Graz University of Technology

Master Thesis

Phase-Aware Performance Evaluation

PAPE

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

Phase Processing Laboratory

by
Andreas Gaich, 0673057

Supervisor:
Mowlaee Beikzadehmahaleh Pejman, Ph.D.

Assessors/Examiners:
Franz Pernkopf, Assoc.-Prof. Dipl.-Ing. Dr.mont.
Matthias Frank, Dipl.-Ing. Ph.D.
Mowlaee Beikzadehmahaleh Pejman, Ph.D.

Graz, August 30, 2015

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, 28.08.2015
date

(signature)

Abstract

The objective assessment of the overall speech quality of a given speech enhancement system is a multi-disciplinary optimization problem where different characteristics as perceived sound quality and speech intelligibility are involved. Although previous studies reported instrumental measures with reasonable correlation to subjective listening tests, the studied distortion measures rely on a computation based on the spectral amplitude only. Based on recent findings on improvement of perceived quality and speech intelligibility by also incorporating spectral phase enhancement in the modification stage of a single-channel speech enhancement algorithm, listening tests were conducted to study the performance of well-known instrumental measures in terms of predicting the subjective scores in a phase-aware framework. Furthermore, new phase-aware measures were proposed to assess the perceived speech quality and intelligibility in order to quantify how well the estimated speech phase spectrum resembles the phase spectrum of the reference signal. By performing a correlation analysis between the objective and subjective scores the phase-aware measures showed the capability to outperform the existing conventional measures in the field of perceived speech quality estimation and revealed a reasonable high correlation related to speech intelligibility prediction.

Kurzfassung

Die objektive Beurteilung der Qualität eines bestimmten Sprachverbesserungssystems ist ein multidisziplinäres Optimierungsproblem in dem verschiedene Eigenschaften, wie wahrgenommene Sprachqualität und Sprachverständlichkeit, beteiligt sind. Obwohl frühere Studien gezeigt haben, dass es objektive Bewertungskriterien gibt, die gut mit den Ergebnissen aus subjektiven Hörtests übereinstimmen, ist zu beachten, dass diese Bewertungskriterien nur einen Vergleich zwischen den spektralen Amplituden der sprachverbesserten Signale mit den Referenzsignalen durchführen. Basierend auf neusten Erkenntnissen zur Verbesserung der Sprachqualität und Sprachverständlichkeit durch zusätzliche Modifikation der spektralen Phase, wurden Hörtests durchgeführt, um die Leistungsfähigkeit von etablierten Bewertungskriterien innerhalb eines phasen-basierenden Testframework in Bezug auf die Vorhersage der subjektiven Ergebnisse zu ermitteln. Außerdem wurden neue Methoden vorgestellt, die ein objektives Bewertungsmaß auf Grundlage der Unterschiede im Phasenspektrum zwischen den sprachverbesserten Signalen und den Referenzsignalen berechnen. Anhand der Durchführung einer Korrelationsanalyse zwischen den objektiven und subjektiven Ergebnissen konnte gezeigt werden, dass die neuen phasen-basierenden Bewertungsmethoden die subjektiven Einschätzungen in Bezug auf die Sprachqualität besser prädizieren als die bestehenden herkömmlichen. Obwohl die neuen Methoden auch eine angemessen hohe Korrelation in Bezug auf die Sprachverständlichkeit aufweisen, konnte eine Verbesserung im Vergleich zu den bestehenden Bewertungskriterien nicht evaluiert werden.

Contents

1 Introduction

Desired speech signals are often corrupted with some background noise where the recording takes place. This emerges the requirement of a single-channel speech enhancement pre-processor for different speech applications, to name a few: robust automatic speech recognition and speech transmission. The problem has extensively been addressed during the last two decades with some satisfactory performance where the focus in signal modification has been on the enhancement of the noisy spectral amplitude. The main reason for the focus on spectral amplitude modification rather than phase is due to the belief that amplitude enhancement contributes more to the improvement of the perceptual quality of speech in noise. While many proposals are dedicated to find an accurate spectral amplitude estimator, the potential of phase spectrum estimation has often been neglected. For example, early studies reported on unimportance of the speech phase spectrum [1]. More recent studies, on the other hand, support the fact that incorporating phase information leads to improved signal quality in speech enhancement [2–6], source separation [7–10], automatic speech recognition [11–13], speech synthesis, [14] and speech intelligibility [15, 16].

The issue of quality estimation of the output of a speech enhancement algorithm is highly important and many previous studies have been dedicated to find a reliable estimator of quality or intelligibility. In particular, reliable estimators are necessary in order to avoid the need of performing the time consuming listening experiments. State-of-the-art single-channel speech enhancement algorithms employ an amplitude estimator (filter) working in the magnitude domain followed by a signal reconstruction stage where the noisy phase is often directly exploited. For magnitude-only techniques, the commonly used metrics are l_2-norm measures including signal-to-noise ratio (SNR)-based measures [17] and perceptual motivated measures, e.g., perceptual evaluation of speech quality (PESQ) [18] where the human perception is taken into account. PESQ was shown to be a reliable estimator for perceived quality for a wide range of different distortions while the short time objective speech intelligibility (STOI) [19] was reported to correlate well with speech intelligibility. On the other hand, if we replace the noisy phase with the clean phase (ideal scenario), the current metrics would not well represent the amount of improvement as they only reflect the similarity in the magnitude spectrum domain, not the complex domain. Therefore, the possible improvement obtained via enhancement of the noisy phase spectrum will not be reflected by the current existing metrics like SNR-based measures.

While some early studies emphasized on the non-usefulness of the phase information in speech enhancement [1], more recent studies reveal that incorporating the phase information in amplitude estimation and signal reconstruction potentially leads to considerable improvement in the enhanced output signal. For example, in [20, 21] it was shown that by replacing the mixture

phase with an estimated phase, it is possible to achieve considerable improvement in the perceived signal quality.

This thesis presents a detailed analysis on the performance evaluation using different conventional instrumental measures for perceived quality and speech intelligibility by comparing them with the subjective listening results in a phase-aware single-channel speech enhancement framework. Furthermore, new phase-aware instrumental measures are introduced and evaluated by their ability to predict the perceived quality and speech intelligibility of phase-enhanced speech signals.

1.1 Phase-Aware Speech Enhancement

A single-channel speech enhancement system usually follows a three step computation known as *analysis-modification-synthesis* shown in Figure 1.1. In the analysis step a noisy speech signal is transformed to a representation that is favourable to be modified. A common operator is the Short Time Fourier Transform (STFT) that converts a time-domain signal into a time-frequency representation. The modification step then tries to find and eliminate the noise components. In the last step the modified signal is synthesized back to its time domain representation to obtain the enhanced (noise suppressed) speech signal.

Figure 1.1: Block diagram of a typical speech enhancement system following the analysis-modification-synthesis procedure.

Considering recent advances in speech enhancement, the above presented analysis-modification-synthesis procedure is split into three different approaches in matters of how the phase information is employed during the modification and synthesis stage. Typical single-channel speech enhancement systems (in the following termed as conventional) apply a frequency-dependent gain function on the noisy STFT representation and employ the noisy phase at signal reconstruction stage, illustrated on the top of Figure 1.2. The gain function is computed based on the information given by a noise estimator, i.e., [22] followed by a decision-directed a priori SNR estimator. Two popular examples are the short-time spectral amplitude (STSA) [23] and the log-spectral amplitude (LSA) [24] estimator proposed by Ephraim and Malah. The noisy phase was derived as the MMSE-optimal clean phase, assuming that the DFT bins are uniformly distributed and independent of the amplitude spectrum. However, this assumption is not correct. As reported in [25] the group delay deviation spectrum, which is computationally based on the phase spectrum, follows the spectral amplitude behaviour, hence there has to be some correlation between the amplitude and phase spectra.

This observation gives rise to use an enhanced phase in single-channel speech enhancement. The block diagram is shown in the middle of Figure 1.2. A phase estimation is done on top of the conventional amplitude estimation. The enhanced phase is then employed at the signal reconstruction stage. This setup was already used in [21] to enhance perceived speech quality and in [20] to enhance both, perceived speech quality and speech intelligibility in comparison to conventional methods. In the following this is referred as the phase-enhanced setup.

The block diagram at the bottom of Figure 1.2 suggests to use the enhanced phase information not only at signal reconstruction, but also as an input to compute a more reliable spectral amplitude estimate. This leads to even better speech quality results reported in [2]. Furthermore the setup can also be used in an iterative way such that the enhanced amplitude is again used

to obtain a better phase estimate and vice versa until a certain convergence criterion is reached [3], termed as the iterative phase-aware approach in the rest of the thesis.

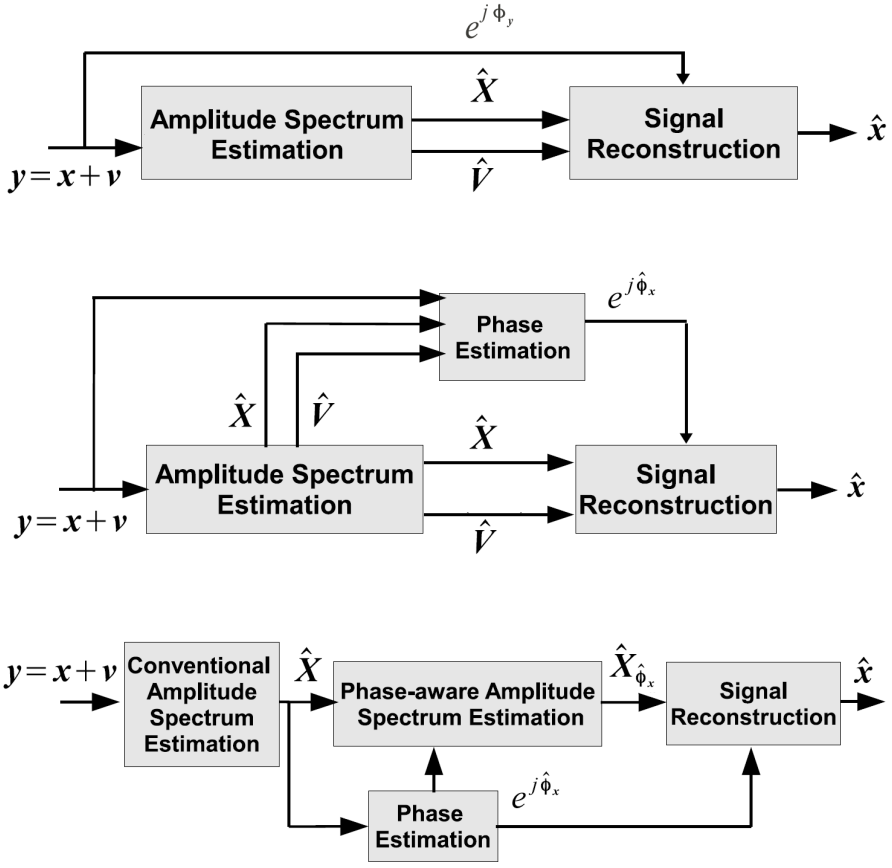

Figure 1.2: Different speech enhancement strategies: Conventional speech enhancement using the noisy phase at signal reconstruction (top); phase-aware enhancement using an estimated phase at signal reconstruction (middle); phase-aware enhancement using an estimated phase for amplitude modification and signal reconstruction (bottom).

An instrumental measure has to work well for all the above mentioned types of single-channel speech enhancement incorporating different phase spectra at the modification and reconstruction stages (completeness of the instrumental measure, mentioned in Chapter 2). That applies to perceived quality measures as well as to intelligibility measures. In the last decades several measures where suggested to have a reliable prediction for either perceived quality or intelligibility when only the spectral amplitude is modified while the noisy phase is directly copied. However, it is unclear if those measures also reasonably perform in a phase-aware framework since this research field is quite new and proper evaluations have not been carried out yet. This work is focused on this issue by taking into account different phase-enhancement methods in different noise scenarios.

1.2 Notations

Let $y(n) = x(n) + v(n)$ be the noisy/degraded signal with $x(n)$ and $v(n)$ denoting the clean and noise signals, respectively. The noisy signal is processed by a speech enhancement algorithm producing the enhanced speech signal $\hat{x}(n)$. Let $Y^c(k,l)$, $X^c(k,l)$, $\hat{X}^c(k,l)$ and $V^c(k,l)$ be the STFT transforms for noisy, clean, enhanced speech and noise signals, respectively, with k and l as the frequency and time indices. The complex spectrum $X^c(k,l)$ consists of spectral

amplitude and spectral phase $X^c(k,l) = X(k,l)e^{j\phi_x(k,l)}$ with $X(k,l)$ as the amplitude and $\phi_x(k,l) = \angle X^c(k,l)$ as the spectral phase.

Figure 1.3 shows the difference between the conventional performance evaluation relying on the spectral amplitude difference ($X(k,l)$ versus $\hat{X}(k,l)$) and the proposed metrics relying on the spectral phase values, i.e., $\phi_x(k,l)$ and $\hat{\phi}_x(k,l)$. Additionally, the proposed measures can also be extended by the use of the spectral amplitudes $X(k,l)$ and $\hat{X}(k,l)$. The motivation behind the use of the spectral phase information for performance evaluation is discussed in the next section.

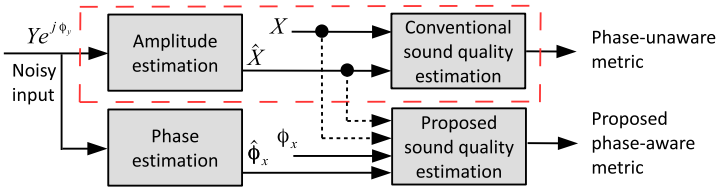

Figure 1.3: Conventional (dashed box) versus the proposed performance evaluation.

1.3 Fundamental Questions

Alsteris and Paliwal in [26] conducted experiments to determine if the spectral amplitude or the spectral phase is more important for speech intelligibility. They used an analysis-modification-synthesis-procedure to produce amplitude-only and phase-only stimuli from a clean speech signal by applying a STFT with a window length of $T_w = 32$ ms and an overlap of $T_w/8$. The clean speech signal STFT is given by

$$X(k,l) = |X(k,l)|\mathrm{e}^{j\phi_x(k,l)}, \tag{1.1}$$

where $|X(k,l)|$ denotes the short-time amplitude spectrum and $\phi_x(k,l) = \angle X(k,l)$ denotes the short-time phase spectrum. The phase-preserved stimuli where generated by setting the amplitude spectrum to unity and the modified STFT is given by

$$\hat{X}(k,l) = \mathrm{e}^{j\phi_x(k,l)}. \tag{1.2}$$

The amplitude-preserved stimuli were obtained by randomizing the phase spectrum with a uniform distribution between 0 and 2π. By performing listening tests on consonant identification, the results suggested that the phase spectrum is as important for the speech intelligibility as the amplitude spectrum [26]. However, it is unclear if the amplitude and phase spectra both contribute to speech intelligibility in an independent fashion.

Furthermore, in 1981 Oppenheim in [27] explored the importance of phase in signals. Among other things he conducted two experiments that emphasized on the impact of phase information to intelligibility, which are reproduced in the following. The first experiment shown in Figure 1.4 illustrates on top an image of two Hollywood actors with the corresponding amplitude and phase spectra in the middle and at the bottom. The first image on the left is the original one followed by the amplitude-only reconstructed (here the phase was set to zero prior reconstruction) and the phase-only reconstructed (with unity magnitude) versions. The last image on the left shows the outcome when the amplitude and phase spectra of the two images of the actors are swapped. It clearly can be seen that the phase contributes more information than the amplitude. In particular the shapes of the images are well preserved that can be interpreted as the *"intelligibility of an image"*. Apparently also the overall image quality is better in the case where the phase is preserved.

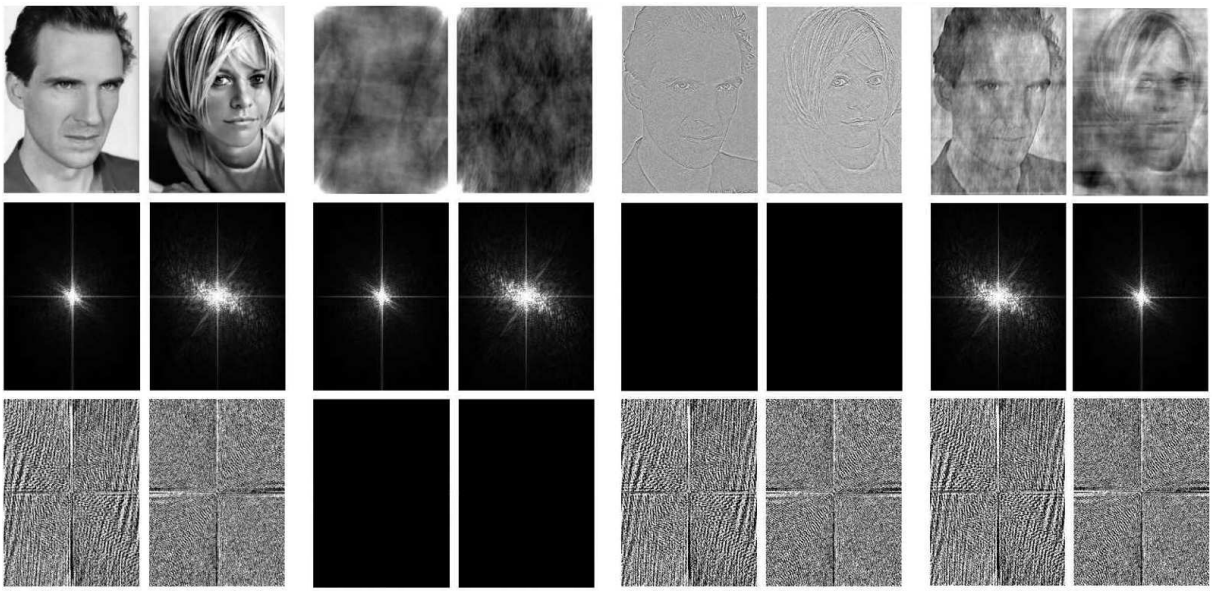

Figure 1.4: An image showing two actors from left to right: Original image; image reconstructed from amplitude spectrum only; image reconstructed from phase spectrum only; image reconstructed from the swapped amplitude and phase spectra.

In the second experiment, the magnitude and phase spectra of a female and male utterance taken from the GRID corpus [28] were swapped. Figure 1.5 shows the spectrograms of the original and reconstructed magnitude-swapped utterances. By comparing the harmonic structures of the original and modified speech samples, it appears that most of the information is carried by its phase. Informal listening tests confirm that the spectral phase information support the intelligibility of the sentence and the gender type while the swapped spectral magnitude is perceived as additive noise.

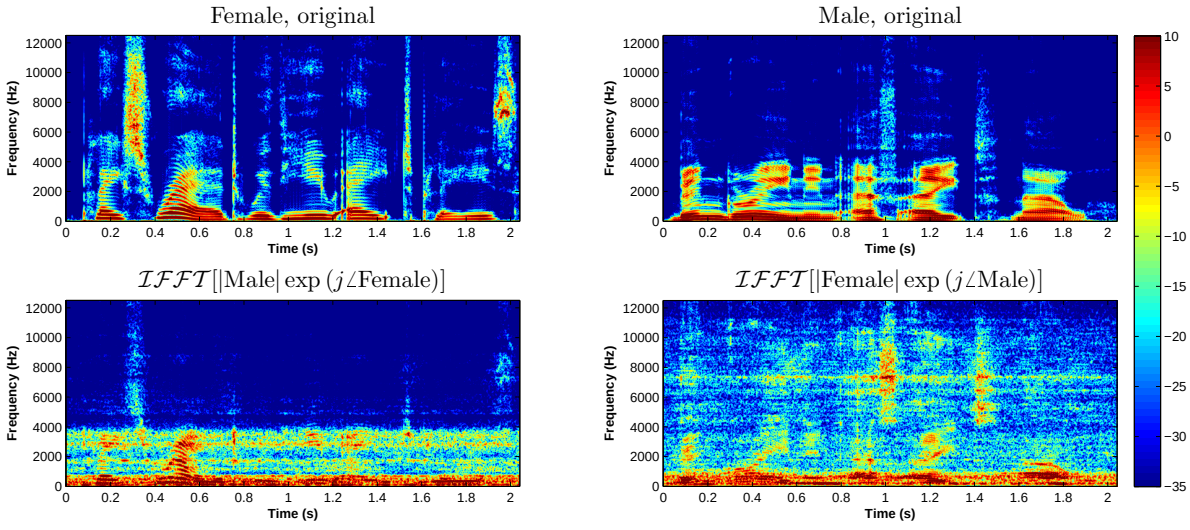

Figure 1.5: Exchanging the Fourier phase and magnitude in voice. (Top left) Female voice spectrogram; (top right) male voice spectrogram; (bottom left) spectrogram of female voice phase and male voice magnitude; (bottom right) spectrogram of male voice phase and female voice magnitude. Both reconstructions are primarily dominated by the Fourier phase, and not the magnitude.

In history the design of a new speech coder or a new single-channel speech enhancement al-

gorithm was always closely related to an evaluation measure. This is not a surprise since an evaluation measure could reveal information about some optimization criterion that leads to new insights for improving the performance of a coder or algorithm, i.e., the start of linear predictive coding (LPC) can be attributed to Itakura and Saito in 1966 [29] leading to the first maximum likelihood approach for automatic phoneme discrimination [30]. In 1970 the same authors introduced the Itakura-Saito distance (ISa) [31] based on the LPC coefficients, which is partially used nowadays to predict speech quality. Yet the types of distortions quantified by this metric are different than those introduced by a speech enhancement method, e.g., quantization errors rather than musical noise.

To this end Figure 1.6 illustrates that existing performance measures can be misleading when used in a phase-aware framework, where on top from left to right the spectrograms of a clean, noisy and three enhanced utterances are shown. The enhancement methods are minimum-mean-square error log-spectral amplitude estimator (MMSE-LSA) [24], STFT phase improvement (STFTPI) [21], and clean phase used at signal reconstruction stage. STFTPI relies on phase reconstruction at harmonics given a fundamental frequency estimate. Strict harmonicity is forced while noise components between two harmonics are entirely removed. The obtained improvement in contrast to the noisy and MMSE-LSA enhanced utterances in PESQ and fwSNR, presented in the title of the figures, does not reflect the introduced distortions perceived and termed as buzzyness by listeners, also reported in [32,33]. The buzzyness in Figure 1.6 is noticed by comparing the harmonic structure of the original clean signal with the STFTPI enhanced speech. A similar tendency is true for the group delay and phase variance plots in the middle and bottom line. In this example, PESQ and fwSNR also suggest that the clean phase information does not provide more towards speech quality in comparison to STFTPI which can be rejected by listening tests. On the contrary, the clean phase information leads to better perceived quality and speech intelligibility following the values of PESQ, fwSNR and STOI. This confirms the earlier results reported by Paliwal in [34].

