Prediction and Optimization of Speech Intelligibility in Adverse Conditions

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op vrijdag 25 januari 2013 om 10:00 uur
door Cornelis (Cees) Harm TAAL
Ingenieur, Media & Kennis Technologie
geboren te Hoogeveen.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. R. L. Lagendijk

copromotor: Dr. ir. R. Heusdens

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. R. L. Lagendijk,	Technische Universiteit Delft, promotor
Dr. ir. R. Heusdens,	Technische Universiteit Delft, copromotor
Prof. dr. ir. W. A. Dreschler,	Academisch Medisch Centrum, Amsterdam
Prof. dr. ir. G. J. T. Leus,	Technische Universiteit Delft
Prof. dr. P. A. Naylor,	Imperial College, London, United Kingdom
Prof. dr. ir. L. J. van Vliet,	Technische Universiteit Delft
Prof. dr. J. Wouters	Katholieke Universiteit Leuven, België

Dr. ir. R. C. Hendriks heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

The work described in this thesis was financially supported by STW and Oticon A/S.

Prediction and Optimization of Speech Intelligibility in Adverse Conditions

Summary

In digital speech-communication systems like mobile phones, public address systems and hearing aids, conveying the message is one of the most important goals. This can be challenging since the intelligibility of the speech may be harmed at various stages before, during and after the transmission process from sender to receiver. Causes which create such adverse conditions include background noise, an unreliable internet connection during a Skype conversation or a hearing impairment of the receiver. To overcome this, many speech-communication systems include speech processing algorithms to compensate for these signal degradations like noise reduction. To determine the effect on speech intelligibility of these signal processing based solutions, the speech signal has to be evaluated by means of a listening test with human listeners. However, such tests are costly and time consuming. As an alternative, reliable and fast machine-driven intelligibility predictors are of interest, since they might replace listening tests, at least in some stages of the algorithm development process.

Two important issues exist with current intelligibility predictors. (1) Many of these methods cannot reliably predict the effect of more advanced nonlinear signal processing algorithms on speech intelligibility. (2) Typically, these measures are based on very complex auditory models or use average statistics of minutes of running speech, which makes it difficult on how to design new (real-time) speech processing solutions in an optimal manner given such a measure. To this end we propose several new measures which show good prediction results with the intelligibility of nonlinear processed speech. The newly proposed measures are of a low computational complexity and mathematically tractable which make them suitable for optimization of new signal processing solutions which aim for improving speech intelligibility.

An important stage in many speech intelligibility predictors is the use of an auditory model. In the first part of this thesis we show that a general sophisticated auditory model can be greatly simplified, while preserving accurate predictions of psycho-acoustic listening experiments. The resulting simplified model facilitates the computation of analytic expressions for masking thresholds while advanced state-of-the-art models typically need computationally demanding adaptive procedures. Its mathematical properties are successfully exploited by optimally redistributing speech energy such that the speech intelligibility is improved when played back in a noisy environment without modifying the

signal-to-noise ratio.

In the design process of new intelligibility predictors we first analyse the strengths and weaknesses of existing measures. In total, 17 different measures are evaluated for intelligibility prediction of time-frequency weighted noisy speech. We show that, despite high correlation with the listening test scores, several measures cannot predict the difference in intelligibility before and after signal processing. We explain that a state-of-the-art measure was not able to predict the intelligibility due to its sensitivity to the DFT-phase components. Issues with existing measures for intelligibility prediction are highlighted and a general normalization procedure as a pre-processing step is proposed which improves their correlation with speech intelligibility.

We propose a new short-time intelligibility measure (STOI) which shows high correlation with the intelligibility of time-frequency weighted noisy speech, including noise-reduced and vocoded speech. In general, STOI shows better correlation with speech intelligibility compared to five other state-of-the-art objective intelligibility models. One important difference between STOI and other measures is its analysis length which is in the order of a few hundreds of ms rather than complete sentences or 20-30 ms length frames. Due to the simple structure of STOI we show in the final part of this thesis that the measure can be interpreted as a mathematical norm, which is applied in the channel-selection technique with cochlear-implant simulations. Several intelligibility predictors indicate large intelligibility improvements with the new method based on STOI compared to a peak-picking algorithm.

Contents

Chapter 1

Introduction

An important goal in speech-communication systems is to record, transmit and playback a speech signal such that it is correctly understood by the receiver. Examples can be found in the field of electronic broadcasting systems like telephony, radio and television but also in public address systems such as used in airports or train stations. This scenario also applies to devices which compensate for a hearing-loss, like hearing aids and cochlear implants, where conveying the message is one of the main goals. In addition to the requirement that the speech should be intelligible, another important aspect is the quality of the speech signal. Preferably, the speech signal should sound pleasant and natural such that the speech signal can be understood with a similar effort as with clean undistorted speech. Moreover, additional information like speaker identity and emotion, should also be correctly interpreted by the listener.

For all these aforementioned examples of speech-communication systems, the speech can get degraded at various stages before, during and after the transmission process from sender to receiver. Typically, these degradations will have a negative effect on speech intelligibility and/or speech quality. Important causes for speech deterioration are environmental factors like background noise at both sides of the communication channel, e.g., the sound of a passing train or a noisy crowd in a restaurant [Fren 47]. But also properties of the communication channel can lead to a speech degradation. For instance, due to the bandwidth constraint in the case of a normal telephone line [Flet 50], the lack of high frequencies in the speech signal might confuse the listener with what was actually said by the speaker. Also for internet-based telephone calls, like a Skype conversation, gaps in the speech signal may occur due to lost digital packets when transmitted over an unreliable and slow internet connection, which negatively affects the intelligibility [Mill 50]. Another important factor is the hearing impairment of the receiver, which might be compensated for by means of an hearing aid or cochlear implant. Unfortunately, even with hearing-aids, environmental sources like background noise and reverberation may still have a strong negative impact on the perceived speech signal in both speech quality and intelligibility [Chun 04].

A large area in the field of speech processing works on developing algorithms which try to compensate for these types of speech degradations, see [Dell 93a, Vary 06] for an overview. These algorithms aim for a restoration of speech quality and/or intelligibility or make the speech signals more robust when transmitted in such adverse conditions. For instance, one straightforward way for improving speech understanding in a noisy environment would be to detect the amount of noise and amplify the speech signal accordingly, before playback in the noisy environment. However, when the noise is already present in the recorded speech signal, more advanced solutions exist like sophisticated noise-reduction algorithms which try to estimate the underlying clean speech given the noisy observation [Loiz 06]. Moreover, relevant for the aforementioned telephony applications, a bandwidth extension method can be used to restore the high frequencies [Iser 08] or lost packets can be revealed with a so-called packet-loss concealment algorithm [Perk 98]. Hearing-loss compensation

algorithms as used in hearing aids, typically restore speech understanding by amplification and try to compress the dynamic range of the speech such that the speech is audible again for the user [Dill 01].

To determine the perceptual consequences of these speech degradations and the effects of the proposed signal processing based solutions, the speech signal has to be evaluated by means of a listening test with human listeners. Many types of listening tests exist where, for example, the change in speech quality or speech intelligibility due to a certain type of processing can be evaluated [Gran 08]. Although a listening test can lead to a judgment as observed by the intended group of users, such tests are costly and time consuming. This may be specifically an issue in the research and development process of a new speech processing algorithm. For example, imagine the situation where one proposes a new noise-reduction algorithm, which has one free parameter controlling the 'amount of noise reduction'. Increasing this parameter leads to more noise-reduction and may therefore improve the speech quality for the end-user. Unfortunately, too much noise reduction also leads to removing unwanted components of the target speech, which can result in a decrease in speech intelligibility. In order to determine the optimal setting for this simple example with respect to an average group of users, many listening tests have to be performed with different settings of the free parameter, which is undesirable. In fact, the truth is that many speech processing algorithms have tens or even hundreds of free parameters, rather than one. Moreover, many other aspects which may change the algorithm behavior should also be taken into account, like the noise type, the amount of noise, speaker gender and speaker type. This variety of possibilities makes it impossible to optimize new algorithms solely based on listening tests.

As an alternative, reliable and fast machine-driven evaluation methods are of interest, since they might replace listening tests, at least in some stages of the algorithm development process. Typically, such an evaluation method is implemented as a computer program and acts like an artificial listener based on some general model of the auditory system. As an output, one number is generally reported as a function of one or more different inputs, like the clean and distorted speech signals. These predictive models aim for a high correlation with the results from an actual listening test like the average percentage of correctly understood words or some kind of speech quality-based ranking score. For the given example of the noise-reduction algorithm, such a measure could be used to explore the space of possible parameter settings in a fast way.

More interestingly would be to use these predictive models for providing hints on how to process the speech in a more fundamental manner, rather than naive *offline* optimization of free parameters of already designed speech processing algorithms. In other words, new signal processing strategies could be designed in an optimal way given such a machine-driven evaluation method. The latter approach will be referred to as *online* optimization in this thesis. One requirement of the predictive model for designing these types of optimal signal processing solutions is that the measure should be transparent and easy

to understand, rather than a 'black box' approach which would be more appropriate for offline parameter optimization. This transparency is needed in order to provide hints on how to process the signal in an optimal manner, e.g., by means of deriving closed form solutions. An additional important property is that the measure should be of low computational complexity such that it can be used in DSP processors, e.g., as in digital hearing instruments.

Many methods exist to predict either the speech quality [Quac 88] or speech intelligibility [Koll 08] for a given type of speech degradation. Especially in the field of speech quality prediction, reliable methods are present which can predict the effect of many types of speech distortions, e.g., [Rix 02], including the effect of signal processing based solutions to compensate for the negative effect of, e.g., background noise. Moreover, several intelligibility measures are developed which can accurately predict the impact of environmental degradations like background noise and reverberation or simple degradations like linear filtering [Koll 08]. Unfortunately, it seems that these intelligibility predictors have more difficulties with predicting the effect of more advanced signal processing algorithms on the speech intelligibility. A clear example can be found in the field of single-channel noise reduction, where there is typically no or only little improvement in speech intelligibility due to the applied noise reduction algorithm [Jens 12, Hu 07a]. In fact, it turns out that several methods even *decrease* the speech intelligibility due to the applied speech enhancement method [Hu 07a, Hilk 12] (see [Kim 09] for an exception). Nevertheless, many intelligibility measures report incorrectly the opposite result and predict that the noise reduction algorithm did a good job and actually *increased* the speech intelligibility [Ludv 93, Dubb 08, Gold 04]. As a consequence, these measures can not be used reliably in the analysis and optimization process of a noise reduction algorithm. Note that, in contrast to speech intelligibility, there is a positive effect on speech quality due to the applied noise reduction algorithm [Hu 07b]. In contrast to intelligibility, this benefit on speech quality is correctly predicted by many speech quality measures [Hu 08a].

Besides the inconsistency between actual and predicted scores for some speech processing conditions, another important difference is present between the fields of speech quality and speech intelligibility prediction. That is, many 'simple', i.e., mathematically tractable, measures with a relatively high correlation with speech quality exist. As a consequence, many researchers develop optimal signal processing algorithms for these measures which therefore improve speech quality. These mathematically tractable measures hardly exist in the field of speech intelligibility. Typically, they are based on very complex auditory models or use average statistics of minutes of running speech, which make it difficult on how to modify the speech signal locally in an optimal manner. Perhaps this could be one of the reasons why in the field of speech processing the focus has been on speech quality rather than speech intelligibility.

1.1 Thesis Scope and Contributions

The focus in this work is on the analysis of existing measures and development of new measures for predicting the effect of signal processing algorithms on speech intelligibility which are applied in adverse conditions. We aim for new measures which are mathematically tractable and of low computational complexity such that they can be used for online optimization. Examples of mathematically tractable measures (sometimes also referred to as 'simple' measures in this thesis) include measures which provide closed-form solutions to signal processing problems and/or may be expressed as a mathematical norm. Although many types of speech degradations can occur, we look at speech processing methods which process speech in noisy conditions. With noisy conditions we assume that the speech is degraded by additive (background) noise. These types of processing include noise reduction algorithms and algorithms which process speech before playback in a noisy environment. We also look at speech vocoders, where pure noise (or noisy speech at extremely low SNRs as we will see in Chapter 4) is processed to simulate the properties of a cochlear implant with normal-hearing listeners, see, e.g., [Wils 08, Dorm 02, Litv 07]. To narrow the scope in this thesis we do not consider reverberated speech or speech processing algorithms specifically meant for this type of environmental degradation like de-reverberation. Moreover, only single-microphone algorithms are considered where multi-microphone methods are not part of this thesis.

The main goal in this thesis can therefore be summarized as follows: the development of new measures for intelligibility prediction of (non)-linear processed speech in noisy conditions, which can be used for online optimization in speech signal processing applications. The main contributions in this thesis can be summarized as follows.

Simplification of auditory model An important aspect of every measure, whether it is for signal detection, audio quality or speech intelligibility prediction, is the use of an auditory model. In Chapter 3 we simplify a general sophisticated auditory model such that it has reduced computational complexity and is mathematically tractable. As a consequence it is suitable for online optimization as we will show in Chapters 3 and 6.

Evaluation of existing measures for intelligibility prediction For many measures it is not known how they perform with intelligibility prediction of processed speech in noisy conditions. An extensive evaluation of 17 different measures is therefore presented in Chapter 4. We show that, despite high correlation with the listening test scores, several measures cannot predict the difference in intelligibly before and after signal processing. We explain that a state-of-the-art measure was not able to predict the intelligibility due to its sensitivity to the DFT-phase components. Issues with existing measures for intelligibility prediction are highlighted and a general normalization procedure as a pre-processing step is proposed which improves their correlation with speech

intelligibility.

Proposal of several new intelligibility measures In Chapter 4 and 5 several new intelligibility measures are proposed which can predict the effect on intelligibly of applied time-frequency weightings to speech degraded by additive noise. Compared to the intelligibility measures evaluated in Chapter 4 higher correlation is obtained with speech intelligibility. The newly proposed short-time objective intelligibility (STOI) measure in Chapter 5 is suitable for online optimization as we will shown in Chapter 7.

Speech energy re-distribution based on auditory model The mathematical properties of the simplified auditory model in Chapter 3 are demonstrated in Chapter 6. Here speech intelligibility is improved in a noisy environment while maintaining good speech quality. Speech energy is redistributed over time and frequency in an optimal manner for the simplified auditory model which now facilitates an analytical solution to this problem.

Channel selection in cochlear implants based on STOI In Chapter 7 we show that STOI can be interpreted as a mathematical norm, which is applied in the channel-selection technique with cochlear-implant simulations.

1.2 Thesis outline

The thesis consists of background information provided in Chapter 2, the presentation of the main results, described in Chapters 3-7, followed by a concluding discussion chapter. The main contributions are presented as a collection of five papers. The first three papers presented in Chapters 3-5 are on the development and evaluation of predictive measures for speech intelligibility, while in the remaining two Chapters 6-7 these measures are used for online optimization. More details per chapter can be found below.

Chapter 2 As an introduction this chapter provides an overview for three different prediction measures, which all have an important role in the remainder of this thesis. Each one of these measures is developed from a different research field and has different applications, which will be explained. The first measure consists of a sophisticated nonlinear auditory model and is meant for predicting the detectability of one sound played in the presence of another sound. In other words, it predicts the amount of masking of one sound due to the presence of another sound. This model is typically used for detailed study of the human auditory system as done by audiologists. The second measure is also based on detectability and masking, however, a more simple auditory model is used. As a consequence, the measure is mathematically tractable and facilitates the use of optimal signal processing algorithms. In the final part of this chapter a measure is explained specifically meant for speech intelligibility prediction for simple degradations based on average speech and noise statistics.

Chapter 3 (based on [Taal 12a]) In this chapter a mathematical expression is given for an advanced spectro-temporal auditory model. We will show that, under certain assumptions, the model can be greatly simplified. As a consequence, the model facilitates the computation of analytic expressions for masking thresholds, while advanced spectro-temporal models typically need computationally demanding adaptive procedures to find an estimate of these masking thresholds. These perceptual models exploiting auditory masking are frequently used in audio and speech processing applications like coding and watermarking. However, conventional models only take into account spectral information in short-time frames and discard time information. As a consequence, these models may introduce undesired audible artifacts in the temporal domain (e.g., pre-echoes). Since the proposed model is based on a more advanced spectro-temporal auditory model it will be shown that these artifacts are not present when the proposed method is used.

Chapter 4 (based on [Taal 11b]) Existing objective speech-intelligibility measures are suitable for several types of degradation, however, it turns out that they are less appropriate in cases where noisy speech is processed by a time-frequency weighting as used in, for example, noise reduction. To this end, an extensive evaluation is presented of objective measures for intelligibility prediction of noisy speech processed with a technique called ideal time-frequency segregation (ITFS) and single channel noise reduction. In total 17 measures are evaluated, including advanced speech-intelligibility measures, speech-quality measures and several more simple frame-based measures. Furthermore, several additional measures are proposed. Several newly proposed algorithms turn out to be good predictors of the listening experiment results. Moreover, a discussion is provided why several algorithms were not able to predict the intelligibility of the noisy and processed noisy speech.

Chapter 5 (based on [Taal 11a]) As a follow-up of the evaluation presented in Chapter 4 a new short-time objective intelligibility measure (STOI) is proposed. The measure shows high correlation with the intelligibility of noisy and time-frequency weighted noisy speech (e.g., resulting from noise reduction) of three different listening experiments. In general, STOI showed better correlation with speech intelligibility compared to five other state-of-the-art objective intelligibility models. In contrast to other conventional intelligibility models which tend to rely on global statistics across entire sentences, STOI is based on shorter time segments (386 ms). Experiments indeed show that it is beneficial to take segment lengths of this order into account.

Chapter 6 (based on [Taal 12d]) In this Chapter an algorithm is presented for intelligibility improvement in noise based on the proposed perceptual distortion measure from Chapter 3. The energy of a speech signal is optimally redistributed over time and frequency given this model and the noise statistics. Since this auditory model takes into account short-time information, transients

will receive more amplification compared to stationary vowels, which is beneficial for improving intelligibility in noise. Note that many other mathematically tractable distortion measures do not take into account short-time envelope information. The proposed method is compared to the noisy unprocessed speech and two reference methods by means of an intelligibility listening test. The results show that the proposed method leads to a statistically significant improved speech intelligibility and improved speech quality compared to the noisy speech, while the reference method with most intelligibility improvement only improves speech intelligibility at the cost of a decreased speech quality.

Chapter 7 (based on [Taal 12b]) The proposed STOI measure from Chapter 7 is simplified such that it can be expressed as a weighted ℓ_2-norm in the auditory domain. Due to the mathematical properties of a norm, STOI can now be used with the matching pursuit algorithm in the *n-of-m* channel selection technique as found in several cochlear implant (CI) coding strategies. With this technique, only a subset of frequency channels (electrodes) are stimulated, such that important channels can be updated more frequently and less significant channels are omitted. Intelligibility predictions with acoustic CI-simulations for normal-hearing listeners indicate that more intelligible speech is obtained with the proposed method compared to a conventional channel selection method based on peak picking.

Chapter 8 In this Chapter a summary and discussion of all the results in this thesis is provided. Moreover, ideas for future research directions are given.

1.3 List of Publications

The author has published the following work during his Ph.D. or are currently under peer review:

Journals

A) C. H. Taal, R. C. Hendriks, R. Heusdens, "A Speech Preprocessing Strategy For Intelligibility Improvement In Noise", Computer Speech and Language (In review), 2012. (reference [Taal 12d]).

B) C. H. Taal, R. C. Hendriks, R. Heusdens, "A Low-complexity Spectro-Temporal Distortion Measure for Audio Processing Applications", IEEE Transactions on Audio, Speech and Language Processing, vol. 20, pp. 1553 - 1564, 2012. (reference [Taal 12a]).

C) C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech", IEEE Transactions on Audio, Speech and Language Processing, vol. 19, pp. 2125-2136, 2011. (reference [Taal 11a]).

D) C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Evaluation of Objective Measures for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech", Journal of the Acoustical Society of America, vol. 130, issue 5, pp. 3013-3027, 2011. (reference [Taal 11b]).

Conference Proceedings

A) C. H. Taal, R. C. Hendriks and R. Heusdens, "Matching Pursuit for Channel Selection in Cochlear Implants Based on an Intelligibility Metric", Eusipco, Bucharest, Romania, 2012. (reference [Taal 12b]).

B) C. H. Taal, R. C. Hendriks and R. Heusdens, "A Speech Preprocessing Strategy For Intelligibility Improvement In Noise Based On A Perceptual Distortion Measure", IEEE Int. Conf. Acoust., Speech, Signal Processing, Kyoto, Japan, 4061-4064, 2012. (reference [Taal 12c]).

C) C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "Intelligibility Prediction of Single-Channel Noise-Reduced Speech", ITG-Fachtagung Sprachkommunikation, Bochum, Germany, 2010. (reference [Taal 10b]).

D) C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "On Predicting the Difference in Intelligibility Before and After Single-Channel Noise Reduction", IWAENC, Tel Aviv, Israel, 2010. (reference [Taal 10c]).

E) C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech", IEEE Int. Conf. Acoust., Speech, Signal Processing, Dallas, TX, pp. 4214-4217, 2010. (reference [Taal 10a]).

F) C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction", Proc. Interspeech 2009, Brighton, UK, pp. 1947-1950, 2009. (reference [Taal 09a]).

G) C. H. Taal, and R. Heusdens, "A Low-complexity spectro-temporal based perceptual model", IEEE Int. Conf. Acoust., Speech, Signal Processing, Taipei, Taiwan, pp. 153-156, 2009. (reference [Taal 09b]).

Invited Talks

A) C. H. Taal, "A speech pre-processing strategy for intelligibility improvement in noise", 4th Workshop on Speech in Noise: Intelligibility and Quality, Cardiff , U.K., 2012.

B) C. H. Taal, "A new objective intelligibility measure for time-frequency weighted noisy speech", 2nd Workshop on Speech in Noise: Intelligibility and Quality, Amsterdam, the Netherlands, 2010.

C) C. H. Taal, "Predicting the effect of single-channel noise reduction on speech intelligibility", IEEE Benelux Signal Processing Symposium: Signal Processing for Digital Hearing Aids, Delft, the Netherlands, 2010.

Patents

A) C. H. Taal, R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "Speech Intelligibility Predictor and Applications Thereof", US Patent 20.110.224.976, 2011.

Chapter 2

Background

2.1 Introduction

This chapter provides background information for reading the remaining chapters of this thesis. One common element in all chapters is the use of measures which try to quantify the perceptual consequences of introduced degradations to audio, like speech, in a mathematical manner. To elaborate on this, three different measures will be treated in detail which predict some kind of perceptual dimension of a modified or degraded speech signal. Throughout the thesis these three methods will be used as a baseline for comparison. Each measure can be interpreted as a representative of a certain family of measures which share an important property relevant to the main goal in this thesis: the development of new measures for speech intelligibility which can be used for optimization in speech signal processing applications. For this we distinguish the following three properties: 1) the use of an auditory model, 2) the mathematical properties of a method which are important for online optimization and 3) the method of signal comparison relevant to speech intelligibility, e.g., rather than speech quality or signal detectability.

The majority of prediction measures treated in this thesis have in common that they use some type of auditory model, where certain stages of the auditory periphery are simulated in order to obtain an internal representation. The model developed by Dau *et al.* [Dau 96a, Dau 96b] contains the most sophisticated nonlinear auditory model of all three measures. Due to the sophisticated level of this auditory model, it can accurately predict the effects of forward and backward masking [Zwic 90, Moor 03] and effects of phase changes in sinusoidal maskers in masking experiments [Dau 96b]. The original application of the Dau-model was to predict the results from psycho-acoustical listening experiments in order to study the human auditory system [Dau 96b]. Hence, it was not necessarily meant for speech signals. However, later studies have shown that the model can also be applied in the field of speech quality and intelligibility prediction [Hans 98b, Hans 97, Kohl 08, Holu 96, Chri 10]. In Chapter 3 we will show that the auditory model can be greatly simplified for predicting the results in masking experiments. This is an important step in making measures more suitable for online optimization as we will see in Chapter 6. In Chapter 4 we reveal that an intelligibility predictor based on the Dau-model outperforms many other methods in intelligibility prediction of time-frequency weighted noisy speech, e.g., as in single-channel noise reduction. Therefore the Dau-model method is used for comparison with a newly proposed intelligibility predictor in Chapter 5 which is of a more simple form.

The second model which is discussed is the perceptual model proposed by Van de Par *et al.* [Par 05] and is inspired on a simple signal detection model as proposed by Green and Swets [Gree 66]. Due to the fact that the Par-model uses a very basic model of the auditory system it has certain mathematical properties that facilitate the use of online optimization algorithms, e.g., least-squares solutions. Note that these properties are not available with the Dau-model. Therefore, many speech and audio processing algorithms have been proposed which optimize for the Par-model, like sinusoidal coding [Heus 06], resid-

ual noise modeling [Hend 04] and speech recognition feature selection [Koni 10]. However, this measure is not necessarily meant for intelligibility prediction. Instead it is based on the detectability of one sound in presence of another sound, i.e., how much speech is audible in the presence of noise. In Chapter 3 we explain that its predictions may be less reliable than the Dau-model in certain situations. Therefore a new method is proposed in Chapter 3 with similar mathematical properties as the Par-model but with prediction results similar to the Dau-model. As a consequence, we can also optimize speech intelligibility in noise for this method as explained in Chapter 6.

The last model is the coherence speech intelligibility index (CSII), which can be considered as a state-of-the-art intelligibility measure [Kate 05]. The CSII can predict the effect on speech intelligibility of signal degradations like background noise and nonlinear processing techniques as typically used in hearing aids [Kate 05]. The good performance of CSII for specific degradations is confirmed in recent evaluative studies [Ma 09, Chen 11]. Moreover, its correlation based comparison between clean and degraded auditory representations of speech signals is an important aspect used in many speech intelligibility predictors as we will see in the Chapters 4 and 5.

2.2 Notation

To describe the three models, the following general notation is used. Let x and y denote two finite length, discrete-time signals representing the original and degraded audio signal, respectively. The majority of methods discussed in this thesis aim for predicting some perceptual dimension, like speech intelligibility, speech quality or detectability, of the degraded signal y. Several methods, like the Par-model, assume an additive type of degradation which will be denoted by $y = x + \varepsilon$, where ε denotes the introduced degradation by the system of interest, e.g., background noise.

The following notation is used to describe an N-point discrete Fourier transform (DFT) of x, say $\hat{x}$, which is defined as,

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi kn/N}, \quad k = 0, ..., N-1, \tag{2.1}$$

where k represents the DFT-bin index, j the imaginary unit and n the time index. Similar definitions hold for $\hat{y}$ and $\hat{\varepsilon}$. A linear convolution between two signals, say x and h, will be denoted by $x * h$, where the outcome for a specific time sample of the convolved signal is defined as,

$$(x * h)(n) = \sum_{m=-\infty}^{\infty} x(m)\, h(n-m). \tag{2.2}$$

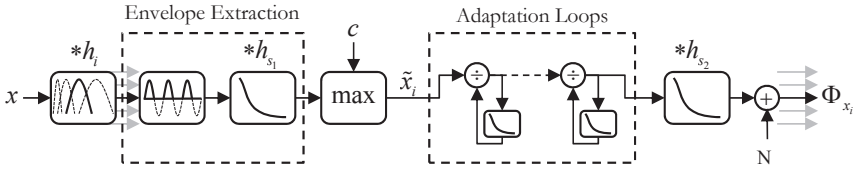

Figure 2.1: How to obtain the internal representation in the Dau-model.

2.3 Dau-model

The perceptual model developed by Dau *et al.* [Dau 96a, Dau 96b] transforms the audio signal which enters the outer ear into a spectro-temporal auditory nerve response by simulating several stages of the auditory periphery. The eventual output of this perceptual model can be interpreted as the information which will be processed by the human brain. By calculating the internal representations for two different signals, a comparison can be made in the auditory domain in order to make a prediction about their perceptual difference.

Originally the Dau-model was proposed to act as an artificial observer to study the auditory system by predicting results from psychoacoustic listening tests as obtained with real listeners [Dau 96b]. However, it has been shown in later studies that the internal representations obtained by the Dau-model can also be used for predicting quality of degraded audio and speech signals [Hans 98b, Hans 97, Kohl 08] and speech intelligibility for normal-hearing [Chri 10] and hearing impaired [Holu 96] people. Although all these approaches use the same perceptual model as proposed in [Dau 96a], they differ in how and which internal representations are compared.

In the next section it will first be explained how to obtain an internal representation following [Dau 96a]. More specific details are given on the so-called adaptation loops which are an important aspect of the Dau-model. These adaptation loops mimic the neural adaptive properties of the auditory periphery. Finally two different ways of comparing internal representations will be highlighted: (1) the original decision device for predicting results from psychoacoustic listening experiments as explained in [Dau 96a] and (2) a method proposed in [Chri 10] for predicting the speech intelligibility of degraded speech.

2.3.1 Internal Representation

The signal processing stages in the Dau-model, which model certain parts of the auditory periphery, are illustrated in Figure 2.1. Roughly six stages can be distinguished in order to obtain an internal representation:

- An auditory filter-bank mimicking the frequency dependent displacement on the basilar membrane.
- An envelope extraction stage which simulates the transformation of the

mechanical oscillations of the basilar membrane into receptor potentials in the inner hair cells.

- A max-operator to introduce an absolute threshold in quiet.
- A series of feedback loops, refered to as adaptation loops, to include the effects of neural adaptation and temporal masking effects.
- A low-pass filter for modeling the temporal integration as present in the auditory system.
- The addition of internal noise to limit the resolution of the internal representation due to, for example, spontaneous firing of neurons.

For the auditory filterbank a gamma-tone filterbank is used, e.g., [Patt 92], where the notation h_i is used to denote the impulse response of the gamma-tone filter with frequency index i. Typically, the filterbank spans a relevant frequency range for hearing, where their center frequencies are linearly spaced on a perceptual relevant frequency scale like equivalent rectangular bandwidths (ERBs) or critical bandwidths [Zwic 90, Moor 03]. After the auditory filterbank, a hair-cell model is applied which consists of halfwave rectification followed by a low-pass filtering. Where the half wave rectifier discards all the negative inputs, the low-pass filter will smooth the nonnegative signal over time and reduces the temporal structure. As a result, the output will tend to follow the envelope structure within each auditory filter.

Let the impulse response of the smoothing low-pass filter be denoted by h_{s_1}, which is implemented as a one-pole IIR low-pass filter with a cutoff frequency equal to 1 kHz. A mathematical description at the output of the hair-cell model for an arbitrary input signal x is then given by,

$$\mathcal{H}\left\{x * h_i\right\} * h_{s_1}, \tag{2.3}$$

where the operator $\mathcal{H}$ represents half-wave rectification defined as,

$$\mathcal{H}\left\{x\right\}(n) = \begin{cases} x(n) & x(n) \geq 0 \\ 0 & x(n) < 0 \end{cases} . \tag{2.4}$$

As illustrated in Figure 2.1, the hair-cell output is limited to a minimum value c. This step can be interpreted as modeling internal sounds caused by, for example, blood streams and muscle activity, which will introduce a mimum hearing threshold. This means that below a certain playback level, the model should not be able to detect the signal anymore. A second, more practical, reason for clipping the hair-cell output is to prevent a division by a small number in the adaptation loops as will be explained in the next section. The notation $\tilde{x}_i$ is used to denoted the clipped signal at the output of the hair-cell model within frequency channel i, that is,

$$\tilde{x}_i(n) = \max\left(\left(\mathcal{H}\left\{x * h_i\right\} * h_{s_1}\right)(n), c\right). \tag{2.5}$$

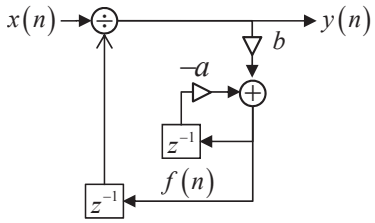

Figure 2.2: Structure of one feedback loop as used in the adaptation loops of the Dau-model.

The neural adaptation stage models the response of a neuron, according to its hair cell input stimuli. The adaptation properties are included such that the model correctly predicts the temporal masking properties of the auditory system, like forward and backward masking [Zwic 90, Moor 03]. This stage is modeled by applying five feedback loops connected in series to the clipped hair-cell output. Due to the fact that five adaptation loops are chosen, the system will have a transform close to the logarithm for stationary input signals, which is roughly in accordance with the auditory system. More details will be given on this property later in this section.

The structure of one feedback loop is illustrated in Figure 2.2 where its output is low-pass filtered and fed back into the system to act as a divisor on its input. The output y in one feedback loop for a given input signal x is given as follows,

$$y\left(n\right) = \frac{x\left(n\right)}{f\left(n-1\right)} \tag{2.6}$$

where f denotes a low-pass filtered version of y given by,

$$f\left(n\right) = by\left(n\right) - af\left(n-1\right) \tag{2.7}$$

and a and b are the filter coefficients for the one-pole IIR low-pass filter. In [Dau 96a] the filter properties are given in terms of a time-constant τ measured in ms which gives the following filter coefficients[1],

$$a = e^{-1000/(\tau f_s)}, b = a - 1, \tag{2.8}$$

where f_s denotes the sample-rate. The time-constant will affect the adaptation time for the system, which is needed to respond to a sudden change in input. This is illustrated in Figure 2.3, where the output is shown for various time constants given an input signal with a sudden on and offset. It is clear that a sudden onset will result in an overshoot while an offset gives an undershoot. As a result the model is more sensitive to signal onsets, e.g., transients in speech,

[1] the relation between the time-constant τ and the cutoff frequency f_c of a one-pole low-pass filter is given by $f_c = 1000/\left(2\pi\tau\right)$, where f_s denotes the sample-rate.

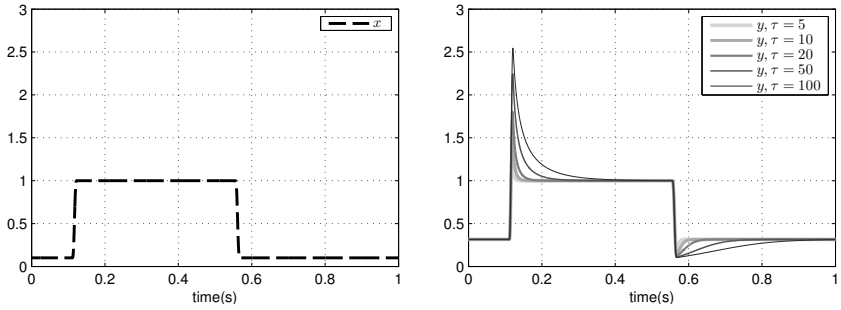

Figure 2.3: Effect of one feedback loop for different values of the low-pass filter time constant τ. The input (x) and output signals (y) are shown on the left and right plot, respectively.

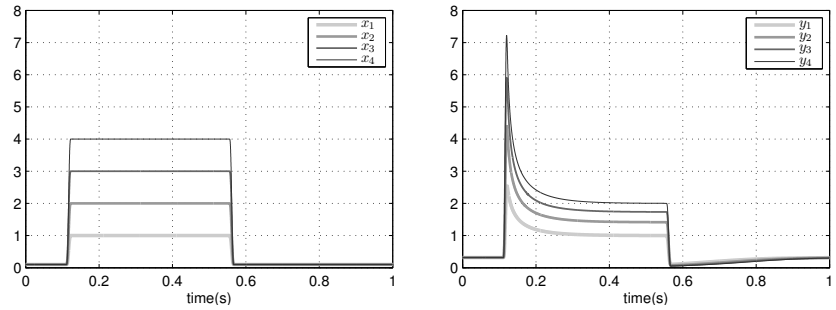

Figure 2.4: Effect of one feedback loop for different input levels. On the left different input signals are shown with corresponding outputs on the right.

and less sensitive to events occurring after a signal offset which may lead to forward masking. When the input $x(n)$ of an adaptation loop is constant, the output of the (normalized) low-pass filter will eventually converge to its input value, i.e., $y(n) = f(n)$. Inspection of Eq. (2.6) reveals that the eventual output of such a feedback loop will then converge to $y(n) = \sqrt{x(n)}$. An illustration is given in Figure 2.4 for four different input levels, where the region 0.4-0.5 seconds illustrates such a convergence. From this example it can also be concluded that at the signal onset the feedback loop results in a more linear transformation rather than taking the square root. In total five adaptation loops are applied connected in series with time constants between 10 and 500 ms [Dau 96a]. If a constant hair-cell output is encountered in Eq. (2.5), e.g., $\tilde{x}_i(n) = c, \forall n$, we obtain the following input output relation,

$$\mathcal{A}\left\{\tilde{x}_i\right\}(n) = \tilde{x}_i(n)^{1/32}, \tag{2.9}$$

where the operator $\mathcal{A}$ denotes the signal transform due to the adaptation loops. As shown in Figure 2.5 this approximates a logarithmic transform to Decibels

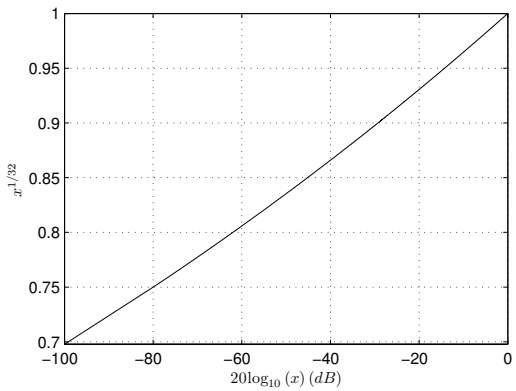

Figure 2.5: Adaptation input versus output and relation with Decibels for stationary input

which equals the human perception of loudness [Zwic 90].

Subsequently, a low-pass filter with impulse response h_{s_2} and $\tau = 20$ ms is applied to the output of the adaptation loops to reduce the effect of high temporal modulation frequencies. Finally, to include a loss of information due to spontaneous neural firing, i.i.d. Gaussian noise with variance σ^2 and zero mean is added denoted by $\mathcal{N}$. The following mathematical description is therefore obtained for the internal representation within one auditory channel,

$$\Phi_{x_i} = \mathcal{A}\{\tilde{x}_i\} * h_{s_2} + \mathcal{N}. \tag{2.10}$$

As an example, a spectro-temporal internal representation is shown for a given input signal in Figure 2.6. For visual clarity the addition of internal noise is omitted in this example. In total 32 gamma-tone filters are used where its center frequencies are linearly spaced on an ERB-scale between 100 and 4500 Hz at a sample-rate of 20000 Hz. The input signal consists of two succeeding windowed sinusoids both with a length of 200 ms. The first sinusoid is windowed with a Hanning window and has a frequency equal to 400 Hz, where the second sinusoid is windowed with a smoothed rectangular function and has a frequency of 1000 Hz. A low amount of white noise is added to the sinusoids in order to reveal the compressive behavior of the adaptation loops. One can clearly observe the under- and overshoots due to the adaptation loops with the windowed sinusoids. From the figure it can also be seen that the low-frequency sinusoid excites a different region than the high-frequency sinusoid.

2.3.2 Defining a Perceptual Difference

Now that an internal representation can be obtained, an important question is how to quantify a perceptual-relevant difference between the two signals. This step would typically involve some kind of simplified model of the cognitive processes occurring in the human brain. It is important to realize that this step

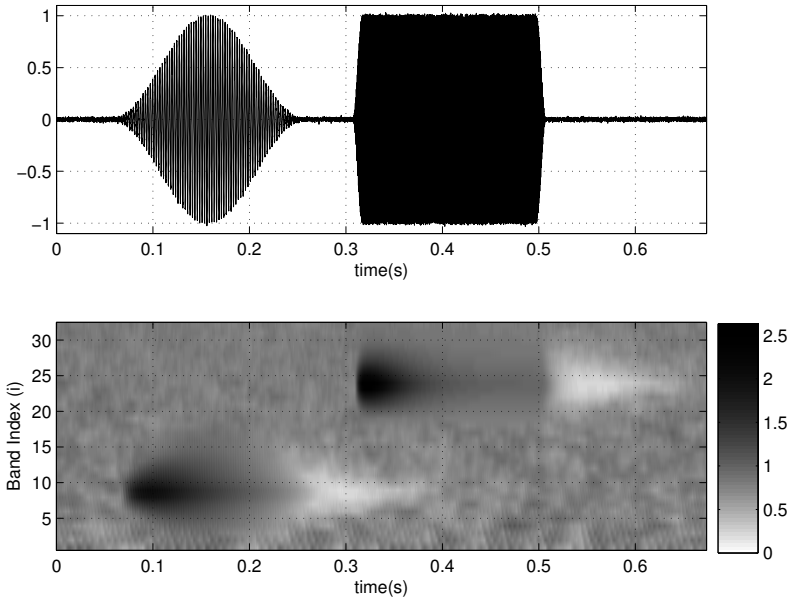

Figure 2.6: Top figure denotes signal x consisting of a 400 Hz windowed sinusoid followed by a 2000 Hz sinusoid with corresponding internal representation Φ_x in the lower figure.

may differ significantly depending on the task. For example, certain types of degradations can be easily detected by a human listener but may not affect the speech intelligibility, which is an important reason why different ways of comparison are needed. An example could be to down-sample a speech signal from 44.1 to 16 kHz which can be easily detected but will not affect the speech intelligibility [Fren 47, Flet 50]. Many ways of signal comparison have been proposed for the auditory model proposed by Dau *et al.* including methods relevant to masking experiments [Dau 96a, Kohl 08, Plas 07], determination of speech quality [Hans 98b, Hans 97] and speech intelligibility [Chri 10, Holu 96]. Two different methods will be explained which are used for comparison throughout this thesis. The first method is based on the original detection based approach [Dau 96a] where subjects have to detect a signal in the presence of a masker. The second approach is the method proposed by Christiansen *et al.* [Chri 10] where the Dau-model is used for speech intelligibility prediction.

Masking

In the original approach, as proposed in [Dau 96a], the Dau-model is used to predict results from psycho-acoustical masking experiments with human subjects. In these listening experiments, the subjects have to detect a signal ε in presence of a masker x. The goal of such an experiment is to find the signal

level such that ε is just not noticeable in the presence of the masker [Dau 96b]. This level is called the masking threshold. Similar types of experiments are conducted in the field of transparent audio coding [Pain 00] or watermarking [Swan 98] where errors are introduced to a signal which should be just not detectable. In this situation, the original clean audio signal acts as a masker on the introduced error, e.g., the quantization error in an audio coder.

As proposed in [Kohl 08], the Dau-model can be used to predict the results of these masking experiments by comparing the internal representations Φ_x and Φ_y, where x is the clean signal and y a degraded version of x with introduced error ε. In order to let the Dau-model predict these masking thresholds correctly, it is assumed that the auditory system has knowledge of the average internal representations, e.g., due to training. Due to the zero mean of the internal noise the average internal representation is simply the earlier described representation in Eq. (2.10) without adding the internal noise. An optimal detector is used [Gree 66] to decide whether an unknown realization corresponds to the clean or degraded signal.

Under these assumptions the average prediction results of the model can be described by the sensitivity index d' [Kohl 08]. The sensitivity index (i.e., distortion detectability) is deterministic and a monotonically increasing function of the probability of correctly detecting the error ε in presence of the masker x [Gree 66]. A higher d' implies a higher probability of correctly detecting the probe in presence of the masker. Let the average internal representations of x and y be denoted by $\bar{\Phi}_x$ and $\bar{\Phi}_y$, respectively. A new distance measure equal to the sensitivity index, say D_{dau}, is then given as follows (see [Kohl 08] for derivations),

$$D_{dau}\left(x,y\right) = d' = \sigma^{-1}\sqrt{\sum_i \left\|\bar{\Phi}_{y_i} - \bar{\Phi}_{x_i}\right\|_2^2}, \tag{2.11}$$

where $\|(\cdot)\|_2$ denotes the ℓ_2 norm, σ the standard deviation of the internal noise as was given in Eq. (2.10) and i the band-index of the auditory filterbank. The calibration of σ is based on the 1-dB criterion in intensity discrimination tasks [Dau 96a].

Many applications are interested in a masking threshold of ε given x, i.e., the maximum level of ε such that it is just not detectable in the presence of x. This threshold can be found by solving $d'\left(x, x+\alpha\varepsilon\right) = 1$ for α, where α is a scalar controlling the level of the introduced error. Due to the complexity of some nonlinear stages in the Dau-model, no closed-form solution exists for this mathematical problem. Instead, a typical approach is to use adaptive procedures similarly to what is done with real listening experiments [Levi 71].

Speech Intelligibility

More recently a method of comparing the internal representations have been proposed by Christiansen *et al.*, where the eventual outcome measure shows a

high correlation with speech intelligibility [Chri 10]. This method first determines the internal representations of the clean and degraded speech signal (as explained in Section 2.3.1), denoted by Φ_x and Φ_y, respectively, where we are interested in the intelligibility of the degraded speech.

Rather than using a squared error in the auditory domain, as used in Eq. (2.11), a sample correlation-coefficient is applied between the clean and degraded internal representations. This is one aspect which is changed in order to make the model more appropriate for intelligibility prediction. Moreover, the auditory filterbank only spans the frequency range between 100 and 8000 Hertz in order to exclude frequencies unimportant for speech intelligibility. In addition, the intelligibility based approach analyzes the signal in short-time (20 ms) segments rather than the complete signal at once, as is the case with the detection based approach.

The sample correlation coefficient is defined as follows for any arbitrary signals x and y,

$$\rho\left(x,y\right)=\frac{\sum\limits_{n}\left(x\left(n\right)-\bar{x}\right)\left(y\left(n\right)-\bar{y}\right)}{\sqrt{\sum\limits_{n}\left(x\left(n\right)-\bar{x}\right)^{2}\sum\limits_{n}\left(y\left(n\right)-\bar{y}\right)^{2}}}. \tag{2.12}$$

where $\bar{x}$ and $\bar{y}$ denote the sample means of x and y, respectively. The use of a correlation based comparison has the advantage that it is insensitive to differences in signal energy. As a consequence, changing the playback level of y independently of x will not change a correlation based outcome measure. This is similar to a real intelligibility listening test where changing the playback level within a certain audible range should not have a large impact on the results.

Another difference between the masking approach and the intelligibility method is that the internal representations are first segmented in 20 ms, 50% overlapping, frames where for each short-time frame a correlation-coefficient is determined. The frame-dependent correlation coefficients are averaged to obtain an overall intelligibility measure. Here only a subset of the time frames are considered which contain a relatively high amount of speech energy. High-level segments are defined here as having a root-mean-square (RMS) level of 0 dB or higher, relative to the overall RMS level of one speech utterance. Let m denote the frame-index and $\mathcal{M}$ the set of high-level segments with cardinality $|\mathcal{M}|$, the outcome measure for one utterance, say $\mathcal{I}_{dau}$, is than given by,

$$\mathcal{I}_{dau}\left(x,y\right)=\frac{1}{|\mathcal{M}|}\sum_{i,m\in\mathcal{M}}\rho\left(\Phi_{x_{i,m}},\Phi_{y_{i,m}}\right). \tag{2.13}$$

Typically, $\mathcal{I}_{dau}$ has to be determined for a large set of sentences for one given degradation type in order to determine an accurate estimate of its average score. This is also the case with real listening tests. This average score is expected to have a monotonic increasing relation with speech intelligibility [Chri 10]. This is also shown in Chapter 4.

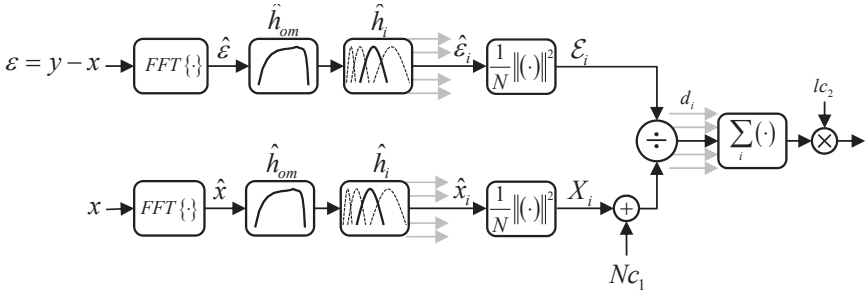

Figure 2.7: Basic structure of the Par-model [Par 05].

2.4 Par-model

As opposed to the Dau-model, the auditory model of the Par-model is of a very simple form such that it leads to a tractable mathematical measure. For example, it facilitates a closed-form solution for masking thresholds as we will see in this Section. One aspect which simplifies the model is that no explicit stage is included to model the nonlinear compressive behavior of the auditory system (for example the adaptation loops in the Dau-model). Another property of the Par-model to accomplish this simple form is that it ignores time-information and considers the signals in the frequency domain only. As a consequence the model is only meant for predicting perceptual differences in short-time segments, i.e. 20-40 ms, where most audio signals like speech are assumed to be stationary and time-information plays a less important role.

