibility based on an auditory preprocessing model", Speech Commun. **52**, 678–692.

Christiansen, S. K., Jepsen, M. L., and Dau, T. (**2014**). "Effects of tonotopicity, adaptation, modulation tuning and temporal coherence in auditory stream segregation of pure-tones sequences", J. Acoust. Soc. Am. **135**, 323–333.

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

Culling, J. F. and Summerfield, Q. (**1995**). "Perceptual segregation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay", J. Acoust. Soc. Am. **98**, 785–797.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**a). "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers", J. Acoust. Soc. Am. **102**, 2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**b). "Modeling auditory processing of amplitude modulation: Ii. spectral and temporal integration", J. Acoust. Soc. Am. **102**, 2906–2919.

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

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers", J. Acoust. Soc. Am. **106**, 2752–2760.

Decorsière, R. (**2013**). *Spectrogram inversion and potential application for hearing research*, chapter 3, 35–59 (Department of Electrical Engineering, Technical University of Denmark, Denmark).

Depireux, D., Simon, J. Z., Klein, D. J., and Shamma, S. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex", J. Neurophysiol. **85**, 1220–1234.

Derleth, R. P. and Dau, T. (**2000**). "On the role of envelope fluctuation processing in spectral masking", J. Acoust. Soc. Am. **108**, 285–296.

Dirks, D. D., Wilson, R. H., and Bower, D. R. (**1969**). "Effect of pulsed masking on selected speech materials", J. Acoust. Soc. Am. **46**, 898–906.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility", J. Acoust. Soc. Am. **97**, 585–592.

Drullman, R., Festen, J., and Plomp, R. (**1994**a). "Effect of reducing slow temporal modulations on speech reception", J. Acoust. Soc. Am. **95**, 2670–2680.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**b). "Effect of temporal envelope smearing on speech reception", J. Acoust. Soc. Am. **95**, 1053–1064.

Dubbelboer, F. and Houtgast, T. (**2007**). "A detailed study on the effects of noise on speech intelligibility", J. Acoust. Soc. Am. **122**, 2865–2871.

Dubbelboer, F. and Houtgast, T. (**2008**). "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility", J. Acoust. Soc. Am. **124**, 3937–3946.

Duquesnoy, A. J. (**1983**). "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons", J. Acoust. Soc. Am. **74**, 739–743.

Duquesnoy, A. J. and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis", J. Acoust. Soc. Am. **68**, 537–544.

Durlach, N. I. (**1963**). "Equalization and cancellation model of binaural unmasking-level differences", J. Acoust. Soc. Am. **35**, 1206–1218.

Egan, J., Lindner, W., and McFadden, D. (**1969**). "Masking-level differences and the form of the psychometric function", Percept. Psychophys. **6**, 209–215.

Egan, J. P. (**1965**). “Masking-level differences as a function of interaural disparities in intensity of signal and of noise”, J. Acoust. Soc. Am. **38**, 1043–1049.

Egan, J. P., Clarke, F. R., and Carterette, E. C. (**1956**). “On the transmission and confirmation of messages in noise”, J. Acoust. Soc. Am. **28**, 536–550.

Elhilali, M.and Xiang, J., Shamma, S. A., and Simon, J. Z. (**2009**). “Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene”, PLoS Biol. **7**, e1000129.

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). “A spectro-temporal modulation index (stmi) for assessment of speech intelligibility”, Speech Commun. **41**, 331–348.

Elliott, T. M. and Theunissen, F. E. (**2009**). “The modulation transfer function for speech intelligibility”, PLoS Comput Biol **5**, pp e1000302.

ETSI EG 202 396-1 (**2008**). “Speech processing, transmission and quality aspects (stq); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database”, (European Telecommunications Standards Institute ,Sophia Antipolis Cedex, FRANCE) .