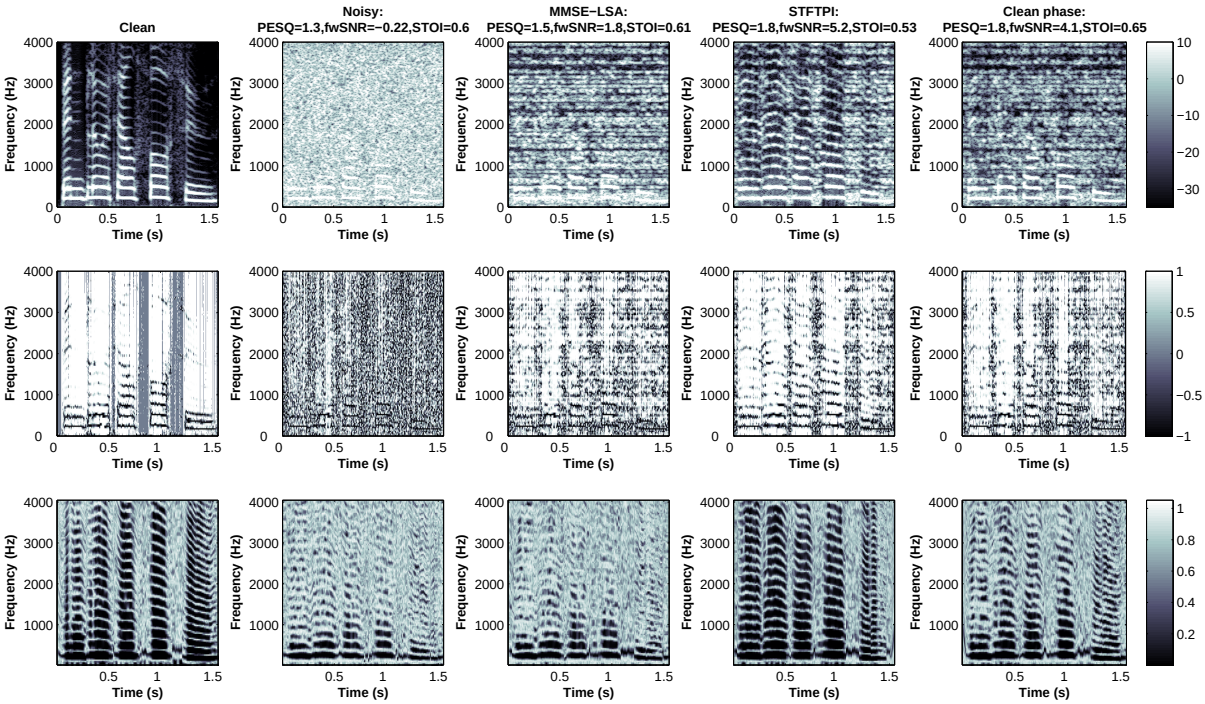

Figure 1.6: Counter example: (Top) spectrogram, (middle) group delay, (bottom) phase variance plots shown for (from left to right) clean signal, noisy unprocessed signal, phase enhanced signal using STFTPI, and phase enhanced using clean phase corrupted with white noise at 0 dB SNR.

As a last point to emphasize on the power of phase it has to be quoted that *"from a particular magnitude spectrogram, it is possible to reconstruct virtually any time-domain signal with a carefully crafted phase. For instance, one can derive a magnitude spectrogram from that of a speech signal such that it yields either a speech signal similar to the original or a piece of rock music, depending on the choice of the phase"* [35].

Considering all aspects mentioned in this section, in this work, the following three research questions will be addressed:

1. how much the existing conventional instrumental measures correlate with subjective results for phase-aware speech enhancement,
2. whether some new phase-aware measures could outperform the existing ones in terms of predicting the subjective listening results,
3. if there exits a measure that reliably predicts both perceived quality and speech intelligibility.

The third point tackles the idea that a measure which predicts perceived quality as well as speech intelligibility could specify a direction for further research on single-channel speech enhancement methods that improve the overall quality (including perceived quality and speech intelligibility).

1.4 Contributions of this Work

Chapter 2 presents the conventional instrumental quality measures. Due to a big variety of the existing speech quality measures a selection of the most important ones had to be made. The decision was based on the intention to reflect the progress of single-channel speech enhancement and the corresponding evaluation over the last decades incorporating signal-to-noise ratio (SNR)-based [17], linear prediction (LP)-based [17] and perceptually-motivated (e.g. perceptual evaluation of speech quality (PESQ) [36]) measures. Furthermore, some new phase-aware metrics are introduced that calculate a distortion metric based on the spectral phase only.

Chapter 3 presents commonly used instrumental intelligibility measures and the proposed phase-aware candidates. The selection out of a big variety of intelligibility measures follows the idea of Chapter 2. In principal those measures can be split into four groups:

- based on the articulation index (AI) [37]
- based on the speech transmission index (STI) [38]
- based on a perceptual model (e.g. DAU [39])
- based on mutual information (MI) [40]

Chapter 4 introduces the speech material, noise scenarios and the phase-aware single-channel enhancement algorithms used in the objective and subjective evaluations. Listening tests were conducted separately for perceived quality and speech intelligibility to assess the human listening results. The subjective listening results are presented and discussed.

Chapter 5 deals with the detailed performance analysis of the conventional and proposed instrumental scores by comparing them with the scores obtained by the subjective listening tests in Chapter 4. The evaluation procedure is done with standardized techniques commonly used in the speech enhancement community.

Chapter 6 concludes on the work and gives an outlook on future work.

2 Quality Measures

Subjective listening tests provide the most reliable method for assessment of speech quality, but as mentioned in Chapter 1 they are time consuming and expensive. Instrumental measures have to be found that correlate well with the perception of human listeners. According to [41] a good instrumental measure is characterized by the following six criteria adapted from [42]:

Completeness	All of the speech processing systems already in-use throughout the world fall within the scope of the model. This criterion shows that the development of speech quality models has been intimately related to the historical evolution of the speech processing systems.
Accuracy	The most widely used criterion. The estimated scores are correlated with human perception.
Credibility	The estimation is easily interpretable.
Extensibility	The scope of the model can increase.
Manipulability	The model is easily employed. The model must be totally self-sufficient: there is no need for fine tuning by the users.
Consistency	The relationship between the estimations and the auditory results is monotonic (internal consistency). The absolute estimated values have approximately the same magnitude as the auditory results (external consistency).

Further according to Figure 2.1 instrumental measures can be split into signal-based intrusive and non-intrusive models. Signal-based means that the model uses physical signals or some representation of them for the model input. Intrusive models then evaluate a score based on the information provided by the clean reference signal $x(n)$ and the degraded/enhanced signal $y(n)/\hat{x}(n)$ while non-intrusive models only use the degraded/enhanced speech signal. In this work only non-intrusive models are utilized.

In the next two Sections 2.1 and 2.2 first some conventional instrumental measures commonly used to evaluate speech quality are introduced and second the new phase-based candidates are presented.

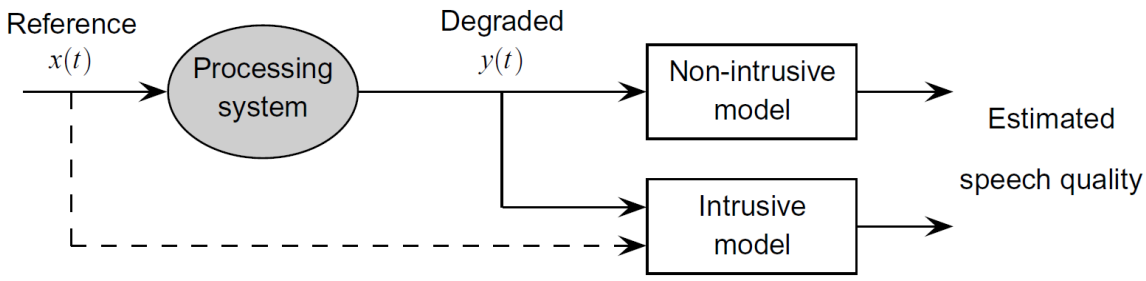

Figure 2.1: Intrusive and non-intrusive speech quality models, [41].

2.1 Conventional Instrumental Measures

Actual conventional quality measures can be divided into three different groups, depending on how they are calculated.

The first group are the SNR-based metrics. These measures are quite simple and calculate the signal-to-noise ratio between the clean/degraded/enhanced speech signal and the noise signal. This can be done either in the temporal or in the frequency domain. Examples for the temporal domain are the global SNR (GSNR) [43] and the segmental SNR (SSNR) [43], described in Sections 2.1.1 and 2.1.2. The frequency based measure is termed as frequency-weighted SNR (fwSNR) [44] and calculates the SNR in temporally segmented bands. The last measure to be accounted for a SNR-based one is the blind source separation evaluation (BSS EVAL) [45]. In a strict sense this measure was invented to evaluate the performance of a separation algorithm applied on audio mixed signals. Since speech enhancement and source separation are very familiar research fields, this measure can also be applied to enhanced speech signals by slightly modifying the metric input.

The second group are spectral distance measures based on linear predictive coding (LPC). These measures assume that speech follows an auto-regressive process within short time frames modelled by linear prediction. Examples are the log-likelihood ratio (LLR) [46], the Itakura-Saito distance (ISa) [31], and the cepstral distance (CEPS) [47], presented in Sections 2.1.5 - 2.1.7.

The metrics of the first two groups are simple to implement and easy to evaluate but their ability to predict perceived quality is limited because they are not close enough to the signal processing happening in the auditory periphery. Much research has been conducted on developing an auditory based measure and after more than ten years of evolution this yielded into a new model, termed Perceptual Evaluation of Speech Quality (PESQ) [36], standardized as the ITU-T Rec. P.862 (2001). More information about PESQ is given in Section 2.1.8.

2.1.1 Global SNR (GSNR)

The GSNR is the oldest known instrumental measure to evaluate quality. It is sample by sample comparison of temporal signals and is calculated using the following equation

$$\mathrm{SNR} = 10 \cdot \log\left(\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N}[x(n) - \hat{x}(n)]^2}\right), \tag{2.1}$$

where $x(n)$ is the clean speech, $\hat{x}(n)$ is the enhanced or degraded speech and N is the total number of samples of the entire utterance.

2.1.2 Segmental SNR (SSNR)

The upper definition in Eq. (2.1) is not well related to different types of speech distortions since it evaluates the speech quality by computing the SNR as an average over the entire utterance. Because speech is known to be quasi-stationary, there happens fluctuations of speech energy over time causing regions, where speech energy is large and noise energy is small, to be masked by regions, where speech energy is low and the noise is audible and vice versa. To overcome this, the SNR can be calculated in short frames and averaged afterwards as defined as the segmental SNR [43]

$$\mathrm{SSNR} = \frac{1}{L}\sum_{l=0}^{L-1} 10 \cdot \log\left(\frac{\sum_{n=Ml}^{Ml+M-1} x^2(n)}{\sum_{n=Ml}^{Ml+M-1}[x(n)-\hat{x}(n)]^2}\right), \tag{2.2}$$

where M is the frame length and L is the number of frames. The length of the frames is chosen to be 32 ms to capture the stationarity of speech. To split the signal into frames a Hamming window with 87.5% overlap is used. The above Eq. (2.2) computes the logarithm before averaging over the frames resulting in an arithmetic average. This causes frames with high SSNRs to be weighted less and emphasis on frames with low SSNRs. One problem with the estimation of the SSNR is the SNR contribution of silent frames. This frames exhibit very low SNR values and therefore bias the overall result. A possible remedy is to bound the SNR values within a certain range. In [48], the authors suggested to limit the values within the range [-10, 35] dB also used in this thesis.

2.1.3 Frequency Weighted SNR (fwSNR)

An extension to the SSNR is to calculate the SNR in the frequency domain to produce the frequency-weighted SNR (fwSNR) [44]

$$\mathrm{fwSNR} = \frac{1}{L}\sum_{l=0}^{L-1} 10 \cdot \frac{\sum_{j=1}^{J} W_j \cdot \log\left(\frac{X^2(j,l)}{[X(j,l)-\hat{X}(j,l)]^2}\right)}{\sum_{j=1}^{J} W_j}, \tag{2.3}$$

where $X(j,l)$ and $\hat{X}(j,l)$ are the clean and enhanced amplitude spectra at time frame l and frequency band j. W_j are band weightings based on the articulation index studies [49] applied to 25 critical bands spanning the frequency range of 50 to 3600 Hz.

2.1.4 Blind Source Separation Evaluation (BSS EVAL)

As mentioned in Section 2.1 BSS EVAL [45] was developed for blind source separation. The implementation in [50] takes the clean and estimated source signals as input. However, fixing the second source to the noisy signal assumes that there is no enhancement on this source signal which is exactly a speech enhancement scenario.

The BSS EVAL metric consists of three different SNR-based metrics: Source-to-Distortion Ratio

$$\mathrm{SDR} = 10 \cdot \log\left(\frac{\sum_{n=1}^{N} s_{target}^2}{\sum_{n=1}^{N}[e_{interf}+e_{noise}+e_{artif}]^2}\right), \tag{2.4}$$

the Source-to-Interference Ratio

$$\mathrm{SIR} = 10 \cdot \log\left(\frac{\sum_{n=1}^{N} s_{target}^2}{\sum_{n=1}^{N} e_{interf}^2}\right), \tag{2.5}$$

and the Source-to-Artefact Ratio

$$\mathrm{SAR} = 10 \cdot \log\left(\frac{\sum_{n=1}^{N} [s_{target} + e_{interf} + e_{noise}]^2}{\sum_{n=1}^{N} e_{artif}^2}\right), \tag{2.6}$$

where s_{target} denotes the true desired source modified by an allowed distortion and e_{interf}, e_{noise}, and e_{artif} are the interference, noise and artefact error terms. These four terms are estimated by orthogonal projections

$$s_{target} = P_{s_j}\hat{s}_j \tag{2.7}$$
$$e_{interf} = P_{\boldsymbol{s}}\hat{s}_j - P_{s_j}\hat{s}_j \tag{2.8}$$
$$e_{noise} = P_{\boldsymbol{s},\boldsymbol{n}}\hat{s_j} - P_{\boldsymbol{s}}\hat{s}_j \tag{2.9}$$
$$e_{artif} = \hat{s}_j - P_{\boldsymbol{s},\boldsymbol{n}}\hat{s_j} \tag{2.10}$$

with the orthogonal projectors defined as

$$P_{s_j} = \prod\{s_j\} \tag{2.11}$$
$$P_{\boldsymbol{s}} = \prod\{(s_{j'})_{1\leq j'\leq n}\} \tag{2.12}$$
$$P_{\boldsymbol{s},\boldsymbol{n}} = \prod\{(s_{j'})_{1\leq j'\leq n}, (n_i)_{1\leq i\leq m}\}, \tag{2.13}$$

where s_j is the j^{th} target source, $s_{j'}$ are the unwanted sources and n_i are the noise sources.

2.1.5 Log-likelihood Ratio (LLR)

The LLR metric as well as the other two metrics below (ISa, CEPS) use the assumption that the production of speech follows a source-filter model represented by the p^{th} order all-pole filter of the form

$$x(n) = \sum_{i=1}^{p} a_x(i)x(n-i) + G_x u(n) \tag{2.14}$$

where $a_x(i)$ are the filter coefficients of the all-pole, G_x is the filter gain, and $u(n)$ describes the excitation signal of the glottis modelled as unit variance white noise. The all-pole coefficients are determined using linear prediction. The LLR measure is computed as the dissimilarity between the all-pole models of the clean and enhanced speech signals $x(n)$ and $\hat{x}(n)$ by the following equation

$$d_{\mathrm{LLR}}(\boldsymbol{a}_x, \boldsymbol{a}_{\hat{x}}) = \ln\left(\frac{\boldsymbol{a}_{\hat{x}}^T \boldsymbol{R}_x \boldsymbol{a}_{\hat{x}}}{\boldsymbol{a}_x^T \boldsymbol{R}_x \boldsymbol{a}_x}\right), \tag{2.15}$$

where $\boldsymbol{a}_x$ and $\boldsymbol{a}_{\hat{x}}$ are the LPC coefficients of the clean and enhanced signals and $\boldsymbol{R}_x$ is the autocorrelation matrix of the clean signal. The LPC coefficients describe the filter of the vocal tract and thus the formants in the amplitude spectrum. An LP model order of 10 is used for speech signals with a sampling frequency below 10 kHz, to be of interest in this evaluation. Eq. (2.15) can be seen in a way that it penalizes differences in the formant location. The distances are calculated at frames of length 32 ms using a Hanning window with 87.5% overlap and the overall score is obtained by simple averaging over the frames. This procedure is also used for the following two measures presented in Sections 2.1.6 and 2.1.7. The LLR measure is limited in the range of [0,2] where 0 belongs to the upper bound that $x(n) = \hat{x}(n)$.

2.1.6 Itakura-Saito Distance (ISa)

The Itakura-Saito distance [31] is very similar to the log-likelihood ratio. The only difference is that it incorporates the filter gains in the calculation of the distance measure

$$d_{\mathrm{ISa}}(\boldsymbol{a}_x, \boldsymbol{a}_{\hat{x}}) = \frac{\sigma_x^2}{\sigma_{\hat{x}}^2} \frac{\boldsymbol{a}_{\hat{x}}^T \boldsymbol{R}_x \boldsymbol{a}_{\hat{x}}}{\boldsymbol{a}_x^T \boldsymbol{R}_x \boldsymbol{a}_x} + \ln\left(\frac{\sigma_x^2}{\sigma_{\hat{x}}^2}\right) - 1, \tag{2.16}$$

where σ_x^2 and $\sigma_{\hat{x}}^2$ are the all-pole gains for the clean and enhanced speech. According to the implementation of Loizou [17] the ISa distance is limited between [0,100].

2.1.7 Cepstral Distance (CEPS)

The Cepstral distance provides an estimate of the difference between the log spectral amplitudes of two different spectra and uses the cepstrum coefficients for computation [47]

$$d_{\mathrm{CEPS}}(\boldsymbol{c}_x, \boldsymbol{c}_{\hat{x}}) = \frac{10}{\ln(10)} \sqrt{2 \sum_{q=1}^{Q} [c_x(q) - c_{\hat{x}}(q)]^2}, \tag{2.17}$$

where $c_x(q)$ and $c_{\hat{x}}(q)$ are cepstrum coefficients of the clean and enhanced signals, respectively and Q is the order of the LPC analysis. The real cepstral coefficients are obtained by taking the ISTFT of the log amplitude spectrum and can be used to separate the excitation signal (described by the higher coefficients) from the vocal tract filter (described by the lower coefficients). Another way to get access to the cepstral coefficients is to calculate them recursively from the LPC coefficients [51]

$$c(m) = a_m + \sum_{q=1}^{m-1} \frac{q}{m} c(q) a_{m-q} \qquad 1 \leq m \leq Q. \tag{2.18}$$

The cepstral distance is limited in the interval of [0,10].

2.1.8 Perceptual Evaluation of Speech Quality (PESQ)

As mentioned in the introduction of this chapter, PESQ is a proposed metric by the ITU-T recommendation. Its main purpose is to evaluate perceived speech quality of coded speech (e.g. CELP) transmitted over a telecommunication channel. These distortions are different than those introduced by a speech enhancement algorithm. However, PESQ was reported to be a good predictor for speech enhancement [52] or speech separation [53]. The computation of the

PESQ score is algorithmically complex and thus out of scope to be explained in detail. A basic overview is shown in Figure 2.2.

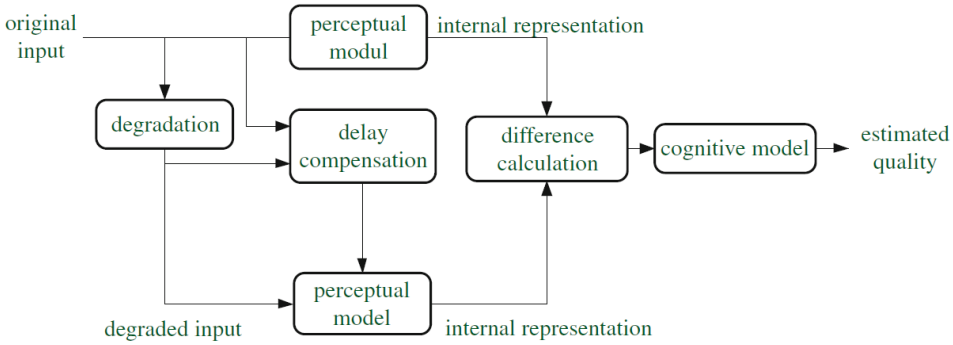

Figure 2.2: Basic structure of PESQ computation, [54].

The original and degraded signals are level aligned to a standard listening level of 79 dB SPL and filtered to model a standard telephone handset. In the next step the two input signals are time aligned assuming piecewise constant delays of the transmission channel. Delay changes in speech are allowed. The inner representations, which are used to calculate the score are in general loudness spectra obtained by auditory filtering. Therefore the incoming signal is split into frames of 32 ms with an overlap of 50% and then a FFT with a Hamming window is performed. The frequency bins are then grouped equally spaced on a modified Bark scale. Further steps are frequency equalization of the clean to the degraded signal, gain variation equalization of the degraded to the clean signal and a loudness mapping (Sone).

The obtained original and degraded loudness spectra are used to calculate the absolute difference, termed disturbances. PESQ treats positive and negative differences differently, because these differences contribute differently to the perceived quality. Distortions by additive noise components are more likely to be audible then distortions that happens by omitting some frequency content. This is considered by two different disturbances: the average disturbance value d_{sym} and the average asymmetrical disturbance value d_{asym} where the averaging over frequency and time is done by different L_r norms

$$L_r = \left(\frac{1}{L} \sum_{l=1}^{L} \text{disturbance}[l]^r \right)^{1/r}, \tag{2.19}$$

where r is separately chosen for d_{sym} and d_{asym} for the frequency as well for the time average. The final score is computed by linear weighting of the disturbances:

$$PESQ = 4.5 - 0.1 \cdot d_{sym} - 0.0309 \cdot d_{asym}. \tag{2.20}$$

The weighting was found by performing a training on a database of 30 subjective listening tests. The score lies between 1.0 (bad quality) and 4.5 (no distortion).

2.2 Proposed Instrumental Measures

As presented in Section 1.3 conventional quality measures show incompleteness in estimating the speech quality outcome of a phase-aware method. In particular they ignore the modifications of the phase at the signal reconstruction stage and calculate a distance based on the amplitude

spectrum.

The proposed measures calculate a distance based on the phase spectrum only, either direct on the phase values termed as mean square error of phase (MSE) described in Section 2.2.4 or on some representations derived from the phase which show correlation to the amplitude spectrum. These representations are the group delay (GD), the instantaneous frequency deviation (IFD), and the phase deviation (PD) introduced in Sections 2.2.1 - 2.2.3. The representations are not new by themselves but so far were not used to predict speech quality.

In general all proposed measures are frame-based with a frame length of 32 ms according to the quasi-stationarity of speech. A FFT is applied on each frame using a Chebyshev window with a dynamic range of 25 dB and an overlap of 87.5%. The use of a Chebyshev window with 25 dB dynamic range is motivated by an investigation of Paliwal in [34]. There he studied the influence of different analysis windows to obtain the phase spectrum. The Chebyshev 25 dB window was experimentally found to be optimal to obtain the clean spectral phase out of the clean reference signal. With this clean phase the speech signal was then reconstructed after the noisy spectral amplitude was enhanced by the MMSE-STSA approach, invented by Ephraim and Malah [23].

2.2.1 Group Delay (GD)

The instantaneous phase spectrum of the STFT is difficult to interpret. It has a random structure due to the mapping of the phase values to the interval of $[-\pi, \pi]$. Other representations are necessary to get useful information out of the phase spectrum. A simple approach is to have a look at the first derivative. Since the STFT is a time-frequency representation of the speech signal, a derivation in time and frequency is possible. The group delay is defined as the negative frequency derivative of the phase spectrum

$$\tau(\omega) = -\frac{\partial\phi(\omega)}{\partial\omega}. \tag{2.21}$$

For discrete-time processing the group delay is approximated by

$$\tau(k,l) = -\Delta_k\phi(k,l) = -(\phi(k,l) - \phi(k-1,l)), \tag{2.22}$$

where k and l denote the frequency bin and frame index. Figure 2.3 shows the amplitude spectrum, the $\cos(-\Delta_k\phi(k,l))$, and the instantaneous phase of the utterance *"bin blue at l four soon"* for a noisy, enhanced (MMSE-LSA [24]), and clean scenario taken from the GRID corpus [28]. While the instantaneous phase exhibits no explicitly visible structure, the group delay shows a clear structure reflecting the harmonics of the amplitude spectrum. To get an accurate estimate of the group delay the phase has to be unwrapped prior to estimation. This is down by the Matlab function *"unwrap"*. The function subtracts multiples of 2π if the jump of consecutive phase values is greater than π or adds multiples of 2π if the jump of consecutive phase values is less than $-\pi$. This unwrapping procedure is also used for the other proposed metrics introduced in the following sections. As summarized in [55], the group delay has been reported useful in various speech processing applications.