The basic structure of the perceptual model proposed by van de Par *et al.* [Par 05] is shown in Figure 2.7. As with most of the models discussed in this thesis, the Par-model will compare two signals, say x and y. However, the Par-model assumes that $y = x + \varepsilon$ where ε is of some additive form and available in isolation. Therefore, rather than calculating an internal representation of x and y, the internal representation of x and ε is used. Note that the signals x, y and ε are taken as short-time frames where we omit the short-time frame-index for notational convenience.

2.4.1 Internal Representations

As illustrated in Figure 2.7 first a discrete fourier transform (DFT) is applied to x and ε, resulting in signals $\hat{x}$ and $\hat{\varepsilon}$, respectively. Subsequently, a filter is applied to simulate the properties of the outer and middle ear, denoted by h_{om}. For reasons which will become clear in Section 2.4.3 the frequency response is taken equal the to inverse of the threshold in quiet, i.e., a frequency-dependent curve which denotes the masking threshold of a sinusoid in quiet [Fast 07, Moor 03]. The following equation can be used which approximates the threshold in quiet [Pain 00],

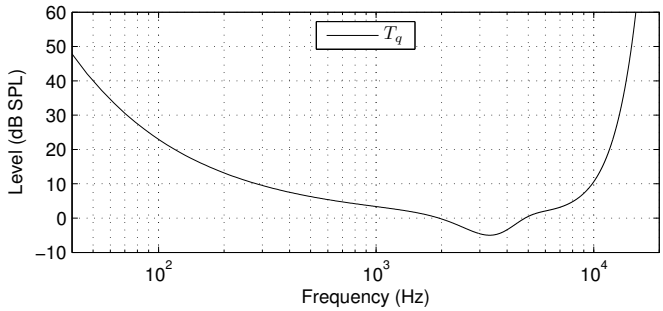

Figure 2.8: Threshold in quiet.

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \quad (2.14)$$

where f denotes frequency in Hz. This curve is shown in Figure 2.8. By compensating for the dB transform the following frequency response is obtained for the outer-middle ear filter,

$$\hat{h}_{om}(k) = \nu 10^{T_q\left(\frac{kf_s}{N}\right)/20}, \quad (2.15)$$

where k denotes the DFT bin index, N the DFT size in samples and f_s the sample rate. The scalar ν is used to normalize the frequency response such that $\hat{h}_{om}(k) = 1$ for k corresponding to 1 kHz. After the outer-middle ear filter an auditory filterbank is applied based on gammatone filters [Patt 92]. Let i denote the frequency band index, the magnitude response of such a filter is well approximated by [Par 05],

$$\hat{h}_i(k) = \left(1 + \left(\frac{kf_s/N - f_c(i)}{\kappa \mathrm{ERB}(f_c(i))}\right)\right)^{-\eta/2}, \quad (2.16)$$

where ERB denotes the transformation of the filter center frequency f_c in Hz to equivalent rectangular bandwidths (ERBs) as defined in [Glas 90], η the filter order which is equal to 4 [Par 05] and κ is a normalization term defined as,

$$\kappa = 2^{\eta-1}(\eta-1)!/\pi(2\eta-3)!!, \quad (2.17)$$

where ! denotes the factorial and !! the double factorial (see [Grad 00, page-xliii] for more details on the double factorial). Both the frequency responses of the outer-middle ear filter and the auditory filterbank are shown in Figure 2.9. The filters are applied to the signals x and ε by means of a point wise multiplication in the frequency domain which gives,

$$\hat{x}_i(k) = \hat{h}_{om}(k)\,\hat{h}_i(k)\,\hat{x}(k), \quad (2.18)$$

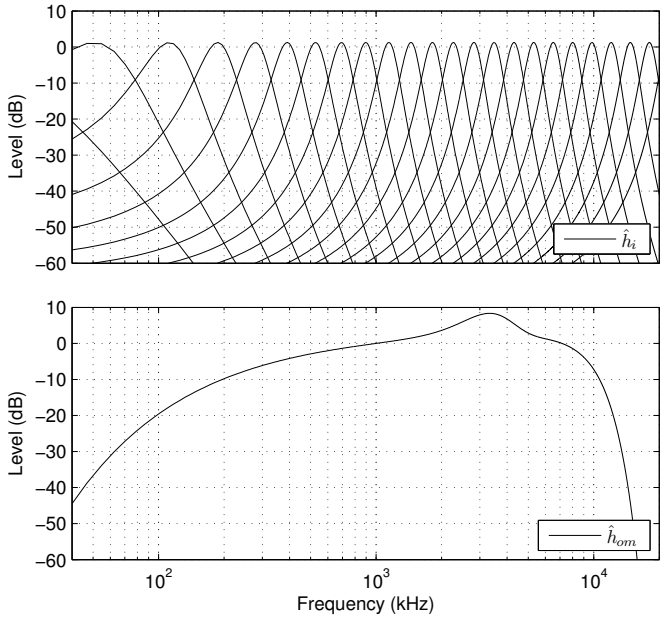

Figure 2.9: Frequency responses of the the auditory filterbank (top) and outer middle ear filter (bottom) as used in the Par-model.

and

$$\hat{\varepsilon}_i(k) = \hat{h}_{om}(k)\,\hat{h}_i(k)\,\hat{\varepsilon}(k)\,. \tag{2.19}$$

Finally the power is determined within each band to obtain an internal representation, which gives,

$$X_i = \frac{1}{N}\sum_k |\hat{x}_i(k)|^2, \tag{2.20}$$

and

$$\mathcal{E}_i = \frac{1}{N}\sum_k |\hat{\varepsilon}_i(k)|^2, \tag{2.21}$$

for the clean and the error signal, respectively. Figure 2.10 shows an example of these internal representations for two short-time frames, where x is a clean speech vowel and ε windowed white noise. Note that because of the power integration in Eqs. (2.20, 2.21) the internal representations are more smoother than the actual power spectra and that the energy of higher and lower frequencies is reduced significantly due to the outer middle ear filter. It is also important to realize that the frequency responses of h_{om} and h_i are defined as being real and positive, while normally there would be some phase component

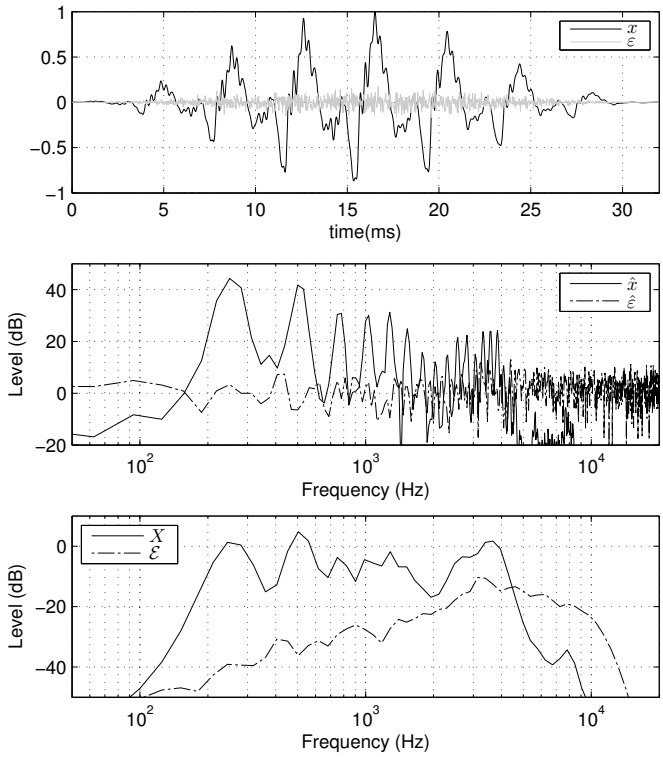

Figure 2.10: A windowed clean speech vowel plus windowed white noise (top) with corresponding spectra (middle) and internal representations (bottom).

involved in a gammatone filter [Patt 92]. This is ignored in [Par 05] due to the eventual integration stage in Eqs. (2.20, 2.21) where the absolute value is taken and phase information is ignored.

2.4.2 Defining a Perceptual Difference

The clean and noise signal are compared in the internal domain where a perceptual difference is determined related to the detectability of ε when played in presence of x (similarly as with the Dau-model in Eq. (2.11)). This perceptual difference in the Par-model is based on the energy detection model from the field of signal detection as proposed by Green and Swets [Gree 66]. The detectability within one band is defined as the noise-to-signal ratio in the internal domain as follows,

$$d_i = \frac{\mathcal{E}_i}{X_i + Nc_1}, \tag{2.22}$$

where Nc_1 is added to introduce an absolute threshold in quiet. This step models internal sounds caused by, for example, blood streams and muscle activity. Note that the outer-middle ear filter is responsible for making this threshold frequency dependent.

Van de Par *et al.* [Par 05] suggested to combine the within-channel sensitivity indices over all auditory bands by means of an additive operation in order to mimic the spectral integration properties of the auditory system (see e.g., [Buus 86, Lang 92]). This gives the following equation for the eventual outcome of the distortion measure,

$$D_{par}(x, \varepsilon) = lc_2 \sum_i d_i = lc_2 \sum_i \frac{\mathcal{E}_i}{X_i + Nc_1} \tag{2.23}$$

where c_2 is a calibration parameter used to modify the sensitivity of the model and,

$$l = \min\left(300, 1000N/f_s\right), \tag{2.24}$$

denotes a factor to include the maximum temporal integration properties of the auditory system. As a consequence, increasing the playback length of a signal will result in a higher predicted detectability, which is in accordance with a human observer up till lengths of approximately 300 ms [Brin 64]. As shown in [Par 05] the outcome of equation Eq. (2.23) is monotonically increasing related to the probability of correctly detecting the noise ε in presence of the signal x (i.e., a higher D_{par} implies a higher probability of correctly detecting the probe in presence of the masker).

The parameters c_1 and c_2 are calibrated such that the model correctly predicts the masking threshold of a 1 kHz tone in silence and the 1 dB just noticeable level difference for a 70 dB SPL, 1 kHz tone. The model is calibrated such that $D_{par} = 1$ corresponds to a distortion at the threshold of detection of ε [Par 05].

2.4.3 Mathematical Properties

Due to the simple structure of the Par-model it has certain mathematical properties which makes the measure very suitable for optimization in audio and speech processing applications. For example, by defining the following weighting function,

$$a(k) = lc_2 \sum_i \frac{\left|\hat{h}_{om}(k)\right|^2 \left|\hat{h}_i(k)\right|^2}{\sum_l |\hat{x}_i(l)|^2 + N^2 c_1}, \tag{2.25}$$

the measure can now be expressed as,

$$D_{par}(x, \varepsilon) = \sum_k |\hat{\varepsilon}(k)|^2 a(k). \tag{2.26}$$

Note that the weighting function a is independent of ε and can be pre-calculated within each short-time frame, stored and reused. The result is that, in order to evaluate Eq. (2.26) for any ε, only one FFT has to be applied followed by a simple linear weighting. Due to the low computational complexity of the Par-model it is therefore very useful in the context of an audio coder where, typically, the model has to be evaluated many times, e.g., in a rate-distortion loop. Another important property is that the weighting function a is real and positive so that, in fact, the perceptual distortion measure defines a norm. This allows incorporating perceptual properties in least-squares optimization algorithms like sinusoidal coding [Heus 06, Heus 02a] and residual noise modeling [Hend 04].

In many applications, one is interested in a masking threshold given x, i.e., the maximum level of ε such that it is just not detectable in the presence of x. This threshold can be found by solving $D_{par}(x, \alpha\varepsilon) = 1$ for α, where α is a scalar controlling the level of the introduced error. Due to the simple mathematical form of the Par-model an analytic solution exists given by,

$$\alpha = \frac{1}{\sqrt{D_{par}(x, \varepsilon)}}. \tag{2.27}$$

Note that this solution does not exist with many complex perceptual models, e.g., the Dau-model. Next to masking thresholds some applications, like [Heus 06, Heus 02a, Swan 98], require knowledge of the masking curve, which describes the masking threshold for a sinusoid as a function of frequency. This masking curve will provide information on how to shape the spectrum of an introduced error such that the perceptual impact of the error is minimized. Such a sinusoid is described by,

$$\varepsilon_p(n) = \alpha_p \cos(2\pi pn/N), \tag{2.28}$$

where N is the DFT-size, p/N the normalized frequency of the sinusoid and α_p its amplitude. In the DFT-domain we obtain the following description,

$$\hat{\varepsilon}_p(k) = \begin{cases} N/2 & k = \{-p, p\} \\ 0 & otherwise \end{cases} \tag{2.29}$$

The masking curve can now be found by solving $D_{par}(x, \alpha_p \varepsilon) = 1$ for all p, which gives,

$$\alpha_p = \frac{\sqrt{2}}{N\sqrt{a(k)}}. \tag{2.30}$$

Hence, the masking curve can be easily obtained by inverting and scaling the weighting function a as was defined in Eq. (2.25). At this point it is also easy to see that the masking curve will be equal to the inverse of the threshold in quiet in the situation that there is no masker, i.e., $X = 0$. First we use the following property of the joint frequency response of the filterbank,

$$\sum_i \left|\hat{h}_i(k)\right|^2 \approx 1, \tag{2.31}$$

which is true when a sufficient number of auditory filters is used. Subsequently we have the following weighting function,

$$a(k) = \frac{lc_2}{N^2 c_1}\left|\hat{h}_{om}(k)\right|^2 \tag{2.32}$$

which gives the following masking curve for the case that $x = 0$,

$$\alpha_p = \sqrt{\frac{2c_1}{lc_2}}\hat{h}_{om}(k)^{-1}. \tag{2.33}$$

Hence, due to the fact that the outer-middle ear filter was taken equal to the inverse of the threshold in quiet in Eq. (2.15) the model indeed correctly predicts the threshold in quiet.

2.5 Coherence Speech Intelligibility Index

The Coherence Speech Intelligibility Index (CSII) was introduced by Kates and Arehart in 2005 [Kate 05] and can be interpreted as an extension of the original speech intelligibility index (SII) standardized in [ANSI 97]. While the original SII was mainly meant for linear degradations like additive background noise and linear time-invariant filtering, the CSII can also be used for nonlinear processing artifacts like peak and center clipping [Kate 05]. These clipping artifacts are related to distortions which, for example, may occur in a hearing aid [Kate 05]. In contrast to the SII, the CSII uses the coherence function [Cart 73, Kate 92] to estimate the speech and noise spectra, hence the name Coherence SII.

First the basics of the SII will be explained followed by the proposed extensions by Kates and Arehart, which are based on the coherence function.

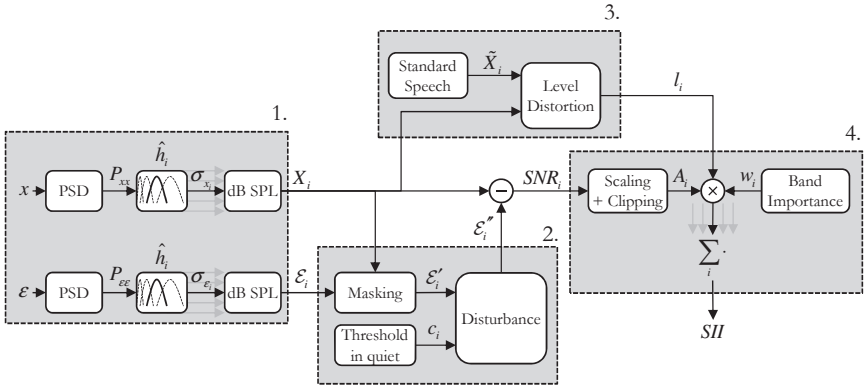

Figure 2.11: Basic structure of the Speech Intelligibility Index (SII)

2.5.1 SII

In the SII standard different versions exist where different types of filter-banks can be used based on, for example, critical bands or octave bands. Here the SII standard based on 21 critical bands will be explained which is the one used in the CSII. The basic structure of the SII is illustrated in Figure 2.11 where four different parts are highlighted by means of the numbered gray boxes. The calculation process of the SII can be summarized by taking a weighted average of clipped SNRs where a high SII implies highly intelligible speech. The rational behind each stage will be explained together with their mathematical details.

In the first part the average long-term spectra of the speech and noise are calculated. This is done in several critical bands and subsequently converted to dB SPL, which can be interpreted as a very simple model of the auditory system. Let x and $y = x + \varepsilon$ denote time-domain signals of the clean and noisy speech, respectively, where ϵ denotes the noise signal. The short-time DFT coefficient will be denoted by $\hat{x}(k)$ with frequency-bin index k. The speech power spectral density (PSD) is then given as follows,

$$P_{xx}(k) = E\left[\frac{1}{N}|\hat{x}(k)|^2\right], \tag{2.34}$$

where N denotes the frame-length in samples and E the expectation operator. In practice this PSD is not available and has to be estimated. Although there are different approaches for determining the average spectra, we use a method based on the periodogram which is also used in the CSII [Kate 05]. This estimator is given by averaging the periodogram over several short-time frames,

$$\underline{P}_{xx}(k) = \frac{1}{MN}\sum_m |\hat{x}(m,k)|^2, \tag{2.35}$$

where M denotes the number of short-time frames, m the short-time frame index and the underbar notation $\underline{(\cdot)}$ denotes the use of an estimator. Similar

definitions hold for estimation of the noise PSD denoted by $P_{\varepsilon\varepsilon}$. The estimated noise and speech PSDs are then weighted and summed in order to get the average power within a critical band. For the speech signal the average power within a critical band is then defined as follows,

$$\sigma_{x_i}^2 = g_i \sum_k P_{xx}(k) \left|\hat{h}_i(k)\right|^2 \tag{2.36}$$

where $\left|\hat{h}_i(k)\right|^2$ denotes the magnitude spectrum for a ro-ex filter which has similar properties as the gammatone magnitude response as in Eq. (2.16) (see [Moor 83] for more details on ro-ex filters). In total 21 filters are used where the filter center frequencies, denoted by $f_{c(i)}$, are given in [ANSI 97] and the bandwidths, denoted by $f_{b(i)}$, follow the critical bandwidths as explained in [Zwic 80]. The term $g_i = \left(N f_{b(i)}\right)^{-1}$ in Eq. (2.36) denotes a normalization factor to compensate for the bandwidths. In order to account for the perception of loudness the powers are converted to dB SPL. Here we assume that the RMS of a digital signal with $|x(n)| < 1$ corresponds to a playback level of 96 dB SPL. This gives,

$$X_i = 10\log_{10}\left(\sigma_{x_i}^2\right) + 96, \tag{2.37}$$

where similar definitions hold for the noise critical band power $\mathcal{E}_i$. Note that in practice the playback level may be unknown of a speech signal. To overcome this a reasonable choice is to assume a playback level of 65 dB SPL.

Examples are shown in the top two plots in Figure 2.12 for noise with a low-pass characteristic (left) and white noise (right) at an SNR of 10 and -5 dB, respectively. The clean speech spectrum is estimated from the complete Timit database [Garo 93]. This database consists of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences [Garo 93].

In the second gray box in Figure 2.11 the noise statistics will be slightly smoothed over frequency in order to incorporate masking effects. Furthermore, the noise spectrum will be clipped to a minimum value to include a threshold in quiet. To include effects of masking the first step in the SII is to determine the amount of self-speech spectrum masking. This means that if the speech energy is relatively large in one band it may mask some speech in a neighboring frequency band. This self-speech spectrum masking is defined by the original speech spectrum minus 24 dB and is only taken into account when it exceeds the noise level which gives the following intermediate outcome,

$$B_i = \max\left(\mathcal{E}_i, X_i - 24\right). \tag{2.38}$$

Subsequently for each band the value C_i is determined which equals the slope per octave of the spread of masking, that is,

$$C_i = 0.6\left(B_i + 10\log\left(f_{bw(i)}\right)\right) - 80. \tag{2.39}$$

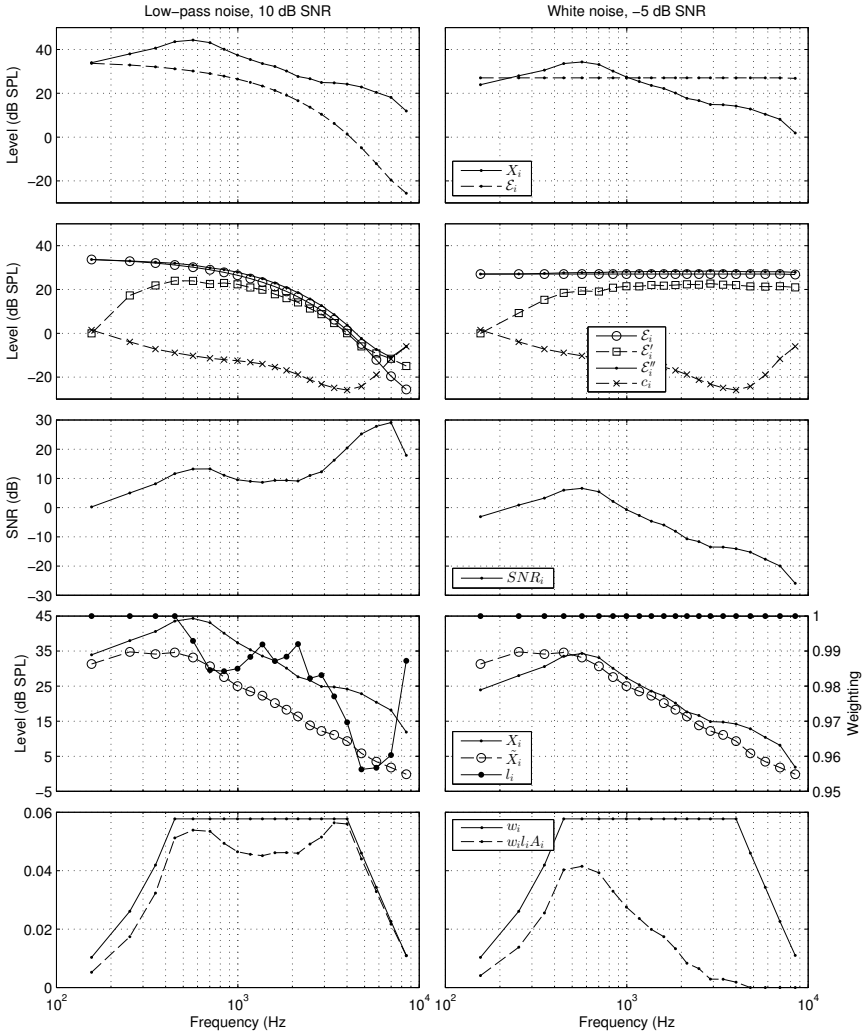

Figure 2.12: Intermediate signals in the calculation process of the SII as illustrated in Figure 2.11. Examples are shown for the noise with a low-pass characteristic, 10 dB SNR (left column plots) and white noise, -5 dB SNR (right column plots). See text for more details.

The values B_i are then smoothed over the different frequency bands by means of a power addition as follows for frequency bands $i = 2...21$,

$$\mathcal{E}'_i = 10\log_{10}\left(\sum_{k}^{i-1} 10^{0.1\left(\mathcal{E}_i + 3.32C_k\log_{10}\left(f_{c(i)}/h_k\right)\right)}\right), \tag{2.40}$$

where $\mathcal{E}'_1 = B_1$. Finally $\mathcal{E}'_i$, which represents the effects of masking, is combined with the noise spectrum by means of a power addition and clipped to include a threshold in quiet denoted by c_i. This gives,

$$\mathcal{E}''_i = \max\left(c_i, 10\log_{10}\left(10^{\mathcal{E}_i/10} + 10^{\mathcal{E}'_i/10}\right)\right). \tag{2.41}$$

The second row of plots in Figure 2.12 illustrates the effects of the masking and clipping stage on the noise spectrum. It is clear that the effect of masking slightly increases the noise level but will only have a small effect. In the high frequency band the noise with the low-pass filter characteristic falls below the threshold in quiet and is clipped to c_i. The adjusted noise spectrum and the clean speech spectrum are then used to calculate an SNR within each frequency band as illustrated in the third row of plots in Figure 2.12.

For the situation that the playback level of the speech is too loud, it is assumed in the SII that this will have a negative impact on the speech intelligibility. This is represented by the third gray box in 2.11 where a weighting function l_i is calculated. When the speech is too loud, l_i will reduce the contribution of this band to the eventual score. This is accomplished by comparing the calculated speech spectrum X_i with the standard speech spectrum level at the normal vocal effort denoted by $\tilde{X}_i$, which can be found in the standard [ANSI 97]. The weighting l_i is given as follows,

$$l_i = \min\left(1, 1 - \left(X_i - \tilde{X}_i - 10\right)/160\right), \tag{2.42}$$

and is clipped between 0 and 1 where $l_i = 1$ implies that the speech signal is not too loud and therefore will not affect the eventual outcome of the SII. Examples are shown in Figure 2.12 in the two plots of the fourth row (note the vertical axis on the right which is used to denote the value of l_i). In the left plot the speech level is too loud where l_i drops below zero, while in the left plot the speech is at a normal level which results in $l_i = 1$ for all frequency bands.

In the last part of the SII, denoted by the fourth gray box in Figure 2.11, the SNRs per frequency are limited between -15 and +15 dB and subsequently normalized between 0 and 1. This gives,

$$A_i = \max\left(\min\left(\frac{X_i - E'' + 15}{30}, 1\right), 0\right). \tag{2.43}$$

In the final stage the clipped and normalized SNRs are summed and weighted with the band importance functions as given in [ANSI 97] and the level distortion as was given in Eq. (2.42) as follows,

$$SII = \sum_i l_i w_i A_i. \tag{2.44}$$

The weighting functions have the property that $\sum_i w_i = 1$. As a consequence, the SII is also limited between 0 and 1 which equals 0% or 100% intelligible speech, respectively. The last two plots at the bottom in Figure 2.12 show the weighting functions together with the eventual, clipped, weighted and normalized SNRs. Clearly, in the left plot the speech is less intelligible than the right plot.

2.5.2 SII based on Coherence

One issue with the SII is that the noise spectrum is needed in isolation in order to estimate the noise PSD. However, in practice this is not always available. For example, in the case when nonlinear distortions are applied to the speech signal, e.g., as in hearing aids, we only have access to y and it is not clear how to determine ε. In the CSII this is solved by estimating the noise and speech PSD by means of the coherence function [Cart 73]. The coherence function is given as a normalized correlation measure in the frequency domain as follows,

$$\gamma(k) = \frac{P_{xy}(k)}{\sqrt{P_{xx}(k) P_{yy}(k)}}, \tag{2.45}$$

where P_{xx} and P_{yy} denote the PSDs of the clean and degraded speech signal, respectively, and P_{xy} equals the cross spectral density between x and y and is defined as,

$$P_{xy}(k) = E\left[\frac{1}{N}\hat{x}(k)\,\hat{y}^*(k)\right]. \tag{2.46}$$

Similarly as in Eq. (2.35) all the spectral densities in Eq. (2.45) are estimated with a periodogram-based estimator. For the cross power spectral density between x and y this gives,

$$\underline{P}_{xy}(k) = \frac{1}{MN}\sum_m \hat{x}(m,k)\,\hat{y}^*(m,k), \tag{2.47}$$

where the asterisk denotes complex conjugation. Subsequently, Kates and Arehart use the coherence function to estimate the speech PSD,

$$\underline{P}_{xx}(k) = |\gamma(k)|^2 P_{yy}(k) \tag{2.48}$$

and the noise PSD,

$$\underline{P}_{\varepsilon\varepsilon}(k) = \left(1 - |\gamma(k)|^2\right) P_{yy}(k). \tag{2.49}$$

These estimated PSDs can then be directly used in the conventional SII procedure as was explained in the previous section. One of the advantages of

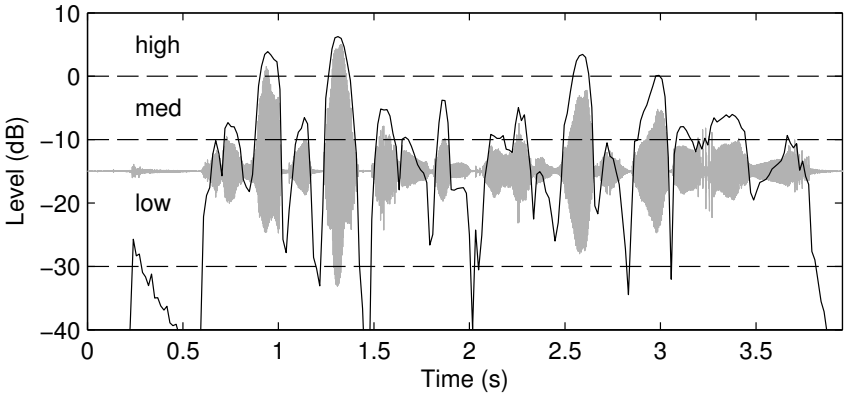

Figure 2.13: The CSII divides the short-time frames in three different level ranges which are denoted by the dashed horizontal lines. The solid line shows the energy within the short-time frames.

this approach is that for the case that the noise and speech are statistically independent this will lead to the same results as the conventional SII.

One additional change which was also proposed in the CSII was to divide the speech signal into three different level ranges and perform a separate SII calculation for each subset of short-time frames. Subsequently, the individual SII scores for each level range are then combined to form a new score. One motivation for this is that some low-level regions, e.g., consonants, may have relatively low energy but should be more important than high-energy regions. The level division process is illustrated in Figure 2.13 where three levels are used with level boundaries of -20, -5 and 5 dB. The level in each short-time frame, denoted by the solid line, is determined by calculating its RMS-value and convert it to a log scale. The results are shown in the figure together with the underlying waveform of the speech signal. When the frames are grouped in the three level ranges, the PSDs are estimated as previously explained in this section and an SII value is calculated for each level-range. Finally, the three scores are combined as follows to calculate the three-level CSII denoted by I_3,

$$I_3 = 0.16CSII_{Low} + 0.84CSII_{Mid} + 0.0CSII_{High}. \tag{2.50}$$

which is expected to have a monotonic increasing relation with the speech intelligibility of the degraded speech y [Kate 05]. Alternatively I_3 can be converted with a logistic function to predict absolute intelligibility scores [Kate 05].

2.6 Relation to Thesis Chapters

Three different models are explained which predict a perceptual impact of introduced degradations to the audio signal like detectability or intelligibility.

The Dau-model [Dau 96a] contains a sophisticated, nonlinear spectro temporal auditory model and is suitable for predicting masking thresholds in psycho-acoustic listening experiments. Due to the details of the auditory model the method can also predict results of temporal masking experiments, e.g., forward masking. However, because of the complexity and the non-linearity of the auditory model, no closed-form solutions exist for, e.g., masking thresholds. In Chapter 3 we propose a new method which shows similar masking predictions as the Dau-model but includes closed-form solutions for masking thresholds. In addition, the computational complexity is reduced significantly.

The Dau-model is also successfully applied for intelligibility prediction of various (non)-linear speech degradations in the method proposed by Christiansen en Dau [Chri 10]. In Chapter 4 we reveal that this method is one of the best performing intelligibility predictors of noisy speech signals which are processed with a time-frequency varying weighing, e.g., as in single-channel noise reduction. The Dau-model is therefore used as a baseline system for comparison with our newly proposed intelligibility predictor in Chapter 5. The Dau-model is also used for analysis of a newly proposed signal processing algorithm relevant for cochlear implant users in Chapter 7.

The Par-model [Par 05] is based on signal detection of introduced errors similarly as with the Dau-model. However, the spectral auditory model in the Par-model is of a more simple form than the Dau-model and is therefore suitable for online optimization algorithms. As a consequence it facilitates closed-form solutions for masking thresholds and masking curves. However, time information within short-time frames is ignored by this model. We will show in Chapter 3 that the masking predictions are therefore less reliable for signals which are non-stationary within short-time frames, e.g., transients in speech signals. We propose a new method in Chapter 3 which has better masking threshold predictions with similar mathematical properties and computational complexity as the Par-model.

The coherence speech intelligibility index (CSII) [Kate 05] is specifically meant for intelligibility prediction of nonlinear distortions. Recent evaluations show indeed that this measure has high correlation with the speech intelligibility of single-channel noise reduced speech, e.g., [Ma 09]. This is confirmed in Chapter 4. The CSII is therefore used as a baseline system for comparison with our newly proposed intelligibility predictor in Chapter 5. However, in this Chapter 4 we also show that CSII is not suitable for intelligibility prediction of vocoded speech due to its phase sensitivity.

Chapter 3

A Low-complexity Spectro-Temporal Distortion Measure for Audio Processing Applications

---

This chapter is published as "A Low-complexity Spectro-Temporal Distortion Measure for Audio Processing Applications", by C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen in the *IEEE Trans. Speech, Audio and Language Processing*, vol. 20, pages 1553 - 1564, 2012.

3.1 Introduction

It is well-known that the properties of the human auditory system play an important role in the development of various audio and speech processing algorithms. One such example is transparent audio coding where, by reducing the bit-rate, errors are introduced to a signal such that the distorted signal is perceptually indistinguishable from the original [Pain 00]. Here, a typical approach is to shape the quantization error in the frequency domain, on a frame-by-frame basis, according to the so-called masking threshold per auditory band. As long as the error signal is below this threshold, the original signal will act as a masker on the error signal. This phenomenon, called auditory masking, is also exploited in the field of watermarking [Swan 98], where some type of information is embedded (the watermark) by means of adding noise in such a way that it is masked by the clean signal.

In order to determine whether an introduced error is audible, the system under test typically uses a perceptual model. A well-known perceptual model is the ISO/IEC 11172-3 (MPEG-1, layer I) psychoacoustic model 1 [Comm 93]. This perceptual model is typically used in the field of audio coding [Pain 00, Pan 95], but is also applied in the field of other audio and speech processing applications like speech enhancement [Jabl 04] and watermarking [Swan 98]. Here, the masking threshold per frequency band is found by first separating the signal in tonal and noise maskers, after which for each of these spectral components a spreading function is defined [Pain 00]. Then, by power addition of these spreading functions, a masking threshold is obtained. This method is based on the assumption that the detectability of a specific frequency component is only determined by the auditory filter centered around that particular frequency. However, this assumption is not in line with various results in literature (e.g. [Buus 86]), where it is suggested that the detectability of a specific frequency component is also determined by off-frequency auditory filters.

Van de Par *et al.* introduced a perceptual distortion measure, which we will refer to as the Par-model, including spectral integration [Par 05]. That is, the detectability of a specific frequency component is also determined by off-frequency auditory filters. This method showed better correspondence with data from psychoacoustic listening tests than the MPEG-1 model. Moreover, it does not need to separate the signal into tonal and noise maskers. It has been shown that the Par-model leads to better coding results compared to the MPEG-1 model for various fixed bit-rates in the field of sinusoidal coding [Par 05]. In addition, the Par-model is defined as a mathematical norm, which allows for incorporating perceptual properties in least squares optimization algorithms. Examples are found in sinusoidal coding [Heus 02b] and residual noise modeling [Hend 04]. Note that in the field of speech processing, mathematical tractable distortion measures are also used, like the log-spectral distance or distortion measures based on linear prediction (see e.g., [Gray 76, Quac 88] for an overview). Although these measures include some perceptual properties they do not account for auditory masking effects.

Many perceptual models, like the Par-model and the MPEG-1 perceptual model, assume that the introduced error occurs simultaneously with the clean signal within one short-time frame (20-40 ms) and, therefore, do not take any temporal information into account. The consequence is that if an error is introduced before an onset of the clean signal in the same frame, these spectral models will consider the error to be masked, which is actually not the case. In fact, this will lead to so-called pre-echoes which are unwanted perceptual artifacts [Pain 00]. Although some backward masking may occur to mask the pre-echo, this is typically not sufficient since backward masking is only present a few milliseconds before the onset of the clean signal [Zwic 90, Moor 03]. A solution to prevent pre-echoes is called temporal noise shaping [Herr 99], which minimizes the squared error by means of frequency domain linear prediction. However, this method is not based on a perceptual model. Other solutions are window switching [Pain 00] and moving transient locations [Vafi 01]. These methods are heuristic in nature and do also not take into account some type of perceptual model.

There are more advanced perceptual models available which do take into account time information. Examples can be found in the field of computational auditory modeling where neural firing patterns are obtained by modeling certain stages of the auditory periphery, e.g., [Lyon 82, Dau 96a]. However, these approaches are not meant for optimization algorithms in (real-time) audio and speech processing applications and, as a consequence, may be computationally demanding. For example, in the advanced auditory model developed by Dau *et al.* [Dau 96a, Dau 96b] (Dau-model) a masking threshold for a given error signal can only be found by using adaptive procedures [Levi 71], as is done in [Dau 96b], and a closed-form analytic expression is not available. This means that when used in a coding environment, for each newly introduced quantization level the model must be applied several times in order to find an estimation of its masking threshold, which is computationally demanding. Another problem with these advanced models is that they are typically not defined for short-time frames, this in contrast to the Par-model and the MPEG-1 model. These properties make it difficult to use these advanced models in the applications we are interested in.

In this article a new distortion measure defined for short-time frames is presented based on a spectro-temporal auditory model. The measure is simplified under certain assumptions valid for the applications of interest in this article (e.g., coding, watermarking). This leads to a more tractable measure in the sense that analytic expressions now exist for masking thresholds. Furthermore, it will be shown that the proposed methods predict similar masking thresholds compared to an advanced spectro-temporal model with a large reduction in complexity.

3.2 Preliminaries

Let x and y denote two finite length discrete-time signals of length N, representing the original and degraded audio signal, respectively. The degraded signal will be written as $y = x+\varepsilon$, where ε can be interpreted as the introduced degradation by the system of interest (e.g., quantization noise). The N-point DFT of x, say $\hat{x}$, is defined as,

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi kn/N}, \quad k = 0, ..., N-1, \tag{3.1}$$

where k represents the DFT-bin index, j the imaginary unit and n the time index. Similar definitions hold for $\hat{y}$ and $\hat{\varepsilon}$. Furthermore, circular convolution will be denoted by $x \circledast y$. The ℓ_p-norm of x is defined as,

$$\|x\|_p = \left(\sum_n |x(n)|^p \right)^{1/p}. \tag{3.2}$$

In this work we assume that all time-domain signals and filters are real valued.

3.3 Proposed Spectro-Temporal Distortion Measure

Fig. 3.1 shows the structure of the proposed method. First, an auditory model, which mimics certain stages of the auditory periphery, is applied to the clean and degraded signal in order to obtain their corresponding internal representations, denoted by $I_{x,i}$ and $I_{y,i}$, respectively, where i denotes the auditory channel. A perceptual difference is then defined by applying a distance measure between the internal representations denoted by 'perceptual distance' in the figure. Note that this approach of modeling stages of the auditory periphery and comparing these signals in a spectro-temporal auditory domain is typically used by more advanced perceptual models, e.g., [Dau 96a, Lyon 82, Rix 02, Thie 00], and not by short-time models used in online optimization algorithms (like the Par-model) due to complexity reasons. However, we will show that under certain assumptions the complexity of such an advanced auditory modeling approach can be greatly reduced.

In Section 3.3.1 more details will be given about the auditory model we use, followed by defining a perceptual distance measure between these internal representations in Section 3.3.2. Then, under certain assumptions, the model will be simplified in order to reduce its complexity in Section 3.3.3, followed by some implementational details in Section 3.3.4.

3.3.1 Auditory Model

The auditory model consists of a filter representing the frequency characteristics of the outer and middle ear, followed by an auditory filter bank resembling

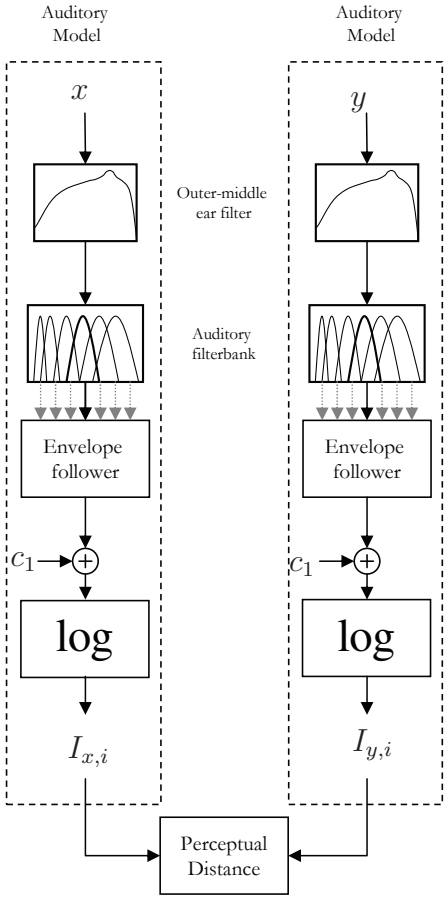

Figure 3.1: Basic structure of the proposed model, which compares the internal representations I_x and I_y of the clean (x) and degraded (y) audio signal, respectively. First an outer middle ear filter is applied followed by an auditory filter bank. The haircell transduction staged is modeled by an envelope follower. Finally, a log-transform is applied to mimic the compressive properties of the outer haircells after which the internal representations are compared by means of applying a distance measure (see text below for more details).

the properties of the basilar membrane in the cochlea. An envelope extraction stage is used to simulate the properties of the hair-cell transduction. Subsequently, a constant is added to represent physiological internal noise (caused by muscle activity, blood streams, etc.) in order to introduce an absolute hearing threshold. Finally, a log transform is applied to resemble the compressive behavior due to the outer hair-cells.

For the outer-middle ear filter a magnitude spectrum equal to the inverse of the threshold in quiet is used to let the model correctly predict the absolute hearing threshold. This threshold describes the playback level of a sinusoid, such that it is just not perceived by an average listener. A mathematical expression approximating the threshold in quiet can be found in [Pain 00]. For the auditory filter bank the same gammatone-based approach as in [Par 05] is used. In total 64 filters are used where the center frequencies are linearly spaced on an ERB-scale between 0 and $f_s/2$ Hz, where f_s denotes the sample rate.

Let h_i denote the joint impulse response of the outer middle ear filter and the i^{th} auditory filter where x filtered by h_i is denoted by $x_i = x * h_i$. Similarly we have $y_i = y * h_i$. Per channel, the envelope extraction stage is included by taking the absolute squared value followed by a low-pass filter, say h_s. With this, a mathematical description of the internal representation of x in the i^{th} auditory filter can then be written as,

$$I_{x,i} = \log\left(|x_i|^2 * h_s + c_1\right), \tag{3.3}$$

where c_1 denotes the constant representing internal noise. Similarly, the internal representation of y can be defined as,

$$I_{y,i} = \log\left(|y_i|^2 * h_s + c_1\right). \tag{3.4}$$

3.3.2 Perceptual Distance between Internal Representations

In order to define a perceptual difference between x and y, their corresponding internal representations $I_{x,i}$ and $I_{y,i}$ should be compared somehow. One procedure is to apply an ℓ_p-norm on the difference between the internal representations of the clean and degraded audio signal, where increasing p will give more importance to high-energy regions in the eventual distance measure, e.g., spectral peaks in vowels. In this paper we choose $p = 1$. As we will show (see Section 3.5), for this choice of p the measure can be simplified into a mathematical tractable distortion measure while predicting results with sufficient accuracy are obtained compared to psychoacoustic listening experiments.

Applying an ℓ_1 norm to the difference between the internal representations gives a within-channel detectability defined by,

$$d_i(x, y) = \|I_{y,i} - I_{x,i}\|_1 . \tag{3.5}$$

These within-channel detectabilities are then combined by means of a summation in order to include the spectral integration properties of the auditory system,

$$\begin{aligned} d(x,y) &= c_2 \sum_i d_i(x,y) \\ &= c_2 \sum_i \|I_{y,i} - I_{x,i}\|_1 \\ &= c_2 \sum_i \left\| \log\left(\frac{|y_i|^2 * h_s + c_1}{|x_i|^2 * h_s + c_1} \right) \right\|_1, \end{aligned} \tag{3.6}$$

where an additional calibration constant c_2 is included in order to set the sensitivity of the model (see Section 3.3.4).

3.3.3 Low-complexity Approximation

Eq. (3.6) can be approximated by a simpler form which leads to an analytical expression for the masking threshold as we will show in Section 3.4. We assume that x and ε are uncorrelated, i.e., $E(X\mathcal{E}) = 0$, which gives the possibility to discard certain cross-terms in the within-channel temporal envelope of y. This assumption is typically valid for quantization noise in audio coders but also in data-hiding applications like watermarking. The within-channel temporal envelope of y can be expressed as,

$$|y_i|^2 * h_s = |x_i + \varepsilon_i|^2 * h_s = |x_i|^2 * h_s + |\varepsilon_i|^2 * h_s + 2(x_i\varepsilon_i) * h_s. \tag{3.7}$$

As a consequence of the averaging properties of the smoothing low-pass filter h_s and the assumption that x and ε are uncorrelated, it holds that,

$$2(x_i\varepsilon_i) * h_s \approx 2E(X_i\mathcal{E}_i) = 0 \tag{3.8}$$

Motivated by this the following approximation is used,

$$|y_i|^2 * h_s \approx \left(|x_i|^2 + |\varepsilon_i|^2 \right) * h_s. \tag{3.9}$$

By combining Eq. (3.9) and Eq. (3.6) we get,

$$d(x,y) \approx c_2 \sum_i \left\| \log\left(1 + \frac{|\varepsilon_i|^2 * h_s}{|x_i|^2 * h_s + c_1} \right) \right\|_1. \tag{3.10}$$

Next, we assume that only small errors are introduced to the clean signal which is typically the case in masking situations. Therefore, a good approximation of each element in the summation of Eq. (3.10) can be obtained by only taking into account the first term of the Maclaurin series expansion of $\log(1+z) \approx z$. That gives us the final expression for the new simplified measure, which will be denoted by D. That is,

$$d(x,y) \approx D(x,\varepsilon) \triangleq c_2 \sum_i \left\| \frac{|\varepsilon_i|^2 * h_s}{|x_i|^2 * h_s + c_1} \right\|_1 . \tag{3.11}$$

For high playback level, i.e., $|x_i|^2 * h_s \gg c_1$, the measure reduces to a spectro-temporal, noise-to-signal ratio per auditory band. For very low playback levels, i.e., $|x_i|^2 * h_s \ll c_1$, it can be observed that the constant c_1 will dominate the denominator and therefore an absolute threshold in quiet is introduced.

3.3.4 Implementation Details

The parameters c_1 and c_2 are calibrated such that the model correctly predicts the threshold in quiet at 1 kHz and the 1 dB just noticeable level difference for a 70 dB SPL, 1 kHz tone (see also [Par 05]). It is assumed that an additive distortion ε is just not detectable when $D = 1$. For this procedure the playback level of the audio signals must be known where we assume that the maximum playback level is 96 dB SPL.

For complexity reasons the outer-middle ear filter, the auditory filter bank and the smoothing low-pass filter are all applied by means of a point-wise multiplication in the DFT-domain, where we assume that all filters have a real-valued, even-symmetric frequency response, i.e., $\hat{h}(k) = \hat{h}(-k)$. This particular choice will lead to time-domain aliasing due to circular convolution, however, proper windowing is used to minimize the effect of these unwanted artifacts. For the smoothing lowpass filter h_s the magnitude response of a one-pole filter is used with cutoff frequency $f_c = 1000$ Hz. The cut-off frequency controls the sensitivity of the model towards the temporal structure of the clean and degraded signals. The particular choice of $f_c = 1000$ roughly simulates the transduction properties of the inner hair cells [Dau 96a]. Let $a = -e^{-2\pi f_c/f_s}$. The frequency response of h_s is then given by,

$$\hat{h}_s(k) = \frac{(1+a)}{\sqrt{1 + a^2 + 2a\cos(2\pi k/N)}}. \tag{3.12}$$

In order to save computational power the denominator in Eq. (3.11), i.e., $|x_i|^2 * h_s + c_1$, can be pre-calculated independent of ε. The measure can then be evaluated for any introduced error by just calculating the spectro-temporal envelope of ε divided by this pre-calculated term. In fact, the following gain-function can be pre-calculated independent of ε,

$$g_i^2 = \frac{c_2}{|x_i|^2 \circledast h_s + c_1} \circledast h_s, \tag{3.13}$$

where the measure can then be expressed as follows (see Appendix 3.A),

$$D(x,\varepsilon) = \sum_i \|\varepsilon_i g_i\|_2^2. \tag{3.14}$$

The measure can now be evaluated for any arbitrary error just by applying the DFT-based filter bank followed by a spectro-temporal gain function.