ETSI TR 102 251 (**2003**). “Speech processing, transmission and quality aspects (stq); anonymous test report from 2nd speech quality test event 2002”, (European Telecommunications Standards Institute) .

Ewert, S. and Dau, T. (**2000**). “Characterizing frequency selectivity for envelope fluctuations”, J. Acoust. Soc. Am. **108**, 1181–1196.

Ewert, S. and Dau, T. (**2004**). “External and internal limitations in amplitude-modulation processing”, J. Acoust. Soc. Am. **116**, 478–490.

Ewert, S. D., Verhey, J. L., and Dau, T. (**2002**). “Spectro-temporal processing in the envelope-frequency domain”, J. Acoust. Soc. Am. **112**, 2921–2931.

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

Fletcher, H. (**1940**). “Auditory patterns”, Rev. Mod. Phys. **12**, 47–65.

Fletcher, H. and Galt, R. (**1950**). “Perception of speech and its relation to telephony”, J. Acoust. Soc. Am. **22**, 89–151.

Fogerty, D. (**2011**). “Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure”, J. Acoust. Soc. Am. **129**, 977–988.

French, N. and Steinberg, J. (**1947**). “Factors governing intelligibility of speech sounds”, J. Acoust. Soc. Am. **19**, 90–119.

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (**2006**). “Masking release for consonant features in temporally fluctuating background noise”, Hearing Res. **211**, 74–84.

Füllgrabe, C., Stone, M. A., and Moore, B. C. J. (**2009**). “Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task (l).”, J. Acoust. Soc. Am. **125**, 1277–1280.

Gallun, F. and Souza, P. (**2008**). “Exploring the role of the modulation spectrum in phoneme recognition”, Ear Hearing **29**, 800–813.

George, E. L. J., Festen, J. M., and Houtgast, T. (**2006**). “Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners”, J. Acoust. Soc. Am. **120**, 2295–2311.

George, E. L. J., Festen, J. M., and Houtgast, T. (**2008**). “The combined effects of reverberation and nonstationary noise on sentence intelligibility”, J. Acoust. Soc. Am. **124**, 1269–1277.

Ghitza, O. (**2001**). “On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception.”, J. Acoust. Soc. Am. **110**, 1628–1640.

Gierlich, H. and Kettler, F. (**2006**). “Advanced speech quality testing of modern telecommunication equipment: An overview”, Speech Commun. **86**, 1327–1340.

Glasberg, B. R. and Moore, B. C. J. (**1990**). “Derivation of auditory filter shapes from notched-noise data”, Hear. Res. **47**, 103–138.

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

Green, D. M. and Birdsall, T. G. (**1964**). "The effect of vocabulary size", In *Signal Detection and Recognition by Human Observers*, edited by John A. Swets (John Wiley & Sons, New York), 609–619.

Green, D. M. and Swets, J. A. (**1988**). *Signal Detection Theory and Psychophysics*, 238–239 (Penisula Publishing, Los Altos California).

Griffin, D. and Lim, J. (**1984**). "Signal reconstruction from short-time fourier transform magnitude", IEEE Trans. Acoust., Speech, and Signal Proc. **32**, 236–243.

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

Hagerman, B. and Olofsson, A. (**2004**). "A method to measure the effect of noise reduction algorithms using simultaneous speech and noise", Acta. Acoust. United Ac. **90**, 356–361.

Hänsler, E. (**1994**). "The hands-free telephone problem: an annotated bibliography update", Annales des Télécommun. **49**, 360–367.

Hawley, M., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer", J. Acoust. Soc. Am. **115**, 883–843.

Hirsh, I., Reynolds, E., and Joseph, M. (**1954**). "Intelligibility of different speech materials", J. Acoust. Soc. Am. **26**, 530–538.

Hohmann, V. and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility", J. Acoust. Soc. Am. **97**, 1191–1195.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an international speech test signal (ists)", Int. J. Audiol. **49**, 891–903.