The group delay instrumental measure is defined as the distance of the group delay spectrum of the clean and enhanced signals $x(n)$ and $\hat{x}(n)$ averaged over time and frequency

$$\boxed{d_{\mathrm{GD}} = \frac{2}{L \cdot K}\sum_{l=1}^{L}\sum_{k=1}^{K/2}\Big(\cos(-\Delta_k\phi_x(k,l)) - cos(-\Delta_k\hat{\phi}_x(k,l))\Big)^2} \tag{2.23}$$

where K and L are the number of frequency bins and time frames, respectively. This metric was first used in [8] and later presented in more detail in [3] to resolve the ambiguity in phase estimation for single-channel speech enhancement. The cosine function in Eq. (2.23) is employed to avoid errors due to the 2π periodicity of the phase. The GD measure has a range between [0,4].

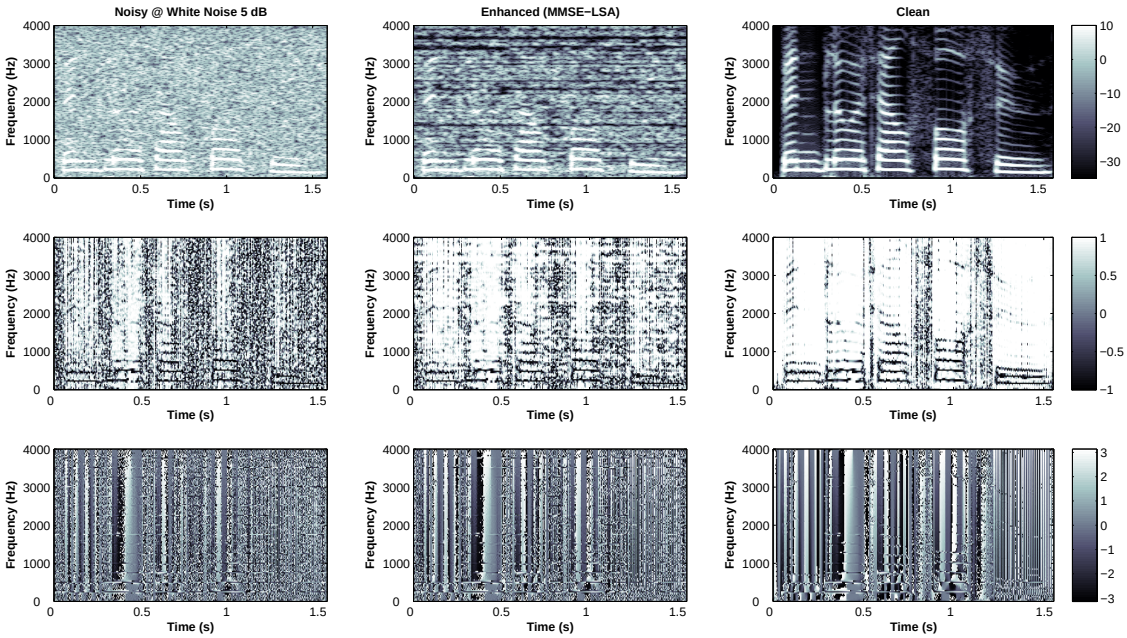

Figure 2.3: Spectrogram (top), group delay (middle), and instantaneous phase (bottom) for the noisy, enhanced, and clean female utterance "bin blue at l four soon".

2.2.2 Instantaneous Frequency Deviation (IFD)

The derivation of the phase spectrum across the time axis is called the instantaneous frequency (IF) [56]. It is calculated from the STFT phase spectrum as follows

$$\nu(\omega, t) = \frac{\partial \phi(\omega, t)}{\partial t}. \tag{2.24}$$

For discrete-time processing the above equation is given as

$$\nu(k, l) = (\phi(k, l) - \phi(k, l-1)). \tag{2.25}$$

Another way to calculate the IF is to use Kay's method [57] that avoids the issues due to the unwrapping problems

$$\nu(k, l) = \angle(X^c(k, l) - X^{c^*}(k, l-1)), \tag{2.26}$$

where $X^c(k, l)$ is the complex STFT spectrum at frequency bin k and time frame l and $*$ denotes the complex conjugate. Eq. (2.26) limits the IF into the range of $[-\pi, \pi]$.

The IF spectrum has been applied to speech recognition [58], pitch extraction [59] and formant

extraction [60]. As discussed in [61], the narrow-band IF spectrum (where the duration of the analysis window is between 20 and 40 ms) contains information of the excitation source but does not reveal the formant frequencies. Therefore they introduced a new representation that displays the pitch as well as the formant structure, called instantaneous frequency deviation (IFD)

$$\mathrm{IFD}_{\phi}(k,l) = \frac{1}{2\pi}\left(\phi(k,l) - \phi(k,l-1)\right) - \frac{2\pi F k}{K}, \tag{2.27}$$

or following Kay's method (Eq. (2.26)) [57]

$$\mathrm{IFD}_{\phi}(k,l) = \angle\left(X^{c}(k,l) - X^{c^*}(k,l-1)\cdot\exp\left(-j\frac{2\pi F k}{K}\right)\right), \tag{2.28}$$

where F denotes the frameshift in samples of consecutive frames. Figure 2.4 shows the $\cos(\mathrm{IFD}_{\phi}(k,l))$ and its relation to the amplitude spectrum, similar to the GD. A Chebyshev 50 dB window was used to generate the phase plots while a Hamming window was used to obtain the spectrograms.

The IFD instrumental measure is defined as

$$\boxed{d_{\mathrm{IFD}} = \frac{2}{L\cdot K}\sum_{l=1}^{L}\sum_{k=1}^{K/2}\left(\cos\left(\mathrm{IFD}_{\phi_x}(k,l)\right) - \cos\left(\mathrm{IFD}_{\hat{\phi}_x}(k,l)\right)\right)^2} \tag{2.29}$$

The measure is bounded in the interval of [0,4] with 0 denoting the best speech quality and was used in [62] to resolve the ambiguity in phase estimation for single-channel speech enhancement.

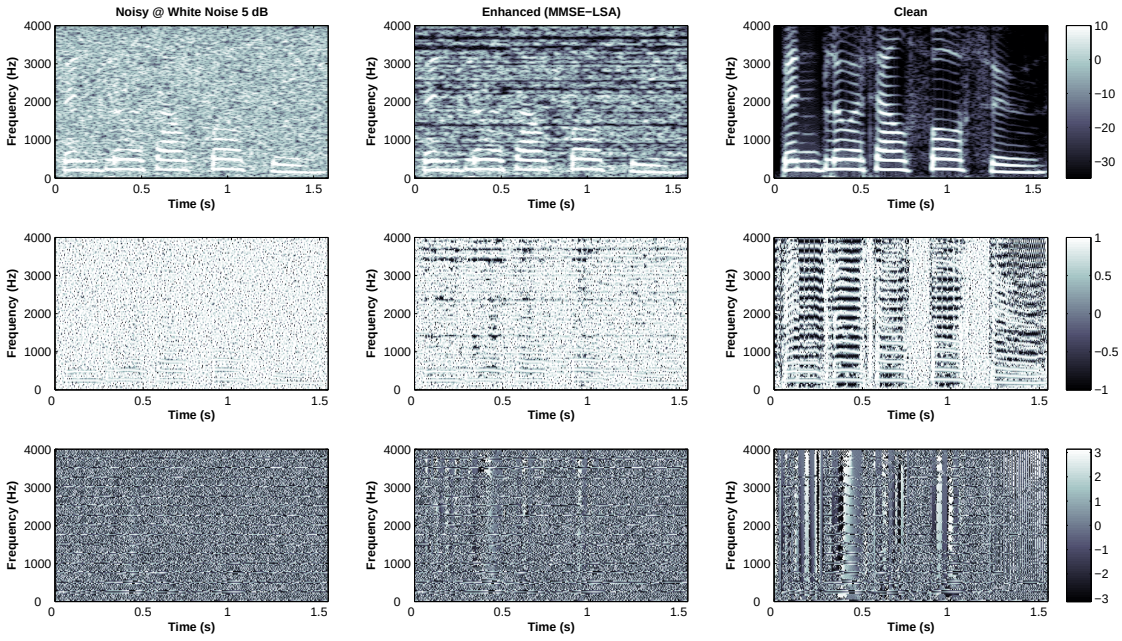

Figure 2.4: Spectrogram (top), instantaneous frequency deviation (middle), and instantaneous phase (bottom) for the noisy, enhanced, and clean female utterance "bin blue at l four soon".

2.2.3 Phase Deviation (PD)

The phase deviation is defined as the difference between the noisy and clean phase spectra

$$\phi_{\text{dev}}(k,l) = \phi_y(k,l) - \phi_x(k,l). \tag{2.30}$$

Its geometric representation is shown in Figure 2.5. In [63], Vary first introduced the concept of PD and determined that it can be used to clarify when phase distortions become perceptually audible in a speech enhancement application. In particular, roughness in synthesized speech was observed when the threshold of $\phi_{\text{dev}} > 0.679$ was exceeded. It was further shown that this value corresponds to a local $SNR \geq 6$dB between the clean and noisy spectral amplitude where the noisy phase provides a reasonable estimate for the clean phase. The PD was employed for joint noise reduction and echo cancellation [64] and for phase estimation [3].

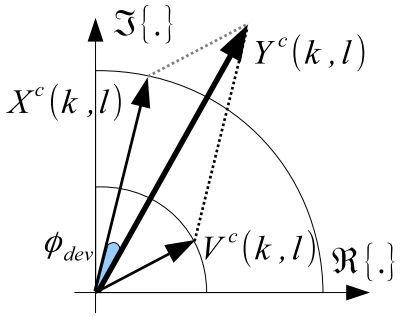

Figure 2.5: Geometric representation for the single-channel speech enhancement problem; showing noisy, clean, and noise complex spectra denoted by $Y^c(k,l)$, $X^c(k,l)$, and $V^c(k,l)$, respectively. The phase deviation ϕ_{dev} is shown as the phase difference between the clean and the noisy speech signal, [65].

Figure 2.6 shows an example of the $\cos(\phi_{\text{dev}}(k,l))$. For high SNR regions, e.g., at the harmonics, $\cos(\phi_{\text{dev}}(k,l)) \to 1$ and hence $\phi_{\text{dev}}(k,l)) \to 0$ while for low SNR regions, e.g., where the low energy fricatives are located, $\cos(\phi_{\text{dev}}(k,l)) \to -1$ and hence $\phi_{\text{dev}}(k,l)) = \pm\pi$.

The $\cos(\phi_{\text{dev}}(k,l))$ has an exact relation to the local *a priori* SNR denoted as $\text{SNR}(k,l) = \frac{X^2(k,l)}{V^2(k,l)}$ and the local *a posteriori* SNR denoted as $\text{SNR}_{post} = \frac{Y^2(k,l)}{V^2(k,l)}$, recently derived in [66]

$$\cos(\phi_{\text{dev}}(k,l)) = \frac{\text{SNR}(k,l) + \text{SNR}_{post}(k,l) - 1}{2\sqrt{\text{SNR}_{post}(k,l)\text{SNR}(k,l)}}. \tag{2.31}$$

In this way the following proposed instrumental measure can be seen as a SNR-based measure defined by the following equation

$$\boxed{d_{\text{PD}} = \frac{2}{L \cdot K} \sum_{l=1}^{L} \sum_{k=1}^{K/2} \left(\cos(\phi_{\text{dev}}(k,l)) - \cos(\hat{\phi}_{\text{dev}}(k,l)) \right)^2} \tag{2.32}$$

where $\hat{\phi}_{\text{dev}}(k,l) = \phi_y(k,l) - \hat{\phi}_x(k,l)$ defines the estimated phase deviation given the estimated phase. As the GD and the IFD the PD measure is bounded in the interval of [0,4].

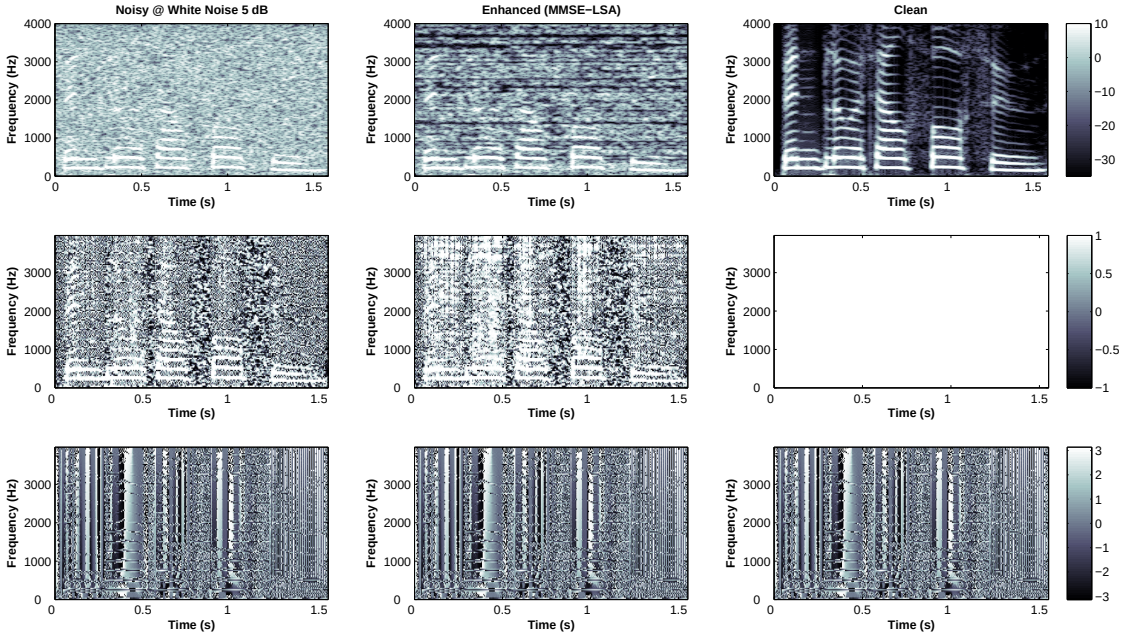

Figure 2.6: Spectrogram (top), phase deviation (middle) and instantaneous phase (bottom) for the noisy, enhanced and clean female utterance "bin blue at l four soon".

2.2.4 Mean Square Error of Phase (MSE)

To complete the set of phase-based metrics the mean square error of phase is introduced. It is a simple metric based on the local phase differences between the clean and enhanced speech signals averaged over all frequency bins and time frames

$$\boxed{d_{\mathrm{MSE}} = \frac{2}{L \cdot K} \sum_{l=1}^{L} \sum_{k=1}^{K/2} \Big(\cos(\phi_x(k,l) - \hat{\phi}_x(k,l)) \Big)^2}. \tag{2.33}$$

The cosine function wraps the values within the interval of [0,1]. In estimation theory, the MSE quantifies the amount of estimation error introduced by an estimator [67]. On the basis of the MSE measure the impact of the instantaneous phase towards quality prediction is studied.

2.2.5 Speech Squared Error (SSE)

The Speech Squared Error (SSE) here is used to measure the distortion between the original and the phase-distorted signals $x(n)$ and $\hat{x}(n)$. This is not the same as looking at the distortion of original and distorted phase spectra only. The SSE was defined in [68] as

$$\eta(\hat{X}(k,l), \Delta\phi(k,l), l) = \sum_{k=1}^{K} |X^c(k,l) - \hat{X}^c(k,l)|^2, \tag{2.34}$$

where $\Delta\phi(k,l)$ is defined as $\phi_x(k,l) - \hat{\phi}_x(k,l)$ and k and l are the frequency bins and time frames, respectively. Assuming that the enhanced speech signal $\hat{x}(n)$ differs only in its phase

spectrum and considering symmetry of the Fourier coefficients, Eq. (2.34) becomes

$$\eta(\hat{X}(k,l),\Delta\phi(k,l),l) = 8\sum_{k=1}^{K/2}\hat{X}^2(k,l)\sin^2\left(\frac{\Delta\phi(k,l)}{2}\right). \tag{2.35}$$

This equation was used in speech coding [68] to find the optimal linear phase when the phase distortion is given. The overall score is computed as the average over the time frames l, given as

$$\boxed{d_{\mathrm{SSE}} = \frac{1}{L}\sum_{l=1}^{L}\eta(\hat{X}(k,l),\Delta\phi(k,l),l)} \tag{2.36}$$

2.2.6 Weighted Speech Squared Error (WSSE)

In a realistic speech enhancement scenario the processed signal $\hat{x}(n)$ will not only contain phase distortions, but also amplitude distortions that have to be accounted. Therefore the Weighted Speech Squared Error (WSSE) is evaluated from Eq. (2.34) as

$$\eta(X(k,l),\hat{X}(k,l),\Delta\phi(k,l),l) = 2\sum_{k=1}^{K/2}(X(k,l)-\hat{X}(k,l))^2 + 4X(k,l)\hat{X}(k,l)\sin^2\left(\frac{\Delta\phi(k,l)}{2}\right), \tag{2.37}$$

where additional emphasis is given to the distortion of the amplitude spectrum. Similar to Eq. (2.36), the final score is computed as the average over time frames

$$\boxed{d_{\mathrm{WSSE}} = \frac{1}{L}\sum_{l=1}^{L}\eta(X(k,l),\hat{X}(k,l),\Delta\phi(k,l),l)} \tag{2.38}$$

3 Intelligibility Measures

There are two aspects of speech quality; the perceived speech quality discussed in Chapter 2 and the speech intelligibility. The speech intelligibility measures the accuracy of how well a message can be understood by a listener. It is a percentage of the correctly identified responses relative to the overall number of responses. It may be evaluated on phones, syllables, words, and sentences. Here we deal with identifying words and letters as a test unit placed in a "command-sentence" like structure explained in Section 4.4.1.

Conventional state-of-the-art instrumental measures are introduced in Section 3.1 followed by the proposed phase-aware candidates in Section 3.2.

3.1 Conventional Instrumental Measures

Conventional intelligibility metrics rely on different concepts to approximate the speech intelligibility.

The first group is based on the Articulation Index (AI) [37], proposed first by French and Steinberg [69] and later refined by Kryter [49]. These measures estimate intelligibility via calculating the speech audibility at frequency bands expressed as SNR under the assumption that the bands carry independent contribution to the total intelligibility. The SNRs are limited to a certain SNR range, normalized between $[0, 1]$ and combined to an overall averaged score by a perceptually motivated weighting. The models based on the AI are the Speech Intelligibility Index (SII) [70], the Coherence Speech Intelligibility Index (CSII) [71], and the SNRloss [72] discussed in Section 3.1.1 - 3.1.3.

The second group is based on the Speech Transmission Index (STI) proposed by Houtgast and Steeneken [38]. Additional to linear only degradation (as originally formulated for the AI), STI also is capable to handle convolutive degradations, e.g. reverberance. It therefore observes the reduction of the temporal amplitude modulation depth of the clean speech signal due to the influence of reverberation and additive noise. An extended version of the STI is the Normalized Covariance Metric [73], described in Section 3.1.4.

While the methods of the first two groups are suitable to predict speech intelligibility for distortion types including additive noise, convolutive noise, clipping, and filtering, they are less appropriate for non-linear filtering distortion types, e.g., ITFS processing [74]. Therefore other intelligibility measures such as the DAU measure [39] or the Short-Time Objective Intelligibility (STOI) [19] were proposed. Unlike STI or SII these measures rely on comparing the envelopes via short-term measures rather than relying on the long-term statistics. DAU and STOI form

the third group, presented in detail in Sections 3.1.5 and 3.1.6.

The last group, historically seen the latest one, comes out of the field of information theory and is based on mutual information. Mutual information between the message transmitted by the talker and the message interpreted by the listener has been widely used as a natural measure to assess the intelligibility [75]. The intelligibility measurement could be applied at the sequence of uttered words, sequence of states of the auditory system or on the message. In [75] it was also shown that the mutual information concept for predicting speech intelligibility turns out to be a generalization of the heuristically derived SII which is a common standard for speech intelligibility prediction and has a long history of development. Sections 3.1.7 and 3.1.8 describe two different implementations of the MI idea: Mutual Information based on KNN (MIKNN) [76] and Speech Intelligibility based on Mutual Information (SIMI) [77].

3.1.1 Speech Intelligibility Index (SII)

The SII is described in the ANSI S3.5-1997 standard [70]. As mentioned above it is an extension of the articulation index (AI) and additionally takes into account spread of masking, vocal effort and hearing loss. Masking becomes an issue when higher energy vowels make lower energy consonants inaudible. The vocal effort takes into account that extreme high sound pressure levels decrease speech intelligibility. The block diagram for SII calculation is shown in Figure 3.1.

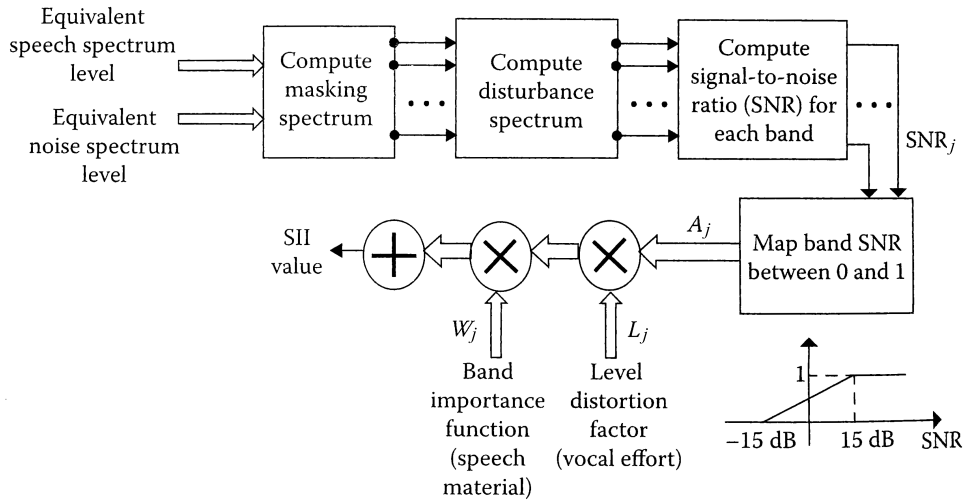

Figure 3.1: Block diagram showing the computation of the SII, [17].

First of all, equivalent speech spectrum levels X_j (in dB) are calculated in 18 one-third octave bands by subtracting the free-field to eardrum transfer functions from the speech spectrum levels measured at the eardrum. The same procedure is done for the equivalent noise spectrum levels V_j. Together with the equivalent hearing threshold levels T_j, these three quantities are the input of the model.