3.4 Masking

3.4.1 Masking Threshold

Many applications are interested in a masking threshold of ε given x, i.e., the maximum level of ε such that it is just not detectable in the presence of x. This threshold can be found by solving $d(x, x + \alpha\varepsilon) = 1$ for α, where α is a scalar controlling the level of the introduced error. Notice that with the distance measure as defined in Eq. (3.6) it is not straightforward to determine a masking threshold. Instead of an analytical solution, a typical approach is to use adaptive procedures similarly to what is done with real listening experiments [Levi 71]. However, many iterations may be needed to determine an estimate of the masking threshold which may be computationally demanding. In addition, depending on the application the procedure has to be repeated for many different error signals ε. Nevertheless, due to the introduced simplifications for the proposed model, as explained in Section 3.3.3, we now have the relation $D(x, \alpha\varepsilon) = \alpha^2 D(x, \varepsilon)$. This gives the following solution for the masking threshold,

$$\alpha = \frac{1}{\sqrt{D(x, \varepsilon)}} \tag{3.15}$$

3.4.2 Masking Curve

In applications like [Par 05, Heus 06] knowledge of the masking curve is required which describes the masking threshold for a (windowed) sinusoid as a function of frequency. This masking curve will provide information on how to shape the spectrum of an introduced error such that perceptual impact of the error is minimized.

Unfortunately, evaluating Eq. (3.15) for all frequencies of interest (from 0 to $f_s/2$) may be computationally demanding. However, due to the introduced simplifications of the model as explained in the previous section an efficient DFT-based expression for the masking curve can be obtained. Let a windowed sinusoid (e.g., Hann window) be denoted by $\varepsilon_k(n) = w(n)\cos(2\pi kn/N)$, where N is the DFT-size and k/N the normalized frequency of the sinusoid. For slowly time-varying windows the output of the auditory filter bank can be approximated as,

$$\varepsilon_k \circledast h_i \approx \hat{h}_i(k)\,\varepsilon_k. \tag{3.16}$$

Note, that the auditory filters were defined such that they have a real-valued spectrum. Hence, no phase shifts and group delays have to be taken into account. Fig. 3.2 shows an example where the actual within channel temporal envelope, i.e., $|\varepsilon_k \circledast h_i|^2 \circledast h_s$, and the estimated within channel temporal envelopes based on Eq. (3.16) are plotted for a 200 Hz and 2000 Hz sinusoid. The plot only shows the auditory filter where its center frequency is closest to

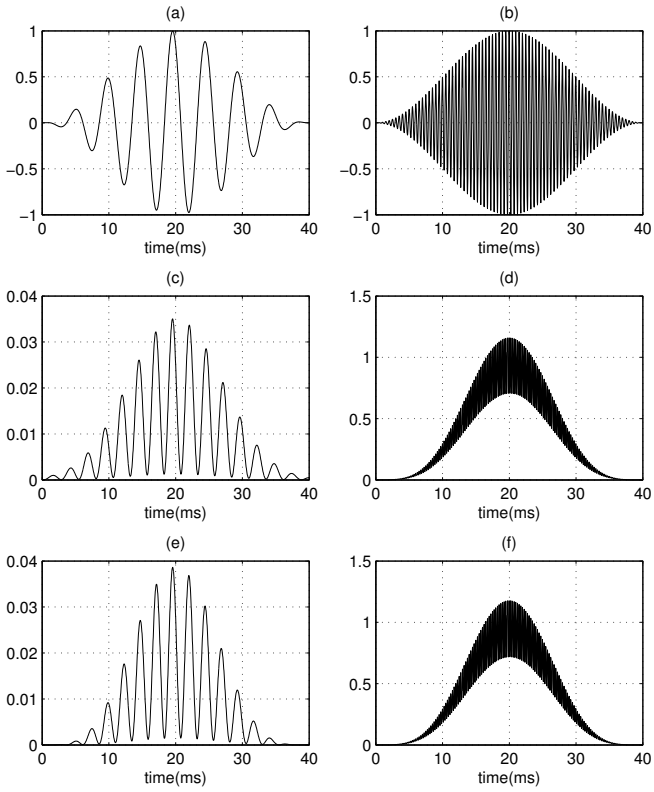

Figure 3.2: (a) Windowed sinusoid of 200 Hz with (c) corresponding temporal envelope as defined in Eq. (3.16) and (e) approximated temporal envelope as explained in Section 3.4.2. Similar plots are shown in (b), (d) and (e) for a 2000 Hz sinusoid. Only the auditory filter is shown where its center frequency is closest to the frequency of the sinusoid.

the frequency of the sinusoid. The figure reveals that a good approximation is obtained of the actual within channel temporal envelope for both frequencies.

In order to define a masking curve we have to solve $D\left(x, \alpha\left(k\right)\varepsilon_k\right) = 1$ for $\alpha\left(k\right)$. By using the approximation in Eq. (3.16) this gives,

$$\frac{1}{\alpha^2\left(k\right)} = \sum_i \hat{h}_i^2\left(k\right) \left\|\varepsilon_k g_i\right\|_2^2, \tag{3.17}$$

which can be rewritten in the following form,

$$\begin{aligned}\frac{1}{\alpha^2(k)} = & \sum_i \hat{h}_i^2(k) \sum_n |(wg_i)(n)|^2 \left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{4\pi kn}{N}\right)\right) \\ = & \frac{1}{2}\sum_i \hat{h}_i^2(k) \times \\ & \left(\|wg_i\|^2 + \sum_n |(wg_i)(n)|^2 \cos\left(\frac{4\pi kn}{N}\right)\right).\end{aligned} \tag{3.18}$$

Eq. (3.18) can be expressed in terms of the DFT of the gain function for each auditory band multiplied with the squared window function, i.e., $\widehat{|wg_i|^2}$. That is,

$$\frac{1}{\alpha^2(k)} = \sum_i \frac{\hat{h}_i^2(k)}{2}\left(\widehat{|wg_i|^2}(0) + \mathrm{Re}\left\{\widehat{|wg_i|^2}(2k)\right\}\right), \tag{3.19}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of any arbitrary complex number. From this equation we can conclude that a complete masking curve can now be obtained by exploiting the (Fast) Fourier transform for $\widehat{|wg_i|^2}$ for each auditory band. Note, that this is a significant reduce in complexity compared to evaluating Eq. (3.15) for each sinusoid individually with frequency $k = 0, 1, ..., N/2$.

3.5 Model Evaluation and Comparison

To evaluate the proposed method, comparisons will be made with a sophisticated spectro-temporal model as proposed by Dau *et al.* [Dau 96a, Dau 96b] and a simpler spectral-only model by van de Par *et al.* [Par 05]. We will demonstrate that the proposed method shares some of the benefits of the complex Dau-model with respect to predicting masking thresholds for non-stationary signals, while it has a similar mathematical tractable form like the Par-model. First both reference models are explained after which comparisons are made by means of predicting masking curves and computational complexity.

3.5.1 Reference models

Par-model

The Par-model is based on the energy detection model from the field of signal detection theory as proposed by Green and Swets [Gree 66], where the task is to detect a probe (e.g., sinusoid) in the presence of some masker (e.g., white noise). For this model it is assumed that at the output of an auditory filter, the signal is absolute squared followed by a temporal integration procedure (note that this model is of a simpler form than the one which is used in the proposed method from Fig. 3.1). As a consequence, the listener observes the stimulus power at the output of an auditory band which is considered to be stochastic

(e.g., due to internal noise). Under the assumption that the stochastic processes are i.i.d. Gaussian and that the auditory system uses an optimal detector to detect the probe in presence of the masker it can be shown that the ratio between the increase in probe power and the standard deviation of the masker is defined as the sensitivity index d' [Gree 66]. The sensitivity index (i.e., distortion detectability) is monotonically increasing related to the probability of correctly detecting the probe in presence of the masker (i.e., a higher d' implies a higher probability of correctly detecting the probe in presence of the masker).

Van de Par *et al.* [Par 05] suggested to combine the within-channel sensitivity indices over all auditory bands by means of an additive operation in order to mimic the spectral integration properties of the auditory system (see e.g., [Buus 86, Lang 92]). Temporal integration is included by multiplying this summation with a factor N. As a consequence, increasing the playback length of a signal will result in a higher predicted detectability, which is in accordance with a human observer up till lengths of approximately 300 ms [Brin 64]. Similar as with the proposed method the auditory filters are implemented by means of a point-wise multiplication in the DFT-domain, hence, a circular convolution in the time-domain. This leads to the following perceptual distortion measure,

$$D_{par}\left(x,\varepsilon\right) = Nc_2 \sum_i \frac{\frac{1}{N}\left\|\varepsilon_i\right\|_2^2}{\frac{1}{N}\left\|x_i\right\|_2^2 + c_1}, \tag{3.20}$$

where c_1 is included in order to introduce a threshold in quiet and c_2 is used to modify the sensitivity of the model. Both parameters are calibrated such that the model correctly predicts the masking threshold of a 1 kHz tone in silence and the 1 dB just noticeable level difference for a 70 dB SPL, 1 kHz tone. The model is calibrated such that $D_{par} = 1$ corresponds to a distortion at the threshold of detection of ε [Par 05].

Note, that the Par-model also has an efficient implementation, where a gain function only depending on x can be pre-calculated (similarly as in Eq. (3.14)). By using Parseval's theorem, i.e., $\left\|x\right\|^2 = \frac{1}{N}\left\|\hat{x}\right\|^2$, the following spectral weighting function can be used,

$$\hat{g}_{par}^2\left(k\right) = \sum_i \frac{h_i^2\left(k\right) c_2}{\frac{1}{N}\left\|\hat{x}_i\right\|^2 + Nc_1}, \tag{3.21}$$

to express the Par-model as an efficient frequency weighted ℓ_2 norm [Par 05],

$$D_{par}\left(x,\varepsilon\right) = \left\|\hat{\varepsilon}\hat{g}_{par}\right\|_2^2. \tag{3.22}$$

Van de Par *et al.* have shown that the masking curve for the Par-model can be directly related to the inverse of this spectral weighting function $\hat{g}_{par}$ [Par 05]. However, the masking curve in [Par 05] is based on rectangular-windowed, normalized complex exponentials rather than sinusoids. By introducing a normalization factor $\sqrt{2}/N$ a full masking curve for rectangular win-

dowed sinusoids is given as follows (an efficient expression for the masking curve for other types of windows is not defined in [Par 05]),

$$\alpha(k) = \frac{\sqrt{2}}{\hat{g}_{par}(k) N} \tag{3.23}$$

Dau-model

The Dau-model acts as an artificial observer and is originally used for accurately predicting masking thresholds for various masking conditions [Dau 96a, Dau 96b]. It has a similar approach as the proposed method in the sense that it compares internal spectro-temporal representations. In order to obtain an internal representation, a 64-channel auditory filterbank is first applied, where the haircell transduction process is modeled by half-wave rectification followed by a 1 kHz low-pass filter. To introduce an absolute threshold, the hair cell output is limited to a minimum value. The auditory model is more advanced in the sense that it also models the non-linear properties of the auditory system due to neural adaptation. This is incorporated by means of the so-called adaptation loops, which will put more emphasis on strong temporal fluctuations, e.g., transients, while more stationary sounds are converted approximately logarithmically [Dau 96a]. Temporal integration of the auditory system is included by means of a 8 Hz low-pass filter per auditory band, followed by addition of internal noise simulated by Gaussian i.i.d. white noise. To let the model correctly predict the threshold in quiet, an outer-middle ear filter is applied before the auditory filterbank, similarly as with the proposed and Par-model.

In [Dau 96a], the perceptual distance between two signals is determined by a correlation based comparison. Due to the addition of internal noise, the internal representations are stochastic and therefore this perceptual distance is also stochastic (similarly as with a real listener). Since we are interested in the average behavior of the model we use the approach from [Kohl 08] and [Plas 07], where it has been shown that the average detectability can be described by summing the squared ℓ_2 norms between the internal representations, per auditory band. Let $\Psi_{x,i}$ and $\Psi_{y,i}$ denote the time-domain signals of the internal representations for the i^{th} auditory band of the clean and degraded signal, respectively. In line with [Kohl 08] its perceptual distance is then defined by,

$$D_{dau}(x, y) = \frac{1}{\sigma} \sqrt{\sum_i \|\psi_{x,i} - \psi_{y,i}\|_2^2}, \tag{3.24}$$

where σ represents the standard deviation of the internal noise. The calibration of σ and the used minimum value to limit the haircell output is done similarly as with the proposed method and the Par-model.

Note that for the Dau-model no analytic expression exists to obtain a masking threshold, in contrast to the Par-model and the proposed model. Instead,

we use the bisection method to estimate the masking thresholds. The iterative procedure was stopped when the error was smaller than 0.1 dB. In order to obtain a masking curve, the masking threshold is determined for a limited set of 30 sinusoids, with frequencies logarithmically spaced between 100 and 10000 Hz. We found that 10-20 iterations was typically sufficient to obtain an estimate of the masking threshold.

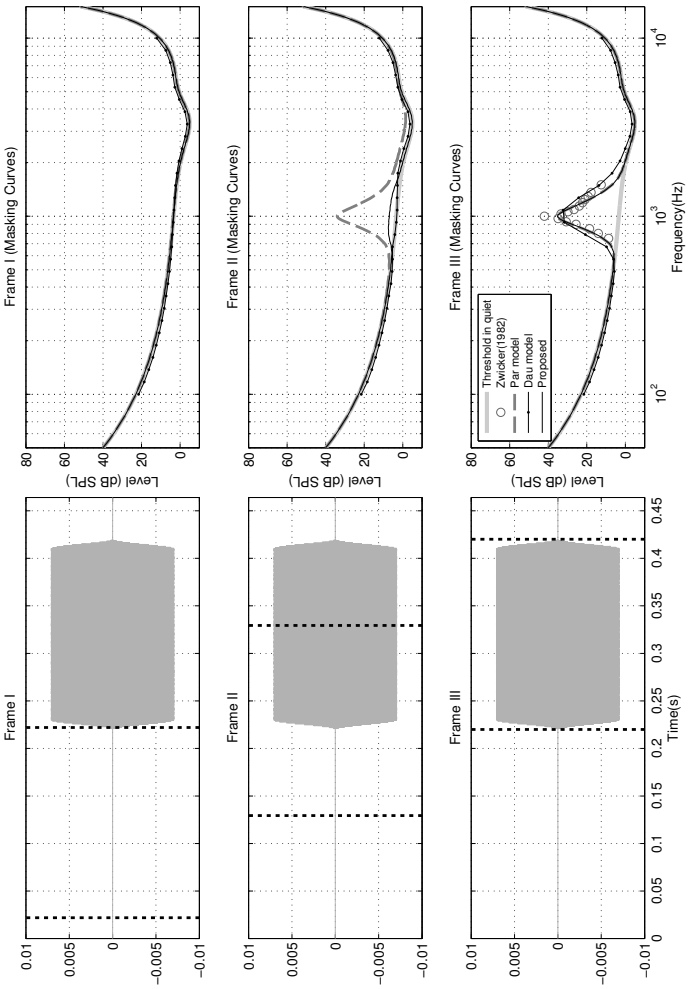

Figure 3.3: Example to illustrate the difference between the proposed method, the spectro-temporal Dau-model and the Par-model [Par 05] which is only based on spectral information. Masking curves are predicted by all models before (Frame I), during (Frame II) and after (Frame III) the onset of a 50 dB SPL, 1 kHz tonal masker with a length of 200 ms (subplots at the left). Their corresponding predicted masking curves are show in the right column plots, where the open circles in the bottom-right plot denote results from psychoacoustic listening experiments [Zwic 82].

3.5.2 Prediction of Masking Curves

To illustrate the correspondences and the differences between the two reference models and the proposed model several masking curves will be predicted. For all models a sample-rate of 44.1 kHz is used.

Masking curves are predicted for a 50 dB SPL, 1 kHz tonal masker with a length of 200 ms including 10 ms ramps. However, in this case three different time segments are analyzed as shown in Fig. 3.3, where masking curves are predicted before, during and after the onset of the tonal masker, denoted by Frame I, II and III, respectively, in the figure. The first frame contains only silence, the second frame partly silence followed by a part of the sinusoid and the last frame is the complete windowed sinusoid. The three plots on the right show the predicted masking curves for all models. The bottom-right plot also contains results from psychoacoustic listening tests [Zwic 82] to evaluate the model predictions.

For the first frame it can be observed that the predictions for all three models are in correspondence, where they correctly predict the masking curve to be equal to the threshold in quiet. However, for the second frame a clear difference is observed for the Par-model. While the proposed method and the Dau-model both predict a masking curve close to the threshold in quiet, the Par-model discards the preceding silence of the masker which leads to a significantly higher masking curve. Since backward masking (see, e.g., [Zwic 90, Moor 03]) is only present from a few milliseconds before the onset of the masker, the masking curve for the second frame should be close to the threshold in quiet. This is in correspondence with the results predicted by the proposed method and the Dau-model. For the third frame, the sinusoidal masker is present in the complete frame, therefore the predicted masking curves for all models are similar. In the bottom right plot results from psychoacoustic listening experiments are shown [Zwic 82] on top of the predicted masking curves, which are in accordance with the predictions for all models.

A similar example is illustrated in Figs. 3.4 and 3.5, which show a short-time segment of speech for a transient and a vowel region, respectively. In both figures the spectrum is downscaled for visual clarity. For the transient region one can clearly see that the masking curve is much higher for the Par-model compared to the proposed method and the Dau-model. Hence, the proposed method detects the sensitivity towards an introduced error before the onset of the transient similarly as the advanced Dau-model. Employing this property in an audio-coding context will lead to, e.g., less pre-echoes or more intelligible consonants. All three models are more in correspondence for the predicted masking curves for the vowel region as is shown in Fig. 3.5. This is due to the fact that the within-temporal envelopes of the vowel have more or less the same temporal structure as the windowed sinusoids which determine the masking curve.

Notice that the masking curves for the Dau-model are slightly lower for lower frequencies compared to the proposed model in Figs. 3.4 and 3.5. A possible cause for this could be the sensitivity of the adaptation loops towards

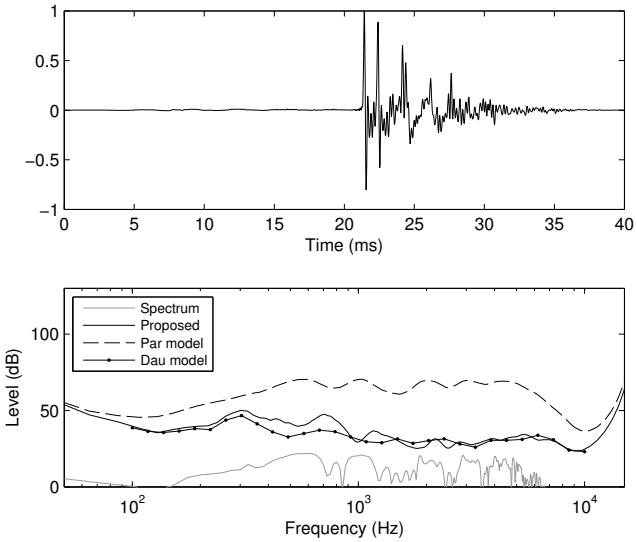

Figure 3.4: A short-time (40 ms) transient region of speech (top plot) with predicted masking curves for the proposed method, the Par-model [Par 05] and the Dau-model [Dau 96a] (bottom plot). The spectrum is down-scaled for visual clarity.

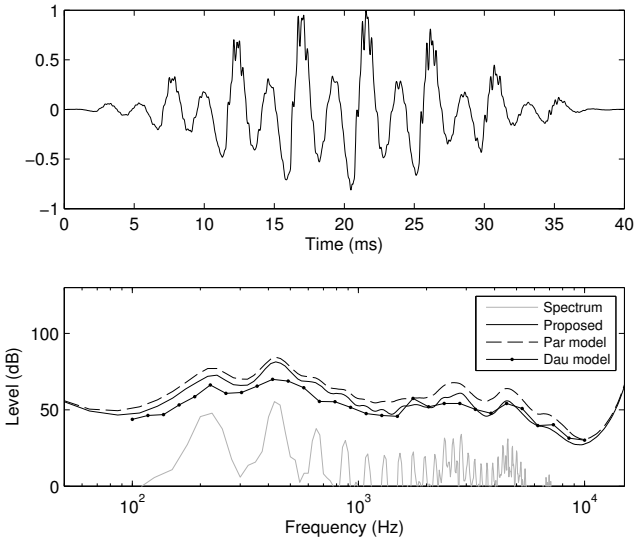

Figure 3.5: A short-time (40 ms) vowel region of speech (top plot) with predicted masking curves (mc) for the proposed method, the Par-model [Par 05] and the Dau-model [Dau 96a](bottom plot). The spectrum is down-scaled for visual clarity.

Table 3.1: Normalized processing-time.

Frame length N	128	256	512	1024	2048
	1) Model evaluation				
Par-model, Eq. (3.20)	0.55	0.69	1.00	1.64	3.99
Proposed, Eq. (3.11)	1.30	2.02	3.22	7.95	16.35
Dau-model, Eq. (3.24)	140.38	142.12	148.22	159.50	184.24
	2) Model evaluation with fixed x				
Par-model, Eq. (3.22)	0.26	0.26	0.28	0.29	0.35
Proposed, Eq. (3.14)	0.51	0.70	1.04	1.94	4.05
Dau-model, Eq. (3.24)	71.00	72.30	76.81	79.83	92.35
	3) Masking curve prediction				
Par-model, Eq. (3.23)	0.15	0.22	0.39	0.91	1.63
Proposed, Eq. (3.19)	0.86	1.34	2.46	9.07	21.48
Dau-model	No analytic expression available				

the preserved phase structure at lower auditory bands. However, the difference between the proposed model and the Dau-model is much smaller compared to the masking curve overestimation for the Par-model for the transient signal. We also would like to add that the Dau-model can also predict masking effects due to neural adaptation, i.e., forward and backward masking [Zwic 90, Moor 03]. This property is not present with the proposed method. However, we believe that for the applications of interest in this work, these masking effects are less important compared to the difference between a spectral-only and a spectro-temporal model.

3.5.3 Complexity

To give an impression of the computational power needed for the proposed method in relation to the two reference models, the computation time is measured for several frame lengths and conditions. All three models are implemented in Matlab. For the Dau-model the IIR-based auditory filterbank in [Patt 92] is used and the complex adaptation loops are implemented in a C++ - based MEX file for computational efficiency. In total three different processing conditions are considered:

1. Evaluation of the perceptual distance for a given x and ε. This refers to Eqs. (3.11), (3.20) and (3.24) for the proposed, Par and Dau-model, respectively.

2. Evaluation of the perceptual distance for a given ε when x is fixed. This is a relevant situation for, e.g., a rate-distortion loop in a coder. This refers to Eqs. (3.14) and (3.22) for the proposed and Par-model, respectively.

For the Dau-model Eq. (3.24) is used where $\psi_{x,i}$ is pre-calculated once and stored.

3. Evaluation of a complete masking curve given x. This refers to Eqs. (3.19) and (3.23) for the proposed and Par-model, respectively. Note that the Dau-model is not included in this test since no analytic expression exists for a complete masking curve. A masking curve is typically used in data-hiding and coding applications to spectrally shape the introduced error in order to perceptually 'hide' the introduced error more efficiently.

For each condition and model, Gaussian i.i.d. vectors of x and ε are generated[1] for $N \in \{128, 256, 512, 1024, 2048\}$. These are typical frame lengths relevant for digital audio and speech processing applications. The performance for each model, condition and frame length N is obtained by taking an average computation time over 100 evaluations. The results are shown in Table 3.1 where the processing times are normalized with respect to the first condition for the Par-model where $N = 512$. Notice that the numbers given in Table 3.1 are rough estimates that are meant as an indication. In general they depend on implementational details.

From the table it is revealed that the proposed method is a factor 10-100 times faster than the Dau-model, depending on the frame length and type of test. The main reason for this difference in performance is most likely the use of a log-transform instead of the sophisticated adaptation loops and the use of an FFT-based filterbank instead of the IIR-based gammatone filters. Despite the fact that the Dau-model has no analytic expression for the masking curve available, an estimation of this curve could be obtained by means of an adaptive procedure per sinusoid (as explained in Section 3.5.1). However, this means that we have to evaluate the Dau-model for each of the $(N/2 + 1)$ sinusoids, multiplied with the number of iterations needed in order to obtain a masking threshold for one sinusoid (10-20 in the experiments from the previous section). Given that the evaluation of a complete masking curve for the proposed model is already much faster than evaluating the Dau-model only *once* (see Table 3.1), one can imagine the large reduction in complexity with the proposed method when one is interested in a masking curve.

Taking into account short-time temporal information comes with a computational cost compared to spectral-only models like the Par-model. This is also what can be concluded from the table where the Par-model is, in general, 3-15 times faster than the proposed model depending on the frame-size and type of test. However, this difference is much smaller than the difference in performance between the proposed model and the Dau-model. Other ways to reduce the computational complexity of the proposed model can be considered by, e.g., reducing the amount of auditory filters.

[1] A more realistic scenario would be to use speech or music for x, however, this will not affect its processing time

3.6 Experimental Results

In this section we demonstrate the properties of the proposed model by means of experimental results and make a comparison with the Par-model. The Dau-model is not included in this comparison since it does not provide the analytical expressions for masking thresholds and masking curves needed in order to generate the signals in the experiment, as will become clear in the remainder of this section.

To illustrate the properties of the proposed model, several audio signals are generated with degradations that are typical for audio and speech processing applications where auditory masking is exploited. A common approach is to spectrally shape the introduced errors according the masking curve in order to perceptually 'hide' the introduced error efficiently. For these applications there is typically a constraint involved which influences the amount of added noise. For example, the total number of bits in an audio coder or the amount of information and robustness of an embedded watermark. For demonstration purposes, these errors are artificially introduced to several clean signals based on the proposed model and the Par-model after which their results are compared.

Clean signals are degraded by i.i.d. Gaussian noise where the noise-only signal is first segmented into short-time (32 ms), 50% overlapping windowed frames and filtered with the predicted masking curve belonging to the corresponding short-time frame of the clean signal. This filtering operation is applied by means of a point-wise multiplication in the DFT-domain, where a square root Hann analysis and synthesis window is used. The total amount of noise that is added to the clean signal is controlled by a constraint on the segmental SNR. The level of the masking-curve filtered noise is adjusted per short-time frame, such that the summation of all individual frame-distortions for the model under consideration is minimized. With this approach it is expected that the proposed method will put less noise in transient regions and add more noise in more stationary frames, in contrast to the Par-model.

Let m denote the frame-index, M the total number of frames, r the segmental SNR constraint in dBs and $\alpha_m \varepsilon_m$ the masking-curve filtered noise for the m^{th} frame. Here α_m is a scalar which controls the level of the noise in that particular frame. The globally optimal distribution of all noise-levels (i.e., α_m for $m = 1, .., M$) is then given by finding the minimum of the following constrained cost function,

$$J\left(\alpha_{1,\ldots,M}, \lambda\right) = \sum_m D\left(x_m, \alpha_m \varepsilon'_m\right) + \lambda \left(\frac{1}{M} \sum_m 20 \log_{10} \left(\frac{\|x_m\|_2}{\|\alpha_m \varepsilon'_m\|_2} \right) - r \right), \quad (3.25)$$

where $\varepsilon'_m = \varepsilon_m \|x_m\|_2 / \|\varepsilon_m\|_2$ denotes a normalized version of ε_m, which implies $\|x_m\|_2 = \|\varepsilon'_m\|_2$. As a consequence of this normalization and using the

relation $D\left(x_m, \alpha_m \varepsilon_m\right) = \alpha_m^2 D\left(x_m, \varepsilon_m\right)$ of Eq. (3.11), the cost function can be expressed as follows,

$$J\left(\alpha_{1,\ldots,M}, \lambda\right) \quad = \quad \sum_m D\left(x_m, \varepsilon'_m\right) \alpha_m^2 \;+\; \lambda' \left(\sum_m \log\left(\alpha_m^2\right) - r' \right). \quad (3.26)$$

where,

$$r' = \frac{-M \log\left(10\right) r}{10}. \quad (3.27)$$

In order to find the optimal distribution of the noise over the frames, given the segmental SNR constraint, the minimum of Eq. (3.26) is found by setting the derivative of the cost function to zero with respect to $\alpha_{1,\ldots,M}$ and λ, that is,

$$\begin{array}{l} \frac{\partial J\left(\alpha_{1,\ldots,M}, \lambda'\right)}{\partial \alpha_m} = 2D\left(x_m, \varepsilon'_m\right) \alpha_m + \frac{2\lambda'}{\alpha_m} = 0 \\ \frac{\partial J\left(\alpha_{1,\ldots,M}, \lambda'\right)}{\partial \lambda'} = \sum_m \log\left(\alpha_m^2\right) - r' = 0 \end{array} . \quad (3.28)$$

Solving this gives,

$$\alpha_l^2 = \frac{\left(e^{r'} \prod_m D\left(x_m, \varepsilon'_m\right) \right)^{1/M}}{D\left(x_l, \varepsilon'_l\right)}, \quad (3.29)$$

where l is used to denote the frame-index of interest. Note, that due to the similarity between the proposed model and the Par-model the derivations for the Par-model in order to distribute the noise is identical. For the proposed model the cutoff frequency of the lowpass filter h_s was lowered to 125 Hz, which resulted in a better noise distribution between transient and stationary frames.

3.6.1 Example

To illustrate the differences in noise distribution between the proposed model and the Par-model, Fig. 3.6 shows the results for the castagnettes excerpt. Here, the segmental SNR was set to 10 dB SNR. In subplot (b) the SNR is plotted per frame, where it can be clearly observed that the proposed method increases the SNR in the frames when a transient is encountered (i.e., the proposed method adds less noise in these frames). The bottom two plots in Fig. 3.6 clearly show that the Par-model adds a lot of noise in the transient regions. The proposed method on the other hand adds more noise in the more stationary regions in order to fulfill the constraint. As will follow from the listening test (see the next section), adding more noise in the transient regions is perceptually more disturbing than the small increment of noise in the non-transient regions.

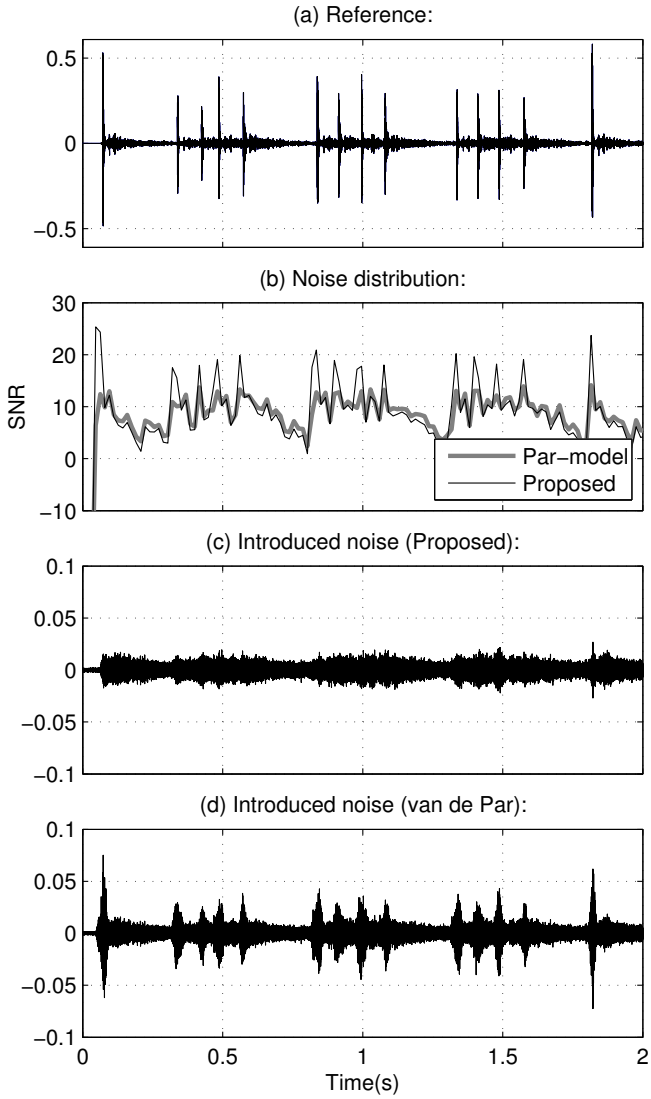

Figure 3.6: Illustration of the noise distribution for the proposed model and the Par-model for the castagnettes excerpt. Subplot (a) shows the clean reference signal, where the distribution of the SNRs per frame for both models is shown in (b). Plots (c) and (d) show the added noise for both models. Notice that the proposed model detects the temporal structure within a short-time frame and puts less noise within transient-frames in contrast to the Par-model.

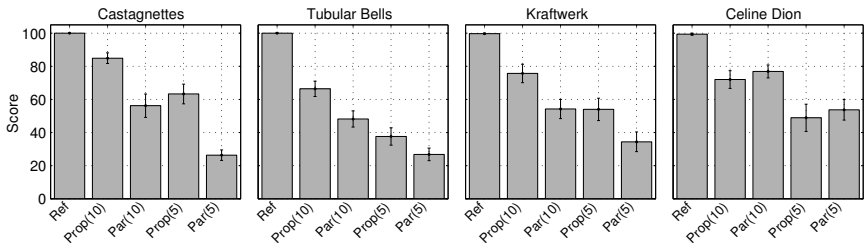

Figure 3.7: Average results and standard errors across all subjects for all of the four excerpts. Noise was added to the reference signals at two different segmental SNRs (5 and 10 dB) for the proposed model (Prop) and the Par-model (Par). Higher scores imply better quality.

3.6.2 Listening Test

The proposed method and the Par-model are compared by means of an informal subjective listening test. Several excerpts are degraded with the noise-distribution procedure as explained in the previous section. A sample rate of 44.1 kHz is used. The excerpts consist of castagnettes, tubular bells, Kraftwerk and Celine Dion which have a length of 7, 12, 12 and 13 seconds, respectively. Here the first three signals have strong transient regions, for which it is expected that the proposed model will show different performance than the Par-model. The Celine Dion fragment contains less transient regions and therefore more similar performance is expected between the two models for this excerpt. The constraints are set to 5 and 10 dB segmental SNR. In total 10 subjects participated in the listening test, which is similar to a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test [ITU 01]. The signals were presented via headphones, where the subjects were able to adjust their volume control to a comfortable level. In total, five different versions for each excerpt had to be ranked on a scale between 0-100 where a higher score denotes better quality. The five signals consist of four degraded versions of the excerpt (2 SNRS for each model) and a hidden reference. The subjects were instructed that a hidden reference was included and were asked to grade this signal with a score of 100. Furthermore, the subjects had access to the clean reference signal for comparison. The participants consisted of employees of Delft University of Technology and have performed in similar listening tests before. They were not connected in any way to this project.

The average scores of the listening test for all subjects are shown in Fig. 3.7 for each excerpt separately. From the results we can conclude that given a segmental SNR, the subjects preferred the proposed method over the Par-model for all signals, except for the Celine Dion excerpt. For Castagnettes and Kraftwerk the proposed model has even similar performance at 5 dB SNR compared to the Par-model at 10 dB SNR. Statistical analysis is performed to verify whether these differences are significant by means of a statistical significance paired t-test for two dependent samples [Shes 04]. The null hypothesis is

Table 3.2: Details on the performed t-tests for the alternative hypothesis that the subjective score for the proposed model is higher than the Par-model.

	Segmental SNR = 10 dB		Segmental SNR = 5 dB	
	Significant?	p-value	Significant?	p-value
Castagnettes	Yes	0.0011	Yes	0.0007
Tubular Bells	Yes	0.0048	Yes	0.0470
Kraftwerk	Yes	0.0106	Yes	0.0227
Celine Dion	No	0.7700	No	0.8785

that both means are equal, while the alternative hypothesis corresponds to the situation that the mean score of the proposed model is higher than the score from the Par-model. Table 3.2 shows the p-values of the likelihood that the null hypothesis is true. The alternative hypothesis is accepted at a significance level of $\alpha = 0.05$. From this analysis it can be concluded that the proposed method shows statistically significant better performance for all excerpts, except Celine Dion. For the Celine Dion fragment the difference between the Par-model and the proposed model was not statistically significant, as was hypothesized.

3.7 Relation Between Proposed Model and the Par-model

In the previous experiments it was shown that the proposed method is more sensitive to transient regions compared to the Par-model. Notice that this sensitivity of the model towards the temporal structure of the signal can be controlled with the cutoff frequency f_c of the smoothing filter h_s. Here, a lower cutoff frequency implies a lower sensitivity towards the temporal structure and hence the model behaves more like a purely spectral distortion measure. In fact, it can be shown that the proposed model and the Par-model are identical when the cutoff frequency f_c of the smoothing low-pass filter h_s is set to 0 Hz in Eq. (3.11). Inspection of Eq. (3.12) shows that for a cutoff frequency of 0 Hz, we get the following magnitude response of h_s,

$$\hat{h}_s(k) = \begin{cases} 1 & k = 0 \\ 0 & \text{otherwise} \end{cases} . \tag{3.30}$$

Recall that the smoothing low-pass filter was implemented as a point-wise multiplication in the DFT-domain. Therefore the output of the within-channel temporal envelope is now equal to its mean squared value,

$$\begin{aligned}\left(|x_i|^2 \circledast h_s\right)(n) &= \tfrac{1}{N}\sum_k \widehat{|x_i|^2}(k)\,\hat{h}_s(k)\,e^{j2\pi nk/N} \\ &= \tfrac{1}{N}\widehat{|x_i|^2}(0)\,\hat{h}_s(0) \\ &= \tfrac{1}{N}\|x_i\|_2^2,\end{aligned} \tag{3.31}$$

Note that the within-channel temporal envelope of x is now a constant value independent of time n. If we follow the same procedure for obtaining the within-channel temporal envelope of the error ε, the distortion measure from Eq. (3.11) can then be expressed as,

$$D(x,\varepsilon) = c_2 \sum_i \left\| \frac{\frac{1}{N}\|\varepsilon_i\|_2^2\, u}{\frac{1}{N}\|x_i\|_2^2\, u + c_1} \right\|_1 \tag{3.32}$$

where $u(n) = 1$ for $n = 0, ..., N-1$. The argument of the ℓ_1 norm is now a constant positive signal, independent of n. Therefore the summation over n in this norm can be replaced by a multiplication with the total signal length N, which, in fact, gives the expression for the Par-model,

$$D(x,\varepsilon) = Nc_2 \sum_i \frac{\frac{1}{N}\|\varepsilon_i\|_2^2}{\frac{1}{N}\|x_i\|_2^2 + c_1} = D_{par}(x,\varepsilon). \tag{3.33}$$

Note that the underlying auditory model of the Par-model is of a simpler form than the auditory model of the proposed spectro-temporal distortion measure (as explained in Section 3.3.1). For example, a hair-cell model and a log-transform are not taken into account. With Eq. (3.33) we can conclude that the Par-model can actually be derived from a more complex auditory model if and only if $f_c = 0$. Also of interest is the multiplication with N in Eq. (3.33), which follows directly from the derivations. In the Par-model this multiplication was artificially introduced in order to include the temporal integration properties of the auditory system [Par 05].

3.8 Conclusions

A new perceptual distortion measure is presented based on a sophisticated spectro-temporal auditory model, which is simplified under certain assumptions valid for auditory masking applications like coding or watermarking. This led to a more tractable distortion measure in the sense that analytic expressions now exist for masking thresholds. This is typically not the case for more advanced spectro-temporal models, which need computationally demanding adaptive procedures to estimate masking thresholds. Furthermore, the distortion measure is of a simpler form since it can be evaluated for any arbitrary error just by applying a DFT-based auditory filter bank, followed by a multiplication with a spectro-temporal gain function. This gain function is only dependent on the clean signal and denotes the sensitivity to errors over time and frequency and can be reused for any arbitrary error. The proposed method

gave similar masking predictions as the advanced spectro-temporal Dau-model with only a fraction of its computational power.

It has been shown that the proposed model can be interpreted as an extended version of the Par-model: a perceptual model based on spectral integration which ignores time-information. The benefits of the proposed method compared to the Par-model are made clear in several experiments, from which it can be concluded that for non-stationary frames (e.g., transients) the Par-model underestimates the audibility of introduced errors and therefore overestimates the masking curve. As a consequence, the system of interest incorrectly assumes that errors are masked in a particular frame which may lead to audible artifacts like pre-echoes. This was not the case with the proposed method which correctly detects the errors made in the temporal structure of the signal.

3.A Derivation of spectro-temporal gain function g_i

In this appendix it will be shown how to rewrite Eq. (3.11) to Eq. (3.14). Recall that the distortion measure was defined as follows,

$$D\left(x,\varepsilon\right)=c_2\sum_i\left\|\frac{\left|\varepsilon_i\right|^2\circledast h_s}{\left|x_i\right|^2\circledast h_s+c_1}\right\|_1. \tag{3.34}$$

Next we use the fact that the argument of the ℓ_1 norm in Eq. (3.34) is positive and the property $\|z\|_1=\left\|z^{1/2}\right\|_2^2$ when $z\geq 0$. By defining the signal,

$$b_i=\frac{c_2}{\left|x_i\right|^2\circledast h_s+c_1}, \tag{3.35}$$

the distortion measure can now be expressed in terms of an inner product,

$$\begin{aligned} D\left(x,\varepsilon\right) &= \sum_i\left\|\left(\left(\left|\varepsilon_i\right|^2\circledast h_s\right)b_i\right)^{\frac{1}{2}}\right\|_2 \\ &= \sum_i\left\langle\left(\left(\left|\varepsilon_i\right|^2\circledast h_s\right)b_i\right)^{\frac{1}{2}},\left(\left(\left|\varepsilon_i\right|^2\circledast h_s\right)b_i\right)^{\frac{1}{2}}\right\rangle \\ &= \sum_i\left\langle\left|\varepsilon_i\right|^2\circledast h_s,b_i\right\rangle. \end{aligned} \tag{3.36}$$

By applying Parseval's theorem we get the following expression in the frequency domain,

$$D\left(x,\varepsilon\right)=\frac{1}{N}\sum_i\left\langle\widehat{\left(\left|\varepsilon_i\right|^2\circledast h_s\right)},\hat{b}_i\right\rangle. \tag{3.37}$$

By using the duality of a circular convolution in the time-domain and a pointwise multiplication in the frequency domain we have,

$$D\left(x,\varepsilon\right)=\frac{1}{N}\sum_{i}\left\langle \widehat{\left(\left|\varepsilon_i\right|^2\right)}\hat{h}_s,\hat{b}_i\right\rangle=\frac{1}{N}\sum_{i}\left\langle \widehat{\left(\left|\varepsilon_i\right|^2\right)},\hat{b}_i\hat{h}_s^*\right\rangle. \tag{3.38}$$

Since $\hat{h}_s$ was defined real (see Section 3.3.4) we have that $\hat{h}_s=\hat{h}_s^*$. Therefore, by applying Parseval's theorem again the following measure in the time-domain is obtained,

$$D\left(x,\varepsilon\right)=\sum_{n}\left\langle \left|\varepsilon_i\right|^2,b_i\circledast h_s\right\rangle \tag{3.39}$$

Now let,

$$g_i^2=b_i\circledast h_s=\frac{c_2}{\left|x_i\right|^2\circledast h_s+c_1}\circledast h_s, \tag{3.40}$$

be defined as a spectro-temporal varying gain function. Due to the fact that $g_i\geq 0$, the proposed method can now be written as a summation of weighted ℓ_2 norms per channel,

$$D\left(x,\varepsilon\right)=\sum_{n,i}\left|\varepsilon_i\left(n\right)g_i\left(n\right)\right|^2=\sum_{i}\left\|\varepsilon_i g_i\right\|_2^2 \tag{3.41}$$

Chapter 4

An Evaluation of Objective Measures for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech

The following article appeared in "An Evaluation of Objective Measures for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech", by C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen in the *Journal of the Acoustical Society of America*, vol. 130, Issue 5, pages 3013 - 3027, 2011, and may be found at (http://link.aip.org/link/?JASMAN/130/3013/1).

4.1 Introduction

Speech processing systems often introduce degradations and modifications to speech signals, e.g. quantization noise in a speech coder or residual noise and speech distortion in a noise reduction scheme. To determine the perceptual consequences of these artifacts, the algorithm at hand can be evaluated by means of a listening test or an objective machine-driven quality assessment. Although a listening test can lead to a judgment as observed by the intended group of users, such tests are often costly and time consuming. Therefore, accurate and reliable objective evaluation methods are of interest since they might replace a listening test, at least in some stages of the algorithm development process. Although it is not straightforward to describe the overall quality of a speech processing system, people tend to divide the evaluation into the attributes of speech quality, (i.e. pleasantness/naturalness of speech) and speech intelligibility. The primary focus of this work is on speech intelligibility.

One of the first objective intelligibility measures was developed at AT&T Bell Labs around 1920 and eventually published by [Fren 47]. [Kryt 62] made the measure better accessible by proposing a calculation scheme, which is currently known as the articulation index (AI). The basic approach of AI is to determine the signal-to-noise ratio (SNR) within several frequency bands; the SNRs are then limited, normalized and subjected to auditory masking effects and are eventually combined by computing a perceptually weighted average. This approach evolved to the speech intelligibility index (SII) and was standardized under S3.5-1997 [ANSI 97]. Since AI is mainly meant for simple linear degradations, e.g., additive noise, [Stee 80] proposed the speech transmission index (STI), which is also able to predict the intelligibility of reverberated speech and non-linear distortions. For this objective measure, a noise signal with the long-term average spectrum of speech is amplitude modulated at several modulation frequencies with a cosine function and applied to the communication channel. The eventual outcome of the STI is then based on the effect on the modulation depth within several frequency bands at the output of the communication channel. While the STI is based on changes in the temporal modulation domain, the spectro-temporal modulation index (STMI) proposed by [Elhi 03] takes into account *joint* specto-temporal modulations. They show that STMI is also applicable for joint spectro-temporal distortions like phase jitter distortions and phase shifts next to additive noise and reverb. The majority of recently published models are still based on the fundamentals of AI [e.g. Rheb 05, Kate 05] and STI [for an overview see Gold 04].

In contrast to speech intelligibility, for speech-quality prediction a wide variety of objective measures are available (see e.g., [Loiz 07a, Dell 93b] for an overview). [Quac 88] evaluated a large amount of objective speech-quality measures for a wide range of degradations and proposed various new objective quality measures. Typically, these quality measures are defined for short time frames (≈25 ms), e.g., based on linear prediction coefficients and/or loudness differences in some time-frequency (TF) representation. More recently, [Beer 02] developed the advanced objective speech quality measure PESQ,

which can be considered as state of the art in the field of speech quality prediction. Several studies are available where PESQ is adjusted in order to assess the intelligibility in stead of speech quality of several signal degradations such like beamforming [Beer 04], low-bitrate vocoders [Beer 05] and speech-enhancement systems [Kita 07, Yama 06]. Recent findings also show that other objective speech-quality measures may be used for speech-intelligibility prediction [Liu 08, Taal 09a, Ma 09].

Although there appears to be a relation between speech quality and speech intelligibility [Prem 95], it is not that obvious that speech-quality measures can be used for speech-intelligibility assessment. For example, [Liu 08] indicated that for SNRs below -10 dB speech may still be partly intelligible, while a lower bound for speech quality (a MOS equal to 1 indicating bad quality) is already reached. Correlation between quality and intelligibility may therefore not be present in these regions. Furthermore, there are still many types of signal degradations for which the relation between quality and intelligibility is not well understood, and perhaps not even present. For example, the quality of noisy speech may be improved by applying a single-channel noise-reduction algorithm [Hu 07b], while the intelligibility is typically not improved or sometimes even decreased [Hu 07a]. Moreover, many objective intelligibility measures still predict incorrectly a significant intelligibility improvement after noise reduction [e.g., Ludv 93, Dubb 08, Gold 04, Taal 10c]. Only recently, new promising intelligibility measures for single-channel noise reduction have been proposed by [Ma 09], which are of great interest for the analysis of existing algorithms. However, for the development of near-future noise-reduction algorithms which aim for intelligibility improvements, these measures should be reliable for a wide variety of TF-varying gain functions applied to noisy speech and not only the ones used in conventional systems. New algorithms may involve different strategies for which it is unknown if the measures from [Ma 09] are reliable.

In this work an evaluation is presented of objective measures for the intelligibility prediction of noisy speech processed with a technique called ideal time frequency segregation (ITFS) [Brun 06]. ITFS is an approach from the field of computational auditory scene analysis (CASA), simulating the remarkable properties of the auditory system to segregate a target speaker from a noisy environment. This technique is particularly of interest, since it delivers a wide variety of applied TF-weightings which can have a much stronger effect on speech intelligibility compared to single-channel noise reduction. An important reason for this difference is that ITFS assumes knowledge of the clean speech signal. Although it can therefore not be used as a practical noise-reduction algorithm (i.e., the clean speech is unknown in practice), large intelligibility improvements can be achieved with ITFS [Kjem 09]. Moreover, the evaluation presented in this work also contains ITFS-settings which decrease the speech intelligibility of noisy speech to a larger extent than conventional noise reduction systems. The variety of signals resulting from ITFS is also demonstrated by the fact that ITFS can be applied to essentially noise-only signals, which gives

fully intelligible speech [Kjem 09] somewhat similar to multichannel vocoded speech [Shan 95]. Objective measures which can correctly predict all these different aspects of ITFS are therefore expected to be robust for a wide variety of applied TF-weightings to noisy speech. Such measures may provide hints on how, and how not to process noisy speech in future algorithms which aim for intelligibility improvements. In addition, intelligibility prediction of the vocoded speech signals in ITFS is of interest in the field of cochlear implants. Namely, presenting vocoded speech to normal-hearing listeners has been a valuable method of simulating listening tests for cochlear implant users [Loiz 98]. Hence, such reliable measures could be used, for example, in the development process of new speech-coding strategies for cochlear implants.