Houtgast, T. (**1989**). “Frequency selectivity in amplitude-modulation detection”, J. Acoust. Soc. Am. **85**, 1676–1680.

Houtgast, T., Steeneken, H., and Plomp, R. (**1980**). “Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics”, Acustica **46**, 60–72.

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

Houtgast, T. and Steeneken, H. J. M. (**1973**). “The modulation transfer function in room acoustics as a predictor of speech intelligibility”, Acustica **28**, 66–73.

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

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

IEC 268-13 (**1985**). “Sound system equipment - part 13: Listening tests on loudspeakers”, (International Electrotechnical Commission, Geneva, Switzerland)) .

IEC60268-16 (**2003**). “Sound system equipment - part 16: Objective rating of speech intelligibility by speech transmission index”, (International Electrotechnical Commission, Geneva, Switzerland) .

ISO389-7 (**2005**). “ISO389-7 acoustics - reference zero for the calibration of audiometric equipment- part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions.”, (International Standardization Organization, Geneva) .

ITU-T P.800 (**1996**). “Methods for subjective determination of transmission quality - series p: Telephone transmission quality; methods for objective and subjective assessment of quality”, (International Telecommunication Union, Geneva, Switzerland) .

ITU-T P.830 (**1996**). “Subjective performance assessment of telephone-band and wideband digital codecs - telephone transmission quality; methods for objective and subjective assessment of quality”, (International Telecommunication Union, Geneva, Switzerland) .

ITU-T P.862 (**2001**). “Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, (International Telecommunication Union, Geneva, Switzerland) .

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). “A computational model of human auditory signal processing and perception”, J. Acoust. Soc. Am. **124**, 422–438.

Jørgensen, S., Cubick, J., and Dau, T. (**2014**a). “Perceptual and model-based evaluation of speech intelligibility in mobile telecommunication systems”, Speech Commun. submitted.

Jørgensen, S. and Dau, T. (**2011**). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing”, J. Acoust. Soc. Am. **130**, 1475–1487.

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

Jørgensen, S., Decorsiere, R., and Dau, T. (**2014**b). “Effects of manipulating the envelope signal-to-noise ratio on speech intelligibility.”, J. Acoust. Soc. Am. submitted.

Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). “A multi-resolution envelope power based model for speech intelligibility”, J. Acoust. Soc. Am. **134**, 436–446.

Jürgens, T. and Brand, T. (**2009**). “Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model”, J. Acoust. Soc. Am. **126**, 2635–2648.

Keidser, G. (**1993**). “Normative data in quiet and in noise for "dantale"-a danish speech material”, Scand. Audiol. **22**, 231–236.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech", J. Acoust. Soc. Am. **126**, 1415–1426.

Kleinschmidt, M. (**2002**). "Methods for capturing spectro-temporal modulations in automatic speech recognition", Acta Acustica united with Acustica **88**, 416–422.

Knudsen, V. O. (**1929**). "The hearing of speech in auditoriums", J. Acoust. Soc. Am. **1**, 56–82.

Kohlrausch, A., Fassel, R., and Dau, T. (**2000**). "The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers", J. Acoust. Soc. Am. **108**, 723–734.

Kondo, K. (**2011**). "Estimation of speech intelligibility using perceptual speech quality scores", in *Speech and Language Technologies*, Edited by Prof. Ivo Ipsic. (InTech, Rijeka, Croatia) Chapter 8. 155–74.

Kowalski, N., Depireux, D. A., and Shamma, S. (**1996**). "Analysis of dynamic spectra in ferret primary auditory cortex: I. characteristics of single unit responses tomoving ripple spectra", J. Neurophysiol. **76**, 3503–3523.

Kruger, B. and Kruger, F. M. (**1997**). "Speech audiometry in the usa", In *Speech Audiometry*, edited by Michael Martin (Whurr Publishers Ltd, London) , Chap. 12, 233–277.

Kryter, K. (**1960**). "Speech bandwidth compression through spectrum selection", J. Acoust. Soc. Am. **32**, 547–556.

Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index", J. Acoust. Soc. Am. **34**, 1689–1697.

Langner, G. and Schreiner, C. E. (**1988**). "Periodicity coding in the inferior colliculus of the cat. i. neuronal mechanisms", J. Neurophysiol. **60**, 1799–1822.

Lavandier, M. and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms.", J. Acoust. Soc. Am. **127**, 387–399.

Lavandier, M., Jelfs, S., and Culling, J. F. (**2012**). “Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources.”, J. Acoust. Soc. Am. **131**, 218–231.

Le Roux, J., Kameoka, H., Ono, N., and Sagayama, S. (**2010**). “Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency”, Proc. 13th Int. Conference on Digital Audio Effects, 397–403.

Lim, J. (**1978**). “Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise”, IEEE T. Acoust. Speech **26**, 471–472.

Liu, W. M., Jellyman, K. A., Mason, J. S. D., and Evans, N. W. D. (**2006**). “Assessment of objective quality measures for speech intelligibility estimation”, Proc. IEEE Int. Conf. Acoust., Speech, Signal Proces. **1**, 1225–1228.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). “Evaluation of a noise reduction method-comparison between observed scores and scores predicted from STI”, Scand. Audiol. Suppl. 38 **22**, 50–55.

Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (**1990**). “Prediction of speech intelligibility of non-linearly processed speech”, Acta Oto-Laryngol., Suppl. 469 190–195.

Lyon, R. and Shamma, S. (**1996**). “Auditory representations of timbre and pitch”, in *Auditory Computation* edited by H. Hawkins, E. T. McMullen, A. Popper, and R. Fay (Springer Verlag, New York) 221–270.

McDermott, J. H. and Simoncelli, E. P. (**2011**). “Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis”, Neuron **71**, 926–940.

McLoughlin, I., Ding, Z., and Tan, E. (**2002**). “Intelligibility evaluation of gsm coder for mandarin speech using cdrt”, Speech Commun. **38**, 161–165.

Mickes, L., Wixted, J. T., and Wais, P. E. (**2007**). “A direct test of the unequal-variance signal detection model of recognition memory”, Psychon B. Rev. **14**, 858–865.

Miller, G. and Licklider, J. (**1950**). “Intelligibility of interrupted speech”, J. Acoust. Soc. Am. **22**, 167–173.

Miller, G. A. (**1947**). “The masking of speech”, Psychol. Bull. **44**, 105–129.

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). “The intelligibility of speech as a function of the context of the test material”, J. Exp. Psychol. **41**, 329–335.

Moore, B. C., Glasberg, B. R., and Schooneveldt, G. P. (**1990**). “Across-channel masking and comodulation masking release”, J. Acoust. Soc. Am. **87**, 1683–1694.

Moore, B. C. J. and Glasberg, B. R. (**1983**). “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns”, J. Acoust. Soc. Am. **74**, 750–753.

Müsch, H. and Buus, S. (**2001**). “Using statistical decision theory to predict speech intelligibility. i. model structure”, J. Acoust. Soc. Am. **109**, 2896–2909.

Nemala, S., Patil, K., and Elhilali, M. (**2013**). “A multistream feature framework based on bandpass modulation filtering for robust speech recognition”, IEEE Trans. Audio Speech Lang. Processing **21**, 416 –426.

Nielsen, J. B. and Dau, T. (**2009**). “Development of a danish speech intelligibility test”, Int. J. Audiol. **48**, 729–741.

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

Noordhoek, I. M. and Drullman, R. (**1997**). “Effect of reducing temporal intensity modulations on sentence intelligibility”, J. Acoust. Soc. Am. **101**, 498–502.

Paliwal, K., Wójcicki, K., and Schwerin, B. (**2010**). “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain”, Speech Commun. **52**, 450–475.