The next step is to compute the equivalent masking spectrum

$$Z_j = 10\log\left(10^{0.1V_j} + \sum_{m=1}^{j-1} 10^{0.1[B_m + 3.32C_m \log\left(\frac{0.89f_j}{f_m}\right)]}\right) \tag{3.1}$$

with f_j being the center frequencies of the one-third octave bands. The parameter C_j defines the slope per one-third octave band of spread of masking and B_j is a function of the self-masking

spectrum, given by:

$$B_j = \max(X_j - 24, V_j). \tag{3.2}$$

The disturbance spectrum level is defined as

$$D_j = \max(Z_j, N_j), \tag{3.3}$$

where N_j is the equivalent internal noise spectrum defined as

$$N_j = R_j + T_j, \tag{3.4}$$

with R_j as the reference internal noise spectrum describing an external masker signal that produces the pure-tone threshold in quiet. The effective band SNR (also called audibility function) in band j is clipped between [-15,15] dB and mapped into the range of [0,1]

$$\mathrm{SNR}_j = \frac{X_j - D_j + 15}{30}. \tag{3.5}$$

The last stage adjusts the audibility function to account for the speech level distortion factor L_j and the band-importance function W_j

$$\mathrm{SII} = \sum_{j=1}^{18} W_j \cdot \mathrm{SNR}_j. \tag{3.6}$$

The band-importance function is responsible to fit different speech data sets, e.g., non-sense syllables and sentences. In this thesis those values are taken corresponding to average speech. The score of SII ranges between [0,1] where a number above 0.75 corresponds to a good speech intelligibility and a value below 0.45 indicates a poor intelligibility.

3.1.2 Coherence Speech Intelligibility Index (CSII)

The CSII is an extension of the SII and is reported in [71] as a reliable speech intelligibility predictor for non-linear distortions such as peak- and center-clipping. The main difference of the CSII to the SII is that it replaces the SNR term by a Signal-to-Distortion ratio (SDR) computed from the coherence between the clean and processed speech signals $x(n)$ and $y(n)$. The magnitude-squared coherence function (MSC) is calculated out of the cross- and autospectra averaged across the windowed data segments. For L data segments the MSC is given by

$$\mathrm{MSC} = \frac{\sum_{l=0}^{L-1} |X_l(k) Y_l^*(k)|^2}{\sum_{l=0}^{L-1} |X_l(k)|^2 \sum_{l=0}^{L-1} |Y_l(k)|^2}, \tag{3.7}$$

where the asterisk denotes the complex conjugate. $X_l(k)$ and $Y_l(k)$ are the spectral amplitudes at frame l and frequency bin k. The speech power spectrum is computed as

$$\hat{P}_Y = \mathrm{MSC} \cdot S_{yy}(k) \tag{3.8}$$

and the noise power spectrum is given by

$$\hat{P}_V = [1 - \text{MSC}]S_{yy}(k) \tag{3.9}$$

with $S_{yy}(k)$ defined as the output power spectral density. The SDR is then estimated using the MSC:

$$\text{SDR}(j) = \frac{\sum_{k=1}^{K} W_j(k)\text{MSC} \cdot S_{yy}(k)}{\sum_{k=1}^{K} W_j(k)[1 - \text{MSC}]S_{yy}(k)}, \tag{3.10}$$

where $W_j(k)$ are simplified ro-ex filters to model the auditory filter bank with center frequencies and bandwidths given by the ANSI S3.5-1997 standard [70]. The general procedure to compute the CSII score is the same as in the SII, but with the SNR term replaced by Eq. (3.10) and the equivalent speech and noise spectrum levels replaced by the terms in Eq. (3.8) and Eq. (3.9), respectively.

The authors in [71] further divided the clean speech signal envelope into three amplitude regions leading to CSII_{low}, CSII_{mid}, and $\text{CSII}_{\text{high}}$. The regions are defined by the relative root-mean-square (rms) level of each frame in comparison to the rms level of the whole utterance. Frames with an relative rms level greater than or equal to the overall rms level are termed as high-level frames. Frames with an rms-level between [-10,0] dB are termed as mid-level and frames between [-30,-10] dB are termed as low-level. The three-level CSII terms are computed separately and are linearly weighted to obtain a complete score

$$I_3 = w_{\text{low}}\text{CSII}_{\text{low}} + w_{\text{mid}}\text{CSII}_{\text{mid}} + w_{\text{high}}\text{CSII}_{\text{high}}. \tag{3.11}$$

The investigation in [71] showed that mainly the mid-level score dominates the prediction of speech intelligibility.

3.1.3 SNRloss

The authors in [72] presented the SNRloss measure as a reliable intelligibility predictor for noisy speech modified by some speech-enhancement algorithm. The basic idea of this measure is to calculate a spectral distortion as the difference between the input Signal-to-Noise ratio $\text{SNR}_X(j,m)$ and the effective Signal-to-Noise ratio $\text{SNR}_{\hat{X}}(j,m)$ of the enhanced signal, termed as SNRloss

$$\begin{aligned} L(j,l) &= \text{SNR}_X(j,l) - \text{SNR}_{\hat{X}}(j,l) \\ &= 10 \cdot \log\left(\frac{X(j,l)^2}{V(j,l)^2}\right) - 10 \cdot \log\left(\frac{\hat{X}(j,l)^2}{V(j,l)^2}\right) \\ &= 10 \cdot \log\left(\frac{X(j,l)^2}{\hat{X}(j,l)^2}\right) \end{aligned} \tag{3.12}$$

where $X(j,l)$, $\hat{X}(j,l)$, and $V(j,l)$ are the spectral amplitudes of the clean, enhanced, and noise signals at frequency band j and frame l, respectively. The STFT analysis length is 20 ms with 75% overlap between adjacent frames using a Hamming window.

The SNRloss can be either positive or negative referring to the presence of spectral attenuation distortions or spectral amplification distortions. According to Figure 3.2, this distortions are

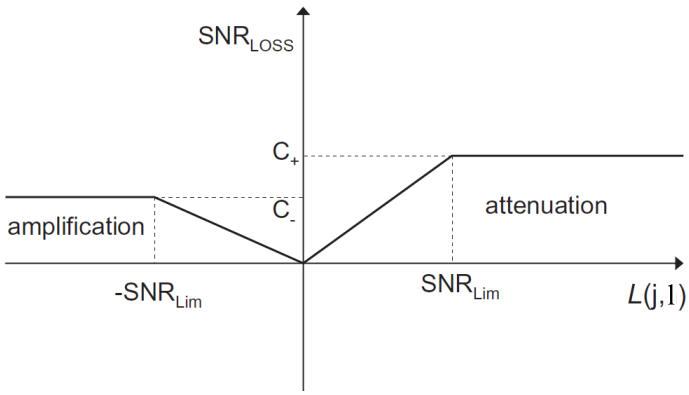

Figure 3.2: Mapping between clean and enhanced signals to SNRloss. The slope of the mapping function is controlled by the parameters C_+ and C_- and the SNR range ($[-SNR_{Lim}, SNR_{Lim}]$), [72].

limited to a range of SNR levels $[-\mathrm{SNR}_{\mathrm{Lim}}, \mathrm{SNR}_{\mathrm{Lim}}]$

$$\hat{L}(j,l) = \min(\max(L(j,l), -\mathrm{SNR}_{\mathrm{Lim}}), \mathrm{SNR}_{\mathrm{Lim}}), \tag{3.13}$$

and mapped to the range of $[0,1]$ using the following equation

$$\mathrm{SNRloss}(j,l) = \begin{cases} -\dfrac{C_-}{\mathrm{SNR}_{\mathrm{Lim}}}\hat{L}(j,l) & \text{if } \hat{L}(j,l) < 0 \\ \dfrac{C_+}{\mathrm{SNR}_{\mathrm{Lim}}}\hat{L}(j,l) & \text{if } \hat{L}(j,l) \geq 0 \end{cases} \tag{3.14}$$

where the parameters C_+ and C_- (defined in the range of $[0,1]$) are used to emphasise differently on attenuation and amplification distortions. However, in [72] the values for the dynamic SNR range were found experimentally to be $[-3,3]$ dB and $C_+ = C_- = 1$. The final SNRloss score is computed by averaging over all frames and frequency bands using a band-importance function denoted by $W(j)$, and given by:

$$\mathrm{SNRloss} = \frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_{j=1}^{J} W(j)\cdot \mathrm{SNRloss}(j,l)}{\sum_{j=1}^{J} W(j)}, \tag{3.15}$$

where L is the total number of frames and J is the total number of bands. The band-importance functions were taken from Table B.1 (short-passage functions) in [70] and linearly interpolated to the 25 critical-band center frequencies used in the implementation. The SNRloss measure is defined in the interval of $[0,1]$. The value of 0 means perfect intelligibility.

3.1.4 Normalized Covariance Metric (NCM)

The NCM measure is a speech-based alternative of the well-known STI measure [38], computed as follows [73]. Instead of using a sinewave-modulated signal, a speech signal is used as a probe signal. The clean and enhanced signal is first band-pass filtered into 20 bands with center frequencies ranging form 300 to 3400 Hz. Then the Hilbert transform is used to calculate the envelope of these signals and downsampled afterwards to 25 Hz to obtain the modulation envelopes $x_j(n)$ and $\hat{x}_j(n)$ of the clean an enhanced speech in each band j. These are used to

calculate the normalized covariance in each band

$$\rho_j = \frac{\sum_n (x_j(n) - \mu_j)(\hat{x}_j(n) - \nu_j)}{\sqrt{\sum_n (x_j(n) - \mu_j)^2 \sum_n (\hat{x}_j(n) - \nu_j)^2}}, \tag{3.16}$$

where μ_j and ν_j determine the mean values of $x_j(n)$ and $\hat{x}_j(n)$ respectively. The SNR is calculated as

$$\mathrm{SNR}_j = 10\log\left(\frac{\rho_j^2}{1-\rho_j^2}\right), \tag{3.17}$$

and limited to the range of [-15,15] dB. A linear mapping scheme is used to compute the transmission index (TI) at frequency band j, given by

$$\mathrm{TI}_j = \frac{\mathrm{SNR}_j + 15}{30}. \tag{3.18}$$

The NCM measure is then given by averaging the transmission indices across all frequency bands with additional weighting factors W_j used in the SII standard [70]

$$\mathrm{NCM} = \frac{\sum_{j=1}^{J} W_j \mathrm{TI}_j}{\sum_{j=1}^{J} W_j}. \tag{3.19}$$

The NCM measure is reported in [17] to be the best predictor for reverberant speech.

3.1.5 DAU Auditory Model (DAU)

Unlike the previous measures which are based on calculations on the physical acoustic signal, the DAU measure predicts speech intelligibility by using an internal representation of the speech signals. The used psycho acoustically validated model of auditory processing shown in Figure 3.3 was first presented by Dau in [78]. The incoming signal is filtered by a fourth-order gammatone filterbank consisting of 32 bandpass filters with center frequencies ranging from 100 to 8000 Hz. Afterwards each channel gets half-wave rectified and low-pass filtered at 1 kHz to preserve the temporal fine structure of the signal for low frequencies and to extract the envelope for high frequencies. This stage roughly simulates the oscillations of the basilar membrane into receptor potentials in the inner hair cells. The next stage consists of a chain of five non-linear adaptation loops with time constants of 5, 50, 129, 253, and 500 ms. This adaptation loops simulate the fire rates of the inner hair cells which depend on the speed of fluctuations of the speech signal. Furthermore masking effects are described with this loops. The 8 Hz modulation low-pass filter extracts the envelope of the pre-processed signal. To account for the limited resolution of the auditory system, an internal noise with constant variance is added at the end.

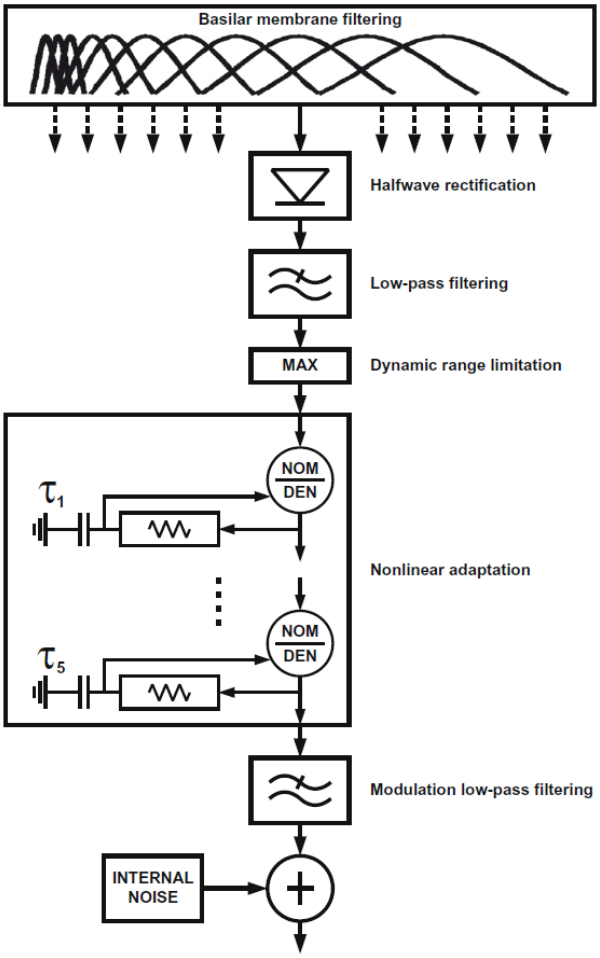

Figure 3.3: The auditory perception model by Dau [78].

The above described model was later used by Christiansen *et al.* to predict speech intelligibility and is described in [39]. An overview of the intelligibility model is shown in Figure 3.4. After processing the reference and degraded signals with the auditory model, the linear cross-correlation coefficient is calculated between the two inner representations in frames of length 20 ms and with an overlap of 50%. At the same time the root-mean-square (rms) level of every frame of the reference signal is compared to the rms level of the whole reference signal and categorized as high- mid- or low-level. The cross-correlation coefficients are then averaged separately for each level and finally are linearly weighted to obtain an overall score

$$\mathrm{DAU} = w_{\mathrm{low}}\bar{\rho}_{\mathrm{low}} + w_{\mathrm{mid}}\bar{\rho}_{\mathrm{mid}} + w_{\mathrm{high}}\bar{\rho}_{\mathrm{high}}, \tag{3.20}$$

where w_{low}, w_{mid}, and w_{high} are the weights and $\bar{\rho}_{\mathrm{low}}$, $\bar{\rho}_{\mathrm{mid}}$, and $\bar{\rho}_{\mathrm{high}}$ are the averaged level scores. High-level segments are defined to have an rms level of 0 dB or higher than the overall rms level. The mid-level segments are limited between -5 to 0 dB and the low-level segments are defined between -15 to -5 dB. This level-based calculation is motivated by [71] but the bounds are a little bit different. In the investigation in [39] the best weighting was found by only taking high level frames into account also used in this thesis.

In the implementation of this measure a logistic function is used to map the objective scores to the subjective scores. This mapping is replaced by another one to fit the data described in Section 4.1.

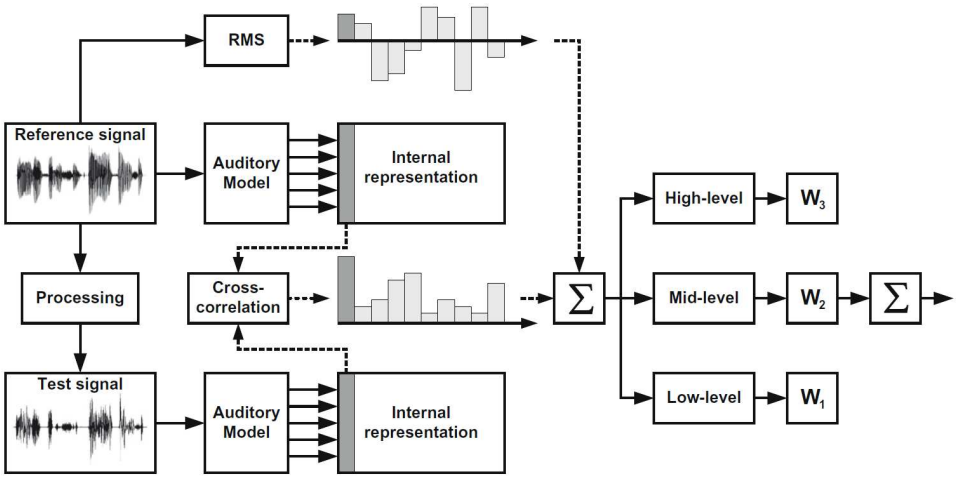

Figure 3.4: Schematic of the DAU measure, [39].

3.1.6 Short-Time Objective Intelligibility (STOI)

The intelligibility measure STOI proposed in [19] is one of the most widely used speech intelligibility metric in the speech enhancement community and is known to have high correlation to noisy and time-frequency weighted noisy speech. STOI compares the temporal envelopes of the clean and degraded speech in short-time regions by means of a correlation coefficient. Its basic principle is shown in Figure 3.5.

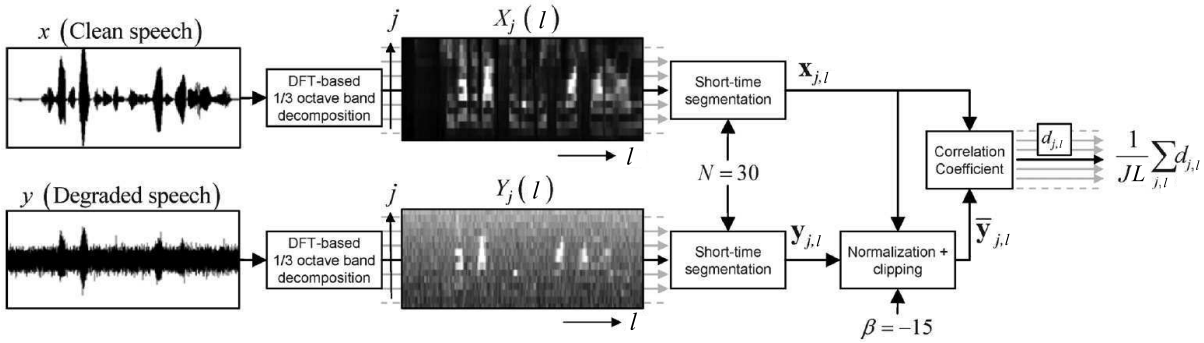

Figure 3.5: Principle of the STOI measure, [19].

After resampling the speech samples to 10 kHz in the first step, the clean and degraded speech, denoted by x and y respectively, are TF-decomposed into frames with a length of 256 samples (25.6 ms), where each frame is zero-padded up to 512 samples. For the STFT a Hann window with 50% overlap is used. Silent frames, which do not contribute to speech intelligibility, are removed by excluding all frames, where the speech energy is lower than 40 dB with respect to the maximum energy of the speech signal. Then DFT-bins are grouped into 15 one-third octave bands with center-frequencies from 150 Hz to 4.3 kHz. The norm of each band is referred as a TF-unit, defined as

$$X_j(l) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,l)|^2}, \tag{3.21}$$

where $\hat{x}(k,l)$ denotes the k^{th} DFT-bin of the l^{th} frame and k_1, k_2 denote the one-third octave band edges of the j^{th} band.

This TF-units are stacked into vectors with $N = 30$ elements yielding an analysis length of 384 ms and is called short-time temporal envelope of the clean speech:

$$\boldsymbol{x}_{j,l} = [X_j(l-N+1), X_j(l-N+2), \cdots, X_j(l)]^T. \tag{3.22}$$

Before comparison, the degraded short-time temporal envelope $\boldsymbol{y}_{j,m}$ is first normalized to the energy of the clean temporal envelope $\boldsymbol{x}_{j,m}$ and is clipped, so that the local Signal-to-Distortion-Ratio between the clean and degraded envelopes does not fall below -15 dB. The normalization compensates for global level differences which should not strongly affect the speech intelligibility while the clipping procedure upper bounds the sensitivity of the model related to one TF-unit.

The intermediate intelligibility measure is then calculated as the correlation coefficient between the clean and modified degraded vector $\hat{\boldsymbol{y}}_{j,m}$

$$d_{j,l} = \frac{(\boldsymbol{x}_{j,l} - \mu_{\boldsymbol{x}_{j,l}})^T (\hat{\boldsymbol{y}}_{j,l} - \mu_{\hat{\boldsymbol{y}}_{j,l}})}{\|\boldsymbol{x}_{j,l} - \mu_{\boldsymbol{x}_{j,l}}\| \|\hat{\boldsymbol{y}}_{j,l} - \mu_{\hat{\boldsymbol{y}}_{j,l}}\|}. \tag{3.23}$$

The overall measure is finally obtained by averaging over all frames and bands:

$$d = \frac{1}{JL} \sum_{j,l} d_{j,l}, \tag{3.24}$$

where L represents the total number of frames and J is the number of one-third bands. The STOI measure is defined in the interval of $[0, 1]$ with a higher number meaning a better speech intelligibility.

3.1.7 Mutual Information based on KNN (MIKNN)

Other than employing some kind of Signal-to-Noise ratio (SNR) or a correlation-based comparison between the spectro-temporal representations of clean and enhanced speech, the method in [76] calculates the speech intelligibility by relying on the estimated mutual information between the clean and enhanced speech at temporal envelopes.

Mutual Information is a general measure of dependence between two random variables X and Y

$$I(X;Y) = \iint P_Z(x,y) \ln\left(\frac{P_Z(x,y)}{P_X(x)P_Y(y)}\right) dxdy, \tag{3.25}$$

where $P_X(x)$ and $P_Y(y)$ are the marginal probability density functions and $P_Z(x,y)$ is the joint density of $Z = (X,Y)$. MI can also be defined by the differential entropy $h(X)$ that expresses the degree of information that the observation of the random variable provides

$$h(X) = -\int P_X(x) \ln(P_X(x))\, dx. \tag{3.26}$$

Together with the joint entropy $h(X,Y)$ the MI can be rewritten as

$$I(X;Y) = h(X) + h(Y) - h(X,Y) \tag{3.27}$$

where

$$h(X,Y) = -\iint P_Z(x,y)\ln(P_Z(x,y))\,dxdy \tag{3.28}$$

The mutual information is always greater than or equal to 0 if X and Y are independent. In the used implementation the MI is estimated by a k-nearest neighbour (KNN) approach by Kraskov *et al.* [40]. This non-parametric statistical approach has the advantage that no prior knowledge about the distributions of the random variables has to be known. The speech intelligibility score is obtained as shown in Figure 3.6.

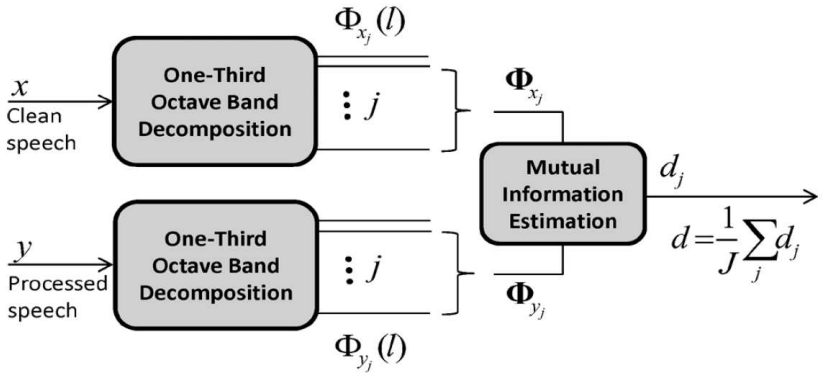

Figure 3.6: Speech intelligibility prediction based on mutual information. The mutual information is estimated by a k-nearest neighbour approach, [76].