In total 17 objective measures are evaluated for the intelligibility prediction of ITFS-processed noisy speech. This study comprises three state-of-the-art measures for single-channel noise reduced speech as proposed by [Ma 09], the Dau auditory model (DAU) [Chri 10] and the normalized subband envelope correlation (NSEC) [Bold 09] which both show high correlation with ITFS-processed speech, the advanced speech-quality measure (PESQ), and several conventional frame-based speech-quality measures, e.g., segmental SNR. We address some differences between quality and intelligibility prediction for ITFS-processed speech and propose a general technique which improves the performance of the frame-based quality measures when used for intelligibility assessment. From the evaluation several new promising measures for intelligibility prediction of ITFS-processed speech are revealed. To demonstrate the robustness of these measures and the generality of ITFS-processed speech, we show that they also show good prediction results for a listening test where several single-channel noise reduction algorithms are evaluated.

4.2 Intelligibility Data

The intelligibility data is obtained from a study by [Kjem 09], where speech is degraded with various noise types at various SNRs followed by ITFS-processing as explained in [Brun 06]. ITFS is similar to conventional noise reduction in the sense that a TF-varying gain function is applied to noisy speech. However, instead of a continuous gain function, a *binary* TF-weighting is applied to the noisy speech called the ideal binary mask (IBM) [Wang 05]. Since details of ITFS systems differ, e.g., in thresholds used to determine the binary TF-weighting, TF-decompositions, gain values used etc., we describe the specific system [Kjem 09] used to generate the speech data underlying our study.

4.2.1 Signal Processing

The IBM has a value equal to one, when the instantaneous SNR within a certain TF unit exceeds a user-defined local criterion (LC) and is zero otherwise. A mathematical description for the IBM is given as follows,

Table 4.1: The different SNRs in dB used for each noise type (taken from [Kjem 09]).

	SSN	bottles	cafeteria	car
20% SRT	-9.8	-18.4	-13.8	-23.0
50% SRT	-7.3	-12.2	-8.8	-20.3

$$IBM\left(t,f\right)=\begin{cases}1 & \text{if } T\left(t,f\right)-M\left(t,f\right)>LC\\ 0 & \text{otherwise}\end{cases}, \tag{4.1}$$

where $T\left(t,f\right)$ and $M\left(t,f\right)$ denote the signal power in dBs, at time t and frequency f, for the target (clean speech) and the masker (noise only), respectively. The TF decomposition is performed at a sample rate of 20 kHz, by means of a gammatone filterbank [e.g., Patt 92] consisting of 64, 2048 tap FIR filters followed by a time segmentation of 20 ms windowed frames with an overlap of 10 ms. The gammatone filters are linearly spaced on an ERB scale between 55 and 7500 Hz The value of each TF unit is then defined as the signal energy within such a time segment. Next, the IBM is calculated, upsampled to the original sample rate, and multiplied with the noisy signal in each band. Finally, the signal is reconstructed by applying the time-reversed gammatone filters and adding the auditory bands.

4.2.2 Test Material

The test signals are taken from the Dantale II corpus [Wage 03], which consists of five-word sentences all spoken by the same Danish female speaker. The sentences are of the grammatical form name-verb-numeral-adjective-noun (e.g. Ingrid owns six old jackets), where each word in the sentence is picked randomly from a list of 10 possible words. Before ITFS-processing, the speech signals are mixed with four noise types: speech shaped noise (SSN), cafeteria noise, noise from a bottling factory hall and car interior noise and mixed at three different SNRs, including the 20% and 50% speech reception threshold (SRT) and an SNR of -60 dB (The x% SRT is the SNR at which the average listener achieves x% intelligibility). The SNR of -60 dB is included for the generation of the vocoded speech signals. [Kjem 09] performed a different listening test to determine the SRTs by finding the psychometric function for each noise type with the adaptive procedure described by [Wage 03], where the noisy signal energy was normalized before playback. The SRTs were then found by sampling the psychometric function where the results are shown in Table 4.1.

Eight different values for LC are chosen, including an unprocessed condition where only the noisy speech is presented, i.e. $LC=-\infty$. LC is chosen such that the percentage of ones in the IBM varies from approximately 1.5% to 80%. In addition, an alternative way of calculating the IBM is included, which is only

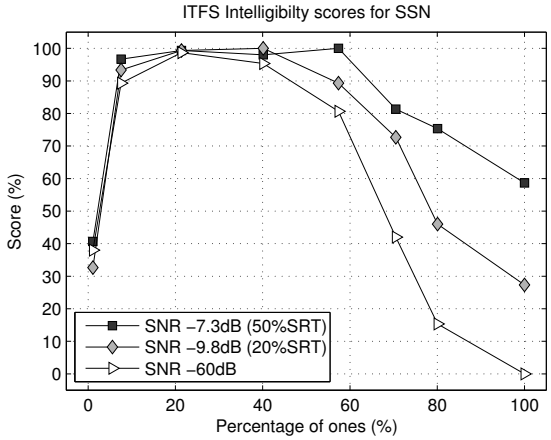

Figure 4.1: Intelligibility of ITFS-processed speech, degraded with speech shaped noise [replotted from Kjem 09]. The percentage of correct words is plotted as a function of the mask density, i.e., the total percentage of ones in the IBM. The mask density of 100% refers to a binary mask with only ones and equals the noisy unprocessed speech.

based on the clean speech. This so called target binary mask (TBM) is obtained by comparing the clean speech power with the power of a signal with the long-term spectrum of the clean speech, within a TF unit. Therefore, the noise itself is not needed in order to determine the binary mask. Note, that the TBM equals the IBM for the case that SSN is used, therefore the TBM is not included for the SSN case. In total, this results in (4*IBM + 3*TBM)*(3*SNR)*(8*LC)=168 conditions to be tested in the listening experiment.

4.2.3 Listening Experiment

For the listening experiment, 15 normal-hearing native Danish speaking subjects participated. The correctly recognized words were recorded by an operator without providing any form of feedback. The average score for all users in each condition was consequently obtained by the average percentage of correct words.

As an example, the results for all SSN conditions are plotted in Figure 4.1. Here, the percentage of correct words is plotted as a function of the mask density, i.e., the total percentage of ones in the IBM excluding noise-only regions [see Kjem 09, for how the noise-only regions are defined]. Note, that the rightmost point refers to a binary mask with only ones, i.e. $LC = -\infty$, which equals the condition where the noisy speech is unprocessed. It can be clearly observed that the speech can be made fully intelligible when the mask density is $\approx 20\%$, independent of the SNR. This is even valid for the -60 dB case, which will be a challenging condition for the objective measures, since all temporal fine

Table 4.2: The evaluated objective measures with their corresponding abbreviations and full names.

Objective measure name	Abbr.
Dau auditory model [Chri 10]	DAU
Normalized subband envelope correlation [Bold 09]	NSEC
Coherence SII [Kate 05]	CSII
Normalized covariance based STI [Gold 04]	CSTI
Perceptual evaluation of speech quality [Beer 02]	PESQ
Log likelihood ratio [Gray 76]	LLR
Itakura saito distance [Itak 70]	IS
Cepstral distance [Gray 76]	CEP
Segmental SNR [Dell 93b]	SSNR
Magnitude spectral distance	MSD
Log spectral distance	LSD
Frequency weighted SSNR [Trib 78]	FWS1
Normalized frequency weighted SSNR [Hu 08a]	FWS2
Weighted spectral slope metric [Klat 82]	WSS
Van de Par auditory model [Par 05]	PAR
Magnitude spectral correlation coefficient	MCC
Log spectral correlation coefficient	LCC

structure is lost. When the mask density is lowered the intelligibility actually decreases, which can even drop below the intelligibility of the unprocessed noisy speech. This is the case for the 50% SRT signals.

4.3 Objective Measures

An overview of the objective measures with their corresponding abbreviations and references can be found in Table 4.2. DAU, NSEC, CSII and CSTI are intelligibility measures, PESQ an advanced quality measure and the measures LLR, IS, CEP, SSNR, MSD, LSD, FWS1, FWS2, WSS and PAR are speech-quality measures based on short-time ($\approx$ 20-40 ms)frames. MCC and LCC are newly proposed measures based on spectral correlation in short-time frames.

4.3.1 Preliminaries

For each of the objective measures evaluated in this study, a general descriptive notation was adopted. The outcome of an objective measure is denoted by $d(x, y)$, where x is the clean speech and y the processed speech. Let m, k and n denote the time-frame, frequency-bin and time-sample index, respectively. The n^{th} sample of the m^{th} Hann-windowed frame of x is then denoted by $x_m(n)$ and its k^{th} DFT bin by $X_m(k)$. Similarly, $y_m(n)$ and $Y_m(k)$ represent the time

frame and the DFT bin of the processed speech, respectively. Furthermore, let M, N and K denote the total number of frames, the frame length and the total number of DFT bins, respectively. For other frequency decompositions (e.g., critical bands), the band index will be denoted by j where J equals the total number bands. For all objective measures, a sample rate of 10 kHz is used with $N = 256$ and $K = 512$, unless noted otherwise.

4.3.2 Intelligibility Measures

Dau auditory model

The advanced auditory model developed by [Dau 96a] (DAU) has been used as an intelligibility predictor by [Chri 10] and shows high correlation with ITFS-processed speech. First, the spectro-temporal internal representations of x and y are determined as described in [Dau 96a], followed by a segmentation in short-time frames within each auditory channel. Subsequently, each frame is compared by means of a correlation coefficient. Let $\Phi_{x,m}(n,j)$ and $\Phi_{y,m}(n,j)$ denote the internal representations of the complete signals x and y, respectively, for the m^{th} frame. The measure is then simply defined as,

$$d_{DAU}(x,y) = \frac{1}{M}\sum_{m}\frac{\sum_{n,j}\left(\Phi_{x,m}(n,j)-\mu_{\Phi_{x,m}}\right)\left(\Phi_{y,m}(n,j)-\mu_{\Phi_{y,m}}\right)}{\sqrt{\sum_{n,j}\left(\Phi_{x,m}(n,j)-\mu_{\Phi_{x,m}}\right)^2\sum_{n,j}\left(\Phi_{y,m}(n,j)-\mu_{\Phi_{y,m}}\right)^2}}, \tag{4.2}$$

where $\mu_{\Phi_{x,m}}$ and $\mu_{\Phi_{y,m}}$ denote the average value of $\Phi_{x,m}$ and $\Phi_{y,m}$, respectively.

Coherence speech-intelligibility index

The coherence speech-intelligibility index (CSII) [Kate 05] is based on the magnitude squared coherence function which is defined as the magnitude squared of the normalized cross-spectral density between x and y, that is,

$$|\gamma(k)|^2 = \frac{|E[X(k)Y^*(k)]|^2}{E\left[|X(k)|^2\right]E\left[|Y(k)|^2\right]}, \tag{4.3}$$

where the asterisk denotes complex conjugation and $E[\cdot]$ denotes the expectation operator. [Kate 05] use a periodogram-based estimator for the spectral densities in Eq. (4.3) (e.g., $\frac{1}{M}\sum_{m} X_m(k)Y_m^*(k)$ estimates the cross-spectral density between $X(k)$ and $Y(k)$). Eq. (4.3) can be used to express the SNR within an auditory filter as follows [Kate 05],

$$SNR(j) = \frac{\sum_{k} W_j(k)|\gamma(k)|^2 E\left[|Y(k)|^2\right]}{\sum_{k} W_j(k)\left(1-|\gamma(k)|^2\right)E\left[|Y(k)|^2\right]}, \tag{4.4}$$

where W_j denotes the frequency weighting of an auditory band by means of a ro-ex filter [Kate 05]. The eventual CSII is then calculated by using the traditional SII [ANSI 97] with the SNR replaced by Eq. (4.4). We use the implementation as proposed by [Ma 09], which shows high correlation with the intelligibility of single-channel noise-reduced speech (referred to as CSII_{mid}, W_4, $p = 1$ by [Ma 09]).

Normalized Covariance Based Speech Transmission Index

The normalized covariance based speech transmission index (CSTI) [Koch 92, Gold 04] shows good results for several types of nonlinear signal degradations, e.g., clipping and spectral subtraction. Let Ψ_x and Ψ_y denote the magnitude envelopes, within an octave band, of the clean and processed speech, respectively. The CSTI is then defined as the correlation coefficient between the band magnitude envelopes within an octave band of the processed and clean speech, that is,

$$r_j = \frac{\sum\limits_m \left(\Psi_x\left(m,j\right) - \mu_{\Psi_x}\right)\left(\Psi_y\left(m,j\right) - \mu_{\Psi_y}\right)}{\sqrt{\sum\limits_m \left(\Psi_x\left(m,j\right) - \mu_{\Psi_x}\right)^2 \sum\limits_m \left(\Psi_y\left(m,j\right) - \mu_{\Psi_y}\right)^2}}, \tag{4.5}$$

This correlation coefficient is then translated to an apparent SNR [Gold 04],

$$aSNR\left(j\right) = \frac{r_j^2}{1 - r_j^2}, \tag{4.6}$$

which is then clipped between -15 and +15 dB and normalized between 0 and 1. Let $\overline{aSNR}\left(j\right)$ denote the clipped and normalized apparent SNR, the overall CSTI is then obtained by a weighted average,

$$d_{csti}\left(x,y\right) = \sum_j \overline{aSNR}\left(j\right) w\left(j\right), \tag{4.7}$$

where we use w as proposed by [Ma 09] to improve its performance with respect to single-channel noise reduced speech (referred to as NCM, $W_i^{(1)}$, $p = 1.5$ by [Ma 09]).

Normalized Subband Envelope Correlation

Similarly as DAU, the normalized subband envelope correlation (NSEC) also shows good correlation with ITFS-processed speech [Bold 09]. First, a 16 channel gammatone filterbank (80 Hz to 8000 Hz, equally spaced on the ERB scale) is applied on the clean and processed speech, after which the normalized, compressed and highpass filtered intensity envelopes $\Lambda\left(m,j\right)$ are extracted. The eventual distance between the clean and processed speech is then defined by the normalized correlation over all time and frequency points, that is,

$$d_{nsec}(x,y) = \frac{\sum_{m,j} \Lambda_x(m,j)\,\Lambda_y(m,j)}{\sqrt{\sum_{m,j} \left(\Lambda_x(m,j)\right)^2 \sum_{m,j} \left(\Lambda_y(m,j)\right)^2}}, \tag{4.8}$$

where Λ_x and Λ_y represent intensity envelopes of the clean and processed speech, respectively.

4.3.3 Speech Quality Measures

PESQ

Perceptual evaluation of speech quality (PESQ) [Beer 02] can be considered as a state of the art speech-quality predictor. Because PESQ is rather complex, we will only briefly describe its main aspects. First, the clean and processed speech are time aligned in order to compensate for any delay differences, after which both signals are processed by a psycho-acoustical model to obtain their internal representations. After global and local normalization these representations are compared resulting in so-called time-frequency dependent disturbance densities. By combining these values a PESQ-score is obtained. In this research, the wide band implementation of PESQ from [Loiz 07a] is used.

Frame-Based Measures

The measures explained in this section are only defined for short-time frames, i.e., $d(x_m, y_m)$. For notational convenience, the frame index m is omitted for these measures and the notation $\hat{d}(x,y)$ is used instead of $d(x,y)$. To obtain for each objective measure one total distance measure, the individual frame distances should be combined somehow. This is done by means of a simple average. However, to eliminate the influence of any outliers, first all individual frame distances are sorted, where the average is only taken over the 5%-95% quantile range [Hans 98a]. This gives,

$$d(x,y) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \hat{d}(x_m, y_m), \tag{4.9}$$

where $\mathcal{M}$ denotes the set of frames in the 5%-95% quantile range and $|\mathcal{M}|$ its cardinality.

Several basic and well-known speech-quality measures are included like the segmental SNR (SSNR) [e.g., Loiz 07a, Dell 93b], where the SNR is determined within short-time frames and combined. The log-likelihood ratio (LLR) [Gray 76], ceptral distance (CEP) [Gray 76] and the Itakura-Saito distance (IS) [Itak 70] are also common speech-quality measures, which assume that speech is an auto-regressive process for short-time segments which can be modeled with linear prediction methods. In contrast to LLR and CEP, IS is also a function of the LPC gains, which implies that a linear scaling applied on the speech will influence the outcome of the IS, which is not the case for the LLR. CEP

is a function of the cepstral coefficients which can be estimated directly from the LPC coefficients [Quac 88]. For more mathematical details for these three measures see, e.g., [Quac 88, Hans 98a, Loiz 07a].

Critical-Band Based Measures Several measures evaluated in this research use a perceptually motivated frequency analysis by means of a DFT-based critical-band decomposition. This is implemented by applying an ℓ_2-norm on the critical-band filtered DFT spectrum, that is,

$$\Gamma_{x_m}(j) = \sqrt{\sum_{k=0}^{K/2} |H_j(k) X_m(k)|^2}, \tag{4.10}$$

where $\Gamma_{x_m}(j)$ denotes the level within the j^{th} critical band of x_m and H represents an approximation of the magnitude spectrum of a 4^{th} order gammatone filter [e.g., Patt 92] as described in [Par 05]. The signal is decomposed into 32 different filter channels equally spaced on an ERB scale ranging from 150 to 4250 Hz to include, approximately, a relevant frequency range for speech intelligibility [Fren 47].

One of the simplest distance measures applied on critical band spectra is the magnitude spectral distance (MSD), where an ℓ_2-norm is applied on the difference between the clean and processed magnitude spectra, that is,

$$\hat{d}_{MSD}(x,y) = \sqrt{\sum_{j=0}^{J-1} |\Gamma_y(j) - \Gamma_x(j)|^2}. \tag{4.11}$$

The same distance measure is also applied on the log spectra, i.e. $20\log_{10}(\Gamma(j))$ denoted by log spectral distance (LSD), which is more in line with how level differences are perceived by the auditory system.

A logical extension of the SSNR is to determine an SNR within a critical band. This approach is proposed in [Trib 78] and is known as the the frequency weighted SNR (FWS) and is given by,

$$\hat{d}_{FWS}(x,y) = \frac{\sum_{j=0}^{J-1} w(j)\, 10\log_{10}\left(\frac{\Gamma_x(j)^2}{(\Gamma_y(j)-\Gamma_x(j))^2}\right)}{\sum_{j=0}^{J-1} w(j)}, \tag{4.12}$$

where w denotes the AI-index weights [Kryt 62] as proposed by [Quac 88]. An adjusted version is also included as proposed by [Ma 09], which has better performance with single-channel noise reduced speech (referred to as fwSNRseg, $p = 1$ by [Ma 09]). Here, before applying the critical band filters in Eq. (4.10), the DFT spectra of the clean and processed speech frames are first normalized to unit length in the ℓ_1-sense. Furthermore, weighting functions based on the clean speech signal are used. We denote the approach with the AI weights by FWS1 and the latter version with FWS2.

Klatt *et al.* defined a distance measure known as the weighted spectral slope metric (WSS) [Klat 82], which is based on the spectral slopes in each band. First, the slope for each log-spectral critical band is calculated as follows,

$$s(j) = 20\log_{10}\Gamma(j+1) - 20\log_{10}\Gamma(j). \tag{4.13}$$

Then, a weighting function per band is used which is based on the level difference between the current band and the band containing the closest peak, and on the level difference between the current band and the band with the maximum peak in the spectrum, that is,

$$w(j) = \frac{c_g}{(c_g + \Gamma_g - 20\log_{10}\Gamma(j))}\frac{c_l}{(c_l + \Gamma_l(j) - 20\log_{10}\Gamma(j))}, \tag{4.14}$$

where Γ_g denotes the global maximum log-spectral magnitude of all critical bands and Γ_l the local log-spectral magnitude of the peak which is nearest to band j. The values c_g and c_l are constants which were set to 20 and 1, respectively [Klat 82]. The final outcome of the WSS is then defined as,

$$\hat{d}_{WSS}(x,y) = \sum_{j=0}^{J-1} w(j)\left(s_x(j) - s_y(j)\right)^2. \tag{4.15}$$

[Par 05] proposed an auditory model based on spectral integration (PAR) and combines the noise-to-signal ratio within the critical bands to determine the eventual distortion outcome. The measure is defined as,

$$\hat{d}_{PAR}(x,y) = Nc_2\sum_{j=0}^{J-1}\frac{\Gamma_{\varepsilon * h_{om}}(j)^2}{\Gamma_{x * h_{om}}(j)^2 + c_1}, \tag{4.16}$$

where $\varepsilon = y - x$, h_{om} denotes the outer-middle ear filter, and the constants c_1 and c_2 are needed for calibration. Here, the constant c_1 can be adjusted to adapt the model sensitivity and c_2 refers to the standard deviation of internal noise responsible for an absolute hearing threshold in the absence of an input signal (masker). The model is calibrated according to [Par 05].

4.3.4 Proposed Measures MCC and LCC

The correlation coefficient is a widely used outcome measure in the field of objective intelligibility assessment. In fact, all of the intelligibility measures explained in Section 4.3.2 are based on this correlation measure. While CSTI and CSII investigate the temporal correlation within one critical band, DAU and NSEC consider the correlation in the joint spectro-temporal domain. However, no measure based only on spectral correlation has been evaluated. Note that FWS2 is perhaps the closest to such a spectral-correlation based measure and shows indeed modest correlation with speech intelligibility [e.g. Taal 09a, Ma 09]. However, FWS2 only normalizes the speech spectra energy before

evaluation and does not compensate for its mean value, which is the case for the correlation coefficient. Motivated by this, a measure based on the spectral magnitude correlation coefficient (MCC) is included,

$$\hat{d}_{MCC}(x,y) = \frac{\sum_{j=0}^{J-1} \left(\Gamma_x(j) - \mu_{\Gamma_x}\right)\left(\Gamma_y(j) - \mu_{\Gamma_y}\right)}{\sqrt{\sum_{j=0}^{J-1} \left(\Gamma_x(j) - \mu_{\Gamma_x}\right)^2 \sum_{j=0}^{J-1} \left(\Gamma_y(j) - \mu_{\Gamma_y}\right)^2}}, \tag{4.17}$$

where μ_{Γ_x} and μ_{Γ_y} denote the sample mean of the clean and processed critical band values. The same TF-decomposition is used as with the critical-band based measures. Similarly as with LSD the same procedure is also applied on the log critical-band spectra (LCC).

4.4 A Critical-Band Based Normalization Procedure

For all the frame-based measures (SSNR, LLR, IS, CEP, MSD, LSD, FWS1, FWS2, WSS, PAR, MCC, LCC), several issues can arise when using them directly for intelligibility assessment. This is caused on one hand by certain differences between speech quality and speech intelligibility prediction, but also by the nature of some of the objective measures.

The first issue is that some of these measures are sensitive to global level differences between the clean and processed speech. This is undesirable, since the intelligibility will not be affected severely when the playback level is adjusted in a listening experiment. Initial results showed indeed that the performance of several measures (e.g., SSNR, IS, FWS1, LSD, MSD) was completely dominated by these large energy difference for certain ITFS-conditions (e.g., TF-weighted noisy speech at -60 dB SNR), which led to very poor correlation with speech intelligibility. Hence, some kind of general normalization procedure is desired. Note, that the more advanced measures DAU, CSTI, NSEC, CSII an PESQ do not have this problem, since there is already some kind of normalization procedure included.

Secondly, some of these frame-based measures are more sensitive for the frequency regions where the speech energy is dominant. This means that the low-energy high frequencies of speech ($\approx$ 2-3 kHz) contribute less compared to lower, more powerful, frequencies ($\approx$ 500 Hz). Although this could make sense in the field of speech-quality assessment, it turns out this is not appropriate for speech-intelligibility prediction. Several studies have shown that these high frequency components are actually of similar importance for the speech intelligibility [e.g., ANSI 97, Stee 80].

The third issue is the fact that certain high ($>$ 5 kHz) and low frequencies ($<$ 200 Hz) are of less importance to speech intelligibility [Fren 47], while they may

be relevant for speech quality. Some measures are sensitive to these frequency ranges, which may bias the results after signal degradation.

To overcome these problems we use a typical procedure from the field of objective intelligibility assessment. This procedure consists of a normalization of the processed and clean critical-band envelopes by its RMS-value before comparison. This approach is used for most of the STI-based [Gold 04] measures and NSEC [Bold 09]. The normalization procedure is applied by pre-filtering the speech signals before evaluation. In this manner, normalization can be applied to any arbitrary objective measure. Let α_j denote the normalization factor for each critical band, which equals the reciprocal of its RMS value,

$$\alpha_j = \left(\frac{1}{KM} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \left| X_m(k) H_j(k) \right|^2 \right)^{-1/2}, \tag{4.18}$$

where H equals the spectrum of one critical band as in Eq. (4.10). The normalized k^{th} DFT bin of the m^{th} frame, say $X'_m(k)$, is then obtained by an addition of all scaled critical bands,

$$X'_m(k) = \sum_{j=0}^{J-1} \alpha_j X_m(k) H_j(k). \tag{4.19}$$

The time-domain signal can now be reconstructed from the weighted short-time DFT bins by means of a simple overlap-add procedure. The processed speech y is normalized with the same procedure. The RMS within each critical band is now fixed, which makes each measure insensitive for global energy differences. Furthermore, each critical band will have an equal contribution to speech intelligibility. Moreover, the total response of the sum of all critical bands will only take into account the frequency range approximately between 150 and 4500 Hz, which is roughly a relevant range for speech intelligibility.

4.5 Evaluation Procedure

For each ITFS condition, 30 five-word sentences are randomly chosen from the corpus, concatenated and ITFS processed. Before applying the objective measures, the silent regions are removed between the five-word sentences. To compare the results of the objective measures and the intelligibility scores, a mapping is needed in order to account for a nonlinear relation. A widely used mapping is the logistic function,

$$f(d) = \frac{100}{1 + \exp(ad + b)}, \tag{4.20}$$

while for some measures a better fit was observed with the following function [Taal 09a],

$$f(d) = \frac{100}{1 + (ad + b)^c}, \tag{4.21}$$

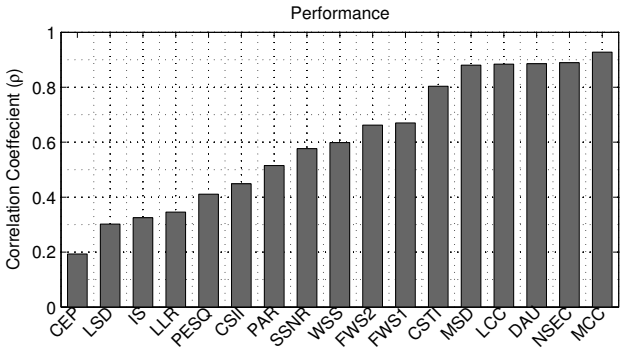

Figure 4.2: Performance with respect to correlation coefficient for all objective measures (higher is better). For all measures except PSQ, CSII, CSTI, DAU, NSEC the speech signals are first subjected to the normalization procedure as explained in Section 4.4.

where a, b and c in Eq. (4.20) and Eq. (4.21) are free parameters, which are fitted to the intelligibility scores with a nonlinear least squares procedure, and d denotes the objective outcome. For each objective measure both mappings are evaluated, where finally the best fit is used. For evaluation we use the correlation coefficient (ρ) and a normalized version of the RMS of the prediction error (σ) (RMSE),

$$\sigma = \frac{1}{100}\sqrt{\frac{1}{S}\sum_{i}\left(s_i - f\left(d_i\right)\right)^2}, \tag{4.22}$$

where s refers to an intelligibility score, S denotes the total number of processing conditions and i runs over all processing conditions. The factor 100 is included to make sure the RMSE is in the same range as the correlation coefficient. The mapping functions may not show a good fit between the intelligibility scores and the objective data for all objective measures. Therefore, the Kendall's Tau [Shes 04] is also included. This outcome measure is independent of the applied (monotonic) mapping and solely tests whether there is a monotonic relation between the intelligibility scores and the objective scores.

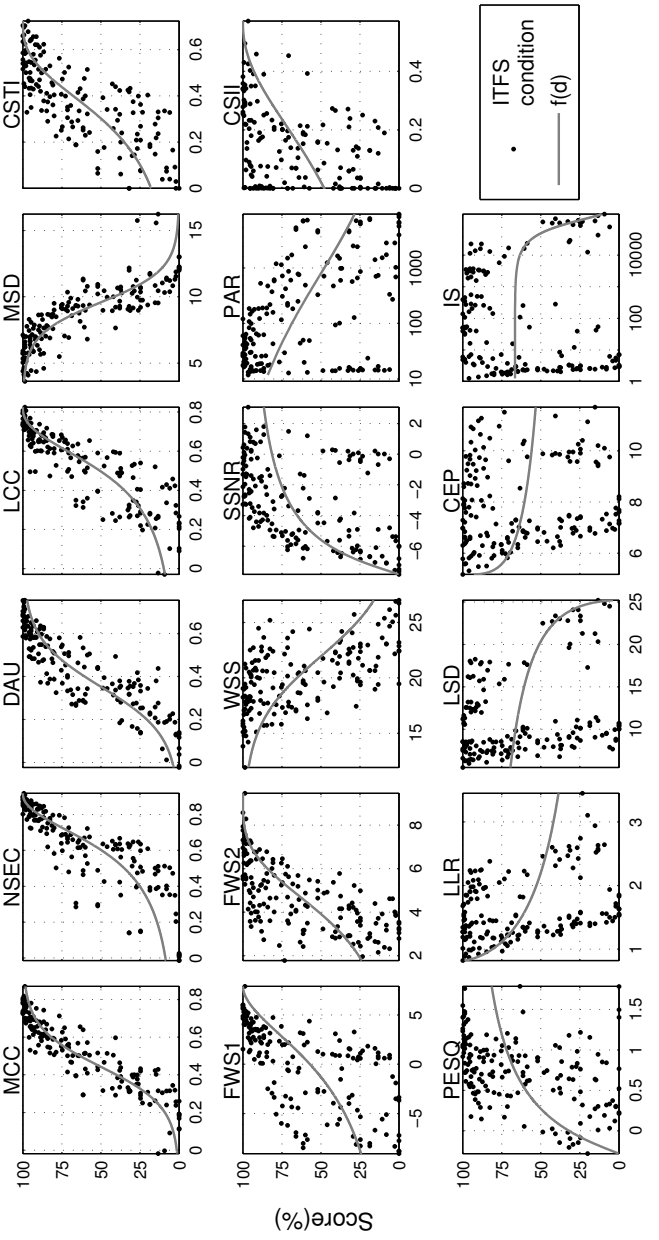

Figure 4.3: Scatter plots for all objective measures together with the fitted mapping function.

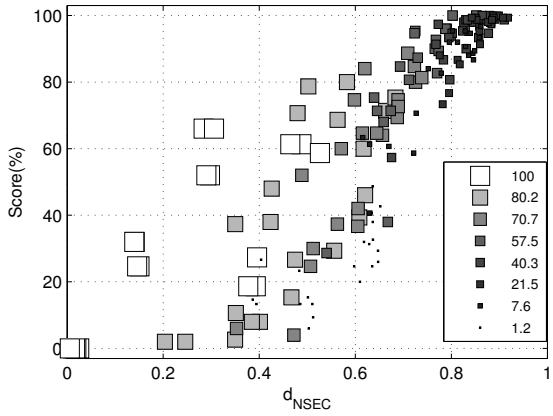

Figure 4.4: Scatter plot for NSEC where the density of the IBM is highlighted by the shading and size of the markers.

4.6 Results and Discussion

For each objective measure, the RMSE, the Kendall's Tau and the correlation coefficient is given in Table 4.3, where, except for DAU, CSII, CSTI, NSEC and PESQ, the signals were first subjected to the proposed critical-band based normalization procedure. To give a clear overview of the differences in performance, the correlation coefficients are ranked in Figure 4.2. Also the scatter plots and the fitted mapping functions are shown in Figure 4.3. We can observe that the proposed measure MCC gave the best results, followed by NSEC, DAU and LCC. The simple MSD correlated better with the intelligibility scores than various other, more advanced objective measures (e.g., CSTI). Remarkably, the more advanced measures CSII and PESQ performed relatively poor.

For the measures CSII, CSTI and FWS2 the new band-importance functions were used as proposed by [Ma 09]. However, we also evaluated the performance with their original implementations (not shown). For CSII and CSTI we did not observe any large changes in performance, while for FWS2 the performance slightly dropped with the version proposed by [Ma 09]. In general, the conclusions made in this work hold for both implementations of each model.

4.6.1 Detailed Evaluation of Intelligibility Measures

Out of the four objective intelligibility measures (DAU, CSII, CSTI, NSEC), the best performance was obtained with DAU and NSEC, which both had similar values for all three outcome measures. In fact, these two measures show the best performance out of all objective measures, except for the proposed measure MCC. CSTI also performed modestly well, while CSII did not perform well.

Table 4.3: RMSE (σ), Kendall's Tau (τ) and correlation coefficient (ρ) for all objective measures.

Name	σ	τ	ρ
PESQ	0.30	0.30	0.41
SSNR	0.27	0.38	0.58
MSD	0.16	0.70	0.88
LSD	0.32	0.19	0.30
FWS1	0.25	0.57	0.67
FWS2	0.24	0.54	0.69
WSS	0.26	0.43	0.60
PAR	0.28	0.34	0.52
MCC	0.12	0.77	0.93
LCC	0.15	0.73	0.88
LLR	0.31	0.24	0.35
IS	0.31	-0.08	0.33
CEP	0.32	0.12	0.19
DAU	0.15	0.73	0.89
CSII	0.29	0.37	0.45
CSTI	0.20	0.63	0.80
NSEC	0.15	0.74	0.89

DAU and NSEC

The good results of DAU and NSEC are in agreement with the results reported in [Bold 09] and [Taal 09a], where it was already observed that both measures appear to be good intelligibility predictors of ITFS-processed speech. Nevertheless, it was observed that both models have a similar weakness and are both more reliable for the ITFS conditions where the intelligibility score is relatively high (90%-100%). To get a better insight in this behavior, an additional scatter plot of NSEC is given in Figure 4.4. Here the IBM density, i.e. the percentage of ones in the binary mask, is denoted by the shading and size of the rectangular markers. A larger and brighter marker indicates a higher density IBM, where the large white squares refer to the mask density of 100%, i.e. the unprocessed noisy speech. The plot clearly illustrates that for these unprocessed conditions, the output of NSEC is much lower compared to the remaining ITFS-processed conditions. This trend is also observed when the density is lowered to 80%, which in general still have a lower objective output. As a consequence, the predicted intelligibility scores for the noisy speech conditions were underestimated. DAU has similar problems, however, from the scatter plot (not shown) it was observed that this problem was only present for the bottles noise.

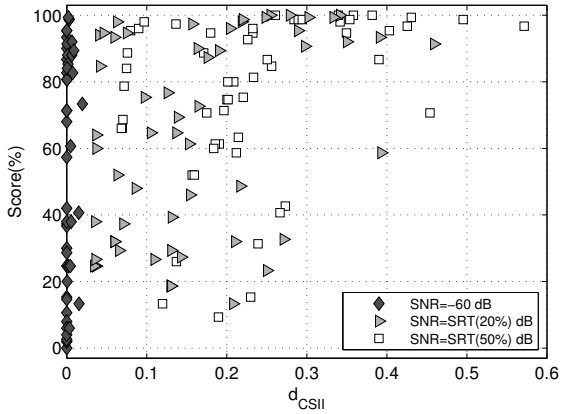

Figure 4.5: Scatter plot of CSII with highlighted SNRs.

CSTI

CSTI yielded a relatively high ranking with respect to all other objective measures. This implies that the promising results of CSTI for clipping and spectral subtraction [Gold 04], are maintained with ITFS-processed speech. Nevertheless, it is clear that the data points are less well fitted by the mapping function than, for example, DAU and NSEC. More specifically, the CSTI turns out to be less reliable for the high intelligible (90-100%) ITFS conditions than DAU and NSEC.

CSII

CSII performed worse than the majority of the evaluated objective measures. Figure 4.5 illustrates that the predicted scores for all -60 dB SNR conditions are underestimated. In fact, most prediction results for these conditions are clipped to 0, i.e., the model predicts the speech to be completely unintelligible. A similar trend occurs for the 20% SRT conditions, which generally show lower objective values than the 50% SRT conditions. This is not in line with the intelligibility scores, where specific settings of *LC* can lead to fully intelligible speech, even at low SNRs.

A possible explanation can be given, by rewriting Eq. (4.3) with an independent phase and magnitude term. Let the polar representations of X and Y with magnitude a and phase θ be denoted by $a_X e^{j\theta_X}$ and $a_Y e^{j\theta_Y}$, respectively. The frequency index k is omitted for notational convenience. This gives,

$$|\gamma|^2 = \frac{\left|E\left[a_X e^{j\theta_X} a_Y e^{-j\theta_Y}\right]\right|^2}{E\left[\left|a_X e^{j\theta_X}\right|^2\right] E\left[\left|a_Y e^{j\theta_Y}\right|^2\right]}, \tag{4.23}$$

A reasonable assumption for speech is that the phase is independently distributed from its magnitude [Erke 07]. Eq. (4.23) can then be rewritten as,

$$|\gamma|^2 = \frac{E\left[a_X a_Y\right]^2}{E\left[a_X^2\right] E\left[a_Y^2\right]} \left| E\left[e^{j(\theta_X - \theta_Y)}\right]\right|^2. \tag{4.24}$$

The right-hand term now indicates the sensitivity for the phase difference, independently of the magnitudes.

For the situation where the clean speech magnitudes are preserved, i.e., $a_X = a_Y$, but a different uniformly distributed phase is used, the right hand term in Eq. (4.24) will be equal to zero. As a consequence, the CSII will report that the clean speech is not intelligible. Since the TF weighting in the ITFS procedure is real valued, the noisy phase will be preserved. Hence, the right term will be very close to zero in Eq. (4.24) for the case that essentially pure noise (-60 dB) is used. This is not in line with the observations described by [Pali 03], where it is reported that, by using a different uniformly distributed phase, the intelligibility is hardly affected.

4.6.2 Detailed Evaluation of Speech Quality Measures

PESQ

The low performance of PESQ was somewhat remarkable. Apparently, its high correlation with speech quality does not guarantee a good correlation with the intelligibility of the ITFS-processed speech signals. This result is different from the observations reported by [Ma 09], where PESQ performed modestly well in terms of predicting intelligibility of single-channel enhanced noisy speech. A possible explanation for this difference is the fact that we used relatively low SNRs, compared to the higher SNRs from [Ma 09], which were set equal to 0 and 5 dB. When lowering the SNR, PESQ will converge to a low value, predicting very poor speech quality; further lowering the SNR will have little effect on speech quality. Nevertheless, in this SNR range a lower bound for speech intelligibility is not necessarily reached yet, as was illustrated in [Liu 08]. This explanation is also motivated by the low PESQ values, which can be observed in its scatter plot in Figure 4.3. Given that PESQ is a reliable predictor of speech quality, it is therefore likely that the intelligibility of ITFS-processed speech does not correlate well with its speech quality.

Frame-Based Measures

Out of all frame-based measures the good performance of MSD was remarkable, since it is probably the simplest measure used in this research. The models FWS1 and FWS2 show modest correlation with intelligibility, which was also reported by [Ma 09]. Poorer results were obtained with WSS and SSNR. The remaining measures in ranking show poor correlation with the intelligibility of ITFS-processed signals.

MSD has approximately the same results as the complex intelligibility models DAU and NSEC. Moreover, MSD shows even better performance than the objective intelligibility measure CSTI. It is hypothesized that the proposed

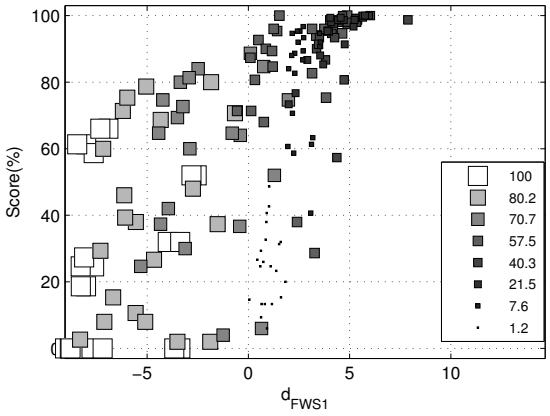

Figure 4.6: Scatter plot for FWS1, where the percentage of ones of the IBM is denoted by the color and size of the markers.

critical-band based normalization plays an important role for these good results (See Section 4.6.3). Rather poor results were obtained with LSD. The main reason for this is that the magnitudes close to zero tend to approach minus infinity due to the log-transform. This situation occurs frequently when the IBM is sparse. This yields a large output value when evaluating the distance between processed and clean speech. The quantile-based procedure which averages all the individual frame distances (See Eq. (4.9)) was not sufficient to take care of these outliers.

Despite their modest correlation, the scatter plots of FWS1 and FWS2 in Figure 4.3 reveal that these measures are mainly reliable for high intelligibility scores. In addition, an oversensitivity is observed for the conditions where an IBM is used with a high percentage of ones as with NSEC. This is clearly illustrated in Figure 4.6, where FWS1 tends to output a lower objective score for most of the noisy unprocessed speech conditions. Figure 4.6 also shows an additional problem, which was present for most of the SNR-based measures. Analyzing the plot reveals that for the lower mask densities (e.g., 1.2% and 7.6%), the output of FWS1 tends to converge to 0 dB. This behavior is even more present in the scatter plot of the SSNR in Figure 4.3, where a cluster of points around 0 dB is observed. Indeed, it is easy to see that Eq. (4.12) is lower bounded by 0 dB for the case where speech information is removed, i.e. $\Gamma_{y(j)} < \Gamma_{x(j)}$. By removing speech information, the speech will eventually become unintelligible. This is not in line with the predictions of the SNR-based measures, which make them less suitable for these types of degradations. Note, that this unwanted behavior is less present with the FWS2. This is due to its normalization procedure, where the DFT spectra of the clean and processed speech frames were first normalized to unit-length in the ℓ_1-sense [Hu 08a]. In principle, the PAR-auditory model can be interpreted as the inverse of the SNR within a critical band. However, the SNRs are not converted to a log scale,

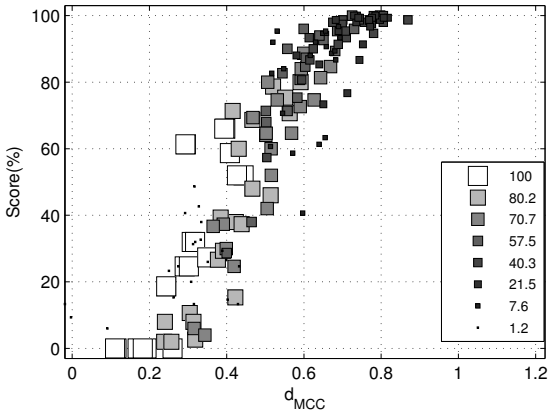

Figure 4.7: Scatter plot for MCC, where the density of the IBM is denoted by the color and size of the markers.

which explains the large range of scores shown in the scatter plot in Figure 4.3 (Notice the log scale on the x-axis). PAR shows similar artifacts as with the SSNR, and FWS measures for the sparse IBM conditions. For PAR, these conditions tend to cluster around $d_{par} \approx 15$.

The last frame-based speech-quality measures according the ranking are LLR, CEP and IS, which all appear to share a similar problem as with the SNR-based models. Where the SNR-based measures converged to a certain value for sparse IBMs, these measures tend to output a large value, when much speech information is removed. Similarly as with LSD, this is caused by the fact that these measures are defined in the log domain.

Additional Proposed Measures MCC and LCC based on Spectral Correlation

From the ranking in Figure 4.2, we see that the relatively simple measure MCC has the best performance out of all objective measures. Despite its simplicity, MCC outperforms both the complex DAU model and NSEC, which makes it a new potential measure for objective intelligibility assessment. As already mentioned, DAU and NSEC are mostly reliable for the ITFS conditions where the intelligibility score is relatively high. As shown in Figure 4.7, this behavior is less present with the MCC, where the mapping shows a better fit with the data over the entire intelligibility range.

Comparable results with DAU and NSEC are obtained with LCC, which is also mainly reliable for the high intelligibility scores. Using the log spectra instead of the magnitude spectra, which is done in the MCC, the correlation with the intelligibility decreases for the evaluated ITFS conditions. Note, that DAU and NSEC also use some kind of compressive nonlinearity. In DAU this is included by means of the adaptation loops, which behave as a log transform for

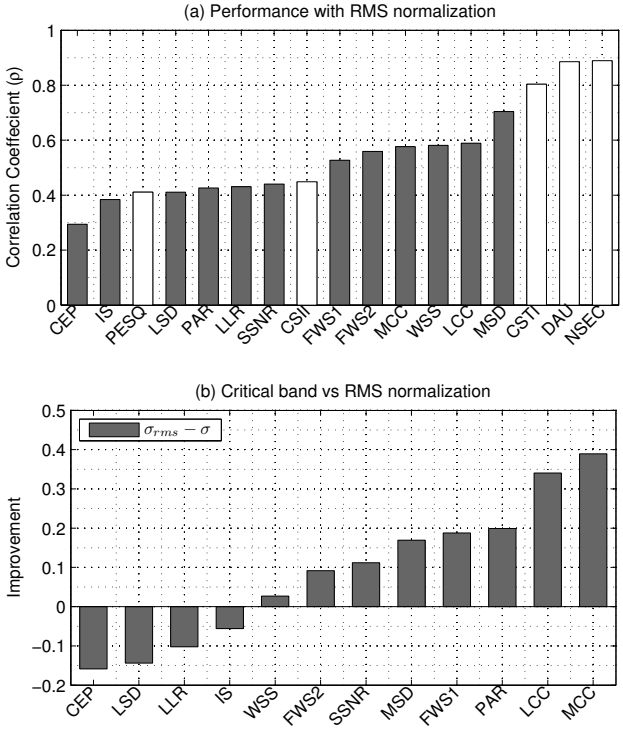

Figure 4.8: (a) Ranking for all frame-based measures when normalization based on RMS is used instead of the critical-band based normalization. The measures not subjected to these normalization procedures are denoted by the white colored bars. (b) The difference in performance between both normalization procedures.

stationary input signals [Dau 96a]. NSEC compresses the band intensity envelopes by raising them to the power 0.15 [Bold 09]. Therefore re-investigating these band-compression stages for intelligibility assessment may be worthwhile.

4.6.3 Influence of Critical-Band Based Normalization Procedure

To determine the influence of the critical-band based normalization procedure, a comparison is made with a normalization procedure based on the RMS, that is $x' = x/RMS(x)$ and $y' = y/RMS(y)$. The RMS-procedure is chosen since it is a straightforward and basic approach often used as an initial stage in more advanced objective measures (e.g., PESQ). Results for the three outcome measures for this experiment can be found in Table 4.4 and in Figure 4.8(a). For comparison reasons, PESQ and the four intelligibility measures are also included, denoted by the white bars (Note that these results are the same as in

Table 4.4: RMSE, Kendall's Tau and the correlation coefficient for all frame-based objective measures when a normalization procedure based on the RMS is applied on the speech signals.

Name	σ_{rms}	τ_{rms}	ρ_{rms}
SSNR	0.30	0.28	0.44
MSD	0.24	0.54	0.70
LSD	0.30	0.24	0.41
FWS1	0.28	0.44	0.53
FWS2	0.26	0.51	0.63
WSS	0.27	0.42	0.58
PAR	0.30	0.25	0.43
MCC	0.27	0.44	0.58
LCC	0.27	0.45	0.59
LLR	0.30	0.32	0.43
IS	0.31	-0.05	0.38
CEP	0.32	0.21	0.29

Figure 4.2, since they were not subjected to the proposed normalization). The difference in performance is shown in Figure 4.8(b), where the measures on the right indicate a stronger improvement due to the proposed critical-band based normalization procedure.

Observing the alternative ranking, none of the outcome measures of the frame-based measures have as good performance as the intelligibility measures CSTI, DAU and NSEC. The only measure which correlates modestly with the intelligibility scores is MSD. Furthermore, as seen in Figure 4.8(b), most of the frame-based measures benefit from the proposed critical-band based normalization procedure, except LSD, CEP, IS and LLR. However, also with the RMS-based normalization procedure these measures turn out to be poor intelligibility predictors.