Pavlovic, C. (**1982**). “Derivation of primary parameters and procedures for use in speech intelligibility predictions”, J. Acoust. Soc. Am. **82**, 413–422.

Payton, K. and Braida, L. (**1999**). “A method to determine the speech transmission index from speech waveforms”, J. Acoust. Soc. Am. **106**, 3637–3648.

Payton, K., Uchanskbi, R. . M., and Braida, L. (**1994**). “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing”, J. Acoust. Soc. Am. **95**, 1581–1592.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). “Speaking clearly for the hard of hearing 1. intelligibility differences between clear and conversational speech”, Journal of Speech and Hearing Research **28**, 96–103.

Piechowiak, T., Ewert, S. D., and Dau, T. (**2007**). “Modeling comodulation masking release using an equalization-cancellation mechanism”, J. Acoust. Soc. Am. **121**, 2111–2126.

Plomp, R. and Mimpen, A. M. (**1979**). “Improving the reliability of testing the speech-reception threshold for sentences”, Audiology **18**, 43–52.

Plomp, R. and Mimpen, A. M. (**1981**). “Effect of the orientation of the speakerŠs head and the azimuth of a noise source on the speech-reception threshold for sentences”, Acoustica **48**, 325–328.

Rennies, J., Brand, T., and Kollmeier, B. (**2011**). “Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet”, J. Acoust. Soc. Am. **130**, 2999–3012.

Rhebergen, K. S. and Versfeld, N. J. (**2005**). “A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners”, J. Acoust. Soc. Am. **117**, 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise”, J. Acoust. Soc. Am. **120**, 3988–3997.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2009**). “The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise”, J. Acoust. Soc. Am. **126**, 3236–3245.

Robles, L. and Ruggero, M. A. (**2001**). “Mechanics of the mammalian cochlea”, Physiol. Rev. **81**, 1305–1352.

Sarampalis, A., Kalluri, S., and Edwards, B.and Hafter, E. (**2009**). "Objective measures of listening effort: Effects of background noise and noise reduction", J. Speech Lang. Hear. Res. **52**, 1230–1240.

Schimmel, S. and Atlas, L. (**2005**). "Coherent envelope detection for modulation filtering of speech", Proc. IEEE Int. Conf. Acoust., Speech, Signal Proces. (ICASSP) **1**, 221–224.

Schreiner, C. and Urbas, V. (**1988**). "Representation of amplitude modulation in the auditory cortex of the cat. ii. comparison between cortical fields", Hearing Res. **32**, 49–65.

Shamma, S., Chadwik, R., Wilbur, W., Morrish, K., and Rinzel, J. (**1986**). "A biophysical model of cochlear processing: Intensity dependence of pure tone response.", J. Acoust. Soc. Am. **80**, 133–145.

Shamma, S., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis", Trends in Neurosciences **34**, 114–123.

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention", Trends in Cognitive Sciences **12**, 182–186.

Smeds, K., Wolters, F., Nilsson, A., Båsjö, S., and Hertzman, S. (**2011**). "Predictive measures of the intelligibility of speech processed by noise reduction algorithms", In *3rd International Symposium on Auditory and Audiological Research: Speech Perception and Auditory Disorders*, editedby Torsten Dau, Morten L. Jepsen, Torben Poulsen, and Jakob C. Dalsgaard (The Danavox Jubilee Foundation, Denmark), pp. 355–362.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception.", Nature **416**, 87–90.

So, S. and Paliwal, K. (**2011**). "Modulation-domain kalman filtering for single-channel speech enhancement", Speech Commun. **53**, 818–829.

Steeneken, H. J. M. and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Am. **67**, 318–326.

Steinberg, J. C. (**1929**). “Effects of distortion upon the recognition of speech sounds”, J. Acoust. Soc. Am. **1**, 121–137.

Strelchyk, O. and Dau, T. (**2009**). “Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing”, J. Acoust. Soc. Am. 3328–3345.