The time-frequency representation of the clean signal x and processed signal y is exactly the same as the one used in the STOI measure described in Section 3.1.6. The intermediate intelligibility score is calculated through comparing the long-term temporal envelopes $\mathbf{\Phi}_{x_j} = [\Phi_{x_j}(1), \cdots, \Phi_{x_j}(L)]^T$ and $\mathbf{\Phi}_{y_j} = [\Phi_{y_j}(1), \cdots, \Phi_{y_j}(L)]^T$ of the clean and processed speech signals at each one-third octave band by means of mutual information with l and j denoting the frame and band index and L and J being the total number of frames and bands, given by:

$$d_j = \hat{I}(\mathbf{\Phi}_{x_j}, \mathbf{\Phi}_{y_j}). \tag{3.29}$$

The final score is the average of the intermediate scores d_j over all sub-bands

$$d = \frac{1}{J}\sum_{j=1}^{J} d_j. \tag{3.30}$$

3.1.8 Speech Intelligibility based on Mutual Information (SIMI)

Similar to Taghia and Martin, at quite the same time Jensen and Taal came up with their idea of predicting speech intelligibility based on mutual information in [77]. In their investigation

they assumed that intelligibility is monotonically related to the mutual information between critical-band amplitude envelopes of the clean signal and the corresponding processed signal. By lower-bounding the mutual information $I(X,Y)$ the resulting model turns out to be a simple function of the mean-square error that arises when estimating a clean critical- band amplitude using a minimum mean-square error (mmse) estimator based on the processed amplitude. The average intelligibility score is computed by the following equation

$$\tilde{I}(X,Y) = \frac{1}{J|Z_X|} \times \sum_{l \in Z_Y \cap Z_X} \sum_{j=1}^{J} \min(\hat{I}(X_j(l); Y_j(l)), I_{max}), \tag{3.31}$$

where $X_j(l)$ and $Y_j(l)$ are the critical-band amplitudes of the clean and processed speech signals at time frame l. The term $J|Z_X|$ denotes the number of speech-active critical-band amplitudes in the clean signal and $l \in Z_Y \cap Z_X$ is the appropriate frame index set. An upper bound $I_{\max} = 0.2$ per critical-band amplitude is introduced to avoid that a single high-information time-frequency unit dominates the overall information score. This value was determined heuristically. For further information about the estimation of MI in the bands please refer to [77]. The results in [77] confirmed that methods with amplitude-only modifications of the critical band amplitudes can not improve the predicted speech intelligibility coinciding with the previous observation in [79].

3.2 Proposed Instrumental Measures

The introduction of some new phase-aware instrumental intelligibility measures follows the same idea as for the quality estimation, that new measures have to be found that reliably predict the speech intelligibility of an enhanced speech signal provided by a phase-aware enhancement method. These are the unwrapped harmonic phase SNR (UnHPSNR) and the unwrapped root-mean-square error (UnRMSE) described in Sections 3.2.1 and 3.2.2 both measuring the phase estimation error in the unwrapped domain. This is motivated by reports where the phase information was shown to have impact on intelligibility or was successfully used in a speech enhancement framework to increase speech intelligibility [20, 26, 80, 81].

Additionally the proposed quality measures in Sections 2.2.1 - 2.2.4 are considered as intelligibility measures as well. As mentioned in Chapter 3, speech quality is subdivided into the parts perceived quality and speech intelligibility. *"The relationship between perceived quality and speech intelligibility is not entirely understood. However, there does exist some correlation between these two. Generally, speech perceived as "good" quality gives high intelligibility, and vice versa. However, there are samples that are rated as "poor" quality, and yet give high intelligibility scores, and vice versa"* [54].

3.2.1 Unwrapped Harmonic Phase SNR (UnHPSNR)

We define the unwrapped harmonic phase SNR as follows:

$$\text{UnHPSNR} = 10 \cdot \log\left(\frac{1}{L}\sum_{l=1}^{L} \frac{\sum_{k=1}^{K/2} X^2(k,l)}{\sum_{k=1}^{K/2} X^2(k,l)(1-\cos(\Psi_x(k,l) - \hat{\Psi}_x(k,l)))}\right), \tag{3.32}$$

where K and L denote the number of frequency bins and time frames, $X(k,l)$ is the clean spectral amplitude and $\Psi_x(k,l)$ and $\hat{\Psi}_x(k,l)$ are the unwrapped phase spectra. According to

[82] the unwrapped phase is obtained through the approach of phase decomposition

$$\psi(h,l) = \underbrace{\angle VT(h,l) + \psi_d(h,l)}_{Unwrapped\ phase\ \Psi(h,l)} + \underbrace{h \cdot \sum_{l'=0}^{l} \omega_0(l')(t(l') - t(l'-1))}_{Linear\ phase\ \psi_{lin}(h,l)}, \tag{3.33}$$

where $\omega_0(l) = 2\pi f_0(l)/f_s$ with $f_0(l)$ being the fundamental frequency at frame l and h denotes the harmonic number. A robust fundamental frequency estimator (PEFAC) [83] is used to extract the fundamental frequencies from the clean, noisy, and enhanced speech signals. The first term $\angle VT(h,l)$ corresponds to the minimum phase spectrum related to the vocal tract filter. The second term is called dispersion phase $\psi_d(h,l)$ and captures the information of the excitation signal. The third term is the linear phase part that wraps the instantaneous phase across time. The unwrapped phase $\Psi(h,l)$ is calculated by

$$\Psi(h,l) = \psi(h,l) - \psi_{lin}(h,l), \tag{3.34}$$

capturing the phase contributions of the excitation signal and the vocal tract.

As seen in Eq. (3.32) the UnHPSNR score is weighted by the clean spectral amplitude which forces the score to impact more on spectral harmonics. Furthermore, the measure is calculated only at voiced frames where the V/UV detector in the PEFAC implementation with a threshold greater 0.99 is used. According to the recommendation of [66] a Blackman window is used to obtain the phase spectrum. Frames are 24 ms long and the frameshift is determined at 3 ms. The UnHPSNR measure has a lower bound of -3 dB and reaches infinity for perfect phase reconstruction. To make the UnHPSNR more credible the upper bound is set to 25 dB. This leads to a dynamic range of 28 dB which is in alignment with other conventional speech intelligibility instrumental measures.

3.2.2 Unwrapped Root Mean Square Error (UnRMSE)

The unwrapped root mean square error is beside the UnHPSNR the second measure that quantifies the estimation error occurring in the unwrapped phase, introduced by the phase modification procedure. It is defined as

$$\text{UnRMSE} = 10 \cdot \log\left(\frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{\sum_{k=1}^{K/2} X^2(k,l)(\Psi_x(k,l) - \hat{\Psi}_x(k,l))^2}{\sum_{k=1}^{K/2} X^2(k,l)}}\right). \tag{3.35}$$

and follows the same computation for the unwrapped phase $\Psi(h,l)$ and the framing, as discussed above in Section 3.2.1. The clean phase attains the minus infinity value and is lower bounded to 0 dB. Together with the theoretical upper bound, UnRMSE attains values between $[0,5]$ dB.

4 Listening Test

Attention has to be drawn to a careful design of the subjective listening test to obtain useful results. This not only involves the test design itself but also the choice of the participants, the used speech database and the used benchmark methods. Following [84] a satisfactory measurement method is required to meet the six following characteristics as listed in [41]:

Objectivity	The test results are reproducible (verifiable) over different listeners (inter-subjectivity).
Reliability	The test results show no large scattering when a stimulus is repeated to the same listener (intra-subjectivity).
Validity	The parameter measured by the test is the one intended to be measured.
Sensitivity	The distinctions enabled by the test are as fine as those made by the listener.
Comparability	The test is applicable to a wide range of benchmark methods and makes possible comparisons between groups of conditions.
Utility	The pieces of information provided by the listening test are useful.

Sections 4.1 and 4.2 introduce the speech database and benchmark methods utilized in the perceived quality and intelligibility listening tests presented in Sections 4.3 and 4.4. Both tests were conducted in a quiet environment at Graz University of Technology using AKG K 601 High-End Stereo Headphones. The participants were supposed to be normal-hearing given the information of the test questionnaire, but it has to be mentioned that no screening of their auditory system had been undertaken. The perceived quality and intelligibility listening tests were held on different days with different participants to some extent.

4.1 Speech Material

The test material was taken from the GRID corpus [28]. It is a free large multitalker audiovisual corpus that supports evaluations in an automatic speech recognition (ASR) and speech perception context. The corpus consists of 34,000 phonetically balanced high-quality audio and video recordings spoken by 34 native-english speaking talkers (18 male, 16 female). Each sentence exhibits a six word command like structure as presented in Table 4.1. The color, letter, and digit are the *"key words"* used for perceptual listening tests. Each talker produced all possible

combinations of these components, leading to 1000 sentences per talker in total. By performing intelligibility tests on the clean speech utterances in [28] it is suggested that the utterances are easily identifiable under quiet testing conditions.

Table 4.1: Sentence structure for the GRID corpus.

command	color	preposition	letter	digit	adverb
bin	blue	at	A-Z excluding W	1-9, zero	again
lay	green	by			now
place	red	in			please
set	white	with			soon

Out of this database, 50 sentences were chosen for the setup of the upcoming quality and intelligibility listening tests in Sections 4.3 and 4.4 including female and male speakers. A detailed list of the utterances used in the listening tests is provided in Appendix A. The utterances were downsampled to 8 kHz.

4.2 Benchmark Methods

The benchmark methods were chosen to be representative for a phase-aware speech enhancement framework. Therefore, all different types of speech enhancement methods incorporating the noisy, some enhanced, and clean phase introduced in Section 1.1 were considered to be in the analysis. Table 4.2 gives an overview over the benchmark methods with the according abbreviation used in the following and marks, whether the method is used in the perceived quality (QT) or intelligibility (IT) test. For details about the different methods please refer to the literature quoted in Table 4.2. A short description is given below.

Table 4.2: List of benchmark methods used in the objective and subjective analysis

Benchmark method	Type	Abbr.	used in QT	used in IT
Clean Utterance	Reference	Ref	x	x
Noisy utterance	Unprocessed	UP	x	x
Minimum mean-square-error log-spectral estimator (MMSE-LSA) [24]	Conventional	C	x	x
MMSE-LSA + phase estimation using phase decomposition [20]	Phase-enhanced	C + PE	x	x
MMSE-LSA + STFT phase improvement [21]	Phase-enhanced	C + STFTPI		x
MMSE-LSA + clean phase	Phase-enhanced	C + clean	x	
MMSE-LSA + iterative closed-loop phase-aware speech enhancement [3]	Phase-aware	C + PA	x	x

The noisy utterances (unprocessed, UP) were produced by mixing the clean utterances with white and babble noise files taken from NOISEX-92 database [85] at specific SNRs. These files were then enhanced by the conventional minimum mean-square error log-spectral estimator (MMSE-LSA) proposed by Ephraim and Malah, denoted as Conventional (C) [24]. The method applies a frequency-dependent gain function on the noisy DFT coefficients given *a priori* and

a posteriori SNR estimates and employs the noisy phase at the signal reconstruction stage. The noise estimate is given by the improved minima controlled recursive averaging (IMCRA) [86] noise estimator.

On top of the conventional approach, the phase-based enhancement algorithms replace the noisy phase by an improved phase. The upper bound is fixed by the conventional method that employs the clean phase, extracted from the clean utterance, at the signal reconstruction (C + clean).

Short time phase improvement (STFTPI) [21] transforms the instantaneous phase to its baseband representation

$$\phi_{\hat{X}_B}(k,l) = \angle \hat{X}_B(k,l) = \angle \hat{X}(k,l) \mathrm{e}^{-j\frac{2\pi k}{N} klF}, \tag{4.1}$$

where $\hat{X}(k,l)$ denotes the enhanced amplitude spectrum at frequency bin k and time frame l and F is the frameshift of two consecutive time frames. Phase reconstruction is done at voiced frames by using a recursive computation of the baseband STFT-phase along time and frequency given a harmonic signal model and assuming that a STFT bin k is dominated only by the closest harmonic. The accuracy of the method depends on a reliable fundamental frequency estimator to obtain the harmonic frequencies. In their implementation they used the YIN estimator [87] for fundamental frequency estimation.

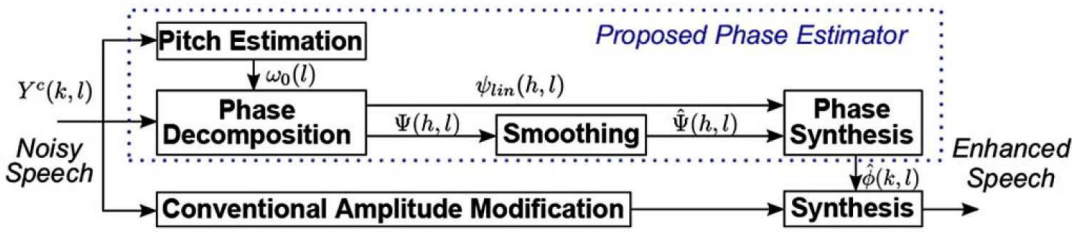

Figure 4.1: Phase estimation using phase decomposition and temporal smoothing, [20].

Figure 4.1 illustrates the block diagram of the phase estimator relying on phase decomposition (C + PE), proposed in [20]. The method decomposes the instantaneous phase $\psi(h,l)$ at harmonics h into an unwrapped and linear phase part as described in Section 3.2.1, using pitch-synchronous segmentation and PEFAC [83] for f_0 estimation. The unwrapped phase $\Psi(h,l)$ is modelled by a von Mises distribution [88] defined as

$$\mathcal{VM}(\mu_c(h,l), \kappa(h,l)) = \frac{\mathrm{e}^{\kappa(h,l)\cos(\Psi(h,l)-\mu_c(h,l))}}{2\pi I_0(\kappa(h,l))}, \tag{4.2}$$

where $\mu_c(h,l)$ and $\kappa(h,l)$ denote the circular mean and concentration. Phase enhancement is achieved by smoothing $\Psi(h,l)$ along time frames to reduce the phase variance at harmonics given by

$$\hat{\Psi}(h,l) = \angle \sum_{l'=l-R/2}^{l+R/2} \mathrm{e}^{j\Psi(h,l')}, \tag{4.3}$$

with R as the number of frames within 20 ms time span. Before reconstruction, the linear phase part is added back to the enhanced unwrapped phase leading to the enhanced instantaneous

phase

$$\hat{\psi}(h,l) = \hat{\Psi}(h,l) + \psi_{\text{lin}}(h,l). \tag{4.4}$$

The iterative phase-aware approach, denoted as (C + PA) [3], is shown in Figure 4.2. The first stage consists of a conventional amplitude estimator [24] followed by a phase estimator, that minimizes the geometric representation of the single-channel speech enhancement problem (see Figure 2.5) at spectral peaks forcing a group delay deviation constraint to resolve the ambiguity in phase, as presented in [8]. The enhanced phase is then used as input to the spectral amplitude estimator given by:

$$\hat{X}_{\hat{\phi}_x}(k,l) = \sqrt{\frac{2}{\beta_1}}\frac{D_{-2}(z)}{D_{-1}(z)}, \qquad \text{where} \quad z = -\frac{2Y\cos(\phi_y - \hat{\phi}_x)}{\sqrt{2\beta_1\sigma_v^2}} \tag{4.5}$$

where $D_{-\nu}(\cdot)$ is the parabolic cylinder function of order ν and $\beta_1 = 1/\sigma_v^2 + 1/2\sigma_{\hat{x}}^2$ where $\sigma_{\hat{x}}^2$ and σ_v^2 denote the enhanced speech and estimated noise power spectral densities with complex Gaussian distribution for the joint distribution of Y and ϕ_Y. A complex spectrogram $\hat{X}^{c,(i)}$ is build by the enhanced amplitude and phase to enter the next iteration i. The iteration is stopped if convergence according to the inconsistency constraint [6], defined in the complex domain as $\text{F}(X^{c,(i)}) = \text{STFT} \circ \text{iSTFT}(X^{c,(i)}) - X^{c,(i)}$, is reached.

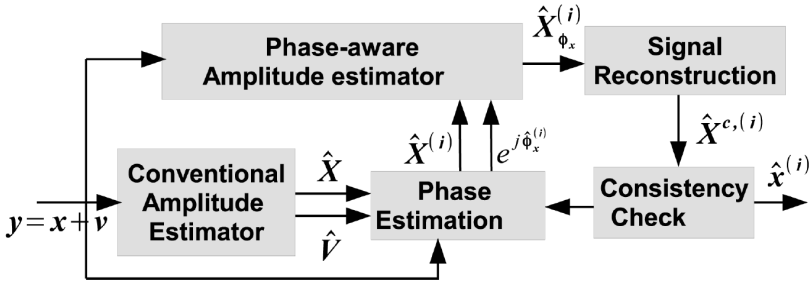

Figure 4.2: Block diagram of the closed-loop single-channel speech enhancement algorithm, [3].

4.3 Quality Listening Test

4.3.1 Setup

A panel of 11 listeners were recruited to participate in the quality test. All participants were students from Graz University of Technology with an age between 19 and 29 years and got paid for their participation. The test database described in Section 4.1 was corrupted with white and babble noise at SNRs of 0, 5, and 10 dB to obtain the unprocessed noisy stimulus (UP) and was processed afterwards by four speech-enhancement algorithms:

- Conventional (C)
- Conventional + clean phase (C + clean)
- Conventional + phase-enhanced (C + PE)
- Conventional + phase-aware (C + PA)

introduced in Section 4.2. Following the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) standard [89], a Hidden Reference as well as an Anchor was included. The Anchor is defined as the low-pass filtered reference signal (Ref) with a cut-off frequency of 2.5 kHz. This choice is different to the original standard that uses an Anchor with $f_c = 3.5$ kHz, because the clean reference signal is already downsampled to 8 kHz. The lower cut-off frequency guarantees a perceptual difference to the original clean reference signal. The filter is designed as a FIR equiripple filter of order 34 with a stopband attenuation of 50 dB and a bandpass ripple < 0.1 dB. Together with the Hidden Reference and the Anchor, seven benchmark methods were used in the listening test.

The graphical user interface is shown in Figure 4.3. Each participant had to evaluate 12 sample-sets chosen out of the drop-down menu. Every sample-set consisted of the open reference and the seven benchmark methods mentioned above. The first six sample-sets were corrupted with white noise and the last six sample-sets were corrupted with babble noise both with increasing SNRs from 0 to 10 dB, leading to two sample-sets per test condition and participant. All utterances were chosen randomly out of the test database and the methods were scrambled independently at each sample-set. Participants were asked to rate the perceived quality on a scale from 0 (bad) to 100 (excellent), where the main attention had to be drawn to the noise reduction and the introduced distortions. To simplify the ranking procedure between the benchmark methods, a *"sort ranking"* button was provided that sorted the rankings from low (left) to high (right). Participants had the possibility to switch forth and back the 12 sample-sets to also adjust the ranking across SNRs and noisy types. The duration of the test was 45 minutes on average. No training phase was provided in this test.

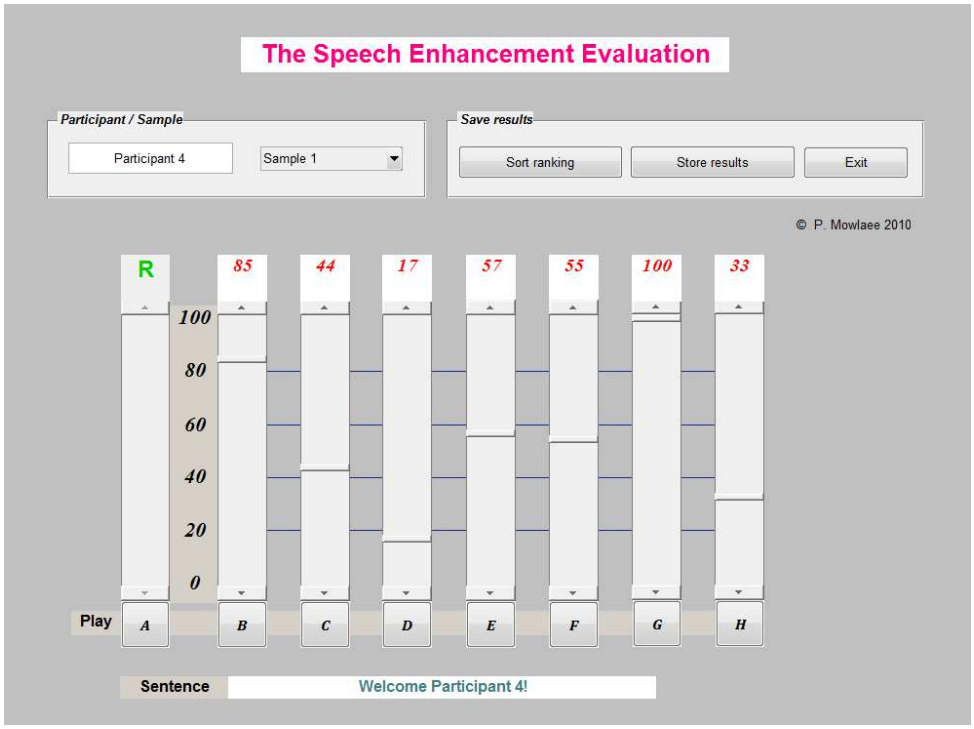

Figure 4.3: Graphical user interface of the perceived speech quality listening test.

4.3.2 Test Results

Figure 4.4 illustrates the Mean Opinion Scores (MOS) and 95% confidence intervals averaged over the 11 participants. The results are differentiated by the test conditions, showing white

noise results at SNRs of 0, 5, and 10 dB at the top and babble noise results accordingly at the bottom. For all noise types and SNRs the same ranking is observed: The unprocessed condition (UP) determines the lower bound followed by the conventional method (C). All methods that incorporate phase enhancement perform better than C, where the iterative phase-aware approach (C + PA) performs on top followed by the phase-enhanced using the clean phase (C + clean) and the phase-enhanced using the estimated phase (C + PE). At this point it has to be mentioned that the iterative approach outperforms the phase-enhanced approach utilizing the clean phase even though it does not have knowledge about the clean phase.

Paired-sample t-tests were conducted to justify the significance of these rankings. Except between C and C + PE, all other rankings were significant with respect to each other with $p < 0.05$. However, C + PE outperforms C significantly for the white noise scenario at SNR = 0 dB as well as at SNR = 5 dB with $p = 0.077$ suggesting that a higher number of participants would also lead to a significant result in this condition. In general, phase enhancement benefits more at low SNR scenarios (0, 5 dB) also visible for babble noise compared to the conventional method. This is confirmed by Vary in [63] where he stated that at local SNRs greater than 6 dB the noisy phase provides a good estimate.

The comparison between white noise and babble noise points out that degradations by babble noise are perceived less annoying by the participants. Babble noise has a lower power spectral density (PSD) than white noise and therefore is percepted at a lower loudness level at equal SNR. On the other hand the overall perceptual quality improvement between UP and C + PE is more pronounced in white noise than babble noise while the improvement between UP and C is approximately the same for both noise types. This is because the phase estimation stage of the PE method [20] relies on f_0 estimation [83] used for phase decomposition which is more accurate in white noise.

Finally the observation along SNRs shows consistent perceived quality improvement for each method and increasing SNRs.