For the MCC and LCC, a clear problem was observed when the proposed critical band based normalization procedure was not included. This is caused by the already present correlation between the average clean and processed long-term spectra. Car noise, SSN, and cafeteria noise have a strong low-frequency content, similar to clean speech, which yields a positive correlation between their average spectra. However, the bottles noise has a strong high-frequency spectra, which shows a negative correlation with the average clean speech spectrum. This is clearly illustrated in the left plot of Figure 4.9, where the noise type is denoted by the marker type. For the conditions where the speech is degraded with the bottles noise the intelligibility is underestimated, while for the remaining noise types the opposite behavior is observed. This problem is not present in the right plot, where the proposed normalization procedure is applied. After normalization the clean and processed long-term average critical-band spectra will be flat and therefore any global correlation

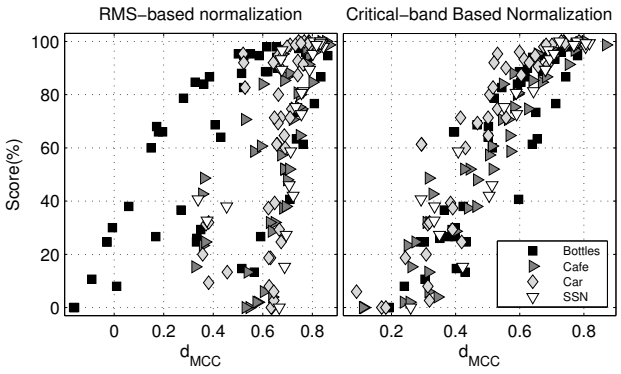

Figure 4.9: Difference in performance between RMS-based normalization (left plot) and critical-band based normalization (right plot) for MCC with respect to noise type.

is removed.

4.7 Generality of Results

From our results, several promising measures are revealed for ITFS-processed speech like MCC, NSEC, DAU and LCC. An interesting conclusion from this evaluation is that the good performing measures are all employing a correlation coefficient in some TF-region. For example, MCC and LCC exploit spectral correlation, while CSTI looks at the correlation between the temporal envelopes within a frequency band. Moreover, DAU and NSEC are based on the correlation in the joint spectro-temporal domain. One important property of the correlation coefficient is its insensitivity to the mean value and the energy of the input signals. This probably also explains the good results obtained with the proposed normalization procedure, which eliminates the effect of the signal energy per critical band.

However, a valid question is if this correlation-based approach will also work with other TF-weighted noisy speech signals other than with ITFS, e.g., single-channel noise reduction. If we compare our findings with the results from the single-channel noise reduction evaluation of [Ma 09], we can conclude that CSTI and FWS2 show reasonable results for both types of processing. As an initial step to indicate the robustness of the promising measures from our study (MCC, NSEC, DAU, LCC and MSD) for TF-weighted noisy speech, an additional listening experiment is conducted where two single-channel noise reduction methods are evaluated. The prediction results from these best measures are compared with the three best performing measures from [Ma 09], that is CSII, CSTI and FWS2, which can be considered state-of-the-art measures for intelligibility prediction for single-channel noise reduced speech. The same evaluation procedure is used as explained in Section 4.5.

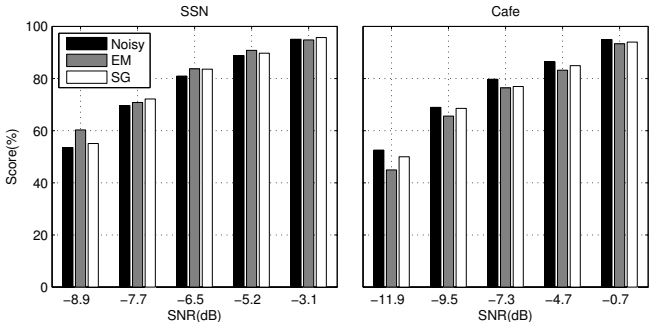

Figure 4.10: Average-user intelligibility scores for unprocessed noisy (UN) speech, and two noise-reduction schemes (EM, SG) for (a) speech shaped noise and (b) cafe noise.

4.7.1 Single-Channel Noise-Reduced Speech

The experiment comprises unprocessed noisy speech and noisy speech processed by two different single-channel noise-reduction algorithms. That is, (1) the standard MMSE-STSA algorithm by [Ephr 84] (EM) which was developed under the assumption that speech and noise DFT coefficients are Gaussian, and (2) an improved version [Erke 07] (SG), which assumes the speech and noise DFT coefficients to be super-Gaussian and Gaussian distributed, respectively. For both algorithms, the *a priori* SNR is estimated with the decision directed approach [Ephr 84] with a smoothing factor of α=0.98. The noise PSD in EM and SG is estimated using Minimum Statistics [Mart 01] and the noise-tracker by [Hend 10], respectively. Maximum attenuation is limited to 10 dB in both algorithms. In SG, the parameters describing the assumed super-Gaussian density of the speech DFT coefficients are γ=1 and ν=0.6 [Erke 07].

The same listening test set-up is used as in Section 4.2. The speech signals are degraded with additive speech-shaped noise (SSN) at a sample rate of 20 kHz. Five different SNRs are considered (-8.9 dB, -7.7 dB, -6.5 dB, -5.2 dB and -3.1 dB), which were chosen such that the psychometric function of clean speech degraded by SSN (based on earlier experiments [Kjem 09]) was sampled approximately between 50% and 100% intelligibility. Fifteen Danish-speaking listeners (normal hearing) were asked to judge the intelligibility of the noisy signals and the two enhanced versions. The three processing conditions (i.e., UN, EM and SG), the two noise types and the 5 SNR values make up 3*2*5=30 conditions. For each of the 30 conditions, each listener is presented with 10 five-word sentences.

The results from the listening experiment are shown in Fig. 4.10. As can be observed, the noise-reduction algorithms have a very small effect on the speech intelligibility compared to the intelligibility of the noisy unprocessed speech. A two-way ANOVA did not showed any significant changes in intelligibility due to each noise-reduction algorithm for each noise type (See p-values in Table 4.5).

Table 4.5: Two-way ANOVA p-values for the hypothesis that there is no effect on intelligibility due to noise reduction for both algorithms (EM, SG) and noise type(SSN, Cafe).

	EM	SG
SSN	0.2470	0.4177
$Cafe$	0.0702	0.4286

Table 4.6: RMSE, Kendall's Tau and the correlation coefficient of the objective measures for intelligibility prediction of the single-channel noise-reduced speech signals.

Name	σ	τ	ρ
MCC	0.06	0.75	0.93
CSII	0.06	0.83	0.92
MSD	0.07	0.74	0.90
LCC	0.07	0.69	0.90
FWS	0.07	0.67	0.89
CSTI	0.07	0.78	0.87
DAU	0.08	0.59	0.84
NSEC	0.10	0.62	0.75

This result is in line with the conclusions from [Hu 08a] where, in general, no noise-reduction scheme could improve the intelligibility of noisy speech.

The prediction results for the objective measures are shown in Table 4.6. From the results we can conclude that the proposed measures MCC, MSD and LCC also have good performance with the single-channel noise reduced signals contained in the listening test next to ITFS-processed speech. In fact, in terms of the correlation coefficient and the RMSE the proposed MCC shows similar performance as the CSII as proposed by [Ma 09], which can be considered as a state-of-the-art intelligibility predictor of noise-reduced speech. Although not as good as the proposed measures from [Ma 09], DAU and NSEC also show moderate correlation with this dataset. Overall it can be stated that the good performing measures for the ITFS-dataset also have good performance with the single-channel noise-reduced set, but not vice-versa.

4.7.2 Other Types of Signal Degradations

We have proposed new objective measures, which show high correlation with the intelligibility of noisy speech signals processed by a TF-varying weighting, like ITFS and single-channel noise reduction. It is not guaranteed that our results are also valid for other degradation types than TF-weighted noisy

speech, e.g., reverberation. For example, [Liu 08] showed that some measures can be very reliable for predicting the effect of speech coders on intelligibility, while the same measures may be unreliable for predicting the intelligibility of noise-reduced speech. This was also demonstrated for the CSTI by [Gold 04], which shows good performance for clipping and spectral subtraction but not for reverberated speech. In future research the promising measures from our research will be evaluated for other types of distortions.

4.8 Conclusions

The focus of this study was the evaluation of various predictive models of intelligibility using ideal-time frequency segregated (ITFS) noisy speech. In total 17 objective measures were evaluated consisting of four advanced objective speech-intelligibility measures (DAU, NSEC, CSII, CSTI), an advanced speech-quality measure (PESQ), and several more conventional frame-based measures (e.g., SSNR). Several of the measures were particularly sensitive to level differences between processed and unprocessed speech. To overcome this problem a general normalization procedure based on equalizing the RMS per critical band was employed. All objective measures were evaluated by means of predicting the intelligibility of 168 different conditions of noisy and ITFS-processed noisy speech signals. From these results the following conclusions can be drawn:

1. Out of all 17 objective measures the highest correlation ($\rho = .93$) with speech intelligibility was obtained with the proposed frame-based measure MCC. This measure was defined as a simple correlation coefficient between the critical-band magnitude spectra of the clean and processed speech.

2. Good results were obtained with DAU and NSEC (both with $\rho = .89$). Nevertheless, these measures turned out to be too sensitive for the noisy unprocessed speech compared to the TF-weighted speech. As a consequence, both measures underestimated the intelligibility for noisy speech compared to TF-weighted noisy speech.

3. LCC and MSD frame-based measures also showed high correlations ($\rho = .88$).

4. The intelligibility measure CSTI gave reasonable results ($\rho = .80$). Therefore, in addition to showing promising results with clipping and spectral subtraction reported by [Gold 04], CSTI is also a reasonable intelligibility predictor for ITFS-processed noisy speech.

5. Poor results were obtained with the CSII, which was not a reliable intelligibility predictor for the ITFS-processed signals used in this research. This was probably due to sensitivity to the DFT phase component.

6. The advanced objective quality-measure PESQ showed a low correlation with speech intelligibility. Since PESQ is a reliable predictor of speech quality, it is therefore likely that the intelligibility of ITFS-processed noisy speech from this study does not correlate with its speech quality.

7. Compared with an RMS-based normalization procedure, the proposed critical-band based normalization improved the correlation with intelligibility for almost all frame-based measures. In particular the measures MCC, LCC and MCD had a large performance improvement due to the proposed critical-band based normalization.

8. The frame-based measures IS, CEP, LSD, LLR, SSNR and PAR showed low correlation ($\rho < .60$) with speech intelligibility. This conclusion holds for both the proposed critical-band based normalization and the RMS-based normalization procedure.

9. The good performing measures in this study (MCC, LCC, DAU, NSEC and FWS2) also showed high correlation with the intelligibility prediction of single-channel noise reduced speech.

Chapter 5

An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech

This chapter is published as "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech", by C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen in the *IEEE Trans. Speech, Audio and Language Processing*, vol. 19, no. 7, pages 2125 - 2136, 2011.

5.1 Introduction

Speech processing systems often introduce degradations and modifications to clean or noisy speech signals, e.g. quantization noise in a speech coder or residual noise and speech distortion in a noise reduction scheme. To determine the perceptual consequences of these artifacts, the algorithm at hand can be evaluated by means of a listening test or an objective machine-driven quality assessment. Although a listening test can lead to a judgment as observed by the intended group of users, such tests are costly and time consuming. Therefore, accurate and reliable objective evaluation methods are of interest since they might replace listening tests, at least in some stages of the algorithm development process. Although it is not straightforward to completely characterize a noisy or processed speech signal, people tend to divide the evaluation into the attributes speech quality, (i.e. pleasantness/naturalness of speech) and speech intelligibility. In this article we focus on speech intelligibility.

One of the first objective intelligibility measures was developed at AT&T Bell Labs around 1920 and eventually published by French and Steinberg [Fren 47]. Kryter [Kryt 62] made the measure better accessible by proposing a calculation scheme, which is currently known as the articulation index (AI). The basic approach of AI is to determine the signal-to-noise ratio (SNR) within several frequency bands; the SNRs are then limited, normalized and subjected to auditory masking effects and are eventually combined by computing a perceptually weighted average. This approach evolved to the speech intelligibility index (SII) and was standardized as S3.5-1997 [ANSI 97]. Since AI is mainly meant for simple linear degradations, e.g., additive noise, Steeneken and Houtgast [Stee 80] proposed the speech transmission index (STI), which is also able to predict the intelligibility of reverberated speech and non-linear distortions. For this objective measure, a noise signal with the long-term average spectrum of speech is amplitude modulated at several modulation frequencies with a cosine function and applied to the communication channel. The eventual outcome of the STI is then based on the effect on the modulation depth within several frequency bands at the output of the communication channel. The majority of recently published models are still based on the fundamentals of AI, e.g., [Rheb 05, Kate 05] and STI (see the work from Goldsworthy and Greenberg [Gold 04] for an overview).

Although the just mentioned objective intelligibility measures are suitable for several types of degradation (e.g., additive noise, reverberation, filtering and clipping), it turns out that they are less appropriate for methods where noisy speech is processed by some type of time-frequency (TF) varying gain function. This includes single-channel noise-reduction algorithms (see the work from Loizou [Loiz 07b] for an overview), but also speech separation techniques like ideal time frequency segregation (ITFS) [Brun 06], where typically a binary TF-weighting is used. For example, STI and various STI-based measures predict an intelligibility improvement when spectral subtraction is applied [Ludv 93, Dubb 08, Gold 04]. This is not in line with the results of listening experiments in literature, where it is reported that single-channel noise-reduction

algorithms generally are not able to improve the intelligibility of noisy speech, e.g., [Hu 07a]. Furthermore, measures like the coherence SII (CSII) [Kate 05] and a normalized covariance-based STI procedure (CSTI) [Gold 04], both show low correlation with the intelligibility of ITFS-processed speech [Taal 09a].

In a recent study, Ma *et al.* [Ma 09] showed that several intelligibility measures could benefit from the use of new (signal-dependent) band-importance functions (BIF). For example, the correlation of CSII and CSTI with the speech intelligibility of single-channel noise-reduced speech increased significantly by the use of these new BIFs [Ma 09]. It is of interest to see if these methods would also work for other types of TF-weighted noisy speech, e.g., ITFS-processed speech. Also two different methods have been proposed lately, which indicate promising results for ITFS-processed speech [Chri 10, Bold 09]. These methods have not been evaluated yet for intelligibility prediction of single-channel noise reduced-speech.

Therefore, a reliable objective intelligibility measure which has high correlation with the speech intelligibility of noisy and various types of TF-weighted noisy speech is of great interest. Such a measure could be used for the analysis of algorithms that process noisy speech. In addition, new algorithms could be developed, which optimize for such an objective measure. To analyze the effect of certain signal degradations on the speech intelligibility in more detail, an objective measure must be of a simple structure. Nevertheless, some measures are based on a large amount of parameters which are extensively trained for a certain dataset. This makes these measures less transparent, and therefore less appropriate for these evaluative purposes.

In this work a simple objective intelligibility measure is proposed which has a strong monotonic relation with the intelligibility scores of various listening tests where noisy speech is processed by some type of TF-weighting[1]. The model has a simple structure in the sense that it is based on only two free parameters. Moreover, it shows better performance than five other reference objective intelligibility measures for these listening tests.

5.1.1 Rationale of Proposed Intelligibility Measure

A general approach in the field of objective intelligibility assessment is to make some type of correlation-based comparison between the spectro-temporal internal representations of the clean and degraded speech signal. For example, CSTI [Gold 04] determines a correlation coefficient between octave-band temporal envelopes and CSII [Kate 05] is based on the coherence function, which is a measure of correlation between complex Fourier-coefficients, over time, as a function of frequency. Another example of a correlation-based measure is the normalized subband envelope correlation (NSEC) proposed by Boldt and Ellis [Bold 09]. In contrast to SNR-based measures (e.g., [ANSI 97, Rheb 05]), the benefit of such a correlation-based approach is the fact that the introduced

[1] An intelligibility model can also predict absolute intelligibility scores (e.g., a percentage of correctly understood words), however, for analysis and/or optimization monotonicity with speech intelligibility is already of great interest.

degradation (i.e., 'the noise') is not needed as a separate signal in isolation from the clean speech. Hence, in addition to speech corrupted by background noise, a correlation-based comparison can also be used for other (nonlinear) types of distortions, e.g., noise-reduced speech, where it is not that straightforward how to separate the clean speech from its introduced distortion.

Several correlation-based measures estimate correlation values for the complete signal of interest at once (e.g., [Bold 09, Gold 04, Kate 05]). Typically, these signals have a length in the order of tens of seconds. A problem which occurs with an analysis length of this order is the fact that a few signal regions with high amplitudes (either from the clean or the degraded speech) may dominate the eventual estimated correlation. There are also measures based on a very short segment size (20-30 ms), e.g., [Chri 10]. However, as a consequence of their poor modulation frequency resolution, certain low temporal modulations are excluded which are important for speech intelligibility. According the results from Drullman *et al.* [Drul 94a] temporal modulations below 2-3 Hz can be removed without affecting intelligibility. Therefore, an analysis window with a length around 333-500 ms would be more appropriate. This is also more in line with the results from van den Brink [Brin 64] which suggest that the temporal integration time of the auditory system has an upper bound of a few hundreds of milliseconds.

Motivated by this we propose a short-time objective intelligibility (STOI) measure, based on a correlation coefficient between the temporal envelopes of the clean and degraded speech, in short-time (384 ms), overlapping segments. Indeed, by experimenting with this segment-length we will show that one actually benefits using segments of this duration.

5.1.2 Further Outline

The remaining part of this article is organized as follows: first more details are given about STOI in Section 5.2. Then, in Section 5.3, three different intelligibility listening experiments are described for different types of processed noisy speech. These results are used to evaluate the intelligibility prediction performance of STOI. Next, more details are given in Section 5.4 about the general evaluation procedure. Finally, the evaluation results are presented together with a discussion in Section 5.5 after which conclusions are drawn.

Figure 5.1: STOI is a function of the clean and degraded speech, which are first decomposed into DFT-based, one-third octave bands. Next, short-time (384 ms) temporal envelope segments of the clean and degraded speech are compared by means of a correlation coefficient. Before comparison, the short-time degraded speech temporal envelopes are first normalized and clipped (see text for more details). These short-time intermediate intelligibility measures ($d_{j,m}$) are then averaged to one scalar value, which is expected to have a monotonic increasing relation with the speech intelligibility.

5.2 STOI

The basic structure of STOI is illustrated in Fig. 5.1. It is a function of the clean and degraded speech, denoted by x and y, respectively. The output of STOI is a scalar value which is expected to have a monotonic relation with the average intelligibility of y (e.g., the percentage of correctly understood words averaged across a group of users). A sample-rate of 10 kHz is used, in order to capture a relevant frequency range for speech intelligibility [Fren 47][2].

First, both signals are TF-decomposed in order to obtain a simplified internal representation resembling the transform properties of the auditory system. This is obtained by segmenting both signals into 50% overlapping, Hann-windowed frames with a length of 256 samples, where each frame is zero-padded up to 512 samples. Before evaluation, silent regions which do not contribute to speech intelligibility are removed. This is done by first finding the frame with maximum energy of the clean speech signal. Both signals are then reconstructed, excluding all the frames where the clean speech energy is lower than 40 dB with respect to this maximum clean speech energy frame. Then, a one-third octave band analysis is performed by grouping DFT-bins. In total 15 one-third octave bands are used, where the lowest center frequency is set equal to 150 Hz and the highest one-third octave band has a center-frequency equal to approximately 4.3 kHz.

Let $\hat{x}(k,m)$ denote the k^{th} DFT-bin of the m^{th} frame of the clean speech. The norm of the j^{th} one-third octave band, referred to as a TF-unit, is then defined as,

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2}, \tag{5.1}$$

where k_1 and k_2 denote the one-third octave band edges, which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly, and is denoted by $Y_j(m)$.

STOI is a function of a TF-dependent intermediate intelligibility measure, which compares the temporal envelopes of the clean and degraded speech in short-time regions by means of a correlation coefficient. The following vector notation is used to denote the short-time temporal envelope of the clean speech,

$$\mathbf{x}_{j,m} = [X_j(m-N+1), X_j(m-N+2), ..., X_j(m)]^T. \tag{5.2}$$

where $N = 30$ which equals an analysis length of 384 ms (see Section 5.5.3 for details on this particular choice). Similarly, $\mathbf{y}_{j,m}$ denotes the short-time temporal envelope of the degraded speech. As illustrated in Fig. 5.1, $\mathbf{y}_{j,m}$ is first normalized and clipped before comparison. The rationale behind the normalization procedure is to compensate for global level differences which should not

[2]Note, that the sample-rate of 10 kHz is not critical. When the window length (in ms) and the frequency-range of the critical bands is preserved the method can be extended to other sample-rates.

have a strong effect on the speech intelligibility (e.g., due to different playback levels of x and y). The clipping procedure makes sure that the sensitivity of the model towards one TF-unit which is severely degraded is upper bounded. As a consequence, further degradation of a speech TF-unit which is already completely degraded (i.e., 'unintelligible') does not lead to a lower intelligibility prediction by the model.

Let $\mathbf{x}(n)$ denote the n^{th} element of $\mathbf{x}$, where $n \in \{1, ..., N\}$ and $||\cdot||$ represent the ℓ_2 norm. The normalized and clipped version of $\mathbf{y}$, say $\bar{\mathbf{y}}$, is then given by,

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{\|\mathbf{x}_{j,m}\|}{\|\mathbf{y}_{j,m}\|}\mathbf{y}_{j,m}(n), \left(1+10^{-\beta/20}\right)\mathbf{x}_{j,m}(n)\right). \tag{5.3}$$

where $\beta = -15$ dB refers to the lower signal-to-distortion (SDR) bound. Indeed, for this case we have that,

$$SDR = 10\log_{10}\left(\frac{\mathbf{x}_{j,m}(n)^2}{\left(\bar{\mathbf{y}}_{j,m}(n) - \mathbf{x}_{j,m}(n)\right)^2}\right) \geq \beta. \tag{5.4}$$

The intermediate intelligibility measure is defined as the sample correlation coefficient between the two vectors,

$$d_{j,m} = \frac{\left(\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}\right)^T \left(\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m}}\right)}{\left\|\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}\right\| \left\|\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m}}\right\|}, \tag{5.5}$$

where $\mu_{(\cdot)}$ refers to the sample average of the corresponding vector. Finally, the average of the intermediate intelligibility measure over all bands and frames is calculated,

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m}, \tag{5.6}$$

where M represents the total number of frames and J the number of one-third octave bands.

5.2.1 Example of Normalization and Clipping Procedure

To illustrate the effect of the normalization and clipping procedure an example is given in Fig. 5.2, where subplot (a) shows a short-time temporal envelope of a clean speech vector together with a noise corrupted version (one frequency band is shown). A corresponding scatter plot is given in Fig. 5.2(b), where $d_{j,m} = 0.81$ denotes the outcome of the intermediate intelligibility measure when clipping would be discarded, i.e., $\bar{\mathbf{y}}$ is replaced with $\mathbf{y}$ in Eq. (5.5) (note, that the applied scaling due to the normalization does not directly affect the correlation coefficient). The normalized and clipped+normalized vectors of the degraded speech are shown in Fig. 5.2(c) together with a scatter-plot in Fig. 5.2(d). From Fig. 5.2(c) it can be observed that the clipping procedure is mainly effective in the noise-only regions (i.e., $n < 11$ and $n > 23$). As a consequence, a higher correlation is obtained ($d_{j,m} = 0.96$) compared to the

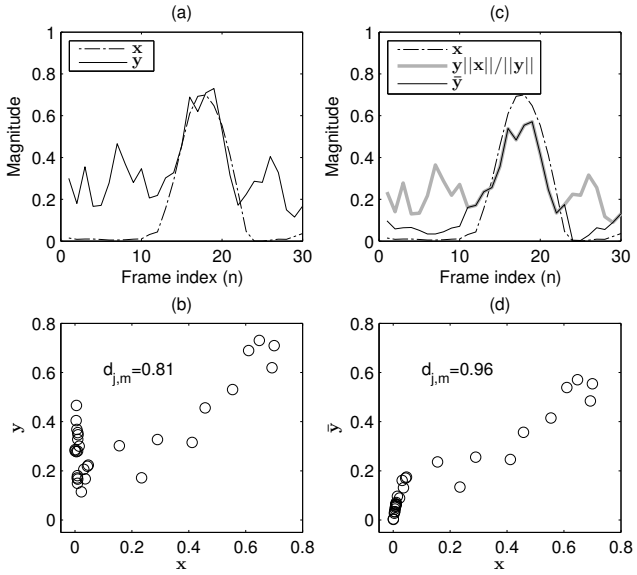

Figure 5.2: Example to illustrate the effect of the normalization and clipping procedure. A clean ($\mathbf{x}$) and noisy ($\mathbf{y}$) speech vector of 30 time-frames (386 ms) is shown in (a) together with a corresponding scatter-plot in (b). Similarly, (c) and (d) show the results for the normalized ($\mathbf{y}||\mathbf{x}||/||\mathbf{y}||$) and clipped+normalized ($\bar{\mathbf{y}}$) degraded vector (See text for more details). Notice that the clipping procedure reduces the effect of the noise in noise-only regions.

situation when clipping would be discarded ($d_{j,m} = 0.81$). This is desired, since it is expected that degrading these regions (where speech is absent within a sentence) will only have a minor impact on speech intelligibility.

5.3 Listening Experiments

In order to evaluate the performance of STOI, its output as described in Eq. (5.6) is compared with the intelligibility scores from three different intelligibility listening experiments. In each of these listening tests noisy speech is processed with different types of TF-weightings. While the first experiment comprises a method where noisy speech signals are ITFS-processed [Kjem 09], the second listening test evaluates the effect on speech intelligibility due to two conventional single-channel noise-reduction schemes. The last experiment evaluates the effect of modifying the applied TF-weighting based on ITFS with artificially introduced errors [Li 08]. Next, more details will be given about these three listening tests.

5.3.1 Ideal Time Frequency Segregation

The intelligibility data from the first experiment is obtained from a listening test conducted by Kjems *et al.* [Kjem 09], where noisy speech signals are ITFS-processed. ITFS is a technique which can improve the intelligibility of noisy speech significantly by applying a binary modulation pattern in a TF-representation [3]. This binary modulation pattern has a value equal to one, when the SNR within a certain TF-component exceeds a user-defined local criterion (LC), and is commonly referred to as the ideal binary mask (IBM). The IBM is given as follows,

$$IBM\left(t,f\right)=\begin{cases}1 & \text{if } T\left(t,f\right)-M\left(t,f\right)>LC\\ 0 & \text{otherwise}\end{cases}, \tag{5.7}$$

where $T\left(t,f\right)$ and $M\left(t,f\right)$ denote the signal power in dBs, at time t and frequency f, for the target (clean speech) and the masker (noise only), respectively. The TF-decomposition is based on a 64-channel gammatone filterbank linearly spaced on an ERB scale between 55 and 7500 Hz. The filterbank is followed by a time segmentation of 20 ms windowed frames with an overlap of 10 ms.

Lowering the LC-parameter in Eq. (5.7) will increase the number of ones in the IBM, where LC= $-\infty$ will result in an IBM with ones only (i.e., the noisy speech is unprocessed). High values for LC will result in sparse IBMs. Kjems *et al.* showed that for certain settings of the LC-parameter as a function of the global SNR, noisy speech can be made fully intelligible. This even holds for the situation that essentially pure noise is modulated with the IBM [Kjem 09]. An alternative IBM is also included which is only based on the clean speech. This so-called target binary mask (TBM) [Kjem 09] is obtained by comparing the clean speech power with the power of a signal with the long-term spectrum of the clean speech, within a TF-component. Therefore, the noise itself is not needed in order to determine the TBM. For more details on the algorithm (e.g., signal reconstruction) the reader is referred to Kjems *et al.* [Kjem 09].

The test signals are taken from the Dantale II corpus [Wage 03], where each excerpt consists of five words, all spoken by the same Danish female speaker. These sentences are degraded by four different types of additive noise: speech shaped noise (SSN), cafeteria noise, noise from a bottling factory hall and car interior noise at three different SNRs: 20% and 50% speech reception threshold (SRT)[4] and an SNR of -60 dB, which represents essentially pure noise. Eight different LC-values are chosen, including an unprocessed condition where only the noisy speech is presented, i.e., LC = $-\infty$ (see the work from Kjems *et al.*[Kjem 09] for more details on the SNR values and LC-parameters).

[3]Note, that here the clean speech is needed separately from the noise source, therefore, large intelligibility improvements are possible. Although this may not seem practical in real-life noisy conditions, this type of processing will deliver a wide variety of processed signals with largely varying intelligibility scores. This is of interest for evaluating STOI.

[4]The x% SRT is the SNR at which the average listener achieves x% intelligibility.

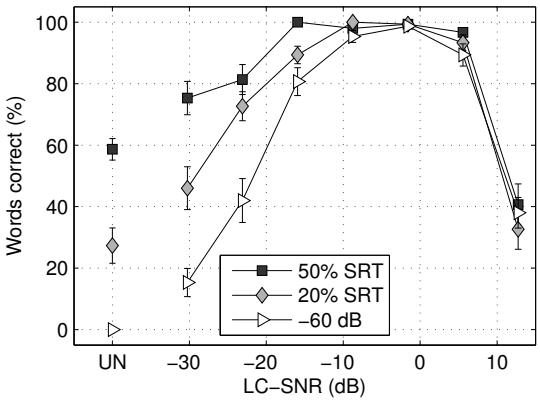

Figure 5.3: Ideal time frequency segregation: average-user intelligibility scores with standard errors of clean speech degraded with speech shaped noise (SSN) at three different SNRs (20%SRT, 50%SRT, -60 dB), followed by ITFS processing (replotted from Kjems et al. [Kjem 09]). The percentage of correct words is plotted as a function of the ITFS algorithm's LC-parameter corrected with the global SNR (See text for more details). The leftmost point refers to an IBM with only ones, i.e., LC = $-\infty$, which equals the noisy unprocessed speech (UN).

For the listening experiment, 15 normal-hearing native Danish speaking subjects participated, where the correctly recognized words are recorded by an operator without providing any form of feedback. Each subject listened to two five-word sentences for each condition. The average score for all users for one condition is then obtained by the average percentage of correct words. In total, this gives (4∗IBM + 3∗TBM)∗(3∗SNR)∗(8∗LC)=168 conditions to be tested in the listening experiment. Only three TBM conditions are included since the TBM equals the IBM for the case that SSN is used, by definition.

As an example, the results for all SSN conditions processed with an IBM are plotted in Fig. 5.3. Here, the percentage of correct words is plotted as a function of the LC-parameter corrected with the global SNR. By subtracting the global SNR from the LC-parameter one can observe from the figure that the noisy speech becomes fully intelligible when the corrected SNR is close to 0 dB. Note, that the leftmost point refers to an IBM with only ones, which equals the condition where the noisy speech is unprocessed (indicated by UN in Fig. 5.3).

5.3.2 Single-Channel Noise Reduction

The second experiment comprises unprocessed noisy speech and noisy speech processed by two different single-channel noise-reduction algorithms. That is, (1) the standard MMSE-STSA algorithm by Ephraim-Malah (EM) [Ephr 84]

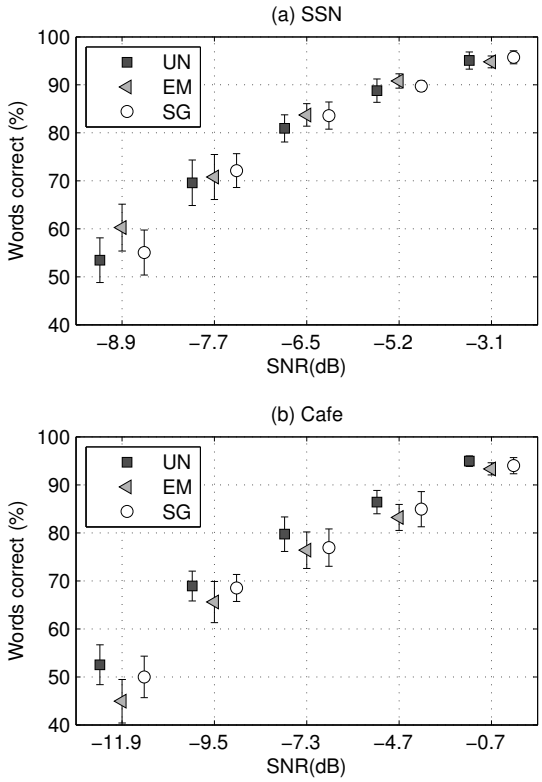

Figure 5.4: Single-channel noise reduction: average-user intelligibility scores with standard errors for unprocessed noisy (UN) speech, and two noise-reduction schemes (EM, SG) for (a) speech shaped noise and (b) cafe noise.

Table 5.1: Two-way ANOVA p-values for the hypothesis that there is no effect on intelligibility due to noise reduction for both algorithms (EM, SG) and noise type(SSN, Cafe).

	EM	*SG*
SSN	0.2470	0.4177
$Cafe$	0.0702	0.4286

which was developed under the assumption that speech and noise DFT coefficients are Gaussian, and (2) an improved version by Erkelens *et al.* (SG) [Erke 07], which assumes the speech and noise DFT coefficients to be super-Gaussian and Gaussian distributed, respectively. For both algorithms, the *a priori* SNR is estimated with the decision directed approach [Ephr 84] with a smoothing factor of α=0.98. The noise PSD in EM and SG is estimated using Minimum Statistics [Mart 01] and the noise-tracker by Hendriks *et al.* [Hend 10], respectively. Maximum attenuation is limited to 10 dB in both algorithms. In SG, the parameters describing the assumed super-Gaussian density of the speech DFT coefficients are γ=1 and ν=0.6 [Erke 07].

As with the previous listening experiment from Section 5.3.1 the speech signals are from the Dantale II corpus [Wage 03], which are degraded by additive speech-shaped noise (SSN) at a sample rate of 20 kHz. Five different SNRs are considered (-8.9 dB, -7.7 dB, -6.5 dB, -5.2 dB and -3.1 dB), which were chosen such that the psychometric function of clean speech degraded by SSN (based on earlier experiments [Kjem 09]) was sampled approximately between 50% and 100% intelligibility.

Fifteen Danish-speaking listeners (normal hearing) were asked to judge the intelligibility of the noisy signals and the two enhanced versions. The three processing conditions (i.e., UN, EM and SG) and 5 SNR values make up 3*5=15 conditions. For each of the 15 conditions, each listener is presented with 10 five-word sentences. The average score for all users and for one condition was consequently obtained by the average percentage of correct words.

The results from the listening experiment are shown in Fig. 4.10. As can be observed, the noise-reduction algorithms have a very small effect on the speech intelligibility compared to the intelligibility of the noisy unprocessed speech. A two-way ANOVA did not showed any significant changes in intelligibility due to each noise-reduction algorithm for each noise type (See p-values in Table 5.1). This result is in line with the conclusions from Hu and Loizou [Hu 08a] where, in general, no noise-reduction scheme could improve the intelligibility of noisy speech.

5.3.3 ITFS with artificially introduced errors

As with Section 5.3.1 the last listening experiment is also based on ITFS. Since the clean speech is needed in Eq. (5.7), the high intelligibility improvements illustrated in Fig. 5.3 are generally not obtained in real-life noisy conditions. In practice one has to estimate the IBM from the noisy speech, which will typically lead to errors [Hu 08b]. In order to find implications for noise reduction, Li and Loizou [Li 08] investigated the effect of artificially introduced errors in the IBM for the case that LC = 0 dB. We regenerated these processed signals as described by Li and Loizou [Li 08], which are used for the evaluation of STOI.

Three types of errors in the IBM were considered by Li and Loizou: (1) A general error, which refers to the procedure where the value of a random selection of TF-units (FFT-based) per time-frame (20 ms) is changed, i.e., a zero in the IBM becomes a one and vice-versa. (2) A type-I error, where a certain percentage of TF-units in the IBM, originally labeled as zero, is changed into a one and (3) a type-II error, where a random selection of TF-units, originally labeled as one, are changed into a zero. For the general errors, five amounts of error in terms of percentage are used (5-40%) and three noise types are considered (SSN, 2-talker babble noise and 20-talker babble noise) all mixed at -5 dB SNR. For the type-I and type-II errors only the 20-talker babble noise is used (also -5 dB SNR) and eight percentages are considered (20-95%). Morover, the unprocessed noisy speech for all three noise types is also included. This gives us a total of 31 conditions: (3 noise types$*$5 error values) + (2 error types)$*$(8 error values) + 3$*$unprocessed.

Seven normal-hearing listeners participated in the listening experiment from Li and Loizou, where all subjects were native American English speakers. The speech material consisted of sentences taken from the IEEE database, see e.g., [Loiz 07b], all produced by the same male speaker, where 20 sentences were used per condition.

The results from Li and Loizou are replotted in Fig. 5.5 [Li 08]. Fig. 5.5(a) illustrates that a general error in the IBM has a similar impact on intelligibility for all noise types. That is, the gain in intelligibility due to the applied IBM drops fast when the percentage of incorrectly TF-units is larger then 10%. In Fig. 5.5(b) it can be clearly observed that a Type-I error has a stronger effect on intelligibility compared to a Type-II error.

5.4 Evaluation procedure

In order to evaluate STOI, 30 sentences are taken from the relevant corpus for each condition of the three listening experiments. That is, the Dantale sentences [Wage 03] for listening experiment 5.3.1 and 4.7.1, and the IEEE sentences (see, e.g., [Loiz 07b]) for listening experiment 5.3.3. These 30 clean and processed sentences are then concatenated and resampled to 10 kHz. We experimented with different values of $N \in$[10, 20, 30, 50, 100, 500] and $\beta \in$[-∞ , -35, -25, -15, -10] only for the intelligibility data originating from the

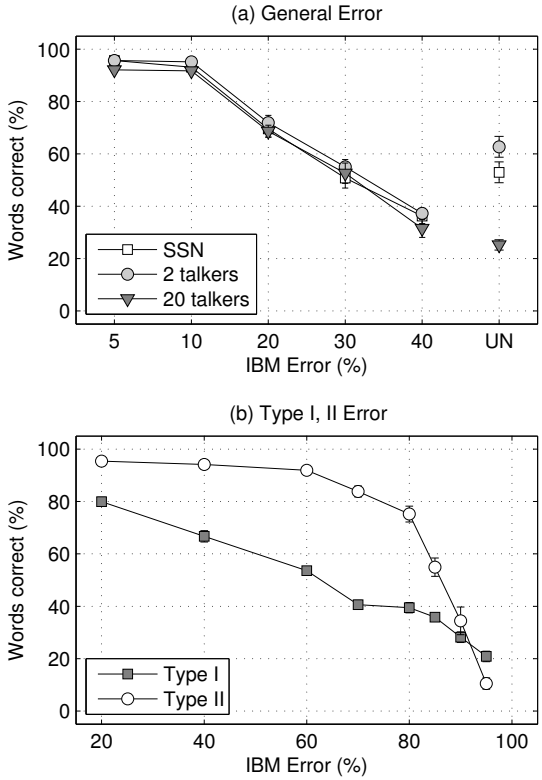

Figure 5.5: ITFS with artificially introduced errors: average-user intelligibility scores with standard errors for artificially introduced errors in the IBM (replotted from the work of Li and Loizou [Li 08]). (a) The effect on speech intelligibility due to a general error for three different noise types at -5 dB SNR. (b) The effect on speech intelligibility due to a type-I or type-II error for 20-speaker babble noise at -5 dB SNR (see text for details). UN indicates the unprocessed noisy speech.

ITFS listening experiment from Section 5.3.1. Note, that $\beta = -\infty$ equals the condition without clipping and therefore without normalization (without clipping the correlation coefficient from Eq. (5.5) is independent of the applied normalization procedure).

Best performance was obtained with $N = 30$ and $\beta =$ -15 dB. These settings were used for evaluating STOI with respect to the remaining two listening tests.

5.4.1 Mapping

We are interested in measuring the monotonic relation between the outcomes of STOI and the actual intelligibility scores. First, a mapping is used in order to account for a nonlinear relation between the STOI outcomes and the intelligibility scores. The main reason for this mapping procedure is to linearize the data such that we can use merits like a linear correlation coefficient. Secondly, with this procedure the STOI scores are mapped to an absolute intelligibility prediction which makes it possible to reveal the distribution of prediction errors amongst all the listening test conditions. For this a logistic function is used,

$$f(d) = \frac{100}{1 + \exp(ad + b)}, \tag{5.8}$$

where a and b are free parameters, which are fitted to the data with a nonlinear least squares procedure. Note, that a logistic function is also monotonic and will therefore not influence the monotonicity between STOI and the intelligibility scores.

While experiments 5.3.1 and 4.7.1 use the Danish Dantale sentences, listening experiment 5.3.3 uses the English IEEE database. In contrast to the IEEE database, the Dantale sentences are taken from a closed set of words. As a consequence, the Dantale sentences are easier to understand for equal adverse listening conditions compared to the IEEE sentences. Objective measures, in general, do not exploit this *a priori* knowledge and will therefore need a different mapping function for each corpus. Motivated by this, the mapping procedure is applied independently for both corpora denoted by $f_{Dantale}$ and f_{IEEE}. Moreover, for the Dantale corpus only the data from the ITFS listening test is used to fit the mapping, which is then reused for the single-channel noise reduction conditions.

5.4.2 Reference Objective Measures

The results of STOI are compared with five other reference objective measures which are all promising candidates for intelligibility prediction of TF-weighted noisy speech. In this section some details will be given for each model.

Dau auditory model

The perceptual model developed by Dau *et al.* [Dau 96a] (DAU) acts as an artificial observer and is originally used for accurately predicting masking thresh-

olds for various masking conditions [Dau 96b]. More recently it is also shown that the model can be used as a good intelligibility predictor for ITFS-processed speech [Taal 09a, Chri 10]. We compare STOI with the intelligibility-model based on the Dau auditory model as proposed by Christiansen *et al.* [Chri 10]. This model is already evaluated with the ITFS intelligibility data from Section 5.3.1 [Chri 10] where good prediction results were obtained. It is of interest to see its performance compared to STOI. First, the spectro-temporal internal representations of x and y are determined as described in [Dau 96a], followed by a segmentation in 20 ms frames within each auditory channel. Subsequently, the internal presentations within each frame of the clean and degraded speech are compared by means of a correlation coefficient jointly over time and frequency. As proposed by Christiansen *et al.* only a subset of frames with high speech energy were considered from which an average correlation coefficient is obtained.

Coherence speech-intelligibility index

The coherence speech-intelligibility index (CSII) [Kate 05] is based on the coherence function which equals the normalized cross-spectral density between the clean and degraded speech. The coherence function is then translated to several frequency-band dependent SDRs, which are combined to one score as in the conventional speech intelligibility index (SII) [ANSI 97]. That is, the SDRs are limited and normalized and are combined by computing a weighted average based on perceptual band-important functions (BIFs). It is shown that CSII can successfully predict the effect on speech intelligibility due to non-linear types of speech distortions like peak-clipping and center-clipping [Kate 05]. Recent results show that by using signal-dependent BIFs instead [Ma 09], the performance of CSII with respect to single-channel noise-reduced speech signals will increase significantly. This CSII variant is also used for the comparison with STOI (referred to as CSII_{mid}, W_4, $p = 1$ by Ma *et al.* [Ma 09]).

Normalized Covariance Based Speech Transmission Index

The normalized covariance based speech transmission index CSTI is based on the correlation coefficient between the band magnitude envelopes within 8 octave bands [Koch 92, Gold 04]. The measure shows good results with respect to various types of signal distortions, e.g., clipping and spectral subtraction [Gold 04]. The correlation coefficients per band are translated to an SNR and combined in a similar way as with the CSII. Also for this measure new BIFs were recently proposed in order to improve its performance with respect to single-channel noise reduced speech [Ma 09]. These BIFs are also used in this article for comparison with STOI (referred to as NCM, $W_i^{(1)}$, $p = 1.5$ by Ma *et al.* [Ma 09]).

Table 5.2: Used values for the free parameters of the nonlinear mappings, for the Dantale and IEEE corpus.

	a	b
$f_{Dantale}(d)$	-14.5435	7.0792
$f_{IEEE}(d)$	-17.4906	9.6921

Frequency-Weighted Segmental SNR

The frequency weighted segmental SNR (FWS) is included for comparison as proposed by Hu and Loizou [Hu 08a]. The measure determines the SNR within several frequency bands in short-time frames (20 ms) which are limited and normalized. Here, the clean and processed speech frames are first normalized to have unit area. This normalization procedure was found to be of critical importance in order to predict the speech quality of enhanced speech [Hu 08a]. In addition, FWS also showed promising results with respect to predicting the speech-intelligibility of single-channel noise-reduced speech [Ma 09]. Again, new BIFs are used as proposed by Ma *et al.* (referred to as fwSNRseg, $p = 1$ by Ma *et al.* [Ma 09]) in order to combine the clipped and normalized SNRs.

Normalized Subband Envelope Correlation

The final intelligibility measure is based on the normalized subband envelope correlation (NSEC) [Bold 09]. This model is already evaluated with the ITFS intelligibility data from Section 5.3.1) [Bold 09] where good prediction results were obtained. Hence, it is of interest to compare its performance with STOI. First, a 64 channel gammatone filterbank is applied on the clean and processed speech, after which the normalized, compressed and highpass filtered intensity envelopes are extracted. The eventual distance between the clean and processed speech is then defined by the normalized correlation over all time and frequency points. Similarly as with DAU, the correlation is determined jointly over time and frequency.

5.5 Results

First the performance of STOI in terms of several correlation measures will be reported for each listening test after which more details are given about how the STOI intelligibility prediction errors are distributed over the various listening tests and processing conditions. Then the effect of the clipping parameter β and the analysis length N is analyzed followed by a comparison with several other intelligibility models.

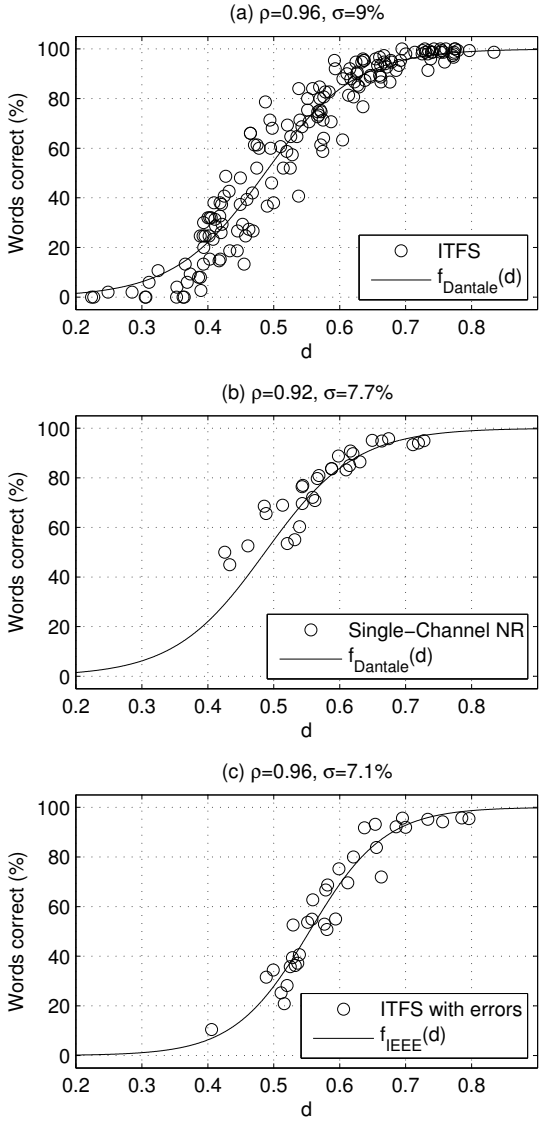

Figure 5.6: Scatter plots between STOI and the speech-intelligibility scores from three different types of TF-weighted noisy speech: (a) ITFS-processed noisy speech (see Section 5.3.1), (b) single-channel noise-reduced speech (see Section 4.7.1) and (c) ITFS-processed noisy speech with artificially introduced errors (see Section 5.3.3). At the top of each plot the correlation coefficient (ρ) and the standard deviation of the prediction error (σ) is denoted.

5.5.1 Correlation between STOI and Intelligibility Scores

The performance of STOI is evaluated by means of the correlation coefficient (ρ) and the standard deviation of the prediction error (σ). A higher ρ denotes better performance while for σ, lower values represent better results. Both merits are applied on the mapped objective scores, i.e., $f(d)$. The scatter-plots for all three listening tests are shown in Fig. 5.6, where their corresponding figures of merit are indicated at the top of each plot. In addition, the applied mapping function $f(d)$is shown. Table 5.2 summarizes the obtained values for the free parameters of the applied mappings.