Sun, D. and Smith III, J. (**2012**). “Estimating a signal from a magnitude spectrogram via convex optimization”, 133rd Conv. Audio Eng. Soc. arXiv:1209.2076.

Taal, C., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). “An algorithm for intelligibility prediction of time-frequency weighted noisy speech”, IEEE Trans. Audio Speech Lang. Processing **19**, 2125–2136.

Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (**1997**). “Speech enhancement based on audible noise suppression”, IEEE Trans. Speech Audio Processing **5**, 497–514.

van de Par, S. and Kohlrausch, A. (**1998**). “Comparison of monaural (cmr) and binaural (bmld) masking release”, J. Acoust. Soc. Am. **103**, 1573–1579.

Verhey, J. L., Dau, T., and Kollmeier, B. (**1999**). “Within-channel cues in comodulation masking release (cmr): Experiments and model predictions using a modulation-filterbank model”, J. Acoust. Soc. Am. **106**, 2733–2745.

Viemeister, N. F. (**1979**). “Temporal modulation transfer functions based upon modulation thresholds”, J. Acoust. Soc. Am. **66**, 1364–1380.

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

Wang, D. L. (**2005**). “On ideal binary masks as the computational goal of auditory scene analysis”, In *Speech Separation by Humans and Machines*, edited by Divenyi, D. (Kluwer Academic, Norwell, MA, USA) Ch. 12 pp 181–197.

Wang, K. and S., S. (**1994**). “Self-normalization and noise-robustness in early auditory representations”, IEEE Trans. Audio Speech Process. **2**, 421–435.

Warren, R. M., Bashford, Jr., J. A., and Lenz, P. W. (**2005**). “Intelligibilities of 1-octave rectangular bands spanning the speech spectrum when heard separately and paired”, J. Acoust. Soc. Am. **118**, 3261–3266.

Westerlund, N., Dahl, M., and Claesson, I. (**2005**). “Speech enhancement for personal communication using an adaptive gain equalizer”, Speech Commun. **85**, 1089–1101.

Wójcicki, K. K. and Loizou, P. C. (**2012**). “Channel selection in the modulation domain for improved speech intelligibility in noise.”, J. Acoust. Soc. Am. **131**, 2904–2913.

Wong, L., Ng., E., and Soli, S. (**2012**). “Characterization of speech understanding in various types of noise”, J. Acoust. Soc. Am. **132**, 2642–2651.

Xiang, J., Poeppel, D., and Simon, J. Z. (**2013**). “Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations”, J. Acoust. Soc. Am. **133**, EL7–EL12.

Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). “Relative contributions of spectral and temporal cues for phoneme recognition”, J. Acoust. Soc. Am. **117**, 3255–3267.

Zhu X., Beauregard G.T., W. L. (**2006**). “Real-time iterative spectrum inversion with look-ahead”, IEEE Int. Conf. Multimedia Expo, 229–232.

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.

The end.

To be continued...

Speech intelligibility depends on various factors related to auditory processing and sound perception, as well as on the acoustic properties of the sound that enters the ear. However, a clear understanding of speech perception in complex acoustic conditions and, in particular, a quantitative description of the involved auditory processes, provides a major challenge in speech and hearing research. This thesis presents a computational model that attempts to predict speech intelligibility performance of normal-hearing listeners in various adverse conditions. The model combines the concept of modulation-frequency selectivity in the auditory processing of sound, with a decision metric for intelligibility that is based on the signal-to-noise envelope power ratio (SNR_{env}). The proposed speech-based envelope power spectrum model (sEPSM) is evaluated in conditions with stationary and fluctuating interferers, reverberation and various types of processed noisy speech, including noisy speech transmitted through mobile phones.
Overall, the results of this thesis support the hypothesis that the SNR_{env} is a powerful objective metric for speech intelligibility prediction. Moreover, the findings suggest that the concept of modulation-frequency selective processing in the auditory system is crucial for human speech perception.

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