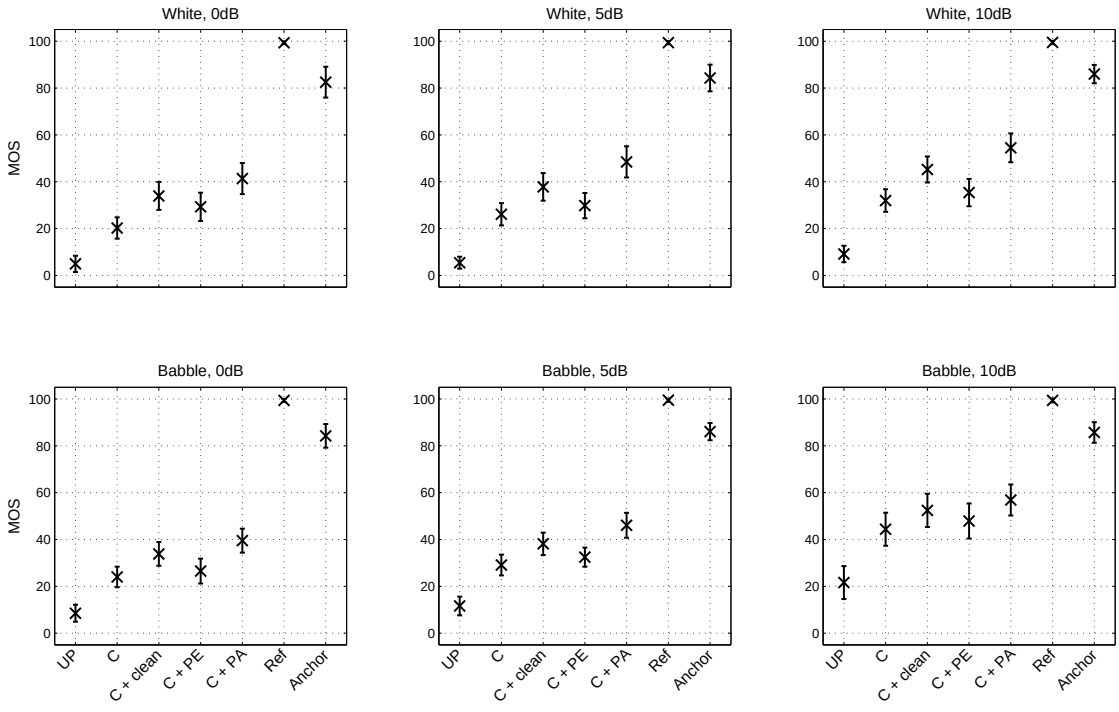

Figure 4.4: Mean Opinion Scores (MOS) of the MUSHRA test for (top) white and (bottom) babble noise scenario shown for eleven participants.

4.4 Intelligibility Listening Test

4.4.1 Setup

A group of 12 listeners participated in the subjective intelligibility test. The group consisted of people that also participated in the quality test as well as some that did not participate in the quality test and were students from Graz University of Technology with an age between 19 and 29 years. Payment was offered to the participants after the test, which had an average duration of 30 minutes. The test database was the same as used in the quality test described in Section 4.1 and was corrupted with white and babble noise at SNRs of 0 and 5 dB to obtain the unprocessed noisy speech signals (UP). Four speech-enhancement scenarios were included similar to the quality test:

- Conventional (C)
- Conventional + STFT Phase Improvement (C + STFTPI)
- Conventional + phase-enhanced (C + PE)
- Conventional + phase-aware (C + PA)

that we already explained in Section 4.2. Together with the clean reference (Ref) and the unprocessed signal (UP), this leads to six benchmark methods used in the test. The test procedure followed the standard described in [90].

Figure 4.5 shows the graphical user interface used to collect the inputs from the participants. The participants were instructed to choose the right "key words" (colour, letter, number) at each presented utterance and had the possibility to play the utterance several times. As an additional aid, the possibilities of each key word as well as the structure of the GRID sentence itself were visualized on the GUI.

The test was organized in four blocks according to noise type and decreasing SNR, shown in Table 4.3. Within each block, for each benchmark method four randomly selected utterances were presented to the participants, where the order of the benchmark methods itself also was randomized. To check the reliability of the participants, at each block two clean reference utterances were included. Participants were rejected if their intelligibility scores of the clean utterances were lower than those of the noisy utterances. This procedure led to ten participants left in the analysis.

Table 4.3: Block organization of the intelligibility listening test.

Block Nr.	Noise type	SNR (dB)
1	White	5
2	Babble	5
3	White	0
4	Babble	0

A training session was provided built up by two steps. In the first step, a small set of clean utterances were presented to the participants to get familiar with the structure of the GRID sentences. The second step was intended to get familiar with the GUI by rating eight examples that were designed to be representative for the noise types and benchmark methods used in the listening test. The training database was independent of the test database and is given in Appendix A.2.

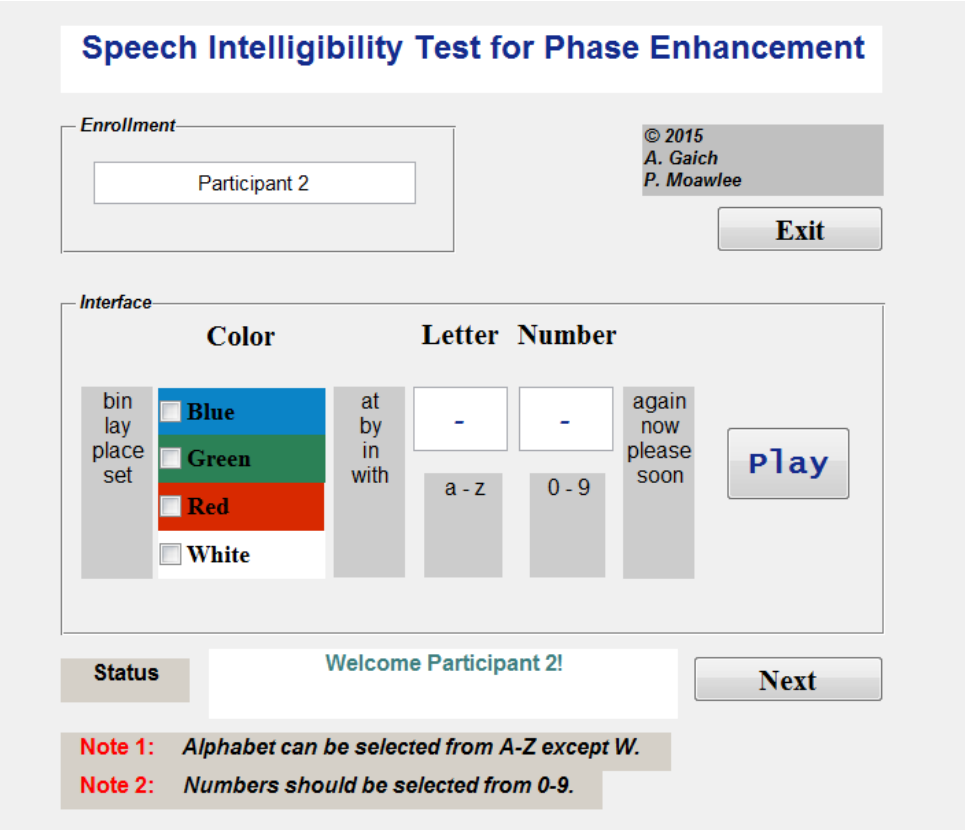

Figure 4.5: Graphical user interface of the intelligibility listening test.

4.4.2 Test Results

The speech intelligibility scores and 95% confidence intervals, obtained from the subjective listening test averaged over ten participants, are shown in Figure 4.6. The scores were evaluated by adding the errors over all three keywords (colour, letter, number) and are differentiated in terms of noise types and SNRs. At same SNR scenarios the intelligibility scores are higher for babble noise rather than white noise except for C + STFTPI and C + PE at SNR = 0 dB. This observation is in alignment with the observation made in the quality test in Section 4.3.2. The benchmark methods C + STFTPI and C + PE replace the noisy phase by an enhanced one using a harmonic representation. Therefore a reliable estimate of the fundamental frequency is necessary which is an erroneous task in an adverse noise condition as babble 0 dB, leading to worse intelligibility scores. This fact also supports the better improvement in intelligibility for white noise rather than babble noise of C + PE compared to C also observed in the quality listening test in Section 4.3.2.

The iterative method (C + PA) outperforms all the other benchmark methods in every test condition. However, this improvement is only significant for babble noise at 0 dB. A two-proportion z-test was conducted to calculate the significance at the 95% confidence level. PA also showed significant improvement in comparison to C except for babble noise at 5 dB.

In addition, observing Figure 4.6, better performance in speech intelligibility is only achieved by benchmark methods that also modify the spectral phase (C + PE, C + PA) in comparison to the unprocessed speech signals (UP) while the conventional method (C) degrades the speech intelligibility, showing the potential of single-channel phase enhancement. The conclusion that conventional speech enhancement algorithms that only modify the spectral amplitude degrades the speech intelligibility is confirmed by observations reported in [91, 92].

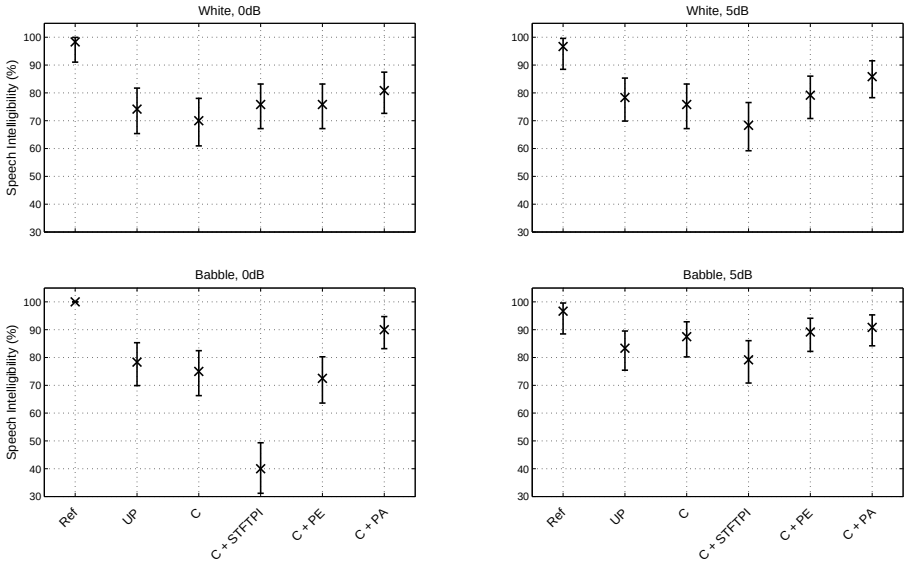

Figure 4.6: Intelligibility scores showing the mean and 95% confidence interval for (top) white and (bottom) babble noise scenario averaged over ten participants.

A similar trend of the intelligibility scores is illustrated in Figure 4.7. Here the results are separated by the three key words, i.e., colour, letter, and number, averaged over all noisy types and SNRs. Colours were most intelligible with scores above 90% followed by the numbers and letters. The low intelligibility scores of the letters can be explained by the misidentification of e.g. /v/ with /b/ and /p/ which differ only by their onsets while showing a very related harmonic structure. The onsets carry less energy and are therefore more likely to be masked by additive noise. The same confusion happened to /m/ and /n/. This observation was also made in [28].

C + PA performs on top showing significance compared to C in each case and showing significance to every other benchmark method for letter and number intelligibility. As stated before, C degrades the speech intelligibility. The performance of C + PE is slightly better than UP, however no significance is determined. Test results differentiated by noise types and SNRs are available in Appendix B.1.

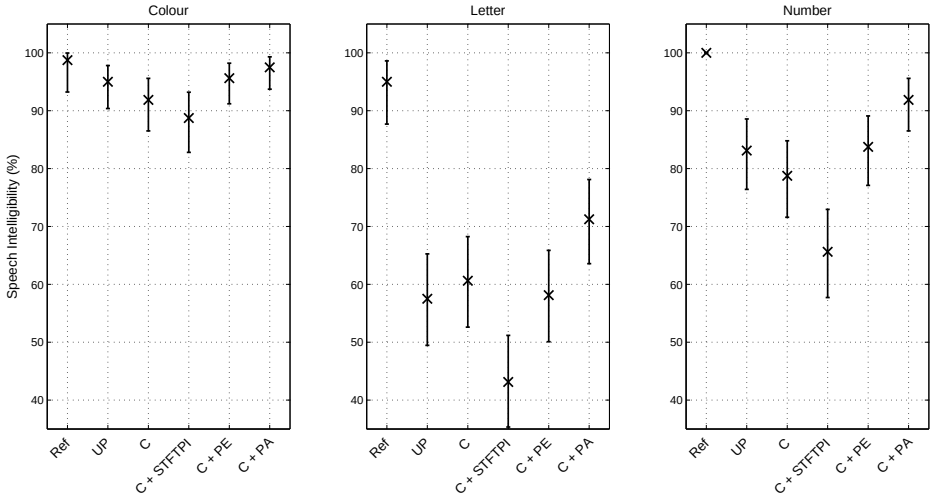

Figure 4.7: Intelligibility scores showing the mean and 95% confidence interval separated by (left) colour, (middle) letter and (right) number averaged over all noise types and SNRs for ten participants.

5 Performance Evaluation

The instrumental measures presented in Chapters 2 and 3 have to be evaluated in terms of how well they reflect the results obtained from the subjective listening tests, presented in Chapter 4. A valid instrumental measure has to reliably predict the subjective scores over a wide range of distortions. To assess the performance of the instrumental measures in a phase-aware single-channel speech enhancement framework, the processing conditions are adapted to be focused on the spectral phase. Therefore, different benchmark methods that modify the spectral phase are included in the evaluation process in addition to the conventional ones. The unprocessed conditions are considered for two different noise types, a stationary noise (white noise) and a non-stationary noise (babble noise), at SNRs of 0, 5, and 10 dB being representative for a speech enhancement scenario in real life.

In the speech enhancement community, a common methodology to evaluate the correlation between subjective scores and the values of the instrumental measures is statistical analysis. We follow the approach described in [93] by computing the normalized correlation coefficient ρ, the root mean squared error σ, and Kendall's tau τ. A normalization and mapping procedure is included to account for the gain variations at the output of the different enhancement algorithms and the non-linear relation between the subjective and objective scores. The evaluation procedure is explained in detail in Sections 5.1 - 5.3.

Perceived quality and intelligibility performance results of the instrumental measures are then presented in Sections 5.4 and 5.5. These results were first published in less detail in [65, 94]. Here, the evaluation is extended to also include the conventional and proposed quality measures in the intelligibility performance assessment as well as to include the conventional and proposed intelligibility measures in the quality performance assessment to get an overall picture of the performance of each instrumental measure.

5.1 Mapping

The evaluation criteria described in Section 5.3, used for the performance analysis, assume that the subjective and instrumental scores are linearly related. This is in general not the case and a fitting function is necessary that describes the relationship between the subjective and the instrumental measurements. In the literature different methodologies are described that provide a reasonable fit, e.g., using a quadratic relationship [95] or a logistic function [96]. We follow the methodology in [93] and apply a logistic function to the average instrumental scores $\bar{d}$, defined

as

$$F(\bar{d}) = \frac{1}{1 + \mathrm{e}^{a\bar{d}+b}}, \tag{5.1}$$

where a and b are two free parameters adapted in order to fit the subjective scores. The average instrumental scores $\bar{d}$ are obtained by averaging the individual distance outcomes over the selected speech database (50 sentences from the GRID corpus) for each processing condition.

Figure 5.1 illustrates an example of the mapping procedure for the PD measure. Subplot (a) shows the mean opinion scores (MOS) gathered from the subjective perceived quality listening test as a function of the processing conditions. The processing conditions, illustrated by stars, include noise types, SNRs, and benchmark methods (without clean and anchor) leading to $2 \times 3 \times 5 = 30$ processing conditions where the MOSs are ranked in an ascending order. According to these 30 processing conditions the appropriate averaged instrumental scores $\bar{d}_{PD}$ are shown in subplot (b) and are mapped to the subjective scores by the logistic function $F(\cdot)$ illustrated in subplot (c). Subplot (d) shows that the instrumental scores are linearised after the mapping and can be used for calculation of the Pearson's correlation coefficient, described in Section 5.3.

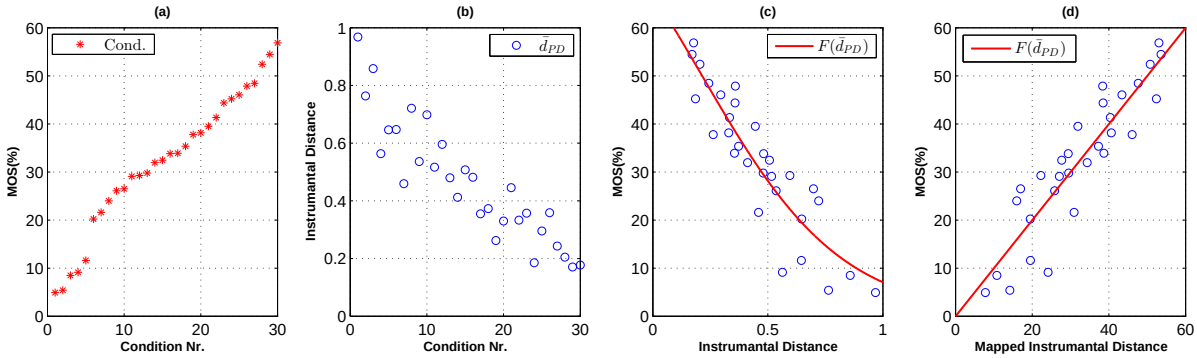

Figure 5.1: Example for the mapping procedure for the PD measure.

Eq. (5.1) can be used if the instrumental measure exhibits a linear relation with the SNR. Some instrumental measures are exponentially related with the SNR and therefore need a different mapping

$$F_{\mathrm{exp}}(\bar{d}) = \frac{1}{1 + \mathrm{e}^{a \cdot \ln(\bar{d}+c)+b}}, \tag{5.2}$$

where c is an extra parameter introduced. To find the free parameters, a non-linear least squares procedure *(MATLAB: nlinfit)* is used for each individual instrumental measure.

5.2 Normalization

Due to the processing of the degraded speech signal the energy of the enhanced speech signal $\hat{x}(n)$ can be significantly different compared to the energy of the clean speech $x(n)$. While this difference will not affect the subjective perceived quality and intelligibility judged by the human listeners, some instrumental measures are sensitive to it. For instance, the SNR-based measures including GSNR and SSNR computed in the time domain are calculated by a sample-by-sample comparison between the clean and the enhanced speech signals. A global difference in the gain of the clean and enhanced speech signal will therefore lead to a lower SNR score.

To avoid this loss in prediction performance, a global normalization scheme is considered where the speech energy of $x(n)$ and $\hat{x}(n)$ are equalized. A normalization factor α is computed using the following equation

$$\alpha = \sqrt{\frac{\sum_{n=1}^{N} x(n)^2}{\sum_{n=1}^{N} \hat{x}(n)^2}}, \tag{5.3}$$

where N denotes the total number of samples. The normalized enhanced speech signal is then given by

$$\hat{x}_{\text{normalized}}(n) = \alpha \cdot \hat{x}(n). \tag{5.4}$$

This normalization is applied to every enhanced speech signal prior to the computation of the instrumental scores. Note, that this normalization procedure does not affect or replace the normalization procedures that are inherent in some existing conventional instrumental measures, i.e., PESQ and STOI.

5.3 Evaluation Criteria

After linearisation, as described in Section 5.1, the correlation between the subjective listening scores and the scores of the instrumental measures can be obtained using the normalized Pearson's correlation coefficient defined as

$$\rho = \frac{\sum_c (S_c - \bar{S})(O_c - \bar{O})}{\sqrt{\sum_c (S_c - \bar{S})^2 \sum_c (O_c - \bar{O})^2}}, \tag{5.5}$$

where S_c and $O_c = F(\bar{d})$ denote the subjective and mapped objective scores at the c^{th} processing condition and $\bar{S}$ and $\bar{O}$ represent the average values computed over all processing conditions, respectively. The numerical value of ρ is defined between $[0, 1]$ where 1 indicates a high correlation.

The second figure of merit is the root-mean-square error (RMSE) between the subjective and the mapped objective scores computed over all conditions, given by:

$$\sigma = \frac{1}{100} \sqrt{\frac{1}{C} \sum_c (S_c - O_c)^2}, \tag{5.6}$$

where C denotes the total number of processing conditions. It is normalized in the range between $[0, 1]$ and provides information how the scores obtained from the instrumental measures scatter around the true subjective scores obtained from the listening tests. Hence it is a measure of accuracy.

Finally, Kendall's tau is included defined as

$$\tau = \frac{N_c - N_d}{\frac{1}{2} C(C-1)}. \tag{5.7}$$

Here N_c and N_d are the number of concordant and discordant pairs in the evaluated set of processing conditions. Let $(S_1, O_1), (S_2, O_2), \cdots, (S_C, S_C)$ be the set of observations of the

subjective and corresponding mapped objective scores. Any pair of observations (S_i, O_i) and (S_j, O_j) are said to be concordant if the ranks for both elements agree; that is, if $S_i > S_j$ and $O_i > O_j$ or if $S_i < S_j$ and $O_i < O_j$. The pairs are discordant, if $S_i > S_j$ and $O_i < O_j$ or if $S_i < S_j$ and $O_i > O_j$. Kendall's tau is a rank correlation coefficient that tests whether there is a monotonic relation between the subjective and objective scores and is independent of the selected mapping. The value lies in the range of $-1 < \tau < 1$ where -1 and 1 define a perfect disagreement/agreement between the two rankings.

The three figures-of-merit introduced above have a direct relation to the criteria discussed in Chapter 2 which define a good instrumental measure. While ρ and σ determine the *Accuracy*, τ reveals the *Consistency* of an instrumental measure.

5.4 Perceived Quality Results

For each instrumental measure, the correlation coefficient, the RMSE, and the Kendall's tau are shown in Figure 5.2. The results are differentiated in the noise type and ranked in terms of the correlation coefficient increasing from left to right. Also the scatter plots of the combined noise scenario between the mapped objective scores and the mean opinion scores are illustrated in Figure 5.3, where in the legend (W-) and (B-) denote white and babble noise results. For further details, the scatter plots of the mapped results separated in terms of the noise type as well as the scatter plots of the unmapped results together with the fitted mapping functions are provided in Appendices B.2.1 and B.2.2, respectively.

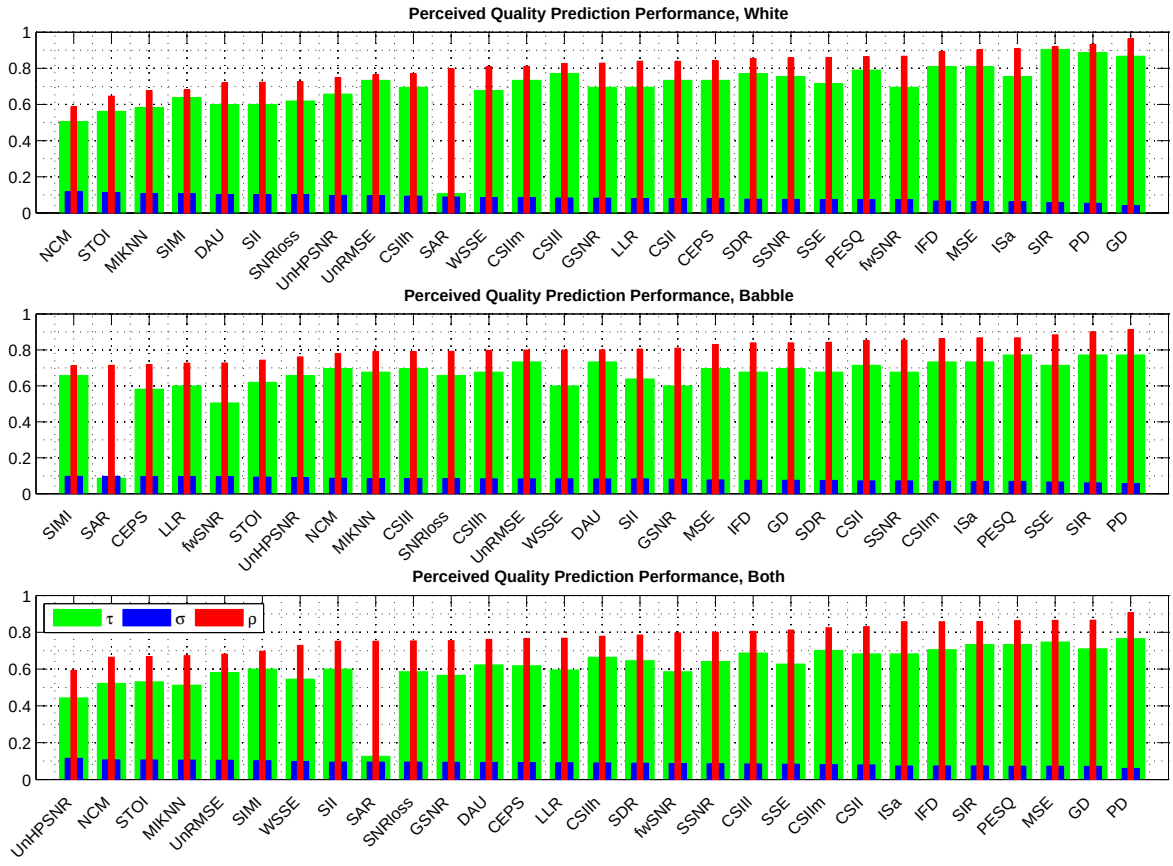

Figure 5.2: Perceived quality performance evaluation of the instrumental measures categorized to (top) white, (middle) babble and (bottom) both noise types.