The plots clearly show good performance by means of a strong monotonic relation between STOI and the speech-intelligibility scores, for all three listening tests. This is reflected in the correlation coefficients which are all above 0.9 and the obtained standard deviations of the predictions errors, which are below 9%. It can be observed from the plots in Fig. 5.6 that the logistic function for the IEEE sentences is shifted more to the right compared to the mapping function for the Dantale sentences: given a STOI score, the actual intelligibility score for the IEEE sentences is slightly lower compared to the Dantale sentences. As hypothesized in Section 5.4.1, this is probably due to the fact that the Dantale sentences are generated from a closed set of words, which makes them more intelligible than IEEE sentences under equal adverse conditions.

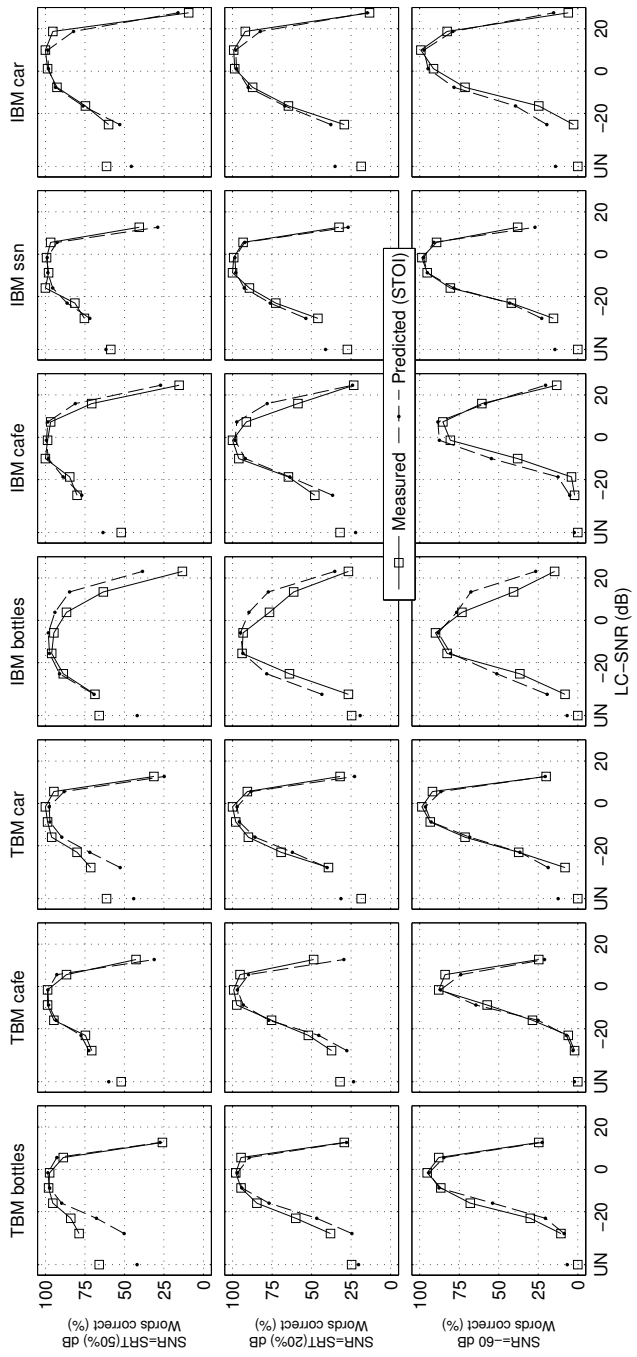

Figure 5.7: Results of STOI predictions for the ITFS-processed speech and the actual measured intelligibility scores from Kjems et al. [Kjem 09] as described in Section 5.3.1. Each row of subplots refers to a certain SNR-value (50% SRT, 20% SRT and -60 dB) where a column is related to a specific noise-type and ITFS-processing setting (TBM, IBM, see Section 5.3.1 for more details).

Figure 5.8: STOI predictions for the single-channel noise-reduced speech conditions from Section 4.7.1. Predicted and measured scores are shown for unprocessed noisy (UN) speech, and two noise-reduction schemes (EM, SG) for speech shaped noise (SSN) and cafe noise.

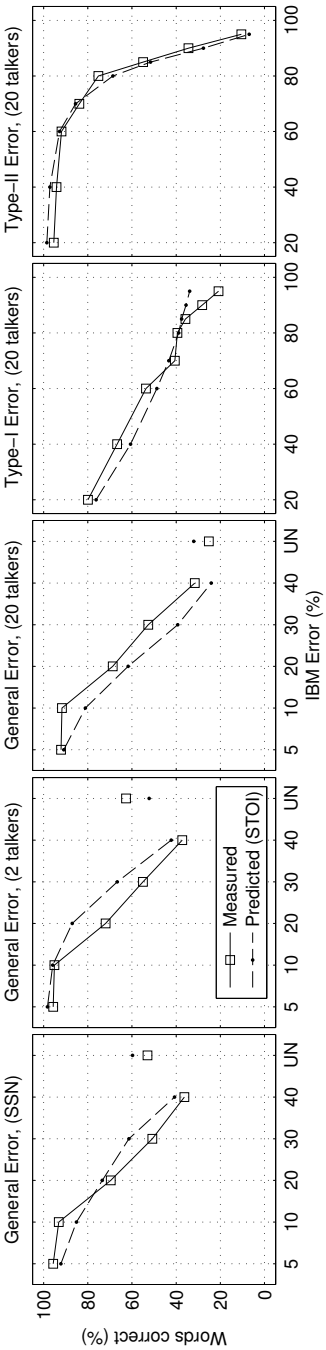

Figure 5.9: Intelligibility scores of ITFS-processed speech with artificially introduced errors (replotted from Li and Loizou [Li 08]) as described in Section 5.3.3 together with the predicted intelligibility scores from STOI.

5.5.2 Analysis of Absolute Intelligibility Predictions

As already mentioned in the introduction, the aim for STOI is to have a monotonic relation with speech intelligibility and not necessarily to predict absolute intelligibility scores. However, by mapping the STOI outcomes using the logistic function $f(d)$, some insight will be gained in the distribution of the prediction errors of STOI. The results are shown in Fig. 5.7, 5.8 and 5.9 for the three listening tests from Sections 5.3.1, 4.7.1 and 5.3.3, respectively.

Fig. 5.7 shows the results for the first listening experiment for all noise-types, SNRs and other specific ITFS-settings (similar to Fig. 5.3). The plots reveal that STOI correctly predicts the effect of the LC-parameter on the speech intelligibility, for almost all cases. This includes the extreme cases where essentially a noise-only signal (-60 dB SNR) is ITFS-processed, resulting in almost 100% intelligible speech for specific LC-values. Note, that for this case all fine-structure of the clean speech is lost and the signals sound rather artificial; a challenging condition. Small problems are observed for both the bottles-noise and the car-noise mixed at 50% SRT for the unprocessed noisy speech (UN) and low SNR-corrected LC values (first, third, fourth and seventh plot of top-row in Fig. 5.7). For these conditions, STOI underestimates the speech intelligibility. An explanation for this could be the fact that these noise types have a significantly different average spectrum compared to the clean speech. Therefore, the errors are distributed in different frequency channels compared to the SSN and cafe-noise conditions. Perhaps these problems can be solved by introducing band-importance functions, see e.g., [Ma 09]. Nevertheless, these problems are rather modest and generally STOI shows good agreement with the data. Note, that STOI was developed with simplicity in mind: the goal was to develop a model with very vew parameters. For this reason we did not include any band-importance functions.

The absolute predicted intelligibility scores for the single-channel noise-reduced speech are shown in Fig. 5.8. From this plot we observe that for low SNRs the intelligibility scores for the SSN-conditions are slightly overestimated. However, these small overestimations are approximately equal for both the unprocessed noisy condition and the noise reduction algorithms EM and SG. By comparing the relative difference in predicted intelligibility scores before and after noise reduction, it can be concluded that STOI correctly predicts no significant effect on the intelligibility. This is in line with the results from the listening test. Similarly, for the cafe-noise no significant change in intelligibility is reported. Note, that several STI-based speech-intelligibility measures report an incorrect intelligibility improvement after noise reduction, e.g., [Ludv 93, Dubb 08, Gold 04].

Fig. 5.9 shows the STOI predictions for the ITFS-processed speech with artificially introduced errors from Section 5.3.3. From the plot we can observe that STOI correctly predicts the effect of the introduced errors in the IBM. Specifically, the Type-I and Type-II error predictions are in strong correspondence with the actual intelligibility scores. For the general error introduced in the IBM, the plots reveal small deviations between the different noise types,

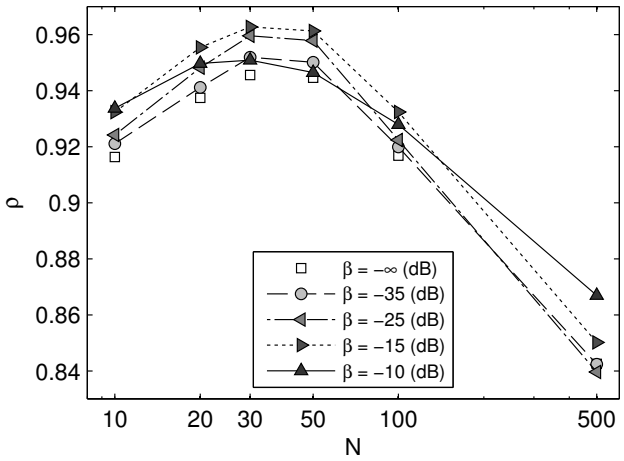

Figure 5.10: Influence of the clipping parameter β and the segment length N for the intelligibility data originating from the ITFS listening experiment from Section 5.3.1.

i.e., the 2-talker noise conditions are slightly overestimated and the 20-talkers noise is slightly underestimated. However, these errors turn out be small.

5.5.3 Effect of parameters N and β

The correlation coefficients obtained for the different values of $N \in [10, 20, 30, 50, 100, 500]$ and $\beta \in [-\infty, -35, -25, -15, -10]$, with respect to the ITFS listening experiment from Section 5.3.1, are shown in Fig. 5.10. From the plot it can be observed that maximum correlation is obtained with $N = 30$ and $\beta = -15$ dB. The same conclusion holds for observing the standard deviations of the prediction errors (not shown). In general, the segment length N has a bigger impact on the results compared to the clipping procedure

The results with respect to N are in line with the rationale behind STOI which was explained in Section 5.1.1. While an estimated correlation coefficient based on very long segments (tens of seconds) may be dominated by outliers, an analysis length which is too short (20-30 ms) may exclude important temporal modulation frequencies. Several listening experiments show that temporal modulations above 2-3 Hz are important for intelligibility [Drul 94a, Arai 99]. For $N = 30$, STOI will be sensitive for temporal modulations of 2.6 Hz and higher which is roughly in accordance with the results of these listening tests. Moreover, the analysis length of $N = 30$ (384 ms) is also more in line with the maximum temporal integration properties of the auditory system, which is in the order of hundreds of milliseconds, e.g., [Brin 64].

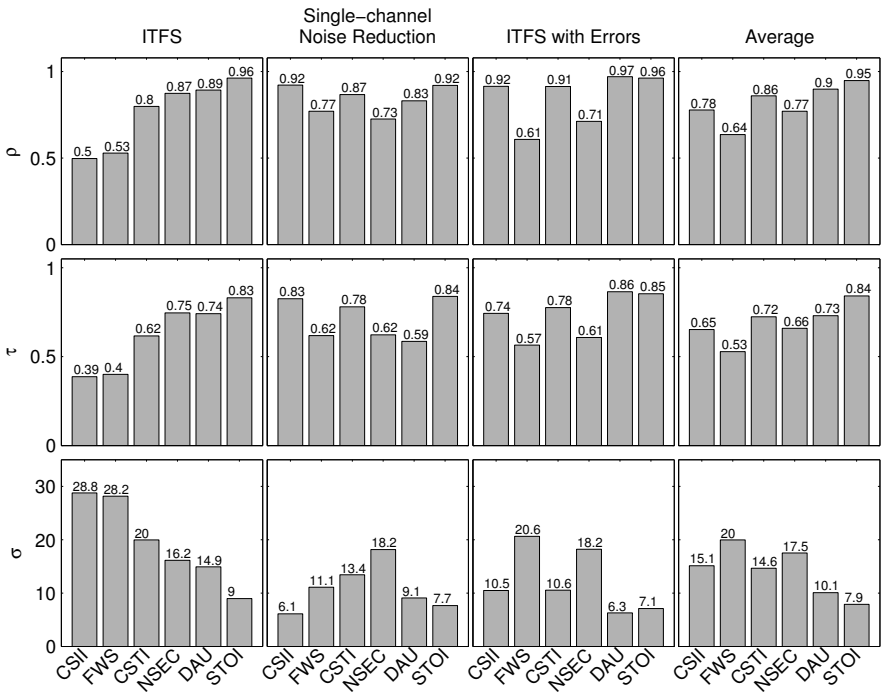

Figure 5.11: Performance of STOI compared with five other reference objective intelligibility models. Each column of all subplots denotes one of the three listening tests as described in Section 5.3, where the last row indicates the average performance measure for all three listening tests. The rows represent the correlation coefficient (ρ), Kendall's Tau (τ) and the standard deviation of the prediction error (σ).

5.5.4 Comparison with Other Intelligibility Models

For the five reference objective measures the same evaluation procedure is used as with STOI (as described in Section 5.4). An additional figure of merit is included: the Kendall's Tau (τ), e.g., [Shes 04], where a higher τ implies better performance. The Kendall's Tau is only based on the ranking and therefore independent of the applied mapping from model output to predicted intelligibility scores, as long as the mapping is monotonic. It is included to make the results more transparent, since the mapping procedure may show a better fit with the data for certain intelligibility models.

The results are shown in Fig. 5.11, where each column of subplots represents one of the three listening tests and each row represents a figure of merit. The average of the three outcomes for each merit is shown in the last column. From these results it can be concluded that STOI has the best average performance for all three listening tests with respect to all figures of merit. Also for the results with respect to each listening test independently, STOI has better performance compared to almost all other measures. Only CSII has similar performance for the 'single-channel noise reduction' listening experiment and DAU shows slightly better results for the 'ITFS with errors' data. Less good results were obtained with FWS which ended up lowest in ranking for the average results for all listening tests. In general, the rankings based on the correlation coefficient are roughly in accordance with the remaining two figures of merit, except for CSTI and NSEC when evaluated for the single-channel noise reduced speech. It turns out that for these two measures, the mapping function $f_{Dantale}$, which was only trained on the ITFS-processed data, did not fit the noise-reduction dataset.

The good results obtained with DAU and NSEC for the first listening experiment (ITFS) are in accordance with the fact that these two measures were also designed and optimized for the ITFS listening experiment by Kjems *et al.* [Kjem 09]. Furthermore, the performance and ranking of CSII, FWS and CSTI for the single-channel noise-reduction intelligibility data is in agreement with the results from Ma *et al.* [Ma 09].

5.6 Discussion

One may argue that STOI has better performance compared to the reference objective measures due to the fact that the parameters β and N have been optimized for. However, instead of extensively tuning these parameters, a limited amount of settings have been tested only with respect to the first listening test. Other settings than $\beta = -15$ dB and $N = 30$ for the last two listening experiments have not been considered. Note, that also NSEC and DAU were designed and optimized for the intelligibility data from Kjems *et al.* [Kjem 09]. Furthermore, the output signals of the single-channel noise-reduction algorithms from the second listening experiment have different types of signal artifacts compared to the ITFS-processed speech. Also the ITFS-processed speech with artificially

introduced errors (the third listening experiment) is based on a different speaker and different noise types. The latter two listening experiments contain a significant amount of 'musical noise' due to their DFT-based approach, in contrast to the gammatone-based approach of the first listening experiment. In summary, STOI has not been optimized for listening tests 2 and 3.

Next to STOI, DAU also showed good performance for all listening tests. However, in contrast to STOI, DAU determines a correlation coefficient in segments of only 20 ms. In line with the results from Section 5.5.3 this could be a reason for their difference in performance with respect to the first two listening tests where STOI shows better performance. However, this is not in agreement with the results for the last listening experiment, where DAU shows slightly better performance than STOI. Maybe this can be explained by the use of the so-called adaptation loops in the DAU-model, which simulate the adaptation properties of the auditory nerve [Dau 96a]. This stage shows a log-compressive behavior for stationary input signals while fast fluctuations are linearly transformed. As a consequence, DAU is more sensitive to transient regions which are of importance for speech intelligibility. This unique property of DAU is not represented in any of the other intelligibility models contained in this research. It would be of interest to investigate the contribution of these adaptation loops with respect to intelligibility prediction (e.g., by excluding them or replacing this stage with a simple log-transform).

Although not as good as STOI and DAU, CSTI also showed good performance with respect to all three listening tests. Note, that without the clipping procedure CSTI and STOI are similar measures in the sense that they are both based on a correlation coefficient per band. However, CSTI determines a correlation coefficient for the complete signal at once instead of the short-time segments used by STOI. In line with the earlier results shown in Fig. 5.10 this difference in analysis window length probably explains their difference in performance. The same holds for NSEC which also considers the correlation for the complete signal at once.

CSII showed good results for the second and third listening experiment, however, poor results were obtained with respect to predicting the intelligibility scores for the ITFS processed speech data. It was observed that CSII predicted incorrectly that all the ITFS-processed noisy speech signals mixed at -60 dB SNR were unintelligible. An explanation for this is the fact that CSII is sensitive for degradations in the temporal fine structure of the clean speech (in contrast to STOI). This is a direct consequence of the coherence function, which takes into account the phase component of the complex DFT coefficients. Note, that for the ITFS-processed noisy signals mixed at -60 dB SNR, the temporal fine structure is completely lost.

FWS is the only measure in this evaluation which is not based on a correlation based comparison between the clean and degraded speech. Instead it uses a conventional SNR per frequency band. This property and the relatively short analysis window of 20 ms probably explains its low ranking compared to all other intelligibility models.

Although STOI is meant for predicting the intelligibility of TF-weighted noisy speech, it would be of interest to investigate its performance with respect to other types of degradations. A recent evaluative study [Schl 10] showed promising results for STOI with respect to envelope thresholding: a nonlinear operation that consists of setting to zero any samples of the original envelope that are below a threshold [Gold 04]. Also CSTI showed good results with respect to envelope thresholding [Gold 04]. As already explained, CSTI and STOI are similar in the sense that they are both based on the correlation coefficient between the temporal envelopes of the clean and degraded speech per frequency band. Goldsworthy and Greenberg concluded that with this correlation-based approach, CSTI was not capable to predict the intelligibility of reverberated speech in quiet and low noise environments [Gold 04]. It could be the case that this conclusion also holds for STOI. However, more research is needed to investigate the effect of the clipping procedure and shorter analysis window length of STOI compared to CSTI. Note, that STOI does work well for additive noise since each of the three different listening tests contain unprocessed noisy speech for different noise types and SNRs.

STOI does not not take into account some type of absolute threshold in quiet. Therefore, its predictions may not be accurate for operations which significantly reduce the level per band and do not have a strong impact on its temporal envelope (e.g., as with lowpass or highpass filtering).

5.7 Conclusions

A short-time objective intelligibility measure (STOI) is presented based on the correlation between temporal envelopes of the clean and degraded speech in short-time (382 ms) segments. This is different from other measures, which typically consider the complete signal at once, or use a very short analysis length (20-30 ms). Experiments with different segment lengths indeed show the benefit of using segment-lengths in the order of hundreds of milliseconds. Further extensive evaluation shows that STOI has high correlation with the speech intelligibility for three different listening tests ($\rho \geq 0.92$ for all listening tests). For each of these three listening tests, noisy speech is processed by some type of TF-varying gain function, including a signal processing technique called 'ideal time frequency segregation' and conventional single-channel noise reduction algorithms. In general, STOI showed better correlation with speech intelligibility compared to five other reference objective intelligibility models. A free Matlab implementation is provided at `http://siplab.tudelft.nl/`.

Chapter 6

Speech Energy Redistribution for Intelligibility Improvement in Noise Based on a Perceptual Distortion Measure

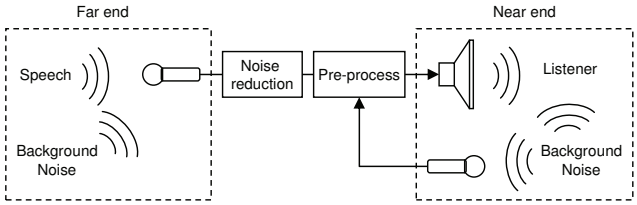

Figure 6.1: The pre-process algorithm improves speech intelligibility for the near-end listener as a function of the near-end noise statistics. It is assumed that a clean speech signal is available with far-end noise successfully removed by noise-reduction.

6.1 Introduction

As illustrated in Fig. 6.1, the speech intelligibility for the near-end listener can be affected by background noise from both sides of the communication channel. That is, the noise can come from both the *far end* and the *near end*. In order to eliminate the negative impact of the far-end noise, one would typically apply a single-channel noise-reduction algorithm (see [Loiz 07b] for an overview). However, the speech can also be pre-processed before playback in order to become more intelligible in presence of the near-end background noise, which is the focus in this work. Here we assume that a clean recording of the speech is available and that the far-end noise is successfully removed with noise-reduction. A relevant application would be a train-station where the intelligibility of an announcement is degraded by a passing train. To improve the speech intelligibility in a noisy environment, one obvious solution would be to increase the level of the speech. However, at a certain point increasing the playback level may not be possible anymore due to loudspeaker limitations. Moreover, unpleasant playback levels may be reached which are close to the threshold of pain. An alternative approach would be to fix the speech energy and redistribute energy within the speech signal over time and/or frequency.

One straightforward and effective approach for improving intelligibility of speech in noise is by boosting high frequencies at a cost of lower frequencies [Grif 68, Nied 76, Hall 10, Skow 06]. For example, Griffiths [Grif 68] derived an optimal linear filter for speech transmission relevant for the articulation index (AI) [Fren 47, Kryt 62]. From this result it was concluded that the speech spectra should be 'whitened' which effectively results in a strong amplification of high frequencies. Similar experimental results were found by Niederjohn and Grotelueschen [Nied 76], where speech was first high-pass filtered followed by fast and severe amplitude compression resulting in a large intelligibility improvement of speech in noise. Also dynamic range compression without any form of high-pass filtering was found to improve speech intelligibility in noise [Rheb 09].

Many other approaches are based on the fact that transient-like parts of

speech signals, e.g., consonants, play an important role in speech intelligibility. For example, Strange *et al.* showed that the center vowel in CVC words is almost fully understandable based on the preceding and succeeding consonant only [Stra 83]. Unfortunately, the energy of consonants is relatively low compared to vowels and therefore, despite their importance, more vulnerable to noise. In line with these findings are the experiments from Gordon-Salant [Gord 86] and Hazan and Simpson [Haza 98], which found significant intelligibility improvements in noise for normal-hearing listeners by amplifying hand-annotated consonants. Similar results were found with hearing-impaired listeners [Kenn 98]. The important perceptual cues of stop consonants are identified in a recent study by Li *et al.* [Li 10]. A follow-up study by Li and Allan [Li 11] shows that by identifying and amplifying these important cues, as found in [Li 10], confusion between consonants can be prevented. The results of these studies could be included in a near-end speech enhancement method, however, similar as with [Gord 86] and [Kenn 98], the method by Li and Allan is not automated [Li 11]. There are methods available which automatically modify the vowel-consonant energy ratio. For example, before amplification, transient regions can be detected with classification rules based on spectral flatness [Skow 06, Huan 09] or the rate of variation of energy and centroid frequency [Jaya 08]. Yoo *et al.* proposed to extract and amplify transient components from the speech based on time-varying filters whose center frequencies and bandwidths were controlled to identify the strongest formant components [Yoo 07]. Note that these methods 1) do not amplify the transients in a frequency-dependent matter and 2) are not taking into account the noise statistics. Based on the results from Li *et al.* [Li 10] it is clear that taking into account these two properties should be more beneficial for improving speech intelligibility.

Although not explicitly based on consonant detection, Sauert and Vary recently proposed several algorithms [Saue 06, Saue 10], which do take into account the noise statistics. These methods improve objective speech intelligibility as predicted by the speech intelligibility index (SII) [ANSI 97]. Other methods exploiting noise statistics exist, for example, based on the masking effects of the auditory system [Brou 08] or a loudness perception model [Shin 07]. Tang and Cooke [Tang 10, Tang 11] investigated several strategies where energy is relocated based on local SNRs. Best results were obtained with a strategy where only the high frequency regions (1800-7500) were amplified, when a local SNR < 5 dB was observed. Interestingly, they also found that redistributing the speech energy, such that the local SNR is made constant either over time, frequency or jointly over time and frequency, may actually decrease the speech intelligibility [Tang 11].

From all the noise-knowledge based methods we conclude that they primarily change the spectrum of the speech and do not use some type of consonant detection strategy. Therefore, the benefits from the earlier mentioned transient-enhancement methods may not be present. One issue with these noise-knowledge based approaches is that their spectral analysis is typically

based on short-time segments (20-40 ms), where the envelope within these frames is completely ignored. However, important information related to consonants may be present within these short-time envelopes. For example, for the identification of important cues of stop consonants in the study by Li *et al.* [Li 10], a computational model was used with a time-resolution of 2.5 ms based on Fletcher's articulation index (AI) [Flet 50].

In this work we present a new method where the speech energy is redistributed as a function of the near-end noise, based on a perceptual distortion measure. The results we present in this article extend existing work due to several reasons: (1) The considered perceptual distortion measure [Taal 09b, Taal 12a] (see Section 6.2.1) takes into account short-time information, which results in a higher sensitivity to transient regions compared to spectral-only models as in, e.g., [Brou 08, Shin 07, Saue 06, Saue 10]. Therefore, the proposed method does not only change the spectrum of the speech to improve speech intelligibility in noise, but also automatically the consonant-vowel ratio as a function of the noise statistics and the speech. (2) We provide an analytic solution to optimally redistribute speech energy relevant for a perceptual distortion measure subject to a power constraint. This is different from the majority of algorithms, which rather normalize the speech signal heuristically after processing which may result in suboptimal solutions. (3) Some algorithms are very effective in improving intelligibility of speech in noise, while they may have poor speech quality (pleasantness or naturalness of speech). For example, aggressive amplitude compression [Nied 76] results in very unnatural speech but the SNR can be lowered down to 15 dB while preserving intelligibility. We will show that the proposed method also has a positive effect on speech quality rather than a negative impact.

The remaining of this article is organized as follows. First we will explain the proposed algorithm and the used perceptual distortion measure in Section 6.2, followed by an evaluation and comparison of other reference methods in Section 6.3. Finally, in Section 6.3.2, a discussion is provided followed by conclusions.

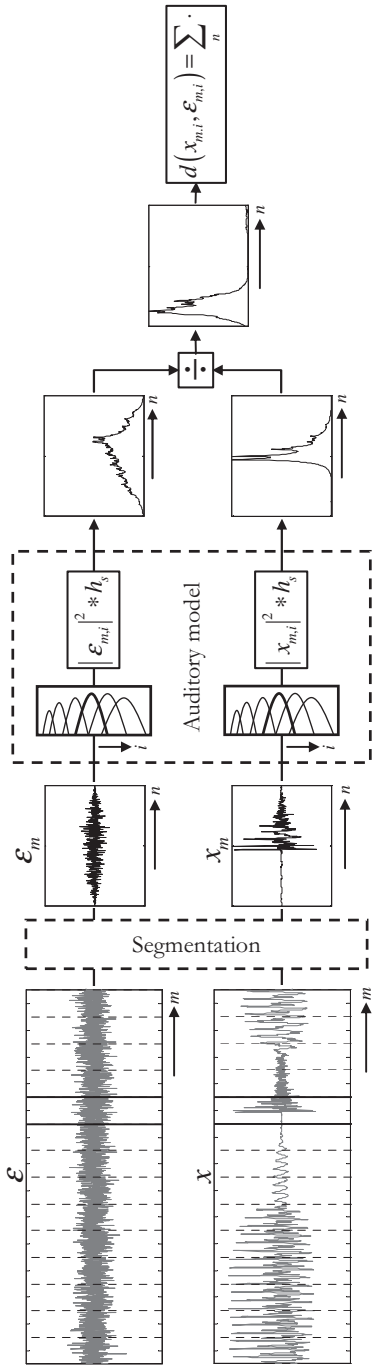

Figure 6.2: Basic structure of the proposed perceptual distortion measure based on the work from [Taal 12a].

6.2 Proposed Speech Pre-Processing Algorithm

Let x denote a time-domain signal representing a speech signal and $x+\varepsilon$ a noisy version, where ε represents background noise. We assume that in isolation, x is fully intelligible and that far-end noise is either absent or successfully removed by a single-channel noise reduction algorithm. The distortion measure considered in this work, denoted by $D(x, \varepsilon)$, will inform us about the audibility of ε in the presence of x. Hence, a lower D value implies less audible noise and therefore more audible speech. Our goal is to adjust the speech signal x such that $D(x, \varepsilon)$ is minimized subject to the constraint that the energy of the modified speech remains unchanged. First, in Section 6.2.1, more details will be given about the considered distortion measure, after which in Section 6.2.2 we will formalize and solve the constrained optimization problem. In Section 6.2.4 some properties of the algorithm are revealed.

6.2.1 Perceptual Distortion Measure

The perceptual distortion measure is based on the work from [Taal 09b] and [Taal 12a]. There are two important motivations why this particular distortion measure is used. (1) The measure takes into account a spectro-*temporal* auditory model and therefore also considers the temporal envelope within a short-time frame (20-40 ms). As a consequence, the distortion measure is more sensitive to transients than spectral-only models, e.g., as used in [Saue 06, Saue 10]. (2) The measure fulfills certain mathematical properties, which make it possible to derive an analytic solution in the eventual constrained optimization problem (See Section 6.2.2).

To guide the reader, we give a brief summary of the perceptual distortion measure presented in [Taal 12a]. The basic structure for the distortion measure is shown in Fig. 6.2. First, a time-frequency (TF) decomposition is performed on the speech and noise by segmentation into short-time (32 ms), 50% overlapping square-root Hann-windowed frames. Then, a simple auditory model is applied to each short-time frame, which consists of an auditory filter bank followed by the absolute squared and low-pass filtering per band, in order to extract a temporal envelope. Here, the filter bank resembles the properties of the basilar membrane in the cochlea, while the envelope extraction stage is used as a crude model of the hair-cell transduction in the auditory system.

Let h_i denote the impulse response of the i^{th} auditory filter and x_m the m^{th} short-time frame of the clean speech. Their linear convolution is denoted by $x_{i,m} = x_m * h_i$. Subsequently, the temporal envelope is defined by $|x_{m,i}|^2 * h_s$, where h_s represents the smoothing low-pass filter. Similar definitions hold for $|\varepsilon_{m,i}|^2 * h_s$. The audibility of the noise in presence of the speech, within one TF-unit, is determined by a per-sample noise-to-signal ratio [Taal 09b]. By summing these ratios over time, an intermediate distortion measure for one TF-unit is obtained denoted by lower-case d. That is,

$$d\left(x_{m,i},\varepsilon_{m,i}\right)=\sum_{n}\frac{\left(\left|\varepsilon_{m,i}\right|^{2}*h_{s}\right)(n)}{\left(\left|x_{m,i}\right|^{2}*h_{s}\right)(n)}, \tag{6.1}$$

where n denotes the time index running over all samples within one short-time frame. As an example, internal representations within one auditory filter are shown in Fig. 6.2 for a windowed noise realization ε_m and a speech transient x_m. Also, the point-wise division in Eq. (6.1) of the internal representations before summation over n is shown in the figure. Due to the fact that the measure uses a per-sample (16 kHz) rather than a frame-based noise-to-signal ratio, the measure is sensitive to the short-temporal structure. The benefit of this will be revealed in Section 6.2.4. Note that the cutoff frequency of the low-pass filter h_s determines the sensitivity of the model towards temporal fluctuations within a short-time frame.

The distortion measure for the complete signal is then obtained by summing all the individual distortion outcomes over time and frequency, which gives,

$$D\left(x,\varepsilon\right)=\sum_{m,i}d\left(x_{m,i},\varepsilon_{m,i}\right). \tag{6.2}$$

6.2.2 Power-Constrained Speech-Audibility Optimization

To improve the speech audibility in noise, we minimize Eq. (6.2) by applying a TF-dependent gain function α which redistributes the speech energy by scaling of the individual (perceptually) filtered frames, i.e., $\alpha_{m,i}x_{m,i}$, where $\alpha_{m,i}\geq 0$. Only TF-units are modified where speech is present. This is done in order to prevent that a large amount of energy would be redistributed to speech-absent regions. We consider a TF-unit to be speech-active when its energy is within a 25 dB range of the TF-unit with maximum energy within that particular frequency band. Note that with the near-end speech enhancement application the clean speech is available and voice activity detection is a relatively easy process (in contrast to the detection of speech already corrupted by noise). The noise is assumed to be a stochastic process denoted by $\mathcal{E}_{m,i}$ and the speech deterministic (recall that the speech signal is known in the near-end enhancement application). Hence, we minimize for the expected value of the distortion measure. Let $\mathcal{L}$ denote the set of speech-active TF-units and $\|\cdot\|$ the ℓ_2-norm, the problem can then be formalized as follows,

$$\begin{aligned}\min_{\alpha_{m,i},\{m,i\}\in\mathcal{L}}\quad & \sum_{\{m,i\}\in\mathcal{L}}E\left[d\left(\alpha_{m,i}x_{m,i},\varepsilon_{m,i}\right)\right]\\ \text{s.t.}\quad & \sum_{\{m,i\}\in\mathcal{L}}\left\|\alpha_{m,i}x_{m,i}\right\|^{2}=r,\end{aligned} \tag{6.3}$$

where $r=\sum_{\{m,i\}\in\mathcal{L}}\left\|x_{m,i}\right\|^{2}$ is the total power measured at the output of the auditory filters and E denotes the expected value. Two important reasons exist

for fixing the speech energy r rather than any other constraint, for example, based on loudness: 1) Typically, algorithms are compared with a listening test by fixing the SNR for which the used global energy constraint is optimal. 2) The used constraint is mathematical tractable in contrast to complex loudness models for which closed-form solutions may not exist, resulting in suboptimal and computational demanding methods, e.g., as in [Shin 07].

By using the method of Lagrange multipliers we introduce the following cost function,

$$J = \sum_{\{m,i\}\in\mathcal{L}} E[d\left(\alpha_{m,i}x_{m,i}, \mathcal{E}_{m,i}\right)] + \lambda\left(\sum_{\{m,i\}\in\mathcal{L}} \left\|\alpha_{m,i}x_{m,i}\right\|^2 - r\right). \tag{6.4}$$

Due to the linearity of the convolution in Eq. (6.1) and the assumption that $\alpha \geq 0$ we have that $d\left(\alpha x, y\right) = d\left(x, y\right)/\alpha^2$. Therefore, in order to minimize Eq. (6.4), we have to solve the following set of equations for α,

$$\begin{array}{rcl} \frac{\partial J}{\partial \alpha_{m,i}} & = & -2\frac{E[d(x_{m,i},\mathcal{E}_{m,i})]}{\alpha_{m,i}^3} + \lambda 2\alpha_{m,i}\left\|x_{m,i}\right\|^2 = 0 \\ \frac{\partial J}{\partial \lambda} & = & \sum\limits_{\{m,i\}\in\mathcal{L}} \alpha_{m,i}^2\left\|x_{m,i}\right\|^2 - r = 0 \end{array} \tag{6.5}$$

The solution is given by,

$$\alpha_{m,i}^2 = \frac{r\beta_{m,i}^2}{\sum\limits_{\{m',i'\}\in\mathcal{L}} \beta_{m',i'}^2\left\|x_{m',i'}\right\|^2}, \tag{6.6}$$

where,

$$\beta_{m,i} = \left(\frac{E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right]}{\left\|x_{m,i}\right\|^2}\right)^{1/4}. \tag{6.7}$$

In order to determine α, we have to compute the expected value $E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right]$, which can be expressed as follows,

$$E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right] = \sum_n \frac{\left(E\left[\left|\mathcal{E}_{m,i}\right|^2\right] * h_s\right)(n)}{\left(\left|x_{m,i}\right|^2 * h_s\right)(n)}, \tag{6.8}$$

Here we used the linearity of the convolution and the summation in order to move the expected value operator inside the distortion measure. To simplify, we assume that the power-spectral density of the noise within the frequency range of an (relatively narrow) auditory band is constant, i.e., has a 'flat' spectrum. As a consequence, the noise within an auditory band can be modeled by $\mathcal{E}_{m,i} = (w_m N_{m,i}) * h_i$, where w_m denotes the window function and $N_{m,i}$ represents a zero mean, i.i.d. stochastic process with variance $E\left[N_{m,i}^2\left(n\right)\right] = \sigma_{m,i}^2, \forall n$. By combining this statistical model and the numerator of Eq. (6.8) we have,

$$
\begin{aligned}
E\Big[|\mathcal{E}_{m,i}|^2(n)\Big] &= E\left[\left|\sum_k h_i\left(k\right) w_m\left(n-k\right) N_{m,i}\left(n-k\right)\right|^2\right] \\
&= \sum_k h_i^2\left(k\right) w_m^2\left(n-k\right) E\left[N_{m,i}^2\left(n-k\right)\right] \\
&= \left(h_i^2 * w_m^2\right)\left(n\right) \sigma_{m,i}^2.
\end{aligned}
\tag{6.9}
$$

where $h_i^2 * w_m^2$ can be calculated offline and reused, and $\sigma_{m,i}^2$ can be estimated with any noise power spectral density (PSD) estimator from the field of single-channel speech enhancement [Loiz 07b] (see next section for more details). Here we use the method from [Hend 10].

6.2.3 Implementation Details

An exponential smoother is applied to $\alpha_{m,i}$ in order to reduce variations which may negatively effect the speech quality, that is,

$$
\hat{\alpha}_{m,i} = \left(1-\gamma\right)\alpha_{m,i} + \gamma\hat{\alpha}_{m-1,i}, \tag{6.10}
$$

where good results were obtained with $\gamma = 0.9$. Note that the applied smoothing in Eq. (6.10) will also prevent that too much energy is distributed to specific TF-units. Hence, large energy differences between TF-units, which may violate some of the motivations for including the power constraint (loudspeaker limitations, unpleasant playback level), are reduced significantly. In rare cases where specific TF-units receive too much amplification, the processed signal is clipped within the available dynamic range.

The filter bank and the low-pass filter are applied by means of a point-wise multiplication in the DFT-domain with real-valued, even-symmetric frequency responses[1]. For the filter bank the approach as presented in [Par 05] is used and for the low-pass filter the magnitude response of a one-pole low-pass filter is used. A total amount of 40 filters are considered spaced according the equivalent rectangular bandwidth (ERB) [Glas 90] between 150 and 8000 Hz. Furthermore, the speech signal is reconstructed by addition of the scaled TF-units where a square-root Hann-window is used for analysis/synthesis.

As mentioned, a noise-tracker from the field of single-channel speech enhancement (i.e., estimating the underlying clean speech given a noisy observation) is used for estimation of the noise PSD. However, three important differences apply when using such a traditional noise-tracker in the field of near-end enhancement:

1. The noise realization which degrades the TF-units in the set $\mathcal{L}$ is a future event. Therefore only the noise PSD from the last known time-frame can be used, which is estimated from previous time-frames. We assume that the noise is stationary during the duration of $\mathcal{L}$.

[1] This particular choice will lead to time-domain aliasing due to circular convolution. However, the applied window function will minimize the effect of these unwanted artifacts. See [Vary 06], page 399 for more details.

2. The noise PSD tracker is applied on the *processed* speech plus noise rather than on the *clean* unprocessed speech plus noise, where the latter equals the situation in single-channel noise reduction.

3. The transfer function from microphone to loudspeaker should be known in order to compensate for any introduced delay and coloration of the signal. In our experiments this is ignored, however, this transfer function can be easily measured offline and included in the algorithm.

Finally it is important to add that noise PSD estimation is significantly easier in the field of near-end enhancement than in single-channel noise reduction, since we have access to the clean speech. Hence, a perfect voice activity detector could also be used where noise statistics are estimated during speech pauses.

The performance of the method depends on the amount of TF-units available in the set $\mathcal{L}$. When this set contains a larger span over time and/or frequency, a better redistribution of energy is possible and a lower final distortion as defined in Eq. (6.2) can be expected. The delay of the proposed algorithm is directly related to the amount of future time-frames in the set $\mathcal{L}$ with respect to the current time-frame. Increasing the lookahead in the set $\mathcal{L}$ will result in a larger delay. Although the delay of the proposed method can be adjusted, depending on the application, we will analyze the following two extreme situations in the remaining of the article: (1) $\mathcal{L}$ contains all TF-units in one entire sentence (say +/- 3 seconds) (PROP1) and (2) $\mathcal{L}$ only contains the set of TF-units in one short-time frame (32 ms) (PROP2). PROP1 is relevant in situations where the speech is pre-recorded and the noise is stationary, e.g., pre-recorded announcements in a car-navigation system, F16 cockpit or safety instructions in a plane. PROP2 is relevant for real-time applications like mobile telephony, or public address systems. For PROP1 the noise PSD is based on averaging estimated noise PSDs over several frames and sentences offline. The delay can be adjusted to anything in between these extreme cases, for example, for mobile telephony where a limited amount of delay is not necessarily an issue [ITU 03].

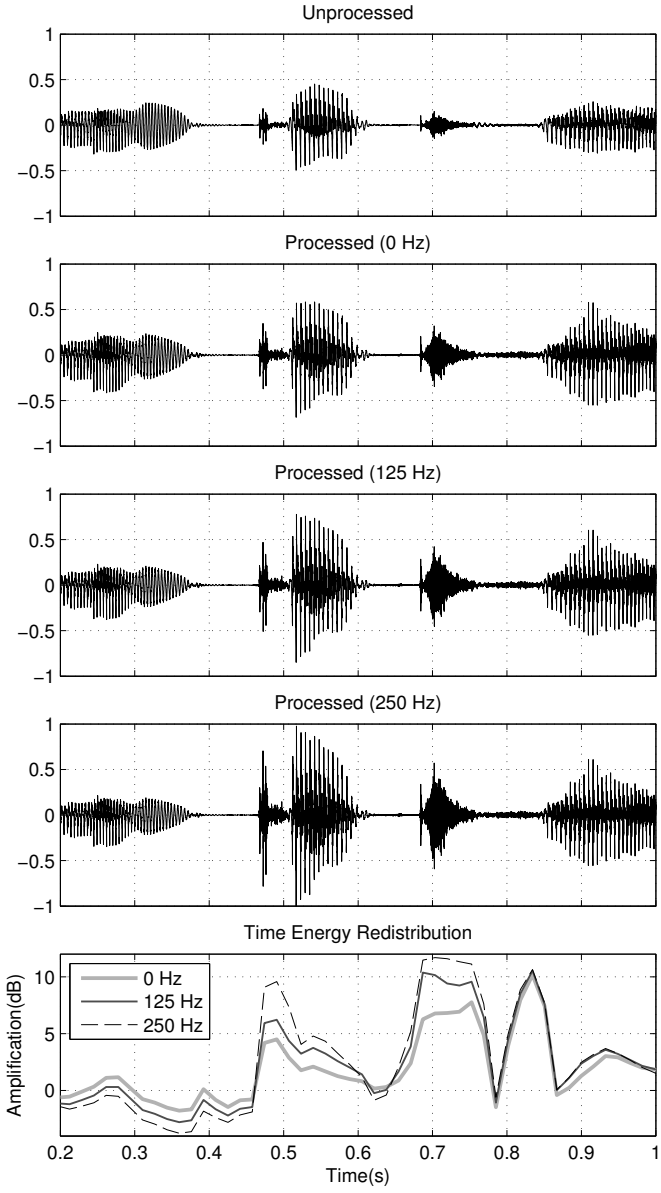

Figure 6.3: Unprocessed and processed (PROP1) speech signal for three different auditory model cutoff frequencies. The bottom plot indicates how the energy is redistributed over time. Notice that transient parts are more amplified when the cutoff frequency is increased.

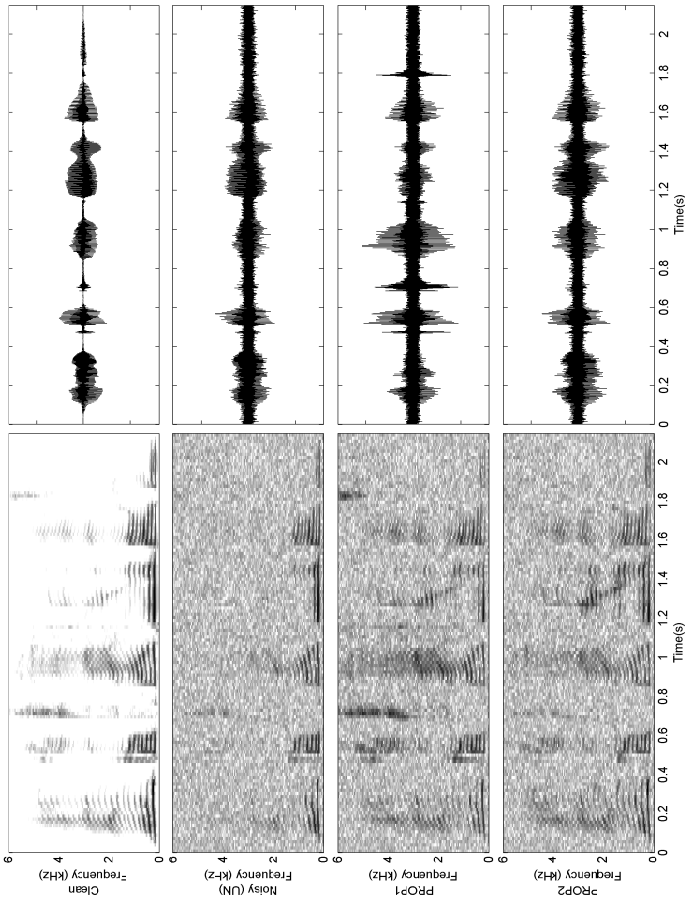

Figure 6.4: Spectrograms and time-domain plots for the clean speech, noisy speech, and noise with added processed speech (PROP1, PROP2). Here, PROP1 redistributes energy over all TF-units and PROP2 only over frequency within one short-time frame.

6.2.4 Properties and Examples

The cutoff frequency of the auditory model low-pass filter h_s (see Section 6.2.1) determines the temporal sensitivity of the distortion measure. A higher cutoff frequency will result in a larger intermediate distortion value for transient signals, and therefore the algorithm will distribute more energy to these regions. We investigated the amount of amplification received by transients as a function of the cutoff frequency. One example is shown in Fig. 6.3 for PROP1 with speech shaped (SSN) noise at -5 dB SNR, where processed clean speech signals are shown for three different cutoff frequencies (0, 125 and 250 Hz). Here, a cutoff frequency of 0 Hz means that the short-time envelope is constant and equals its average value. The bottom plot indicates how the energy is redistributed over time where energy differences are calculated within individual short-time frames independently of frequency. Note that only 0.8 seconds of the entire 2.5 seconds long speech signal is shown. Although transients are also amplified with a cutoff frequency of 0 Hz, this only results in small amplifications in the range of 3-6 dB. In contrast, when we use a cutoff frequency of 125 Hz, transients are amplified in the order of 6-12 dB. This is more in line with results based on earlier research which found better results in this range [Gord 86, Haza 98, Kenn 98, Skow 06].

As an example in Fig. 6.4, the time-domain signals are plotted together with spectograms for the clean and noisy speech and the proposed algorithms PROP1 and PROP2. All spectrograms show the same dynamic range of around 45 dB where the energy of all speech signals (before noise addition) is equal. For this particular example one sentence of speech is used with a length of around 2.5 seconds which is degraded with white noise at 5 dB SNR. From the noisy speech spectrograms we can conclude that the high frequencies are almost fully masked by the noise. For example, the transients at approximately 0.7 and 1.8 seconds are hardly visible anymore and can be expected to be inaudible due to the negative effects of the noise. For PROP1 it can be observed that the transients are almost fully recovered, both in the time domain and spectogram plots. Beside transient regions, the plots also reveal that PROP1 increases the high frequencies of the vowel sounds at, e.g., 0.2 and 1 second(s). For PROP2, this amplification of high frequency vowel sounds is also observed. However, the amplification of transient regions is not present with PROP2, since only energy could be redistributed within one short-time frame. For both PROP1 and PROP2 low frequency regions (around 100-250 Hz) are attenuated in order to accomplish the amplification of high frequencies.