As expected, most of the instrumental measures designed for predicting speech intelligibility showed a poor performance ($\rho < 0.8$) in predicting the perceived quality. These measures include SII, CSIIh, SNRloss, NCM, DAU, STOI, MIKNN, and SIMI in the set of the conventional measures and UnHPSNR and UnRMSE in the set of the proposed phase-aware measures. The rationale behind this is a significant overprediction of the perceived quality of the unprocessed speech signals as observed in Figure 5.3. This is as a consequence of that the unprocessed (noisy) speech signals corrupted by some noise at low SNRs are perceived as low perceived quality while they can be still quite intelligible. The CSII-based measures, except CSIIh, performed at a moderate level ($0.8 < \rho < 0.85$) showing a similar performance supported by the fact that they correlate well to each other as reported in [71].

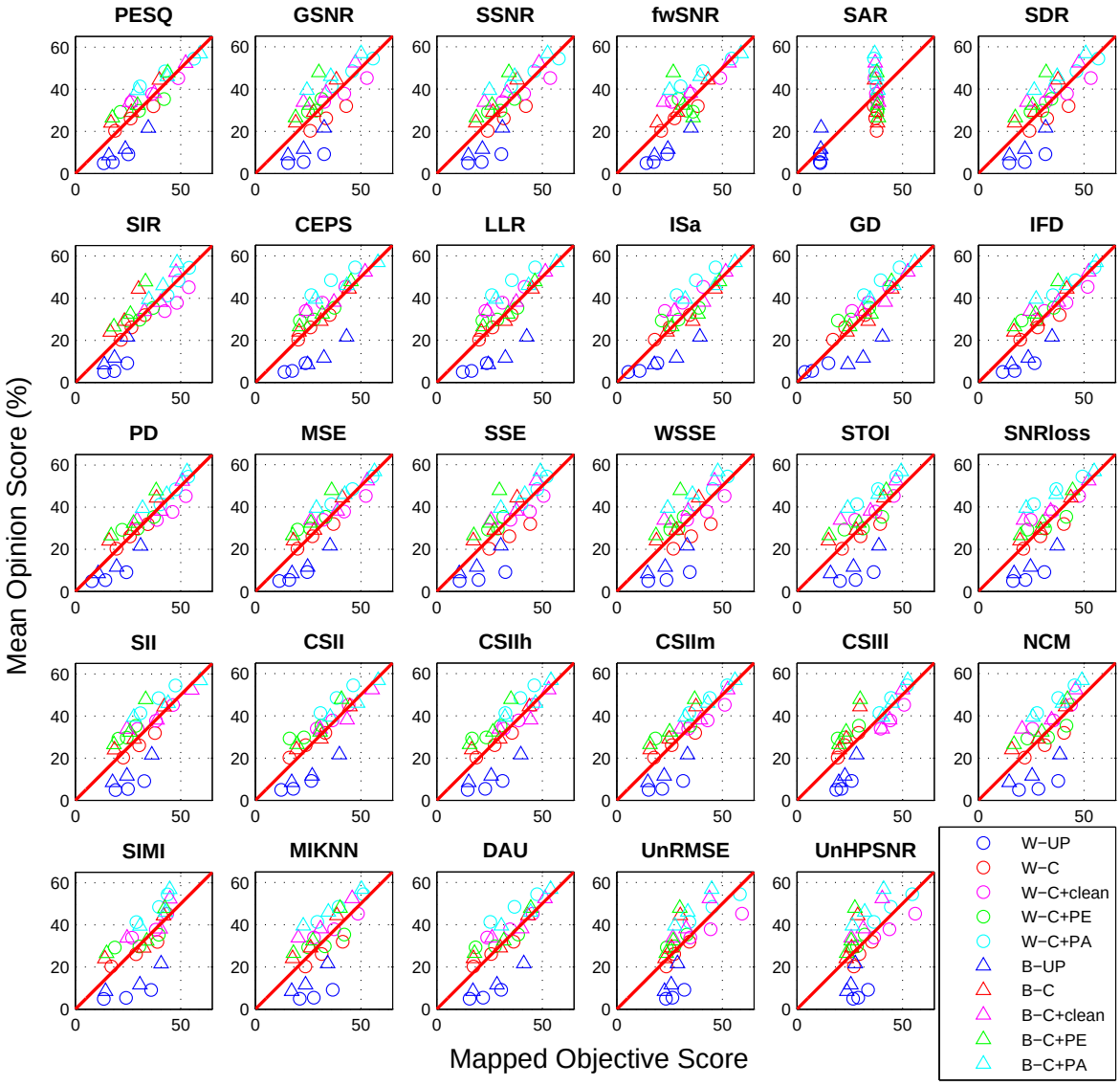

Figure 5.3: Mapped scatter plots of the perceived quality predictions for all instrumental measures in a combined white and babble noise scenario at 0, 5, 10 dB.

A reasonable performance was obtained by the proposed phase-aware quality measures GD, IFD, PD and MSE which all performed better than the conventional ones: GSNR, fwSNR, SAR, LLR and CEPS in every noise scenario. The PD measure exhibited high correlation across noise types ($\rho > 0.9$) and outperformed all the conventional measures, including PESQ, in all noise scenarios. It predicted the unprocessed conditions with better accuracy than any other measure. In general, the unprocessed conditions were overestimated by all instrumental measures in comparison to the enhanced speech conditions. The GD measure was the most reliable predictor in white noise but showed significant degraded performance in babble noise. This could be explained due to the robustness of the group delay representation against additive white noise, as reported in [97]. The proposed quality measures SSE and WSSE that also incorporate spectral amplitude information showed a moderate performance, where the simpler SSE model obtained a higher correlation. Out of the SNR-based measures the SIR showed the highest correlation supported by the observation in [98] with similar performance to ISa and PESQ.

Table 5.1 summarizes the measures that revealed a correlation $\rho \geq 0.8$ in each noise scenario. The mean values of the instrumental scores in the last column of the table were computed by averaging over the three different noise scenarios. The PD measure obtained the highest correlation ($\rho = 0.92$) on average followed by SIR and GD ($\rho = 0.89$). The intelligibility measures CSII and CSIIm were at the bottom of this ranking. A reliable metric should also perform stable across different conditions. Stable performance was observed for PD, PESQ and CSII where only PD showed a reasonable high correlation. Furthermore, the PD measure exhibited the best performance in terms of σ and τ. All together, PD showed reasonable performance in terms of accuracy and consistency.

Table 5.1: Statistical analysis of the top performing measures showing $\rho \geq 0.8$ for each noise scenario.

	White noise			Babble noise			Both noises			**Mean**		
	ρ	σ	τ	ρ	σ	τ	ρ	σ	τ	ρ	σ	τ
PESQ	0.86	0.07	0.79	0.87	0.07	**0.77**	0.86	0.07	0.73	0.86	0.07	0.77
SSNR	0.86	0.08	0.75	0.85	0.07	0.67	0.80	0.09	0.64	0.84	0.08	0.69
SIR	0.92	0.06	**0.90**	0.90	**0.06**	**0.77**	0.86	0.07	0.73	0.89	**0.06**	0.8
ISa	0.91	0.06	-0.75	0.87	0.07	-0.73	0.86	0.07	-0.68	0.88	0.07	-0.72
GD	**0.96**	0.04	-0.87	0.84	0.07	-0.70	0.87	0.07	-0.71	0.89	**0.06**	-0.76
IFD	0.89	0.07	-0.81	0.84	0.07	-0.68	0.86	0.07	-0.71	0.86	0.07	-0.73
PD	0.93	**0.05**	-0.89	**0.91**	**0.06**	**-0.77**	**0.91**	**0.06**	**-0.76**	**0.92**	**0.06**	**-0.81**
MSE	0.90	0.06	0.81	0.83	0.08	0.70	0.86	0.07	0.75	0.87	0.07	0.75
SSE	0.86	0.07	-0.71	0.88	**0.06**	-0.71	0.81	0.08	-0.63	0.85	0.07	-0.68
CSII	0.84	0.08	0.73	0.85	0.07	0.71	0.83	0.08	0.68	0.84	0.08	0.71
CSIIm	0.81	0.09	0.73	0.86	0.07	0.73	0.82	0.08	0.70	0.83	0.08	0.72

5.5 Intelligibility Results

Figure 5.4 illustrates the performance of the speech intelligibility prediction of each instrumental measure differentiated in noise types where the top, middle, and bottom subfigures show the white, babble, and combined noise scenarios. The results are ranked by means of the correlation coefficient increasing from left to right.

A similar observation to the results in the previous section was made in terms of that most of the conventional perceived quality measures showed a low correlation ($\rho < 0.8$) in predicting the subject intelligibility results due to underestimation of the unprocessed conditions, as illustrated in the scatter plots in Figure 5.5. In particular the LPC-based measures CEPS, LLR, Isa, and the proposed phase-aware measure GD were affected to this. The other proposed perceived quality measures IFD, PD, and MSE showed a moderate correlation ($0.8 < \rho < 0.9$) but were

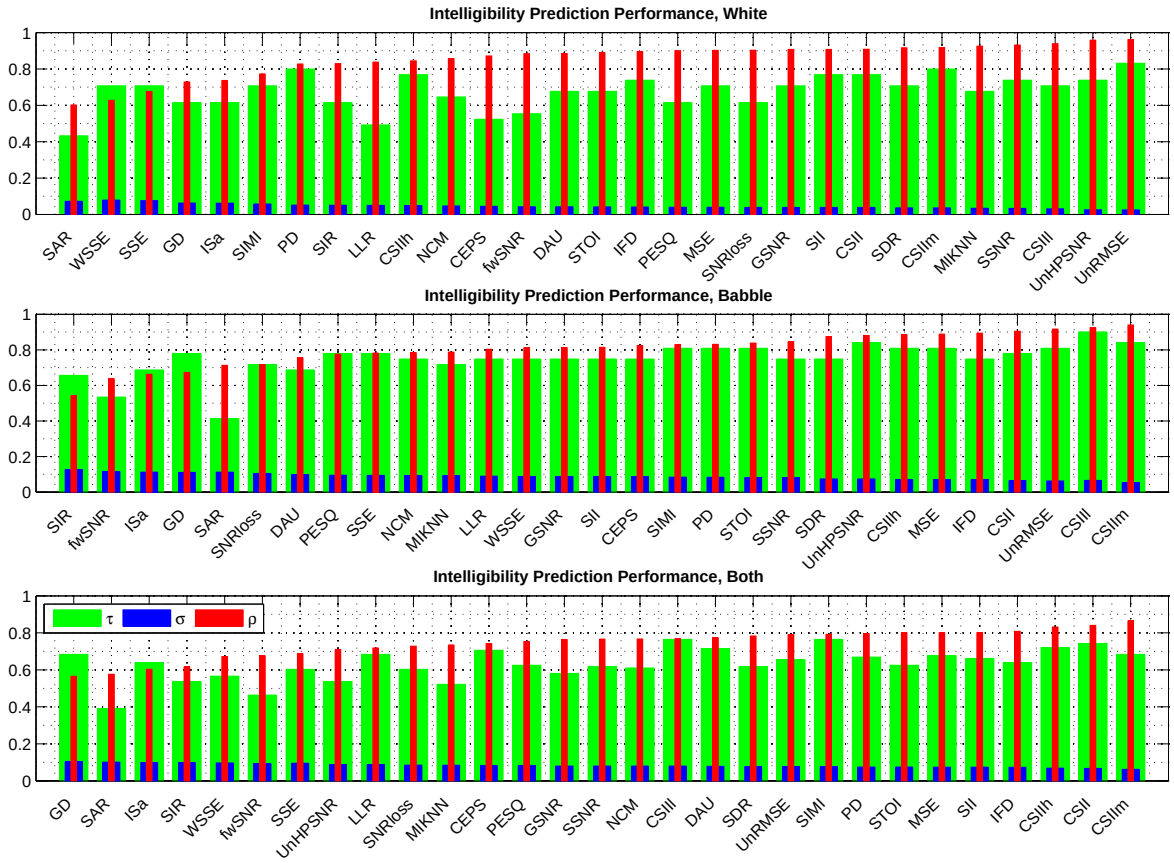

Figure 5.4: Intelligibility performance evaluation of the instrumental measures categorized to (top) white, (middle) babble and (bottom) both noise types.

not able to reach the top-performing intelligibility measures. The rationale behind this is that the quality measures compute a score based on the STFT representation without a weighting function in contrast to the intelligibility measures that use weighted band representations known to be more suitable for intelligibility prediction.

The proposed intelligibility measures UnRMSE and UnHPSNR showed the highest correlation for white noise and a reasonably high correlation for babble noise. Although UnHPSNR is calculated in the same domain as UnRMSE, it showed less correlation. This could be explained by two facts: First, in the unwrapped domain the exact difference between the unwrapped phases $(\psi - \hat{\psi})$ is more reliable than a metric which uses an additional cosine term and second, UnHPSNR computes scores only at voiced segments and neglects frames capturing spectral transitions argued to be important for speech intelligibility in [99]. Both metrics performed better for white noise than babble noise which is a result of the inaccuracy in f_0 estimation, required for phase decomposition, in the babble noise scenario. Although UnRMSE and UnHPSNR revealed a reliable performance in the separated noise scenarios, they were not able predict the subjective results in the combined noise scenario due to a consistent underprediction of the babble noise conditions and a clearly visible overprediction of the STFTPI method at 0 dB for babble noise. As illustrated in Figure 5.5 this significant overestimation was observed for every instrumental measure and explains the degraded performance for the combined noise scenario in comparison to the performance obtained in the separated noise scenarios.

The CSII-based measures exhibited reliable prediction in all the three noise scenarios. For babble noise and the overall scenario, CSIIm was the top performing metric. This is supported by the earlier observation by Kates [71] demonstrating that the mid-level CSII contains much

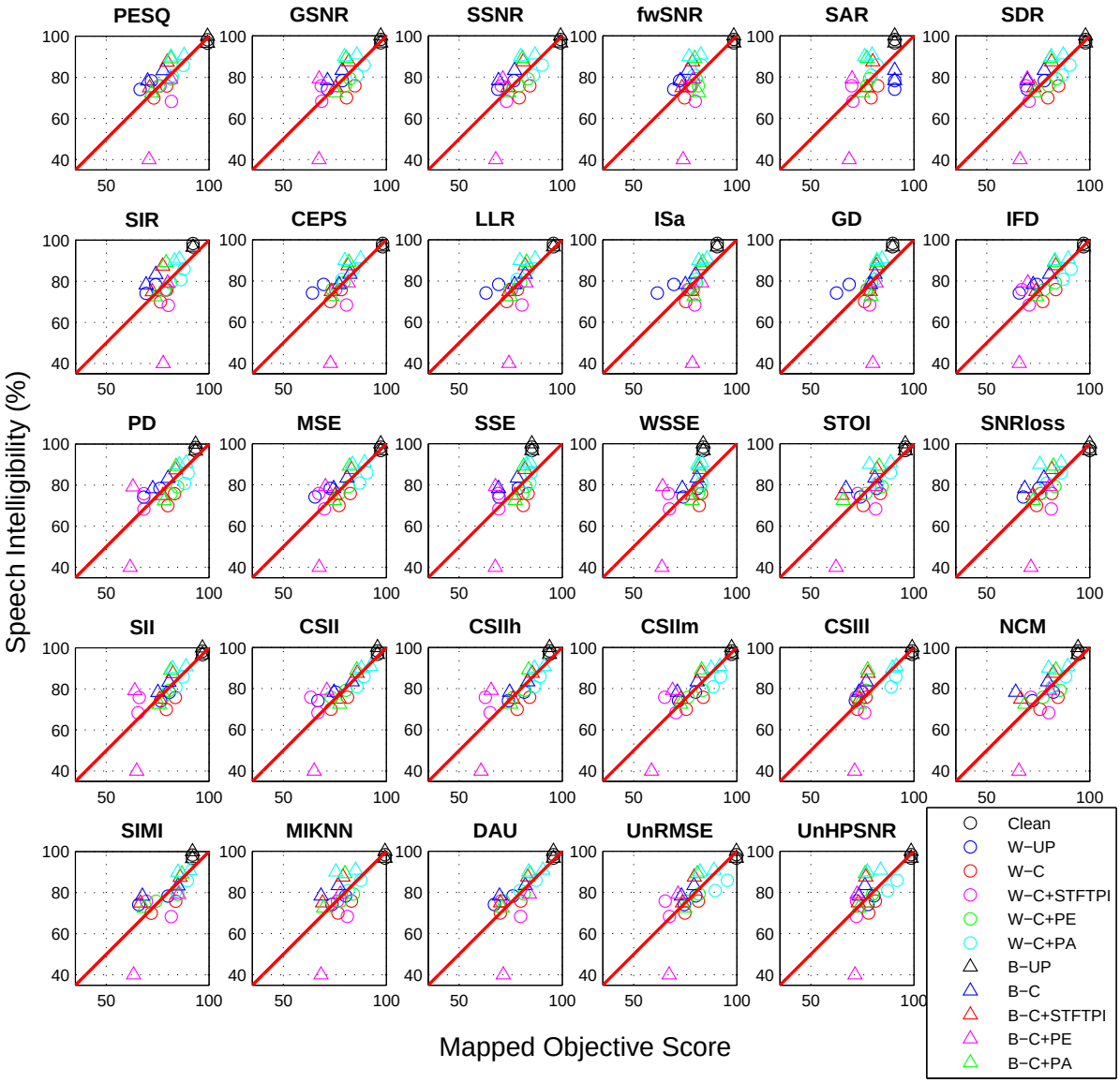

Figure 5.5: Mapped scatter plots of the intelligibility prediction for all instrumental measures in a combined white and babble noise scenario at 0, 5 dB.

information on the envelope transients and spectral transitions as shown in Figure 5.6. These indicate the place and manner of articulation [99]. The CSIIh score mainly represents vowel nuclei and does not distinguish between e.g. /v/ and /b/ which differ in their onsets but share the same vowel phoneme resulting in a lower correlation than CSIIm.

The mutual information based measures MIKNN and SIMI, reported to be reliable predictors for speech intelligibility in a single-channel speech enhancement framework [76, 77], here showed a moderate performance. Note, that these measures compute a score on long term statistics which may cause the inaccurate results using the short sentences ($\sim$ 1.5 s) of the GRID corpus as the speech database. Finally, the DAU measure also performed at a moderate correlation similar to the results presented in [19, 74]

Table 5.2 summarizes the top performing instrumental measures that exhibited $\rho \geq 0.8$ for intelligibility prediction in each noise scenario. Three of the proposed phase-aware quality measures (IFD, PD, MSE) together with STOI, the conventional SII, CSII, CSIIh, and CSIIm constitute the subset. On average the highest correlation was obtained by CSIIm ($\rho = 0.91$)

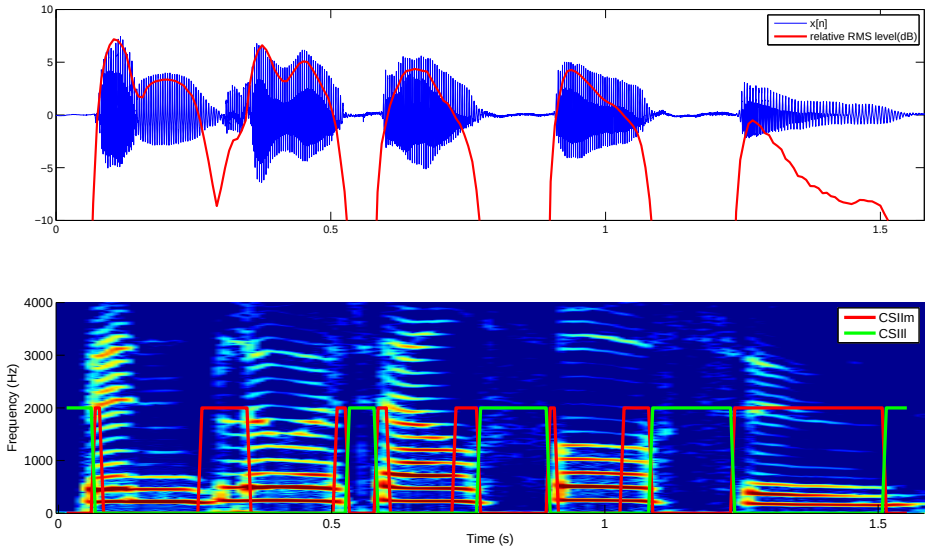

Figure 5.6: Time-domain (top) and spectral (bottom) representation of the GRID sentence "bin blue at l four soon" showing the relative RMS level (dB) and the according regions used to compute CSIIl (green), CSIIm (red) and CSIIh (no colour).

which outperformed all other measures followed by the CSII and IFD measures. High performance of CSII-based measures for single-channel enhanced speech was already reported in [19,74,100]. The IFD and MSE measure showed better performance than STOI in all scenarios. Although the PD measure exhibited the lowest correlation coefficient in this subset of the top performing intelligibility measures, it showed a comparable performance to CSIIm in terms of τ and furthermore revealed the most stable results across noise types, a property that was already observed in the perceived quality evaluation in Section 5.4.

Table 5.2: Statistical analysis of the top performing measures showing $\rho \geq 0.8$ for each noise scenario.

	White noise			Babble noise			Both noises			**Mean**		
	ρ	σ	τ	ρ	σ	τ	ρ	σ	τ	ρ	σ	τ
IFD	0.90	**0.04**	-0.74	0.89	0.07	-0.75	0.81	0.07	-0.64	0.87	0.06	-0.71
PD	0.83	0.05	**-0.80**	0.83	0.08	-0.81	0.80	0.08	-0.67	0.82	0.07	-0.76
MSE	0.90	**0.04**	0.71	0.89	0.07	0.81	0.80	0.07	0.68	0.86	0.06	0.73
STOI	0.89	**0.04**	0.68	0.84	0.08	0.81	0.80	0.07	0.63	0.84	0.07	0.70
SII	0.91	**0.04**	0.77	0.81	0.09	0.75	0.80	0.07	0.66	0.84	0.07	0.73
CSII	0.91	**0.04**	0.77	0.90	0.06	0.78	0.84	0.07	**0.74**	0.88	0.06	0.76
CSIIh	0.84	0.05	0.77	0.89	0.07	0.81	0.83	0.07	0.72	0.85	0.06	**0.77**
CSIIm	**0.92**	**0.04**	**0.80**	**0.94**	**0.05**	**0.84**	**0.86**	**0.06**	0.68	**0.91**	**0.05**	**0.77**

5.6 Combined Performance of Perceived Quality and Intelligibility Prediction

The title of this section may be misleading in terms of to think that there exists a measure that reliably predicts both, perceived quality and speech intelligibility. In fact, it is not clear how the perceived quality and speech intelligibility are related to each other. As an example, Jensen and Taal in [77] reported that speech intelligibility cannot be improved by any processing of the

noisy critical-band amplitudes that on the other hand were used to enhance the perceived speech quality while the phase-based method by Kulmer and Mowlaee in [20] showed joint improvement for both as verified by the listening tests in Sections 4.3.2 and 4.4.2.