6.3 Experimental Evaluation

6.3.1 Speech Intelligibility

The proposed methods PROP1 and PROP2 are compared with two reference methods by means of an intelligibility listening test. This includes the method as proposed by Skowronski and Harris based on changing the vowel-consonant

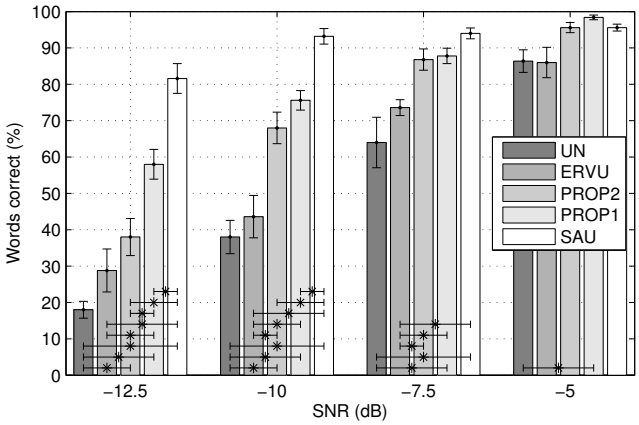

Figure 6.5: Average user intelligibility scores with standard errors of unprocessed (UN) speech degraded with speech shaped noise (SSN) and processed speech plus noise for the proposed algorithms PROP1 and PROP2 and the reference methods ERVU [Skow 06] and SAU [Saue 06].

ratio referred to as energy redistribution voiced/unvoiced (ERVU) [Skow 06]. This particular method detects transients based on the spectral flatness measure where the transient part is amplified by 7.4 dB. After transient amplification the signal is normalized such that it has the same energy as the original speech signal. This method is independent of the noise statistics.

Secondly, the "maximal power transfer" method is included as proposed by Sauert *et al.* (SAU) [Saue 06]. This method is based on a simple model of hearing where it is assumed that its noise reduction pre-processing stage in the human brain acts at least as intelligent as a Wiener filter. Let $\hat{x}_{m,k}$ denote the k^{th} DFT-bin of the m^{th} short-time frame for the clean speech and $\sigma_n^2(m,k)$ the noise PSD. The gain function applied to each DFT-bin is then given as follows,

$$\alpha_{m,k}^2 = \frac{K_1|\hat{x}_{m,k}|^2}{K_1|\hat{x}_{m,k}|^2 + \sigma_n^2(m,k)}, \tag{6.11}$$

where $K_1 = 0.01$ to deliver the best possible speech intelligibility. The power-constraint is included by normalizing the processed speech per short-time frame such that it equals the energy of the unprocessed speech per short-time frame. For SAU we use the noise-tracker as proposed in [Saue 06] which equals a recursive average of the noise periodogram with an adaptively chosen noise floor. Hence SAU assumes access to the noise realization. In order to make a fair comparison the noise tracker in the proposed method is therefore also applied on the noise only rather than the noisy speech.

Table 6.1: p-values for comparing intelligibility scores between algorithms.

SNR	Method	> UN	> ERVU	> PROP2	> PROP1
-12.5	ERVU	0.035	-	-	-
	PROP2	0.002	0.062	-	-
	PROP1	0	0	0	-
	SAU	0	0	0	0.001
-10	ERVU	0.217	-	-	-
	PROP2	0	0.007	-	-
	PROP1	0	0	0.086	-
	SAU	0	0	0	0
-7.5	ERVU	0.094	-	-	-
	PROP2	0.01	0.001	-	-
	PROP1	0.005	0	0.346	-
	SAU	0.001	0	0.037	0.046
-5	ERVU	0.549	-	-	-
	PROP2	0.01	0.021	-	-
	PROP1	0.002	0.007	0.022	-
	SAU	0.015	0.038	0.5	0.967

Listening Test

Ten Dutch-speaking listeners were asked to judge the intelligibility of the unprocessed noisy signals and processed speech signals plus noise. The speech signals were taken from the Dutch Matrix-test [Koop 07], which consists of 5-word sentences spoken by a female speaker. The sentences are of the grammatical form name-verb-numeral-adjective-noun (e.g. Ingrid owns six old jackets) as proposed by Hagerman [Hage 82], where each word in the sentence is picked randomly from a list of 10 possible words. This means that there is a probability of 10% that the correct word is chosen in the case that the speech unintelligible. The subject had access to the closed set of words by means of a 10-by-5 matrix on a computer screen, such that the i^{th} column contains exactly the 10 possible alternatives for the i^{th} word. The task of the listener is to select via a graphical user interface the understood words. For each test sentence, one word from each column must be selected where the sentence was played once only. Signals are sampled at 16 kHz and degraded with SSN at the SNRs of -12.5, -10, -7.5 and -5 dB and processed with PROP1, PROP2, ERVU and SAU. The unprocessed noisy speech is also included in the test (UN). For each condition the listener is presented with five, five-word sentences through headphones (Sennheiser HD 280 pro) where each sentence was used only once. As a consequence, each subject listened to a total of 5 sentences * 5 algorithms * 4 SNRs = 100 sentences in total. The order of presenting the different algorithms and SNRs was randomized. The score per user and for one condition

was consequently obtained by the average percentage of correct words.

Results

The average user scores together with standard errors for all conditions are shown in Fig. 6.5. We found that the differences between subject responses were small, where between-subject correlations were found in the range of 0.76-0.95. Statical analysis is performed per SNR condition by means of multiple paired one-sided t-tests, where in total p-values are determined for ten hypotheses. All four algorithms are tested whether they significantly improved intelligibility compared to the noisy speech. Furthermore, we compared whether PROP1 performed better than ERVU, PROP2 better than ERVU and PROP1, and SAU better than PROP2, PROP1 and ERVU. A statistical significance level of $\alpha = 0.05$ is used with Holm-Bonferroni correction for testing multiple hypotheses [2] [Holm 79]. Significant differences are denoted in Fig. 6.5 by a connection with asterisk marker between the two corresponding bars. The p-values can be found in Table 6.1. Note that for the -5 dB conditions the p-values are relatively high in two situations (ERV>UN and SAU>PROP). This is due to the fact that the ranking is the opposite than the one tested.

From the results we can conclude that for the lowest three SNRs the algorithms have the same ranking in performance. That is, all algorithms improve speech intelligibility compared to the noisy speech where SAU showed the best performance followed by PROP1, PROP2 and ERVU. For the highest SNR of -5 dB results were slightly different, which is probably due to ceiling effects, i.e., most of the listeners had scores close to 100 percent. For the SNRs of -12.5 and -10 dB all algorithms significantly improve speech intelligibility compared to the noisy speech, except for ERVU where the improvements where not statistically significant. The fact that ERVU has a smaller effect on intelligibility compared to the other three methods is expected, since it is only limited to changing the consonant-vowel ratio and not the spectrum of the speech as a function of the noise, which is case with the other methods. Furthermore, as hypothesized in Section 6.2.2, we found that in general PROP1 performs better than PROP2, where a significant improvement was found for -12.5 dB SNR. This is a direct consequence of the fact that the energy can be distributed over the complete signal in PROP1 rather than only within one short-time frame as with PROP2. Best performance was obtained with SAU which showed better performance than all methods for the lowest three SNRs. As we will show in the discussion in Section 6.3.2, the good performance of SAU comes with a cost in speech-quality.

[2] The Holm-Bonferroni test is a sequentially rejective version of the more conservative Bonferonni test, which can be applied in the same situations where the classical Bonferroni test is usually applied [Holm 79].

6.3.2 Speech Quality

From the listening test results we can conclude that the proposed methods PROP1 and PROP2 both lead to intelligibility improvements of the speech, when corrupted by SSN. However, we also found that one of the reference methods (SAU) performed somewhat better than the proposed methods in terms of speech intelligibility. This result is remarkable since SAU includes a power constraint within one short-time frame which implies a much lower algorithmic delay than PROP1, which distributes energy over a complete sentence. One would expect better results with PROP1 since the energy could be redistributed more efficiently over the complete signal rather than only one short-time frame. It is hypothesized that one possible reason for this difference is the fact that the considered distortion measure, as defined in Section 6.2.1, is based on audibility rather than intelligibility [Taal 12a]. Although audibility shares some aspects of intelligibility, it is typically used as a feature for speech quality predictive models, see, e.g., [Quac 88, Loiz 07b]. Perhaps our method optimizes more for speech quality rather than intelligibility, while SAU may be suboptimal in terms of speech quality despite its good intelligibility performance.

The relation between quality and intelligibility, for the considered type of processing artifacts in this article, is therefore further investigated in Section 6.3.2, where we indeed find a mismatch in algorithm ranking with respect to quality and intelligibility. Furthermore, state-of-the-art intelligibility predictors are analyzed whether they correctly predict the intelligibility listening test results and could therefore be used for providing hints on why SAU performs better than PROP1 in terms of intelligibility and not in quality. In the final section in this discussion the algorithmic delay is inspected as a function of speech intelligibly by one of the best performing intelligibility predictors.

Objective PESQ scores

Additional tests are performed to investigate the speech quality of the different methods. Speech quality is predicted by PESQ [Rix 02] for 6 different noise types at SNRS within the range of -15 and 10 dB. The wideband version of PESQ is used, which is standardized as ITU-T recommendation P.862.2 [ITU 05] and is suitable many different degradations as typically encountered in telephony applications. This includes (non-)linear degradations which share similar properties as the proposed algorithm, e.g., the addition of background noise, filtering [Beer 02] and applying TF-varying gain functions as used in noise reduction [Hu 08a] and source separation algorithms [Mowl 12]. Moreover, the use of PESQ with the type of speech processing used in this journal is validated with an additional listening test, where the results are in line with the PESQ predictions.

The speech quality is predicted for UN, PROP1, PROP2, ERVU and SAU. The noise types include babble, F16 cockpit, factory, white, speech shaped and bottling factory noise. An average PESQ score is calculated based on 50

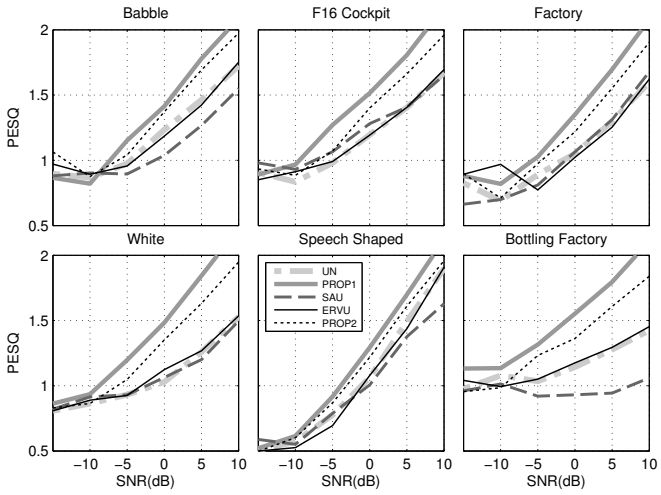

Figure 6.6: PESQ based speech quality predictions for the different algorithms and unprocessed noisy speech for six different noise types.

random sentences from the Timit database [Garo 93], which originates from different types of speakers. The results are shown in Fig. 6.6. For lower SNRs around -10 dB the results show that a lower bound for speech quality is reached and the PESQ scores are more or less random which is in line with the findings reported in [Liu 08]. For higher SNRs it can be observed that the proposed methods PROP1 and PROP2 have a positive effect on speech quality for all noise types and most of the SNRs. ERVU does not strongly affect the speech quality and SAU even has a negative effect on speech quality for some noise types.

Listening Test

As an initial step to see whether these PESQ predictions are in line with real listening tests, the methods PROP1, SAU and UN are compared with each other by means of an AB-preference test. Ten subjects listened to two versions of the same speech sentence and were asked which sentence they preferred in terms of speech quality, e.g., pleasantness and/or naturalness of speech. These subjects were different than from the first experiment. We compared UN with PROP1, UN with SAU and PROP1 with SAU. The order of the two sentences was randomized where the subject listened to five different sentences per algorithm comparison. Thus, in total each subject listened to a total of 3 algorithm pairs * 2 sentences per comparison * 5 sentences = 30 sentences. Random sentences were taken from the Timit database [Garo 93] at a sample rate of 16 KHz and corrupted by SSN at an SNR of 5 dB.

From the results of the listening test, as shown in Fig. 6.7, we can see that the ranking is similar as with the PESQ predictions for SSN. That is, SAU

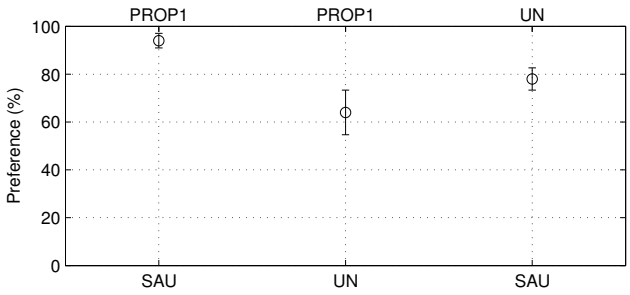

Figure 6.7: Average users score of AB-preference listening test between the unprocessed noisy speech UN and the algorithms PROP1, SAU for speech shaped noise degraded at 5 dB SNR. The error bars denote the standard error of the average user preference.

actually decreases speech quality compared to the unprocessed noisy speech and PROP1 has better speech quality than SAU. Statistical analysis by means of a Wilcoxon rank sum test (significance level of $\alpha = 0.05$) indicates that the comparisons of PROP1>SAU and UN>SAU are statistically significant with $p < 0.0005$. The speech quality of PROP1 was better than UN, however, this was not significant with $p = 0.08$.

6.4 Discussion

6.4.1 Speech Quality versus Intelligibility

One possible explanation for the difference in speech quality and speech intelligibility for SAU in the SSN case, is its strong amplification of higher frequencies. The amplification of high frequencies is also present with PROP1, however, to a less extent. To get a better insight in these properties, the average spectra are plotted in the top figure in Fig. 6.8 for the unprocessed speech, the processed speech for SAU and PROP1 and the noise. Signals are mixed at -10 dB where 50 sentences are used for estimating the spectrum. We observe that indeed the average spectrum of PROP1 is closer to the original speech and therefore may sound more natural and has therefore better quality. However, these high frequencies may be responsible for the good performance in terms of intelligibility.

Besides SSN, which has a low-pass characteristic, also white noise and noise from a bottling factory hall are included which contain more high frequencies as shown in the bottom two plots of Fig. 6.8. From these spectra we can clearly see that SAU tends to give the speech spectrum the shape of the inverse noise spectrum. This is indeed in line with Eq. (6.11) where for low SNRs the gain function will be dominated by the inverse of the noise PSD. This is most visible for the bottling factory noise where the strong high frequencies present in the noise (1000-2000 Hz) are attenuated in the speech. This probably

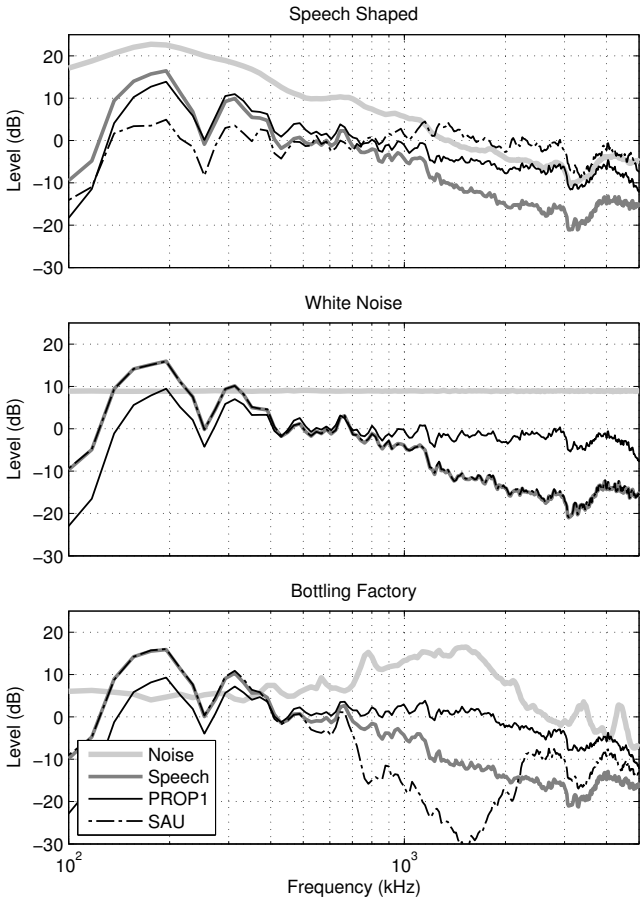

Figure 6.8: Average processed speech spectra for SAU and PROP1 plus unprocessed speech and noise spectra for different noise types.

explains its negative effect on speech quality of SAU for this noise type as was predicted by PESQ. From the spectra we also observe that SAU does not change the spectrum for the white noise, since the inverse of this noise spectrum results in a flat gain function. Therefore, the benefits of SAU with SSN are not expected with white noise. With PROP1 we observe a different effect for white noise, where the high frequencies of speech are amplified instead. We know that amplifying high frequencies in the case of white noise improves speech intelligibility [Grif 68, Nied 76, Skow 06], therefore it is expected that PROP1 will also improve intelligibility and therefore will show better performance than SAU for this particular noise type. In the future, additional tests will be performed to test the algorithms for other noise types.

Based on the experiments we can conclude that intelligibility is more relevant for lower SNRs and quality for higher SNRs. For example, the PESQ scores in Figure 6.5 show a lower-bound convergence around -5 dB. Here, the speech quality is probably dominated by the low SNRs and the added noise, rather than the applied speech pre-processing algorithms. Moreover, a ceiling effect can be observed with respect to speech intelligibility in Figure 6.5 around -5 dB, where most of the conditions result in almost fully intelligible speech. We would like to add that this SNR is relatively low because the speech material is based on a closed set of words and is therefore more easy to understand. In the case of more realistic sentence-based material the intelligibility can still be harmed at 5 dB [Hu 07a] and in the case of non-native speakers this could even go up to 15 dB [Wijn 02]. Note that announcements by non-native speakers is very likely to happen at (international) airports and train stations.

In many applications a large range of SNRs can be expected. In a train station, for example, speech can become unintelligible due to a passing train, while little far-end noise will be present outside rush hour. As a consequence, the algorithm should have good performance in both quality and intelligibility over a wide range of SNRs like the proposed method. For applications where speech is only presented at lower SNRs, or where speech quality is of minor importance like military applications, one could argue that SAU should be preferred over the proposed methods. However, as previously explained, this argument may not be valid for all noise types since it is expected that the proposed method will result in higher intelligibility than SAU in the case of white noise.

6.4.2 Predicted Intelligibility versus Algorithmic Delay

In the experiments performed in this article we included two extreme cases of the algorithm, that is PROP1 and PROP2, referring to a low and high algorithmic delay, respectively. As hypothesized, we found that increasing the lookahead, and therefore the delay, leads to more intelligible speech in noise. An interesting question is how much lookahead is needed in order to reach maximum intelligibility. To answer this question, an initial experiment is performed where the speech intelligibility is predicted with STOI [Taal 11a] as a function of algorithmic delay. STOI is used for prediction since it gave high

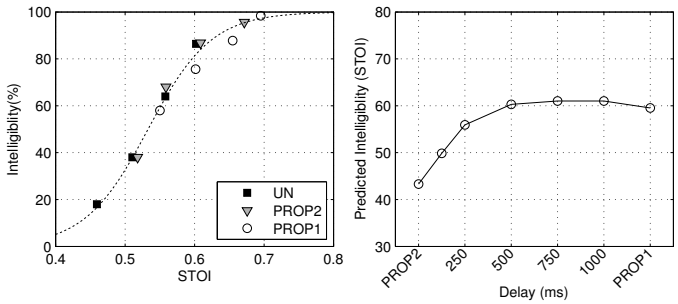

Figure 6.9: STOI predictions versus listening test results (left) and predicted intelligibility by STOI for SSN, -12.5 dB SNR as a function of algorithmic delay (right).

correlation for the conditions PROP1, PROP2 and UN in isolation ($\rho = 0.96$) as shown in the left plot in Figure 6.9. Here a scatter plot is shown between the STOI predictions and the actual intelligibility scores of the listening test. A logistic function is fitted which will be used to map the STOI predictions to intelligibility scores for other algorithmic delays.

In total 50 sentences of the Dutch matrix test [Koop 07] are used, which are degraded with SSN at an SNR of -12.5 dB, where we found the largest difference in intelligibility between PROP1 and PROP2 in the listening test. We considered the following block lengths in which energy could be redistributed by the proposed method: 125, 250, 500, 750, and 1000 ms. Furthermore, the versions PROP1 (approximately 2-3 seconds) and PROP2 (32 ms) were also included in the experiment. To prevent fast changing fluctuations between consecutive blocks, a 50% block overlap together with a Hann window is used. The results are shown in Fig. 6.9, from which we can conclude that from around 500 ms the predicted intelligibility tends to converge to the performance of PROP1. These predictions indicate that almost maximum intelligibility can be achieved for certain voice applications, e.g., international telephone connection, when taking into account the maximum tolerated network delay of around 400 ms [ITU 03].

6.4.3 Algorithm Performance in Far-End Noisy Conditions

In the proposed method we assume that we have access to a clean recording of the far-end speech which has good quality and intelligibility. In the case that the speech is still corrupted by (residual) background noise, the algorithm will increase the audibility of the far-end *noisy* speech, rather than the far-end *clean* speech. In this situation issues may occur, especially when the noisy speech contains many noise-only TF-units. The algorithm will consider those TF-units to be speech rather than noise and therefore energy may be distributed to the

wrong TF-units. Nevertheless, initial (informal) experiments in far-end noisy conditions do suggest that the intelligibility is still improved, though, to a less extent compared to clean ideal conditions. Additional listening tests have to be performed in order to confirm this. However, it is suggested to apply a single-channel noise reduction algorithm, e.g., [Erke 07] to the far-end noisy speech in order to improve the performance of the proposed algorithm.

6.5 Conclusions

A speech pre-processing algorithm is presented to improve the speech intelligibility in noise for the near-end listener without adjusting the speech energy. This was accomplished by optimally redistributing the speech energy over time and frequency based on a perceptual distortion measure. Due to the fact that the distortion measure takes into account short-time information, transient signals, which are more important for speech intelligibility than vowels, receive more amplification. The lookahead of the algorithm can be adjusted to the specific application. To verify the effect of this, two extreme versions were considered of the proposed method: one with maximum lookahead, where energy is distributed over time and frequency jointly for a complete sentence (PROP1), and one with minimum lookahead where energy is redistributed over frequency within a short-time frame (PROP2). From the results we can conclude that the proposed methods result in a large intelligibility improvement compared to the noisy unprocessed speech. PROP1 performed better than PROP2 due to the fact that PROP1 contains a larger time-span where a better redistribution of energy is possible. However, this results in a larger algorithmic delay. The proposed methods were compared with a method where transients were amplified (ERVU) and a method which redistributes energy over frequency within one short-time frame (SAU). PROP1 and PROP2 resulted in higher intelligibility scores than ERVU. Best performance in terms of speech intelligibility was obtained with SAU. However, additional tests reveal that the good performance of SAU comes with a decrease in speech quality in contrast to PROP1, where next to intelligibility, also a positive effect on speech quality was found. Matlab code of PROP1 and PROP2 is provided at `http://www.ceestaal.nl/nrgredist.zip`.

Chapter 7

Matching Pursuit for Channel Selection in Cochlear Implants Based on an Intelligibility Metric

This chapter is published as "Matching Pursuit for Channel Selection in Cochlear Implants Based on an Intelligibility Metric", by C. H. Taal, R. C. Hendriks and R. Heusdens, at *EUSIPCO*, Bucharest, Romania, 2012. This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699.

7.1 Introduction

Reliable machine-driven predictors of speech intelligibly are of great interest in the design process of new speech processing algorithms, e.g., as used in mobile telephony, hearing aids or cochlear implants (CIs). They might replace costly and time consuming listening tests, at least in some stages of the algorithm development process. The drawback of many intelligibility predictors is that they are complex [Chri 10, Chen 11] and do not have certain (mathematical) properties in order to derive optimal signal processing solutions, e.g., least-squares solutions. In previous work we proposed a short-time objective intelligibility (STOI) measure which can accurately predict the effect of background noise and various (non-)linear speech processing algorithms on speech intelligibility [Taal 11a]. We will show that STOI can be simplified to a weighted ℓ_2 norm in the auditory domain which makes the measure mathematically tractable. Since STOI shows high correlation with the intelligibility of vocoded speech [Taal 11a], as typically used in acoustic CI-simulations, the norm will be applied in the channel-selection technique with CI simulations [Wils 08, Dorm 02].

The channel-selection technique is also referred to as the *n-of-m* strategy where n channels of the available m frequency channels (electrodes) are stimulated, such that important channels can be updated more frequently and less significant channels are omitted. Several strategies exist to select those channels, e.g., based on peak-picking [Seli 95], psychoacoustic models [Nogu 05] and other techniques [Wils 08]. However, these techniques optimize for certain (psychoacoustic) criteria which exclude important properties relevant for speech intelligibility [Taal 11b]. For example, criteria relevant for speech intelligibility should take into account temporal modulation frequencies important for intelligibility (4-32 Hz) [Drul 94b] and correlation based comparisons should be used rather than comparisons based on squared errors [Taal 11b]. The proposed norm based on STOI takes into account these aspects.

Due to the mathematical properties of a norm, the channel selection can now be solved in an optimal manner for STOI with the matching pursuit algorithm [Mall 93]. Within this framework the electrical spread per electrode can also be easily taken into account, which is typically not part of the optimization process in existing *n-of-m* strategies. It will be shown that the proposed method leads to more intelligible speech compared to a general peak-picking algorithm by means of acoustical CI-simulations with normal-hearing listeners.

7.2 Derivation of Intelligibility Metric

We will first introduce a general notation and explain the auditory model as used in STOI. Let $x(n)$ and $y(n)$ denote a clean and degraded speech signal, respectively, with time-sample index n, where y is a vocoded version of x. A basic auditory model is applied to both signals in order to obtain an internal representation. Here, we only explain the notation for the internal representation of x. Similar definitions hold for y. Let $\hat{x}_m(k)$ denote the k^{th} DFT-bin of

xw_m, where w_m denotes a Hann-window function with frame-index m. Here, a frame length of 16 ms is used with 50% overlap. The short-time DFT spectrum is converted into auditory bands as follows:

$$X_{i,m} = \sum_k \left| \hat{h}_i(k) \hat{x}_m(k) \right|^2, \tag{7.1}$$

where i denotes the auditory band index and $\hat{h}_i$ represents an approximation of the magnitude response of a 4^{th} order gammatone filter as described in [Par 05]. The value $X_{i,m}$ will be referred to as a time-frequency (TF) unit. In total, 32 filters are used with center frequencies linearly spaced on an ERB scale between 150 and 5000 Hz. STOI compares the clean and degraded speech in the auditory domain in blocks of approximately 400 milliseconds (see next section for more details). The following vector notation is used to denote such a block within one auditory band,

$$\mathbf{x}_{i,m} = \left[\, X_{i,m-M+1} \quad X_{i,m-M+2} \quad \cdots \quad X_{i,m} \, \right]^T, \tag{7.2}$$

where M can be used to control the length of such a speech segment, depending on the sample rate and window size. In this work, a sample rate of 16 kHz is used where $M = 48$. Vectors are concatenated over all auditory bands to denote a complete TF-block as:

$$\mathbf{x}_m = \left[\, \mathbf{x}_{1,m}^T \quad \mathbf{x}_{2,m}^T \quad \cdots \quad \mathbf{x}_{I,m}^T \, \right]^T, \tag{7.3}$$

where $I = 32$ denotes the total amount of auditory filters. The operator notation $\mathbf{x}_m = \mathcal{I}_m\{x\}$ is used to denote the complete transform from the time-domain to one TF-block in the auditory domain.

7.2.1 STOI Background and Simplification

As proposed in STOI [Taal 11a], an intermediate measure relevant for speech intelligibility of one TF-unit is defined as the sample correlation coefficient between the clean ($\mathbf{x}_{i,m}$) and degraded ($\mathbf{y}_{i,m}$) speech temporal band envelopes in one block. Blocks of a few hundreds of milliseconds are used to include important modulation frequencies for intelligibility [Drul 94b]. The correlation coefficient is used, rather than, e.g., a squared error, to make sure that the measure is insensitive to band-level differences between x and y, which should not have a strong impact on speech intelligibility [Taal 11b]. To simplify, the correlation coefficient is defined on the magnitude squared envelopes rather than the magnitude envelopes, as was originally proposed in STOI [Taal 11a]. The benefit of this choice will become clear in Section 7.4. This gives:

$$\rho_{i,m}(x,y) = \frac{\left\langle \mathbf{x}_{i,m} - \mu_{\mathbf{x}_{i,m}}, \mathbf{y}_{i,m} - \mu_{\mathbf{y}_{i,m}} \right\rangle}{\sigma_{\mathbf{x}_{i,m}} \sigma_{\mathbf{y}_{i,m}}}, \tag{7.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product with $\|\cdot\|$ as its induced ℓ_2-norm, $\mu_{\mathbf{x}_{i,m}}$ the sample mean of $\mathbf{x}_{i,m}$ and $\sigma_{\mathbf{x}_{i,m}} = \left\| \mathbf{x}_{i,m} - \mu_{\mathbf{x}_{i,m}} \right\|$. Similar definitions hold for

the degraded speech. The correlation coefficients $\rho_{i,m}(x,y)$ are then combined into one number by computing its average over all TF-units:

$$D = \frac{1}{\mathcal{M}} \sum_{i,m} \rho_{i,m}\,(x,y), \tag{7.5}$$

where $\mathcal{M}$ denotes the total number of TF-blocks. It is expected that D is a monotonically increasing function of the speech intelligibility of y. In computing D only those TF-blocks are considered in the summation where speech is present, see [Taal 11a] for more details. An additional clipping procedure in STOI, which was included to limit the intermediate intelligibility range, is discarded in this work for simplicity.

7.2.2 Interpretation as weighted ℓ_2 norm

To rewrite the intelligibility measure as a norm we first express (7.4) as an inner product:

$$\rho_{i,m}\,(x,y) = \langle \bar{\mathbf{x}}_{i,m}, \bar{\mathbf{y}}_{i,m} \rangle\,, \tag{7.6}$$

where a general normalization procedure is denoted by $\bar{(\cdot)} = \big((\cdot) - \mu_{(\cdot)}\big)/\sigma_{(\cdot)}$. Hence, the inner product $\langle \bar{\mathbf{x}}_{i,m}, \bar{\mathbf{y}}_{i,m} \rangle$ can be used to induce the following norm:

$$\begin{aligned} \|\bar{\mathbf{x}}_{i,m} - \bar{\mathbf{y}}_{i,m}\|^2 &= \|\bar{\mathbf{x}}_{i,m}\|^2 + \|\bar{\mathbf{y}}_{i,m}\|^2 - 2\,\langle \bar{\mathbf{x}}_{i,m}, \bar{\mathbf{y}}_{i,m} \rangle \\ &= 2 - 2\rho_{i,m}\,(x,y). \end{aligned} \tag{7.7}$$

It can now be observed that maximizing $\rho_{i,m}$ implies minimizing the norm $\|\bar{\mathbf{x}}_{i,m} - \bar{\mathbf{y}}_{i,m}\|^2$. However, its minimizing argument only determines the optimal $\mathbf{y}_{i,m}$ up to a scaling $\sigma_{\mathbf{y}_{i,m}}$ and amplitude shift $\mu_{\mathbf{y}_{i,m}}$. In this work we aim for the solution where the clean speech is the target, with the assumption that $\mu_{\mathbf{x}_{i,m}} \approx \mu_{\mathbf{y}_{i,m}}$ and $\sigma_{\mathbf{x}_{i,m}} \approx \sigma_{\mathbf{y}_{i,m}}$. This is motivated by the fact that we are working in blocks of a few hundreds of milliseconds, and it is expected that the errors introduced to $\mathbf{y}_{i,m}$ will average to a minimal impact when summing over all its elements in the calculation of the scaling $\sigma_{\mathbf{y}_{i,m}}$ and amplitude shift $\mu_{\mathbf{y}_{i,m}}$. This gives:

$$\|\bar{\mathbf{x}}_{i,m} - \bar{\mathbf{y}}_{i,m}\|^2 \approx \|a_{i,m}\,(\mathbf{x}_{i,m} - \mathbf{y}_{i,m})\|^2 \tag{7.8}$$

where $a_{i,m} = \sigma^{-1}_{\mathbf{x}_{i,m}}$. By vector concatenation as in (7.3) the summation over frequency i in (7.5) can be replaced by defining a new norm over a complete TF-block. First, a diagonal weighting matrix is defined as:

$$\mathbf{A}_m = \mathrm{diag}\,\big(\; a_{1,m}\mathbf{I}_M \quad a_{2,m}\mathbf{I}_M \quad \cdots \quad a_{I,m}\mathbf{I}_M \;\big)\,, \tag{7.9}$$

where $\mathbf{I}_M$ is the identity matrix of size M. A weighted norm for one TF-block is then given as follows:

$$\|\mathbf{A}_m\,(\mathbf{x}_m - \mathbf{y}_m)\|^2 = \sum_i \|a_{i,m}\,(\mathbf{x}_{i,m} - \mathbf{y}_{i,m})\|^2. \tag{7.10}$$

These weighted norms are then combined by a summation over time, where for optimization purposes the averaging constant $\mathcal{M}$ in (7.5) can be discarded. Note that $\mathbf{A}_m$ is only a function of the clean speech $\mathbf{x}_m$. As a result, it only has to be calculated once for each frame after which the norm can be evaluated for any arbitrary $\mathbf{y}_m$.

7.3 Application to CI channel selection

The proposed intelligibility metric will be used in the CI channel-selection technique with the matching pursuit algorithm [Mall 93]. With this algorithm, a signal x is synthesized as a weighted sum of functions (sometimes called atoms or elements) which are chosen from a dictionary [Mall 93]. The algorithm is iterative, where for each iteration p the best matching function g from the dictionary $\mathcal{D}$ is chosen and subtracted from the residual at the previous iteration. Since only one element is considered per iteration, the algorithm is greedy. The eventual synthesized speech signal can be described by:

$$x \approx \sum_{p} \alpha^{(p)} g^{(p)}, \tag{7.11}$$

where the selection of the best dictionary element and weighting coefficient α is based on minimizing some norm of the eventual residual r. For the $(p+1)^{th}$ iteration this residual is given as follows:

$$r^{(p+1)} = r^{(p)} - \alpha^{(p)} g^{(p)}, \tag{7.12}$$

where for the first iteration the residual is taken equal to the target signal, i.e, $r^{(1)} = x$. The optimal solution for the weighting coefficient and selection of the dictionary element in each iteration is given by [Mall 93],

$$\begin{aligned} \alpha^{(p)} &= \frac{\left\langle g^{(p)}, r^{(p)} \right\rangle}{\left\| g^{(p)} \right\|^2} \\ g^{(p)} &= \arg\max_{g \in \mathcal{D}} \frac{\left| \left\langle g, r^{(p)} \right\rangle \right|}{\|g\|} \end{aligned} \tag{7.13}$$

7.3.1 Intelligibility Relevant Matching Pursuit

Since all diagonal elements of $\mathbf{A_m}$ in (7.10) are real and positive a new norm relevant for speech intelligibility can be defined, say $\|\cdot\|_{\mathbf{A}_m}$, which is induced from the following inner product:

$$\langle \mathbf{x}_m, \mathbf{y}_m \rangle_{\mathbf{A}_m} = \langle \mathbf{A}_m \mathbf{x}_m, \mathbf{A}_m \mathbf{y}_m \rangle . \tag{7.14}$$

Now we can insert the proposed norm and inner product based on STOI in (7.12) and (7.13). Here, the dictionary will be defined by $\mathcal{D} = \mathbf{g}(\gamma)_{\gamma \in \Gamma}$, where Γ denotes the set of CI frequency channel indices. Each element represents the internal representation of a short-time pulse within a specific CI channel

and will be used to model $\mathbf{x}_m$. One can choose the dictionary according to the properties of the CI and include aspects like the pulse duration, channel center frequencies or the amount of current spread. To imply low algorithmic delay no future time-samples are taken into account in these internal representations for a given pulse.

For the first iteration where no channel selection has been made yet, the residual is set to $\mathbf{r}_m^{(1)} = \mathbf{x}_m$, where for the next iterations we have:

$$\mathbf{r}_m^{(p+1)} = \mathbf{r}_m^{(p)} - \alpha^{(p)}\mathbf{g}^{(p)}. \tag{7.15}$$

The solution for the best dictionary element and optimal weighting for each iteration relevant for the proposed metric is then given by:

$$\begin{aligned} \alpha^{(p)} &= \frac{\left\langle \mathbf{g}^{(p)}, \mathbf{r}_m^{(p)} \right\rangle_{\mathbf{A}_m}}{\left\| \mathbf{g}^{(p)} \right\|_{\mathbf{A}_m}^2} \\ \mathbf{g}^{(p)} &= \arg\max_{g \in \mathcal{D}} \frac{\left| \left\langle \mathbf{g}, \mathbf{r}_m^{(p)} \right\rangle_{\mathbf{A}_m} \right|}{\|\mathbf{g}\|_{\mathbf{A}_m}}. \end{aligned} \tag{7.16}$$

After the channels have been selected, the eventual residual $\mathbf{r}_m$ is stored and shifted one time-frame over m for the initial residual $\mathbf{r}_{m+1}^{(1)}$. In this manner, past channel selections are also taken into account for the decisions of the current time-frame.

7.4 Vocoder Details

CI simulations are performed with a vocoder based on sinusoidal carriers similar to [Dorm 02]. In this vocoder 20 channels are used with logarithmically spaced frequencies between 150-5000 Hz. Each sinusoid is segmented into 8 ms length, 50% overlap Hann-windowed frames, which implies a channel simulation rate of 250 Hz. Note that these settings simulate the properties of the CI-processor and are chosen independently of the auditory model from Section 7.2.

First we will show that the time-domain additivity of the TF-spaced sinusoids in the vocoder can be preserved in the auditory domain, which validates the use of (7.11) in the auditory domain. Let a TF-spaced sinusoid be described as follows:

$$s_\gamma(n) = \cos(\omega_\gamma n + \phi)\, w_s(n), \tag{7.17}$$

where ω_γ denotes the angular frequency for channel γ, respectively, and w_s its window function (the subscript s of this vocoder window is used to denote its difference with the auditory model window w_m from Section 7.2). For readability, the vocoder relevant frame-index is omitted and we assume that w_s represents the current frame of interest. Since the phase is of minor importance for intelligibility in these short time frames [Liu 97], ϕ is assumed to be i.i.d. uniformly distributed between 0 and 2π and only the average internal representation is considered. The expected value of s_γ for one TF-unit in the auditory domain, as in (7.1), equals:

$$E_\phi\left[(S_\gamma)_{i,m}\right] = \frac{1}{2}\sum_k \left|\hat{h}\,(k)\,\widehat{(w_s e^{jn\omega_\gamma})}_m\,(k)\right|^2. \tag{7.18}$$

Recall the operator notation $\mathcal{I}_m\{\cdot\}$ defined in (7.3), which denoted the complete transform from the time-domain to one TF-block in the auditory domain. The expected value of the internal representation of a sum of weighted sinusoids weighted is then given by:

$$E_\phi\left[\mathcal{I}_m\left\{\sum_\gamma a_\gamma s_\gamma\right\}\right] = \sum_\gamma |a_\gamma|^2 E_\phi\left[\mathcal{I}_m\{s_\gamma\}\right], \tag{7.19}$$

where α_γ denotes a real and positive weighting function for channel γ and the cross terms between the weighted sinusoids in the auditory domain are zero due to the i.i.d. assumption. This is a direct consequence of taking into account squared magnitudes in (7.1) rather than the squared root of this term. Hence, the weighted sum of sinusoids results in a squared weighted sum of average functions in the auditory domain. Note that a realization of this internal representation is expected to be close to its expected value, since the proposed metric discards all DFT-phase information in (7.1). Motivated by this, each element in the dictionary $\mathcal{D} = \mathbf{g}(\gamma)_{\gamma\in\Gamma}$ is defined as $\mathbf{g}\,(\gamma) = E\left[\mathcal{I}_m\{s_\gamma\}\right]$. The frame index m is taken equal to the last frame which still overlaps with w_s. This means that the dictionary depends on the alignment between w_s and the chosen m. Since the support of w_m (16 ms) is double the support of w_s (8 ms), two possible alignments exist for which the dictionaries, say $\mathcal{D}_1$ and $\mathcal{D}_2$, can be pre-calculated and stored. Two example dictionary elements are shown for both dictionaries in Figure 7.1. This figure also illustrates how m is chosen given w_s by highlighting the windows of the auditory model. The eventual vocoded speech signal for time-frame w_s is then synthesized as[1] $x \approx \sum_p \sqrt{\alpha^{(p)}}\, s_{\gamma^{(p)}}$.

7.5 Experimental Results

The proposed matching pursuit (MP) algorithm is compared with the peak-picking (PP) algorithm which is currently still the basis of several existing coding strategies in CIs [Wils 08]. Signal processing details of the peak-picking algorithm can be found in [Dorm 02].

Three intelligibility predictors are used to assess the intelligibility of MP and PP where the number of selected channels is varied between 1 and 5. These predictors consist of STOI [Taal 11a] (the model which was simplified in Section 7.2), a model developed by Christiansen and Dau (DAU) [Chri 10] and the normalized covariance metric (NCM) [Chen 11]. These measures are recently proposed and can be considered as state-of-the-art for intelligibility

[1] In rare cases it may occur that the optimal α for a specific iteration is negative. Since a negative amplitude in the auditory domain does not have a meaning in the time-domain these channels are discarded.

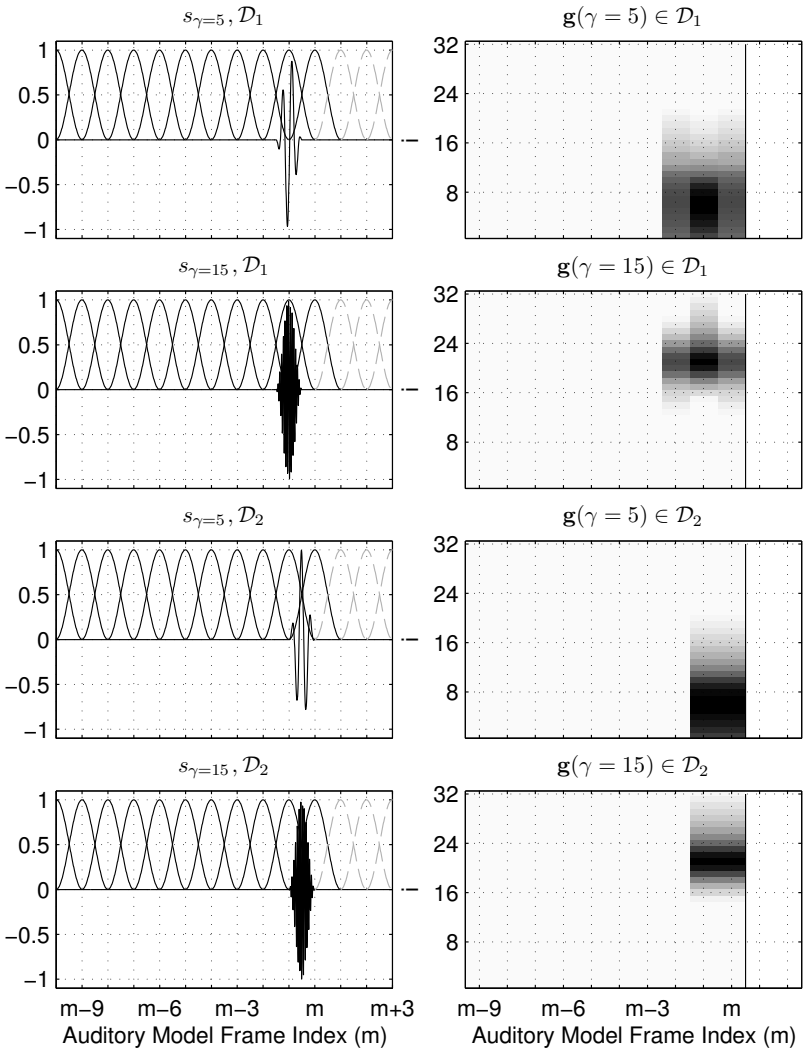

Figure 7.1: Two example elements for each dictionary $\mathcal{D}_1$ and $\mathcal{D}_2$ where $\gamma = \{5, 15\}$. Left plots show realizations of s_γ and right plots the average internal representations.

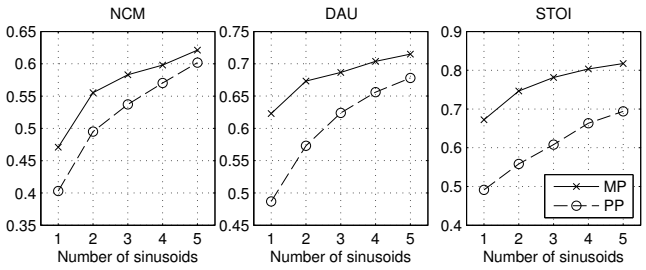

Figure 7.2: Prediction results for proposed matching pursuit (MP) and peak picking (PP) algorithm (a higher score denotes more intelligible speech). The predictors STOI [Taal 11a], DAU [Chri 10] and NCM [Chen 11] are all known to be reliable with vocoded speech.

prediction of vocoded speech. The results are shown in Fig. 7.2 from which we can conclude that all three measures predict that the intelligibility of MP is higher than PP. A result which is in line with informal listening tests. Largest improvements are predicted with STOI, which is not that surprising since this is the measure initially used for optimization. NCM and DAU predict that the speech intelligibility for MP with 1 sinusoid is roughly equal to the intelligibility with PP for 2 and 3 sinusoids, respectively. In the near-future real listening tests will be performed to quantify the absolute difference between MP and PP.

The main differences between MP and PP are illustrated in Figure 7.3, where one TF-block of clean speech is used and only one channel was selected per time instant. For comparison, the clean internal representation is shown together with the internal representations for both methods, denoted by $\mathbf{y_m}$, and their corresponding channel selections. From the plots it is clear that PP tends to select the same channel independently of the previous selected channel. As a result the two formants between 0.1-0.2 seconds and channel 16-24 are completely discarded with PP, which is not the case with MP. There are two important reasons for this different behavior: (1) The proposed metric has a longer integration time such that channels selections from the past are taken into account for the current channel selection. (2) The weighting matrix $\mathbf{A}_m$ 'whitens' the speech and will therefore give a similar importance to high frequencies compared to low frequency content. Another important difference is the fact that the proposed method considers the spread over time and frequency of the sinusoids. Therefore, MP will less often select neighboring channels compared to PP.

Note that the channel stimulation rates in real CI-processors can be much higher than the rate of 250 Hz as used in the vocoder from [Dorm 02]. In a real CI also the channels are typically stimulated sequentially in an interleaved manner, rather than simultaneously, in order to avoid electrical field interactions [Wils 08]. These properties of the CI cannot be included in a vocoder since no acoustical signals exist with such short-time duration and narrow fre-

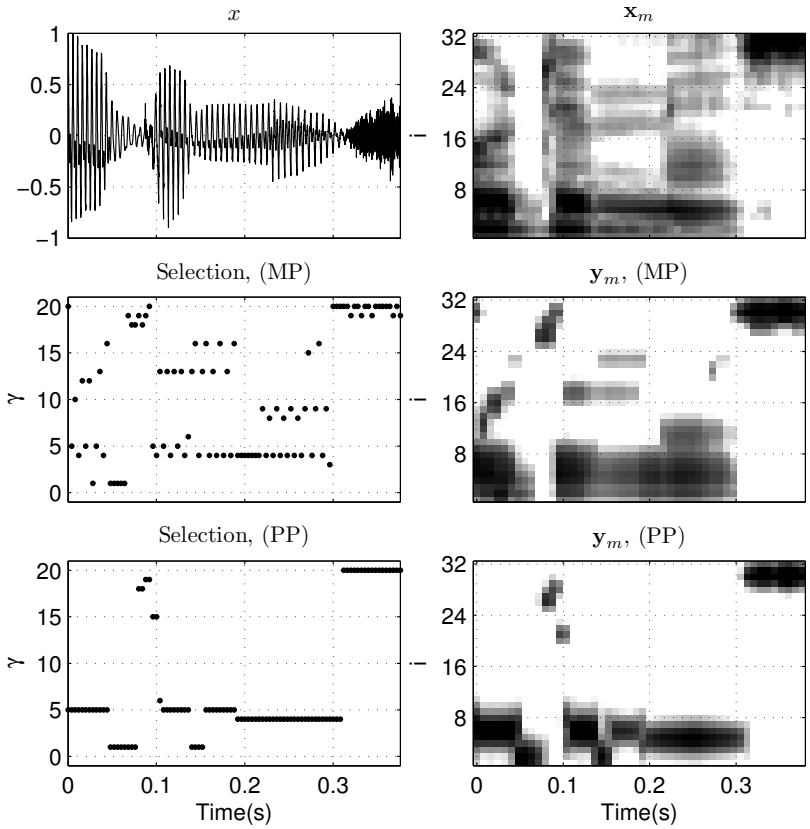

Figure 7.3: Auditory representations of clean and vocoded speech and channel selection for MP and PP. One channel is selected per time-instant for both algorithms.

quency support. It is important to add, however, that these are constraints of the use of any vocoder and not of the proposed channel selection method. Namely, the dictionary can be easily extended to shorter pulse durations in a real CI environment.

7.6 Concluding Remarks

In this paper it is shown that the existing short-time objective intelligibility (STOI) measure can be expressed as a weighted ℓ_2 norm in the auditory domain. Due to the mathematical properties of this norm it facilitated the use of the matching pursuit algorithm in the channel selection technique in cochlear implants (CIs). Acoustic CI simulations are generated based on a sinusoidal vocoder where a large intelligibility improvement was found by three state-of-the-art intelligibility predictors compared to a peak-picking algorithm.

Chapter 8

Discussion and Conclusions

In this thesis the focus was on the development of new machine-driven measures for intelligibility prediction of (non)-linearly processed speech in noisy conditions. An important aspect was the aim for new measures which are mathematical tractable and can therefore be used for online optimization. With online optimization we mean developing new signal processing strategies in an optimal way given such a machine-driven evaluation method, rather than naive offline optimization of free parameters of already designed speech processing algorithms. To this end we successfully proposed several new measures in Chapter 3-5 which show good prediction results in line with subjective listening tests and state-of-the-art objective measures. The newly proposed measures are of low computational complexity and mathematical tractable which make them suitable for online optimization as demonstrated in Chapters 3, 6 and 7.

8.1 Results

In Chapter 3 we analyzed a general procedure of modeling the auditory system, e.g., [Lyon 82, Dau 96a], which is an important aspect of every predictive measure, whether it is for signal detection, audio quality or speech intelligibility prediction. We showed that, under certain assumptions, the model can be greatly simplified when predicting results from psychoacoustic masking experiments. The resulting model facilitates the computation of analytic expressions for masking thresholds and masking curves, while advanced spectro-temporal models, like the Dau-model [Dau 96a], typically need computationally demanding adaptive procedures [Levi 71] to find an estimate of these masking thresholds. We showed that the proposed method gives similar masking predictions as the advanced spectro-temporal Dau-model (with maximum errors around 10 dB), while being a factor 10-100 times faster, depending on the frame length and type of test. An important property of the proposed method compared to existing measures which are suitable for online optimization like the Par-model [Par 05], is its sensitivity to the temporal envelope within short-time frames (20-40 ms). As a consequence, the measure is very sensitive to introduced errors in transient parts of speech which are of great importance in speech intelligibility. From our results we concluded that the proposed model can be interpreted as an extended version of the Par-model [Par 05], which is a mathematical tractable perceptual model based on spectral integration only.

In Chapter 3 the simplified auditory model is used for online optimization where a fixed amount of noise was redistributed over time and frequency. The redistribution is done such that the distortion measure based on the simplified auditory model was minimized, i.e., the audibility of the noise was minimized. This is a typical scenario in the field of audio coding (compression) or data-hiding like audio watermarking. A comparison was made with the Par-model. It can be concluded that for non-stationary frames (e.g., transients) the Par-model underestimates the audibility of introduced errors and therefore overestimates the masking curve. As a consequence, the system of interest incorrectly assumes that errors are masked in a particular frame which may lead

to audible artifacts like pre-echoes. This was not the case with the proposed method which correctly detects the errors made in the temporal structure of the signal.

In Chapter 4 an extensive evaluation is presented of objective measures for intelligibility prediction of noisy speech processed with a time-frequency (TF) varying gain function. Two speech processing techniques are used. In one case binary gain functions are applied in the time-frequency domain with a method called ideal time-frequency segregation (ITFS). In the second case several single-channel noise reduction algorithms are evaluated. Out of all measures, the proposed frame-based measure based on the correlation between the critical-band magnitude spectra (MCC) of the clean and processed speech gave the best results with $\rho = .93$. Good results were also obtained with the Dau-model [Dau 96a] (DAU), namely $\rho = .89$. Poor results were obtained with the coherence speech intelligibility index [Kate 05] (CSII), which turned out to be an unreliable intelligibility predictor for the ITFS-processed signals used in this research. This was probably due to sensitivity to the DFT phase component. We also showed that the correlation predictions between simple predictive measures and speech intelligibility can be improved significantly by applying a normalization procedure independently of the predictive measure. An important conclusion from this evaluation is that the complexity of the auditory model is of minor importance. Some measures based on simple TF-representations performed better than sophisticated nonlinear auditory models. Moreover, two important aspects are found which matter in how the internal representations are compared. (1) Correlation-based comparisons perform better than SNR-based measures or measures based on squared errors. (2) Using a longer temporal integration time than short-time (20-30 ms) frames tends to give better performance.

A new short-time objective intelligibility (STOI) measure is proposed in Chapter 5 which is based on the previously mentioned evaluative study of objective measures in Chapter 4. STOI shows high correlation with the intelligibility of noisy and time-frequency weighted noisy speech (e.g., resulting from noise reduction) of three different listening experiments. In general, STOI showed better correlation with speech intelligibility compared to five other reference objective intelligibility models. Several follow-up studies are published independently of our work where the good performance of the STOI is confirmed [Mowl 12, Schl 10, Cass 11, Xia 12, Gmez 12]. In contrast to other conventional intelligibility models, which tend to rely on global statistics across entire sentences, STOI is based on shorter time segments in the order of a few hundreds of milliseconds. Experiments indeed show that it is beneficial to take segment lengths of this order into account. Moreover, STOI uses a very simple linear DFT-based auditory model which makes the method more suitable for online optimization.

In Chapter 6 the simplified auditory model from Chapter 3 was used for optimizing a speech pre-processing algorithm for speech intelligibility improvement in noise for the near-end listener. The algorithm improved the intelligi-

bility by optimally redistributing the speech energy over time and frequency for the proposed simplified auditory model. Since this auditory model takes into account short-time information, transients will receive more amplification compared to stationary vowels, which is beneficial for improving intelligibility in noise. Note that the proposed distortion measure based on the simplified auditory model from Chapter 3 is based on audibility. Although audibility shares some aspects of intelligibility, it is typically used as a feature for speech quality predictive models, see, e.g., [Quac 88, Loiz 07b]. This could be a reason why our method also improved speech quality in addition to intelligibility. As a consequence, the proposed method may be suboptimal in terms of speech intelligibility optimization. Another important aspect of audibility is the fact that the measure needs the noise and speech in isolation. This makes the method suitable for near-end enhancement but less suitable for intelligibility prediction of single-channel noise reduced speech where speech and noise are not isolated anymore.

We showed that the STOI predictions have high correlation with the intelligibility of vocoded speech, as typically used in acoustic cochlear implant (CI) simulations. In Chapter 7, STOI is therefore used for online optimization in the *n-of-m* channel selection technique as found in several cochlear implant (CI) coding strategies [Seli 95]. With this technique only a subset of frequency channels (electrodes) are stimulated, such that important channels can be updated more frequently and less significant channels are omitted. STOI is further simplified such that it can be expressed as a weighted ℓ_2 norm in the auditory domain. Due to the mathematical properties of a norm, STOI can now be used with the matching pursuit algorithm in the *n-of-m* channel selection technique. Intelligibility predictions with acoustic CI-simulations for normal-hearing listeners indicate that more intelligible speech is obtained with the proposed method compared to a conventional channel-selection method based on peak picking. Reasons for this difference in performance are: (1) STOI considers an analysis window of a few hundreds of milliseconds in order to account for low temporal modulations which are important for speech intelligibility and (2) spectral leakage per channel is accounted for in the mathematical optimization process. It is important to add that these results are based on the validity of the speech vocoder which predicts the results for CI users with normal-hearing users. Although speech vocoders have shown to be valuable in predicting trends of average user results, they do not predict individual results for CI users. However, the basis functions used in the matching pursuit algorithm could be adjusted to include user specific behavior.

In order to simplify STOI we found that the measure could be expressed as a weighted ℓ_2 norm in the auditory domain. Here the weighting function is based on the reciprocal of the total energy within one auditory band of the clean speech. This finding probably explains the good performance of the simple measure based on the magnitude squared difference (MSD) in combination with the proposed critical-band based normalization procedure in Chapter 4. The measure MSD was based on the ℓ_2 norm in the auditory domain where

the proposed critical-band based normalization procedure also normalizes the energy within each auditory band based on the reciprocal of the total energy within one auditory band.

The analysis length of STOI equals a few hundreds of milliseconds which is in line with results of several listening experiments where temporal modulations above 2-3 Hz are important for intelligibility [Drul 94a, Arai 99]. With the current analysis length which is close to 400 milliseconds, STOI is sensitive for temporal modulations of 2.6 Hz and higher which is roughly in accordance with the results of these listening tests. Moreover, the analysis length is also more in line with the maximum temporal integration properties of the auditory system, which is in the order of hundreds of milliseconds, e.g., [Brin 64]. Note that the analysis length of a few hundreds of milliseconds in STOI is an important difference with the simplified auditory model proposed in Chapter 3 which was based on short-time segments (20-40 ms). Therefore, in Chapter 6 where we optimized for this simplified measure an heuristic smoother had to be applied to the gain function over time. This is not necessary when optimizing for STOI.

8.2 Directions of Future Research

Based on our results we have the following recommendations for future research:

Application of Critical-Band Based Normalization Procedure It has been shown that the critical-band based normalization procedure from Chapter 4 improves the correlation with speech intelligibility of simple measures based on squared differences. Many speech processing algorithms in speech communication systems are based on squared differences due to its mathematical tractability. One could simply transform the speech signal with the proposed normalization procedure before transmission in a speech communication system. The magnitude spectrum of errors introduced by the system, e.g., due to quantization in a speech coder, will be shaped in such a way that intelligibility is expected to be less harmed. At the receivers side an inverse filter should be applied to restore the original speech spectrum.

Single-Channel Noise Reduction In the field of single-channel noise reduction, there is typically no or only little improvement in speech intelligibility due to the applied noise reduction algorithm [Jens 12, Hu 07a]. As mentioned in the introduction, many intelligibility measures report the opposite result and predict that the noise reduction algorithm did a good job and actually *increased* the speech intelligibility [Ludv 93, Dubb 08, Gold 04]. We showed that is not the case with STOI [Taal 10c]. Also another study showed excellent results with STOI for speech intelligibility prediction of single-channel noise reduced speech [Xia 12]. It seems logical to derive an optimal noise reduction scheme for STOI (or the weighted ℓ_2 norm based on STOI). However, this may be challenging since STOI is a function of the clean speech signal which is not

available. A typical approach is to assume some type of statistical model where the underlying speech signal must be estimated, see, e.g., [Loiz 07b].

Understanding of Speech Processing Mechanism in the Auditory System In this thesis we showed that speech intelligibility can be successfully predicted with STOI. However, it does not explain the actual underlying mechanism of speech understanding in the auditory system in full detail. For example, the clipping procedure in STOI is a heuristic approach to improve the performance of STOI and may not be directly explained with actual processing going on in the auditory periphery. However, we believe that the good performance of STOI, which is confirmed in several different studies, should be reproducible with a more accurate modeling and motivation of the auditory system. This could be of interest to answer more fundamental questions on how speech is perceived and why, for example, current noise reduction algorithms are not able gain large improvements in speech intelligibility.

Binaural Intelligibility Prediction Model All measures treated in this work are monaural models and are based on the assumption that the left and right ear receive the same speech signal. However, it is well known that human listeners can benefit from using the spacial configuration where the signals are perceived binaurally [Bron 00]. It would be of interest to extend STOI such that it can also handle binaural input, for example, based on interaural cross-correlation [Lyon 83]. We found that using a segment length of a few hundreds of milliseconds is of importance. One could investigate whether this conclusion is also relevant in a binaural intelligibility predictor.

References

[ANSI 97] ANSI. "Methods for calculation of the speech intelligibility index". *S3.5-1997, (American National Standards Institute, New York)*, 1997.

[Arai 99] T. Arai, M. Pavel, H. Hermansky, and C. Avendano. "Syllable intelligibility for temporally filtered LPC cepstral trajectories". *J. Acoust. Soc. Am.*, Vol. 105, No. 5, pp. 2783–2791, 1999.

[Beer 02] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier. "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II-psychoacoustic model". *J. Audio Eng. Soc*, Vol. 50, No. 10, pp. 765–778, 2002.

[Beer 04] J. G. Beerends, E. Larsen, N. Iyer, and J. M. van Vugt. "Measurement of speech intelligibility based on the PESQ approach". In: *Proc. of the Workshop Measurement of Speech and Audio Quality in Networks*, 2004.

[Beer 05] J. G. Beerends, S. van Wijngaarden, and R. van Buuren. "Extension of ITU-T recommendation P.862 PESQ towards measuring speech intelligibility with Vocoders". Tech. Rep., T.N.O., 2005.

[Bold 09] J. B. Boldt and D. P. W. Ellis. "A Simple Correlation-Based Model Of Intelligibility For Nonlinear Speech Enhancement And Separation". In: *Proc. EUSIPCO*, pp. 1849–1853, 2009.

[Brin 64] G. van den Brink. "Detection of Tone Pulse of Various Durations in Noise of Various Bandwidths". *J. Acoust. Soc. Am.*, Vol. 36, No. 6, pp. 1206–1211, 1964.

[Bron 00] A. Bronkhorst. "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". *Acta Acustica united with Acustica*, Vol. 86, No. 1, pp. 117–128, 2000.

[Brou 08] H. Brouckxon, W. Verhelst, and B. Schuymer. "Time and frequency dependent amplification for speech intelligibility enhance-

ment in noisy environments". In: *Proc. Interspeech*, pp. 557–560, 2008.

[Brun 06] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang. "Isolating the Energetic Component of Speech-on-Speech Masking with Ideal Time-Frequency Segregation". *J. Acoust. Soc. Am.*, Vol. 120, No. 6, pp. 4007–4018, 2006.

[Buus 86] S. Buus, E. Schorer, M. Florentine, and E. Zwicker. "Decision rules in detection of simple and complex tones". *J. Acoust. Soc. Am.*, Vol. 80, No. 6, pp. 1646–1657, 1986.

[Cart 73] G. Carter, C. Knapp, and A. Nuttall. "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing". *IEEE Transactions on Audio and Electroacoustics*, Vol. 21, No. 4, pp. 337–344, 1973.

[Cass 11] S. K. Cassia Valentini-Botinhao, Junichi Yamagishi. "Can Objective Measures Predict the Intelligibility of Modified HMM-based Synthetic Speech in Noise?". In: *Proc. Interspeech*, 2011.

[Chen 11] F. Chen and P. Loizou. "Predicting the Intelligibility of Vocoded Speech". *Ear and Hearing*, Vol. 32, No. 3, p. 331, 2011.

[Chri 10] C. Christiansen, M. S. Pedersen, and T. Dau. "Prediction of speech intelligibility based on an auditory preprocessing model". *Speech Communication*, Vol. 52, pp. 678–692, 2010.

[Chun 04] K. Chung. "Challenges and recent developments in hearing aids". *Trends in Amplification*, Vol. 8, No. 3, p. 83, 2004.

[Comm 93] I. Committee. "Coding of Moving Pictures and Associated Audio for Storage at up to about 1.5Mbit/s, part 3: Audio". *ISO/IEC 11172-3*, 1993.

[Dau 96a] T. Dau, D. Püschel, and A. Kohlrausch. "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure". *J. Acoust. Soc. Am.*, Vol. 99, No. 6, pp. 3615–3622, 1996.

[Dau 96b] T. Dau, D. Püschel, and A. Kohlrausch. "A quantitative model of the" effective" signal processing in the auditory system. II. Simulations and measurements". *J. Acoust. Soc. Am.*, Vol. 99, No. 6, pp. 3623–3631, 1996.

[Dell 93a] J. Deller Jr, J. Proakis, and J. Hansen. *Discrete time processing of speech signals*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1993.

[Dell 93b] J. Deller Jr, J. Proakis, and J. Hansen. *Discrete time processing of speech signals*, pp. 580–593. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1993.

[Dill 01] H. Dillon. *Hearing aids*. Thieme Medical Pub, 2001.

[Dorm 02] M. F. Dorman, P. C. Loizou, A. J. Spahr, and E. Maloff. "A Comparison of the Speech Understanding Provided by Acoustic Models of Fixed-Channel and Channel-Picking Signal Processors for Cochlear Implants". *J Speech Lang Hear Res*, Vol. 45, No. 4, pp. 783–788, 2002.

[Drul 94a] R. Drullman, J. Festen, and R. Plomp. "Effect of reducing slow temporal modulations on speech reception". *J. Acoust. Soc. Am.*, Vol. 95, No. 5, pp. 2670–2680, 1994.

[Drul 94b] R. Drullman, J. Festen, and R. Plomp. "Effect of temporal envelope smearing on speech reception". *J. Acoust. Soc. Am.*, Vol. 95, No. 2, pp. 1053–1064, 1994.

[Dubb 08] F. Dubbelboer and T. Houtgast. "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility". *J. Acoust. Soc. Am.*, Vol. 124, No. 6, pp. 3937–3946, 2008.

[Elhi 03] M. Elhilali, T. Chi, and S. Shamma. "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility". *Speech communication*, Vol. 41, No. 2-3, pp. 331–348, 2003.

[Ephr 84] Y. Ephraim and D. Malah. "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator". *IEEE Trans. on Acoust., Speech, Signal Process.*, Vol. 32, No. 6, pp. 1109–1121, 1984.

[Erke 07] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 15, No. 6, pp. 1741–1752, 2007.

[Fast 07] H. Fastl and E. Zwicker. *Psychoacoustics: facts and models*. Springer-Verlag New York Inc, 2007.

[Flet 50] H. Fletcher and R. H. Galt. "The Perception of Speech and Its Relation to Telephony". *J. Acoust. Soc. Am.*, Vol. 22, No. 2, pp. 89–151, 1950.

[Fren 47] N. R. French and J. C. Steinberg. "Factors Governing the Intelligibility of Speech Sounds". *J. Acoust. Soc. Am.*, Vol. 19, No. 1, pp. 90–119, 1947.

[Garo 93] J. Garofolo. "TIMIT: Acoustic-phonetic Continuous Speech Corpus". *National Institute of Standards and Technology (NIST)*, 1993.

[Glas 90] B. R. Glasberg and B. C. Moore. "Derivation of auditory filter shapes from notched-noise data". *Hearing Research*, Vol. 47, No. 12, pp. 103 – 138, 1990.

[Gmez 12] A. M. Gmez, B. Schwerin, and K. Paliwal. "Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio". *Speech Communication*, Vol. 54, No. 3, pp. 503 – 515, 2012.

[Gold 04] R. L. Goldsworthy and J. E. Greenberg. "Analysis of Speech-Based Speech Transmission Index Methods with Implications for Nonlinear Operations". *J. Acoust. Soc. Am.*, Vol. 116, No. 6, pp. 3679–3689, 2004.

[Gord 86] S. Gordon-Salant. "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing". *J. Acoust. Soc. Am.*, Vol. 80, No. 6, pp. 1599–1607, 1986.

[Grad 00] I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series, and Products (Seventh Edition)*. Academic Pr, 2000.

[Gran 08] V. Grancharov and W. B. Kleijn. *Handbook of Speech Processing*, Chap. Speech Quality Assessment, pp. 83–99. Springer, 2008.

[Gray 76] A. H. Gray Jr and J. D. Markel. "Distance measures for speech processing". *IEEE Trans. on Acoust., Speech, Signal Process.*, Vol. 24, No. 5, pp. 380–391, 1976.

[Gree 66] D. Green and J. Swets. *Signal Detection Theory and Psychophysics*. Wiley New York, 1966.

[Grif 68] J. D. Griffiths. "Optimum Linear Filter for Speech Transmission". *J. Acoust. Soc. Am.*, Vol. 43, No. 1, pp. 81–86, 1968.

[Hage 82] B. Hagerman. "Sentences for testing speech intelligibility in noise". *Scandinavian Audiology*, Vol. 11, No. 2, pp. 79–87, 1982.

[Hall 10] J. L. Hall and J. L. Flanagan. "Intelligibility and listener preference of telephone speech in the presence of babble noise". *J. Acoust. Soc. Am.*, Vol. 127, No. 1, pp. 280–285, 2010.

[Hans 97] M. Hansen and B. Kollmeier. "Using a quantitative psychoacoustical signal representation for objective speech quality measurement". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1387–1390, 1997.

[Hans 98a] J. H. L. Hansen and B. L. Pellom. "An effective quality evaluation protocol for speech enhancement algorithms". In: *Proc. Fifth Int. Conference on Spoken Language Processing*, 1998.

[Hans 98b] M. Hansen. *Assessment and prediction of speech transmission quality with an auditory processing model*. PhD thesis, Univ. Oldenburg, 1998.

[Haza 98] V. Hazan and A. Simpson. "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise". *Speech Communication*, Vol. 24, No. 3, pp. 211 – 226, 1998.

[Hend 04] R. C. Hendriks, R. Heusdens, and J. Jensen. "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

[Hend 10] R. C. Hendriks, R. Heusdens, and J. Jensen. "MMSE based noise PSD tracking with low complexity". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4266–4269, 2010.

[Herr 99] J. Herre. "Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction". In: *Audio Engineering Society Convention 17*, pp. 312–325, 1999.

[Heus 02a] R. Heusdens, R. Vafin, and W. Kleijn. "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits". *IEEE Signal Processing Letters*, Vol. 9, No. 8, pp. 262–265, 2002.

[Heus 02b] R. Heusdens and S. Van De Par. "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1809–1812, 2002.

[Heus 06] R. Heusdens, J. Jensen, W. B. Kleijn, V. Kot, O. A. Niamut, S. van der Par, N. H. van Schijndel, and R. Vafin. "Bit-rate scalable intraframe sinusoidal audio coding based on rate-distortion optimization". *J. Audio Eng. Soc*, Vol. 54, No. 3, pp. 167–188, 2006.

[Hilk 12] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale. "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios". *J. Acoust. Soc. Am.*, Vol. 131, No. 1, pp. 531–539, 2012.

[Holm 79] S. Holm. "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp. 65–70, 1979.

[Holu 96] I. Holube and B. Kollmeier. "Speech Intelligibility Prediction in Hearing Impaired Listeners Based on a Psychoacoustically Motivated Perception Model". *J. Acoust. Soc. Am.*, Vol. 100, No. 3, pp. 1703–1716, 1996.

[Hu 07a] Y. Hu and P. C. Loizou. "A comparative intelligibility study of single-microphone noise reduction algorithms". *J. Acoust. Soc. Am.*, Vol. 122, No. 3, pp. 1777–1786, 2007.

[Hu 07b] Y. Hu and P. C. Loizou. "Subjective comparison and evaluation of speech enhancement algorithms". *Speech communication*, Vol. 49, No. 7-8, pp. 588–601, 2007.

[Hu 08a] Y. Hu and P. C. Loizou. "Evaluation of objective quality measures for speech enhancement". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 16, No. 1, pp. 229–238, 2008.

[Hu 08b] Y. Hu and P. C. Loizou. "Techniques for estimating the ideal binary mask". In: *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[Huan 09] D.-Y. Huang, S. Rahardja, and E. P. Ong. "Biologically inspired algorithm for enhancement of speech intelligibility over telephone channel". In: *IEEE International Workshop on Multimedia Signal Processing*, pp. 1–6, 2009.

[Iser 08] B. Iser, W. Minker, and G. Schmidt. *Bandwidth extension of speech signals*. Springer Publishing Company, Incorporated, 2008.

[Itak 70] F. Itakura and S. Saito. "A statistical method for estimation of speech spectral density and formant frequencies". *Electronics and Communications in Japan*, Vol. 53, pp. 36–43, 1970.

[ITU 01] ITU. "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems". *ITU-R BS. 1534-1*, 2001.

[ITU 03] ITU. "One-Way Transmission Time". *ITU-T Recommendation G.114*, 2003.

[ITU 05] ITU. "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs". *ITU-T Recommendation P.862.2*, 2005.

[Jabl 04] F. Jabloun and B. Champagne. "Incorporating the human hearing properties in the signal subspace approach for speech enhancement". *IEEE Trans. on Speech and Audio Process.*, Vol. 11, No. 6, pp. 700–708, 2004.

[Jaya 08] A. Jayan, P. Pandey, and P. Lehana. "Automated detection of transition segments for intensity and time-scale modification for speech intelligibility enhancement". In: *IEEE International Conference on Signal Processing, Communications and Networking*, pp. 63–68, 2008.

[Jens 12] J. Jensen and R. C. Hendriks. "Spectral Magnitude Minimum Mean-Square Error Estimation Using Binary and Continuous Gain Functions". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 20, pp. 92 – 102, 2012.

[Kate 05] J. M. Kates and K. H. Arehart. "Coherence and the Speech Intelligibility Index". *J. Acoust. Soc. Am.*, Vol. 117, No. 4, pp. 2224–2237, 2005.

[Kate 92] J. Kates. "On using coherence to measure distortion in hearing aids". *J. Acoust. Soc. Am.*, Vol. 91, p. 2236, 1992.

[Kenn 98] E. Kennedy, H. Levitt, A. C. Neuman, and M. Weiss. "Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners". *J. Acoust. Soc. Am.*, Vol. 103, No. 2, pp. 1098–1114, 1998.

[Kim 09] G. Kim, Y. Lu, Y. Hu, and P. Loizou. "An algorithm that improves speech intelligibility in noise for normal-hearing listeners". *J. Acoust. Soc. Am.*, Vol. 126, p. 1486, 2009.

[Kita 07] N. Kitawaki and T. Yamada. "Subjective and objective quality assessment for noise reduced speech". In: *Proc. ETSI Workshop on Speech and Noise in Wideband Communication*, pp. 1–4, 2007.

[Kjem 09] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang. "Role of Mask Pattern in Intelligibility of Ideal Binary-Masked Noisy Speech". *J. Acoust. Soc. Am.*, Vol. 126, No. 3, pp. 1415–1426, 2009.

[Klat 82] D. Klatt. "Prediction of perceived phonetic distance from critical-band spectra: A first step". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1982.

[Koch 92] R. Koch. *Auditory sound analysis for the prediction and improvement of speech intelligibility (in German).* PhD thesis, Universität Göttingen, 1992.

[Kohl 08] A. Kohlrausch, J. Koppens, W. Oomen, and S. van de Par. "A New Perceptual Model for Audio Coding Based on Spectro-Temporal Masking". In: *Audio Engineering Society Convention 124*, 5 2008.

[Koll 08] B. Kollmeier, T. Brand, and B. Meyer. *Handbook of Speech Processing*, Chap. 4. Perception of Speech and Sound, pp. 61–82. Springer, 2008.

[Koni 10] C. Koniaris, M. Kuropatwinski, and W. B. Kleijn. "Auditory-model based robust feature selection for speech recognition". *The Journal of the Acoustical Society of America*, Vol. 127, No. 2, pp. EL73–EL79, 2010.

[Koop 07] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure. "Development of a speech in noise test (Matrix)". In: *8th EFAS Congress, 10th DGA Congress*, Heidelberg, Germany, June 2007.

[Kryt 62] K. D. Kryter. "Methods for the Calculation and Use of the Articulation Index". *J. Acoust. Soc. Am.*, Vol. 34, No. 11, pp. 1689–1697, 1962.

[Lang 92] A. Langhans and A. Kohlrausch. "Spectral integration of broadband signals in diotic and dichotic masking experiments". *J. Acoust. Soc. Am.*, Vol. 91, pp. 317–326, 1992.

[Levi 71] H. Levitt. "Transformed up-down methods in psychoacoustics". *J. Acoust. Soc. Am.*, Vol. 49, No. 2, pp. 467–477, 1971.

[Li 08] N. Li and P. C. Loizou. "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction". *J. Acoust. Soc. Am.*, Vol. 123, p. 1673, 2008.

[Li 10] F. Li, A. Menon, and J. B. Allen. "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech". *J. Acoust. Soc. Am.*, Vol. 127, No. 4, pp. 2599–2610, 2010.

[Li 11] F. Li and J. Allen. "Manipulation of consonants in natural speech". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 19, No. 3, pp. 496–504, 2011.

[Litv 07] L. M. Litvak, A. J. Spahr, A. A. Saoji, and G. Y. Fridman. "Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners". *J. Acoust. Soc. Am.*, Vol. 122, No. 2, pp. 982–991, 2007.

[Liu 08] W. M. Liu, K. A. Jellyman, N. W. D. Evans, and J. S. D. Mason. "Assessment of Objective Quality Measures for Speech Intelligibility". In: *Proc. Interspeech*, pp. 699–702, 2008.

[Liu 97] L. Liu, J. He, and G. Palm. "Effects of phase on the perception of intervocalic stop consonants". *Speech Communication*, Vol. 22, No. 4, pp. 403 – 417, 1997.

[Loiz 06] P. Loizou. "Speech processing in vocoder-centric cochlear implants". *Advances in otorhinolaryngology*, Vol. 64, No. R, p. 109, 2006.

[Loiz 07a] P. C. Loizou. *Speech enhancement: theory and practice*, pp. 502–527. CRC, Boca Raton, FL, 2007.

[Loiz 07b] P. C. Loizou. *Speech enhancement: theory and practice*. CRC, Boca Raton, FL, 2007.

[Loiz 98] P. Loizou. "Mimicking the human ear". *IEEE Signal Processing Magazine*, Vol. 15, No. 5, pp. 101 –130, 1998.

[Ludv 93] C. Ludvigsen, C. Elberling, and G. Keidser. "Evaluation of a noise reduction method - comparison between observed scores and scores predicted from STI". *Scandinavian Audiology. Supplement.*, Vol. 38, pp. 50–55, 1993.

[Lyon 82] R. Lyon. "A computational model of filtering, detection, and compression in the cochlea". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1282 – 1285, 1982.

[Lyon 83] R. Lyon. "A computational model of binaural localization and separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1148–1151, 1983.

[Ma 09] J. Ma, Y. Hu, and P. Loizou. "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions". *J. Acoust. Soc. Am.*, Vol. 125, No. 5, pp. 3387–3405, 2009.

[Mall 93] S. Mallat and Z. Zhang. "Matching pursuits with time-frequency dictionaries". *IEEE Transactions on Signal Processing*, Vol. 41, No. 12, pp. 3397–3415, 1993.

[Mart 01] R. Martin. "Noise power spectral density estimation based on optimal smoothing and minimum statistics". *IEEE Trans. on Speech and Audio Process.*, Vol. 9, No. 5, pp. 504–512, 2001.

[Mill 50] G. A. Miller and J. C. R. Licklider. "The Intelligibility of Interrupted Speech". *The Journal of the Acoustical Society of America*, Vol. 22, No. 2, pp. 167–173, 1950.

[Moor 03] B. Moore. *An introduction to the psychology of hearing*. Emerald Group Pub Ltd, 2003.

[Moor 83] B. Moore and B. Glasberg. "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". *J. Acoust. Soc. Am.*, Vol. 74, pp. 750–753, 1983.

[Mowl 12] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin. "Subjective and Objective Quality Assessment of Single-Channel Speech Separation Algorithms". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 69–72, 2012.

[Nied 76] R. Niederjohn and J. Grotelueschen. "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression". *IEEE Trans. on Acoust., Speech, Signal Process.*, Vol. 24, No. 4, pp. 277 – 282, 1976.

[Nogu 05] W. Nogueira, A. Büchner, T. Lenarz, and B. Edler. "A psychoacoustic NofM-type speech coding strategy for cochlear implants". *EURASIP Journal on Applied Signal Processing*, Vol. 2005, pp. 3044–3059, 2005.

[Pain 00] T. Painter and A. Spanias. "Perceptual coding of digital audio". *Proc. of the IEEE*, Vol. 88, No. 4, pp. 451–515, 2000.

[Pali 03] K. K. Paliwal and L. Alsteris. "Usefulness of phase spectrum in human speech perception". In: *Proc. Interspeech*, pp. 2117–2120, 2003.

[Pan 95] D. Pan. "A tutorial on MPEG/audio compression". *IEEE Multimedia*, Vol. 2, No. 2, pp. 60–74, 1995.

[Par 05] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen. "A perceptual model for sinusoidal audio coding based on spectral integration". *EURASIP J. on Appl. Signal Processing*, Vol. 2005, No. 9, pp. 1292–1304, 2005.

[Patt 92] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. "Complex sounds and auditory images". *Auditory physiology and perception - proceedings of the 9th Int. Symposium on Hearing*, Vol. 83, pp. 429–446, 1992.

[Perk 98] C. Perkins, O. Hodson, and V. Hardman. "A survey of packet loss recovery techniques for streaming audio". *Network, IEEE*, Vol. 12, No. 5, pp. 40–48, 1998.

[Plas 07] J. H. Plasberg and W. B. Kleijn. "The sensitivity matrix: Using advanced auditory models in speech and audio processing". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 15, No. 1, pp. 310–319, 2007.

[Prem 95] J. Preminger and D. Tasell. "Quantifying the relation between speech quality and speech intelligibility". *Journal of Speech, Language, and Hearing Research*, Vol. 38, No. 3, p. 714, 1995.

[Quac 88] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality*, pp. 1–377. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[Rheb 05] K. S. Rhebergen and N. J. Versfeld. "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.*, Vol. 117, No. 4, pp. 2181–2192, 2005.

[Rheb 09] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler. "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise". *J. Acoust. Soc. Am.*, Vol. 126, No. 6, pp. 3236–3245, 2009.

[Rix 02] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends. "Perceptual evaluation of speech quality (PESQ): the new ITU standard for end-to-end speech quality assessment part I-time-delay compensation". *J. Audio Eng. Soc*, Vol. 50, No. 10, pp. 755–764, 2002.

[Saue 06] B. Sauert, G. Enzner, and P. Vary. "Near end listening enhancement with strict loudspeaker output power constraining". In: *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[Saue 10] B. Sauert and P. Vary. "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations". In: *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2010.

[Schl 10] A. Schlesinger and M. M. Boone. "The characterization of the relative information content by spectral features for the objective intelligibility assessment of nonlinearly processed speech". In: *Proc. Interspeech*, pp. 1309–1312, 2010.

[Seli 95] P. Seligman, H. McDermott, *et al.* "Architecture of the Spectra 22 speech processor". *Annals of Otology, Rhinology and Laryngology*, Vol. 104, No. suppl 166, pp. 139–141, 1995.

[Shan 95] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. "Speech recognition with primarily temporal cues". *Science*, Vol. 270, No. 5234, p. 303, 1995.

[Shes 04] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition.* Chapman & Hall/CRC, Boca Raton, FL, 2004.

[Shin 07] J. Shin and N. Kim. "Perceptual reinforcement of speech signal based on partial specific loudness". *IEEE Signal Processing Letters*, Vol. 14, No. 11, pp. 887–890, 2007.

[Skow 06] M. Skowronski and J. Harris. "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments". *Speech Communication*, Vol. 48, No. 5, pp. 549–558, 2006.

[Stee 80] H. J. M. Steeneken and T. Houtgast. "A Physical Method for Measuring Speech-Transmission Quality". *J. Acoust. Soc. Am.*, Vol. 67, No. 1, pp. 318–326, 1980.

[Stra 83] W. Strange, J. Jenkins, and T. Johnson. "Dynamic specification of coarticulated vowels". *J. Acoust. Soc. Am.*, Vol. 74, No. 3, pp. 695–705, 1983.

[Swan 98] M. Swanson, B. Zhu, A. Tewfik, and L. Boney. "Robust audio watermarking using perceptual masking". *Signal Processing*, Vol. 66, No. 3, pp. 337–355, 1998.

[Taal 09a] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems. "An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction". In: *Proc. Interspeech*, pp. 1947–1950, 2009.

[Taal 09b] C. H. Taal and R. Heusdens. "A low-complexity spectro-temporal based perceptual model". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 153–156, 2009.

[Taal 10a] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214 – 4217, 2010.

[Taal 10b] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "Intelligibility Prediction of Single-Channel Noise-Reduced Speech". In: *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, 2010.

[Taal 10c] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "On Predicting the Difference in Intelligibility Before and After Single-Channel Noise Reduction". In: *Int. Workshop Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.

[Taal 11a] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 19, No. 7, pp. 2125–2136, 2011.

[Taal 11b] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech". *J. Acoust. Soc. Am.*, Vol. 130, No. 5, pp. 3013–3027, 2011.

[Taal 12a] C. H. Taal, R. C. Hendriks, and R. Heusdens. "A Low-complexity Spectro-Temporal Distortion Measure for Audio Processing Applications". *IEEE Trans. Audio Speech Lang. Process.*, Vol. 20, No. 5, pp. 1553 – 1564, 2012.

[Taal 12b] C. H. Taal, R. C. Hendriks, and R. Heusdens. "Matching Pursuit for Channel Selection in Cochlear Implants Based on an Intelligibility Metric". In: *Proc. EUSIPCO*, pp. 504 – 508, 2012.

[Taal 12c] C. H. Taal, R. C. Hendriks, and R. Heusdens. "A Speech Preprocessing Strategy For Intelligibility Improvement In Noise Based On A Perceptual Distortion Measure". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4061 – 4064, 2012.

[Taal 12d] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "A Speech Preprocessing Strategy For Intelligibility Improvement In Noise (in review)". *Computer Speech and Language*, 2012.

[Tang 10] Y. Tang and M. Cooke. "Energy reallocation strategies for speech enhancement in known noise conditions". In: *Proc. Interspeech*, pp. 1636–1639, 2010.

[Tang 11] Y. Tang and M. Cooke. "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints". In: *Proc. Interspeech*, pp. 345–348, 2011.

[Thie 00] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, *et al.* "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality". *J. Audio Eng. Soc*, Vol. 48, pp. 3–29, 2000.

[Trib 78] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere. "A study of complexity and quality of speech waveform coders". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 586–590, 1978.

[Vafi 01] R. Vafin, R. Heusden, and W. Kleijn. "Modifying transients for efficient coding of audio". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3285–3288, 2001.

[Vary 06] P. Vary and R. Martin. *Digital speech transmission: enhancement, coding and error concealment.* Wiley, 2006.

[Wage 03] K. Wagener, J. L. Josvassen, and R. Ardenkjaer. "Design, optimization and evaluation of a Danish sentence test in noise". *Int. J. of Audiology*, Vol. 42, No. 1, pp. 10–17, 2003.

[Wang 05] D. Wang. “On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis”. In: P. Divenyi, Ed., *Speech Separation by Humans and Machines*, pp. 181–197, Springer US, 2005.

[Wijn 02] S. van Wijngaarden, H. Steeneken, and T. Houtgast. “Quantifying the intelligibility of speech in noise for non-native talkers”. *J. Acoust. Soc. Am.*, Vol. 112, pp. 3004–3013, 2002.

[Wils 08] B. Wilson and M. Dorman. “Cochlear implants: a remarkable past and a brilliant future”. *Hearing research*, Vol. 242, No. 1-2, pp. 3–21, 2008.

[Xia 12] R. Xia, J. Li, M. Akagi, and Y. Yan. “Evaluation of Objective Intelligiblity Prediction Measures for Noise-Reduced Signals in Mandarin”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[Yama 06] T. Yamada, M. Kumakura, and N. Kitawaki. “Word intelligibility estimation of noise-reduced speech”. In: *Proc. Interspeech*, pp. 169–172, ISCA, 2006.

[Yoo 07] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman. “Speech signal modification to increase intelligibility in noisy environments”. *J. Acoust. Soc. Am.*, Vol. 122, No. 2, pp. 1138–1149, 2007.

[Zwic 80] E. Zwicker and E. Terhardt. “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency”. *J. Acoust. Soc. Am.*, Vol. 68, pp. 1523–1525, 1980.

[Zwic 82] E. Zwicker and A. Jaroszewski. “Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels”. *J. Acoust. Soc. Am.*, Vol. 71, No. 6, pp. 1508–1512, 1982.

[Zwic 90] E. Zwicker and H. Fastl. *Psychoacoustics. Facts and Models.* New York: Springer-Verlag, 1990.

Samenvatting

In digitale spraak-communicatie systemen zoals mobiele telefoons, publieke omroepsystemen en gehoorapparaten is het overbrengen van de boodschap één van de belangrijkste doelen. Helaas kan de verstaanbaarheid van de spraak aangetast worden voor, tijdens en na het verzenden van de boodschap van de zender naar de ontvanger. Belangrijke oorzaken hiervoor zijn bijvoorbeeld achtergrondruis, een slechte internet verbinding tijdens een Skype gesprek of een gehoorbeschadiging van de ontvanger van de boodschap. Om hiervoor te compenseren bevatten veel spraak-communicatie systemen signaalverwerkingsalgoritmes die proberen de verstaanbaarheid van de boodschap te herstellen. Om het effect van een dergelijk algoritme op de spraakverstaanbaarheid te bepalen is het gebruikelijk om een luistertest af te nemen met een groep gebruikers. Het nadeel van deze luistertesten is echter dat ze veel tijd kosten en daarvoor kostbaar zijn. Als een alternatief, kan een computer algoritme gebruikt worden om de resultaten te voorspellen van een echte luistertest. Op deze manier kan het ontwikkelingsproces van nieuwe signaalverwerkingsalgoritmes aanzienlijk versneld worden.

Veel van de huidige maten die de verstaanbaarheid voorspellen van verstoorde spraak kennen twee belangrijke nadelen: (1) Ze zijn niet accuraat in het voorspellen van het effect van geavanceerde niet-lineaire signaalverwerkingsalgoritmes en (2) ze zijn veelal gebaseerd op ingewikkelde, rekenintensieve modellen van het auditieve systeem. Deze twee aspecten maken het moeilijk om nieuwe signaalverwerkingsalgoritmes te ontwikkelen welke wiskundig optimaal zijn voor een gegeven verstaanbaarheidsmaat. In deze thesis introduceren we daarom een aantal nieuwe maten welke succesvol de verstaanbaarheid kunnen voorspellen van verschillende niet-lineaire signaalverwerkingsalgoritmes. Deze nieuwe maten vergen weinig computer rekenkracht en zijn op een wiskundig handelbare manier uitgedrukt. Als gevolg hiervan zijn deze nieuwe methoden zeer geschikt voor het afleiden van nieuwe signaalverwerkingsoplossingen welke zich richten op de verbetering van spraakverstaanbaarheid.

Een belangrijk onderdeel in veel spraakverstaanbaarheidsmaten is het auditieve model waarbij verschillende onderdelen van het gehoor worden gesimuleerd. In het eerste deel van deze thesis wordt daarom een algemeen complex auditief model vereenvoudigd met behoud van goede voorspellingen van psychoakoestische luisterexperimenten. Door deze vereenvoudiging is het mogelijk

om maskeerdrempels analytisch uit te drukken terwijl *state-of-the-art* modellen vaak rekenintensieve adaptieve procedures nodig hebben voor het vinden van een maskeer drempel. De wiskundige eigenschappen van het vereenvoudigde model worden succesvol toegepast bij het verbeteren van spraakverstaanbaarheid in ruis. Dit wordt bereikt door de energie van het spraaksignaal her te verdelen over tijd en frequentie, optimaal voor de gegeven maat zonder verandering van de signaal-in-ruis verhouding.

Een uitgebreide evaluatie heeft plaatsgevonden van 17 verschillende maten voor de verstaanbaarheidsvoorspelling van tijd-frequentie gewogen ruizige spraak. Een voorbeeld hiervan is spraak welke bewerkt is met een ruisonderdrukkingsalgoritme. We laten zien dat, ondanks hoge correlatie, verschillende maten niet geschikt zijn voor het voorspellen van het effect van signaalverwerkingsalgoritmes op de spraakverstaanbaarheid. Daarnaast wordt aangetoond dat een *state-of-the-art* methode niet geschikt is voor het voorspellen van bepaalde tijd-frequentie gewogen ruizige spraaksignalen. Een mogelijke verklaring hiervoor is de gevoeligheid van de maat voor fase informatie. Problemen met huidige maten worden uitgelicht en een nieuwe normalisatie procedure is ontwikkeld die toegepast kan worden als een pre-processing stap om de prestaties van bestaande maten te verbeteren.

We presenteren een nieuwe spraakverstaanbaarheidsmaat gebaseerd op de analyse van korte tijdssegmenten genaamd STOI (*short-time objective intelligibility measure*). De voorspellingen van STOI hebben hoge correlatie met de spraakverstaanbaarheid van tijd-frequentie gewogen ruizige spraak inclusief ruisonderdrukte spraak en spraaksignalen afkomstig van een vocoder. Over het algemeen geven de STOI voorspellingen een hogere correlatie met de spraakverstaanbaarheid vergeleken met vijf andere *state-of-the-art* spraakverstaanbaarheidsmaten. Een belangrijk verschil tussen STOI en andere maten is de signaal analyse lengte welke in de orde is van een aantal honderden milliseconden in plaats van complete zinnen of 20-30 milliseconden wat vaak het geval is bij bestaande methodes. Door de simpele vorm van STOI laten we in het einde van deze thesis zien dat de maat uitgedrukt kan worden in een wiskundige norm. Deze norm is succesvol toegepast in de kanaal-selectie techniek in cochleaire implantaten door middel van simulaties met normaalhorenden. Verschillende verstaanbaarheidsmaten laten een grote verbetering zien met de op STOI-gebaseerde techniek vergeleken met een veel gebruikte peak-picking techniek.

Acknowledgements

Many people have contributed in the process of finishing this PhD thesis. I would like to acknowledge a few of them in particular.

A lot of appreciation goes out to both Richard's, that is Richard Heusdens and Richard Hendriks, who supervised me during my PhD-journey in Delft. Without their guidance and encouragement this thesis would never have been finished. In addition, I would like to thank Inald Lagendijk for being my promotor and his helpful comments and suggestions on the thesis.

I owe a lot of gratitude to my fellow PhD-student Jorge Martinez for sharing the office with me for four years. Next to all our interesting work-related discussions, we had a lot of fun. The visit of Margriet and me to your wedding in Mexico was an amazing experience.

Thanks go out to all the other members of the Signal and Information Processing Lab. More specifically, I would like to mention Jan Erkelens, Christian van Bijleveld, Yuan Zeng and Guoqiang Zhang. Many of the research described in this thesis has been presented to and discussed with these particular colleagues which was very helpful. I had many other inspiring colleagues within the department of Mediamatics who made my stay in Delft a memorable one. These include Zeki Erkin, Ahmed Mahfouz, Jan Bot, Alessandro Ibba, Sepideh Babaei, Behnaz Pourebrahimi, David Tax, Marco Loog, Stevan Rudinac and many others.

I also would like to thank my colleagues at Oticon A/S, Smorum, Denmark for all the fruitful discussions we had during our meetings. This includes Ulrik Kjems, Jan-Mark de Haan and Meng Guo. I especially would like to thank Jesper Jensen for his help during my PhD. Additional credits go out to the 'Jensen's Inn', Ballerup, Denmark for its great service and hospitality.

I would like to thank all my friends in Utrecht including Hubert, Martijn, Ramon, Roald, Roelof and Thomas. It is always a relief to make music and hang out with you guys and forget about all the trivial work-related issues.

Many admiration goes out to my parents and my brother Eric. I am extremely thankful for your ongoing interest, support and encouragement during my PhD. I am happy to be part of such a close family.

Most of my gratitude goes out to Margriet. We both experienced that the PhD-ride can be a bumpy one sometimes. I am therefore very grateful for your love and support.

Curriculum Vitae

Cornelis (Cees) Harm Taal was born was in Hoogeveen, the Netherlands, on June 22^{nd}, 1981. He obtained his havo-diploma from O.R.S. Lek en Linge, Culemborg in 1999. Subsequently, Cees studied at the Utrecht School of Arts, Hilversum and received the Bachelor of Art and Technology in Audio Design (Cum Laude) and the M.A. degree in European Media in 2004. In 2007 he obtained the M.Sc. degree in Media and Knowledge Engineering from the Delft University of Technology (DUT), Delft. From 2008 to 2012, he was a Ph.D. Researcher in the Multimedia Signal Processing Group, DUT, under the supervision of Richard Heusdens and Richard Hendriks in collaboration with Oticon A/S hearing aids.

In the beginning of 2012 he started as a Postdoctoral Researcher in the field of audiology and digital signal processing at the Sound and Image Processing Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden headed by Arne Leijon. Currently, he is working as a scientific researcher in the field of cochlear implants at the Leiden University Medical Center, Leiden, in collaboration with Advanced Bionics.