The intention here is to find an instrumental measure that guides the way towards an upgraded single-channel speech enhancement algorithm that improves both perceived quality and speech intelligibility within a phase-aware framework. Therefore Table 5.3 presents those measures that showed $\rho \geq 0.8$ in each noise scenario in the perceived quality and intelligibility evaluation. The mean performances of the measures, computed as the average over the six performance conditions, were almost the same showing a slight benefit for PD and CSIIm. In general the proposed phase-aware quality measures (IFD, PD, MSE) exhibited higher correlation in the quality evaluation termed as (-q) while the conventional intelligibility measures (CSII, CSIIm) showed a better correlation to intelligibility (-i), as to be expected. The benefit of PD and CSIIm against the other measures should not be drawn-out too much. Analysing Tables 5.1 and 5.2 concludes that PD was the best performing measure for perceived speech quality and CSIIm the best performing one for speech intelligibility respectively, but they also showed the lowest correlation in the opposite evaluation field within the subset of the better performing measures, presented in Tables 5.1 and 5.2. Only in terms of the rank correlation coefficient τ, the PD measure revealed a reasonable greater correlation. The CSII measure showed the most stable results across all conditions in terms of ρ and τ.

Table 5.3: Statistical analysis of the combined perceived quality and intelligibility performance of the instrumental measures showing $\rho \geq 0.8$ for each scenario.

	White noise			Babble noise			Both noises			**Mean**		
	ρ	σ	τ	ρ	σ	τ	ρ	σ	τ	ρ	σ	τ
IFD-q	0.89	0.07	-0.81	0.84	0.07	-0.68	0.86	0.07	-0.71	0.86	0.07	-0.72
IFD-i	0.90	0.04	-0.74	0.89	0.07	-0.75	0.81	0.07	-0.64			
PD-q	0.93	0.05	-0.89	0.91	0.06	-0.77	0.91	0.06	-0.77	**0.87**	**0.06**	**-0.78**
PD-i	0.83	0.05	-0.80	0.83	0.08	-0.81	0.80	0.08	-0.67			
MSE-q	0.90	0.06	0.81	0.83	0.08	0.70	0.86	0.07	0.75	0.86	0.07	0.74
MSE-i	0.90	0.04	0.71	0.89	0.07	0.81	0.80	0.07	0.68			
CSII-q	0.84	0.08	0.73	0.85	0.07	0.71	0.83	0.08	0.68	0.86	0.07	0.74
CSII-i	0.91	0.04	0.77	0.90	0.06	0.78	0.84	0.07	0.74			
CSIIm-q	0.81	0.09	0.73	0.86	0.07	0.73	0.82	0.08	0.70	**0.87**	**0.06**	0.75
CSIIm-i	0.92	0.04	0.80	0.94	0.05	0.84	0.86	0.06	0.68			

5.7 Properties with Regard to Additive Noise

A good instrumental measure has to be credible. This means that the score it produces has to be easy interpretable. The following experiment clarifies how the top performing measures of Table 5.3 evolve over SNRs for speech corrupted by additive white and babble noise. For the sake of completeness the conventional state-of-the-art perceived quality and intelligibility measures PESQ and STOI are also included in the analysis as well as the proposed intelligibility measure UnRMSE that showed reasonable performance in the intelligibility evaluation.

Figure 5.7 shows the scores and 95% confidence intervals for each measure averaged over the 50 sentences from the test data set. The boundaries of the SNRs were determined to -20 and 30 dB where the lower bound refers to a speech signal that is completely masked by the additive noise and hence is not intelligible. The upper bound of 30 dB is perceived as the clean speech signal with a slightly audible background noise and leads to a completely intelligible speech signal. All measures except PESQ and UnRMSE reveal a smooth sigmoid function within the defined dynamic range. These two measures are more sensitive to slight distortions in high

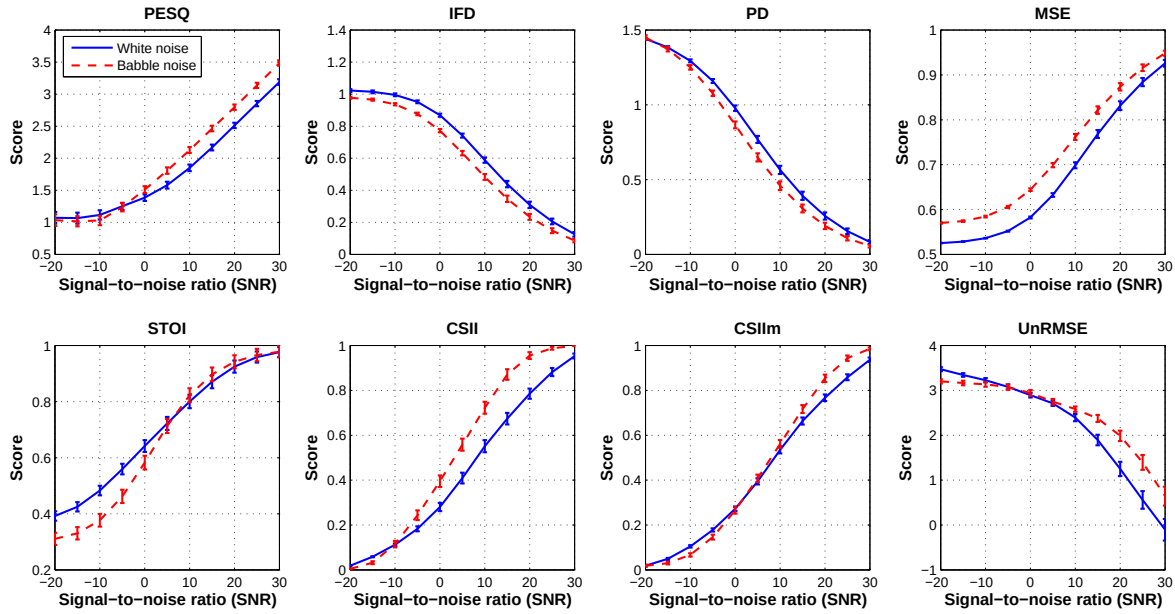

Figure 5.7: Mean objective scores of the best performing instrumental measures evaluated over 50 sentences from the GRID corpus corrupted by additive noise.

SNR regions showing a more linear relationship to the SNR. According to the noise types each measure except UnRMSE shows a better score for babble noise rather than white noise which is in line with the observations of the listening tests. The UnRMSE underestimates the scores in babble noise, an observation already made in Section 5.5. The CSII-based measures are the only one that exploit their whole definition range of [0,1] to evaluate the noisy utterances from not intelligible to intelligible. The proposed measures IFD, PD, and MSE smoothly reach their upper bound towards a clean speech signal defined as zero for IFD and PD and one for MSE. However, comparing the subplots of IFD, PD, and MSE in Figure 5.7 with their definitions in Eq. (2.29), (2.32), and (2.33) it is clear that the bound of four for IFD and PD and zero for MSE will not be reached. Those measures compute a score in the STFT domain based on different phase representations only. The representations show a clear structure at harmonics while revealing a random structure between harmonics and at noise only frames thus leading to a bound somewhere below the theoretical value. The PD measure saturates for low SNRs to a value about 1.5. Good speech quality is interpreted by values less than 0.1 while a score greater than 1 indicates bad speech quality. For the IFD the same lower bound for good speech quality can be observed. Bad speech quality is determined by a value greater than 0.8 where the measure overall saturates to approximately 1.

5.8 Properties with Regard to Phase Modifications

The second property to be explored is the influence of pure phase distortions to the scores of the instrumental measures. Let $\phi_x(k,l)$ be the STFT phase of the clean speech signal $x(n)$ at frequency bin k and frame index l. The phase information $\psi_x(h,l)$ at the harmonics h are obtained by linear interpolation of the spectral phase $\phi_x(k,l)$ at the harmonics. Following the phase decomposition in [20] the unwrapped phase is computed by removing the linear phase

part, given as

$$\Psi_x(h,l) = \psi_x(h,l) - \psi_{x_{\text{lin}}}(h,l). \tag{5.8}$$

The unwrapped phase is modified by adding a random phase with uniform distribution in the interval of $[-\pi, \pi]$ according to the following equation

$$\Psi_{x_{\text{mod}}}(h,l) = \Psi_x(h,l) + \alpha \cdot \mathcal{U}[-\pi, \pi], \tag{5.9}$$

where α lies in the range of [0,1] and determines the variance of the distortion. The distorted unwrapped phase is transformed back to the STFT domain by first adding the linear phase part back to the distorted unwrapped phase. Then the frequency bins are modified within the width of the main-lobe of the analysis window denoted by N_p:

$$\phi_{x_{\text{mod}}}(\lfloor h\omega_0 K \rfloor + i, l) = \psi_{x_{\text{mod}}}(h,l), \qquad \forall i \in [-N_p/2, N_p/2]. \tag{5.10}$$

with K as the DFT size and $\omega_0 = 2\pi f_0/f_s$. The complex spectrum is then constructed using the clean spectral amplitude together with the modified phase given as

$$X^c_{\text{mod}}(k,l) = |X^c(k,l)| \mathrm{e}^{j\phi_{x_{\text{mod}}}(k,l)}, \tag{5.11}$$

and is reconstructed to the time-domain modified speech signal by applying an inverse STFT. The impact of the added random phase to the clean speech signal is illustrated in Figure 5.8.

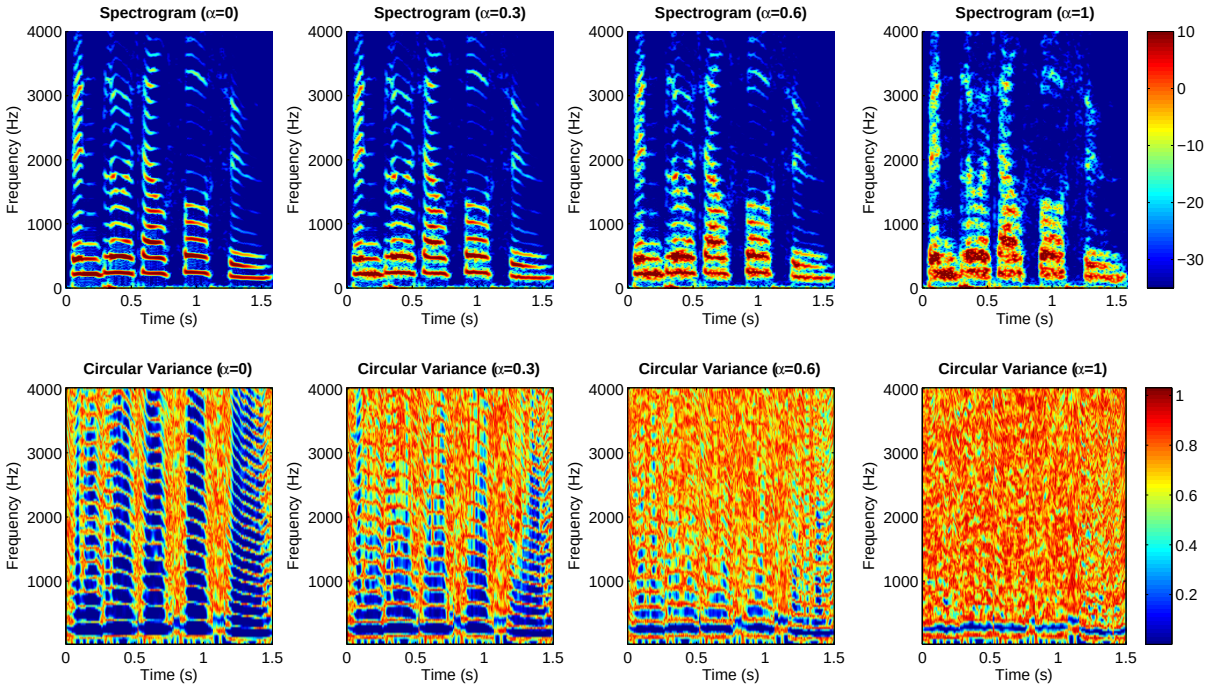

Figure 5.8: Spectrograms (top) and circular variances (bottom) of the sentence "bin blue at l four soon" corrupted by a random phase at different α values.

The subplots at the bottom show the circular phase variance which is a measure of the uncertainty of a phase value along time. Phase variance has been reported as a reliable measure for voice quality assessment in [101]. At voiced frames, the circular variance is zero and reflects the harmonics given a clean speech utterance. With increasing α, the structure observed in the circular variance gets destroyed and for the extreme case, i.e. $\alpha = 1$ only the fundamental harmonic is visible. This residual structure is due to the pitch-synchronous phase decomposition that employs some phase information of the fundamental frequency at the signal reconstruction stage. The spectrograms are illustrated on the top row, revealing that the harmonics scatter corresponding with increasing α. While the formant structure is hardly affected by the added random phase, the harmonic content decreases for increasing α and is totally lost for $\alpha = 1$, perceived as roughness. Since the formants are preserved to a certain extent, the modified speech signals remain quite intelligible, confirmed by informal listening.

The performance of the instrumental measures is shown in Figure 5.9. Because of an imperfect analysis-synthesis framework of the pitch-synchronous phase decomposition, no measure obtained its best score for $\alpha = 0$. In comparison to the experiment in the previous section and Figure 5.7, all measures obtained better scores indicating better perceived quality/intelligibility for phase distortions at $\alpha = 1$ than the low SNR additive noise scenario at -20 dB, verified by informal listening. Remarkable was the insensitivity of STOI to the introduced phase distortions which resulted in an overprediction of speech intelligibility for high α values. The PD and CSII-based measures were most sensitive spanning a wide range with respect to the scores and showed no saturation towards $\alpha \to 1$. Over-sensitiveness of the CSII-based measures with respect to phase was also reported in [74].

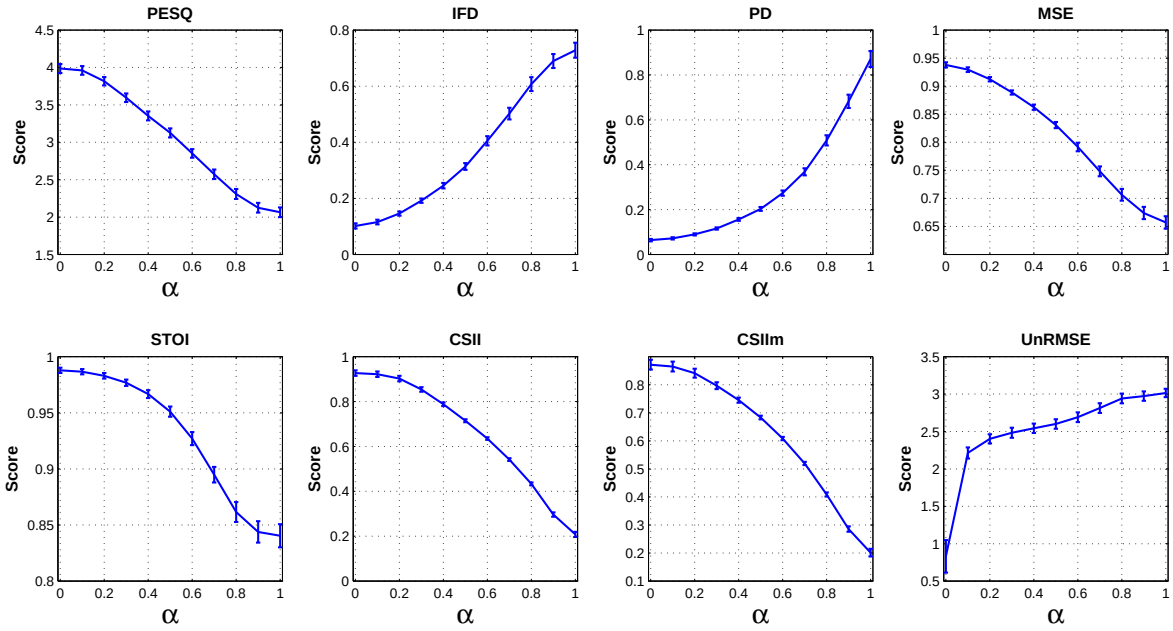

Figure 5.9: Mean objective scores of the best performing instrumental measures evaluated over 50 sentences from the GRID corpus corrupted by phase distortions controlled by α.

6 Conclusion

6.1 Performance of the Instrumental Measures in a Phase-Aware Framework

In this investigation, we addressed the following research questions: First, which existing instrumental measures reliably assess the perceived quality and speech intelligibility of the enhanced speech when both, spectral amplitude and spectral phase are modified. Second, whether a new phase-aware measure outperforms the existing ones in terms of predicting the subjective results and third, if any instrumental measure is capable to determine both, perceived quality and speech intelligibility.

A MUSHRA test and a intelligibility test, following the recommendation by Barker and Cooke, were conducted to collect the subjective quality and intelligibility scores in white and babble noise conditions. A test database consisting of 50 sentences from the GRID corpus was used. The listening test results suggest that the incorporation of a phase modification stage in the enhancement procedure leads to improved perceived quality and speech intelligibility in comparison to the conventional MMSE-LSA method, proposed by Ephraim and Malah, that only modifies the spectral amplitude while copying the noisy phase at signal reconstruction. These results are pronounced more at low SNRs for white noise rather than babble noise due to the inaccurate fundamental frequency estimation in babble noise, an essential task for phase decomposition applied in the benchmark method C + PE.

A statistical analysis was performed to quantify the correlation between the scores of the instrumental measures and the subjective listening results using the Pearson's correlation coefficient ρ, the root-mean-square error (RMSE) σ, and Kendall's tau τ. The well-known measures PESQ and STOI, widely used for perceived quality and speech intelligibility assessment, exhibited a reasonable high correlation but are not the top performing measures in the evaluated test setup.

In particular, the proposed quality measures GD, PD, and MSE outperformed PESQ on average where the PD measure revealed the highest performance ($\rho = 0.92, \sigma = 0.06, \tau = -0.81$) for perceived quality evaluation and was the top performing measure in the babble and combined noise scenario. The GD measure was an accurate predictor in white noise ($\rho = 0.96, \sigma = 0.04, \tau = -0.87$) but significantly overestimated the unprocessed speech signals in babble noise.

The proposed intelligibility measures UnHPSNR and UnRMSE showed a reliable estimation of speech intelligibility in white noise ($\rho > 0.95$), though they were poor predictors in the com-

bined noise scenario. The rationale behind this is the consistent underestimation of the scores in babble noise as a result of the inaccurate f_0 estimation in non-stationary noise. The f_0 tracker is part of the phase decomposition to extract the unwrapped phase used in UnHPSNR and UnRMSE calculation. On average, the CSII-based measures showed a reasonable high correlation to the subjective speech intelligibility scores where CSIIm exhibited the best performance ($\rho = 0.91, \sigma = 0.05, \tau = -0.77$) and outperformed all other measures in the babble and combined noise(where the mapping between the subjective and objective scores was applied to babble and white conditions) scenarios.

At least, no measure was capable to reliably predict perceived quality and speech intelligibility jointly. This observation might be straightforward since an enhanced speech signal perceived with better speech quality not necessarily has to be more intelligible and vice versa. However, a moderate average correlation of $\rho = 0.87$ across perceived quality and intelligibility performance evaluation of the PD and CSIIm measures indicate that there has to be some connection between these two characteristics.

6.2 Outlook

The results showed that the proposed phase-aware quality measures GD, IFD, and PD revealed a high correlation to the subjective listening results. However, the analysis conditions were isolated to phase-aware single-channel enhancement algorithms operating in white and babble noise. Further investigations should study the reliability of the proposed measures for other noise types and processing conditions, i.e., Ideal Time Frequency Segregation (ITFS). The performance of these measures in terms of predicting the speech intelligibility was at a lower level but nevertheless exhibited a moderate correlation for IFD and PD. Since the proposed measures are easily extendible, more research could be done on employing band-importance functions known to be an important part in the computation of well-known state-of-the-art intelligibility measures and analyse if this increases the speech intelligibility prediction.

The proposed phase-aware speech intelligibility measures UnHPSNR and UnRMSE showed a high performance in white noise while they revealed a significant worse performance in the combined noise scenario, where the objective scores were fitted to the subjective scores including white and babble noise conditions. This is due to the inaccurate fundamental frequency estimation in the adverse noise scenario. These two measures could benefit from a more reliable f_0 tracker.

The presented measures are simple and compute a score based on the pure difference between some phase representation of the enhanced and clean speech signals. A recently proposed single-channel speech enhancement system [20], reported to enhance perceived speech quality and speech intelligibility, operates on circular statistics at harmonics. In this sense a new more sophisticated measure could be thought of a distance measure between the enhanced and reference probability distributions of the spectral phase at harmonics with an additional amplitude weighting.

A Speech Database

A.1 Test Database

audio file	utterance	colour	letter	number
bbal4s.wav	bin blue at l four soon	blue	l	4
bbal5n.wav	bin blue at l five now	blue	l	5
bbap9s.wav	bin blue at p nine soon	blue	p	9
bbbs4n.wav	bin blue by s four now	blue	s	4
brap4n.wav	bin red at p four now	red	p	4
brap7a.wav	bin red at p seven again	red	p	7
brav9s.wav	bin red at v nine soon	red	v	9
brbj6p.wav	bin red by j six please	red	j	6
brbj7a.wav	bin red by j seven again	red	j	7
brbq1a.wav	bin red by q one again	red	q	1
brbx2n.wav	bin red by x two now	red	x	2
brbx3s.wav	bin red by x three soon	red	x	3
brbx4p.wav	bin red by x four please	red	x	4
bric2n.wav	bin red in c two now	red	c	2
lgal1s.wav	lay green at l one soon	green	l	1
lgal3a.wav	lay green at l three again	green	l	3
lgap1p.wav	lay green at p one please	green	p	1
lgwm4p.wav	lay green with m four please	green	m	4
lrii2p.wav	lay red in i two please	red	i	2
lrii3a.wav	lay red in i three again	red	i	3
lriizn.wav	lay red in i zero now	red	i	0
lrio5s.wav	lay red in o five soon	red	o	5
lrio7a.wav	lay red in o seven again	red	o	7
lriv1a.wav	lay red in v one again	red	v	1
lrivzp.wav	lay red in v zero please	red	v	0
lrwj3s.wav	lay red with j three soon	red	j	3
lrwp6n.wav	lay red with p six now	red	p	6
lrwp7s.wav	lay red with p seven soon	red	p	7

lrwp9a.wav	lay red with p nine again	red	p	9
lrwx1s.wav	lay red with x one soon	red	x	1
lrwx2p.wav	lay red with x two please	red	x	2
lrwxzn.wav	lay red with x zero now	red	x	0
pbiu3n.wav	place blue in u three now	blue	u	3
pgin2p.wav	place green in n two please	green	n	2
sbam9n.wav	set blue at m nine now	blue	m	9
sban1p.wav	set blue at n one please	blue	n	1
sban2a.wav	set blue at n two again	blue	n	2
sbanzs.wav	set blue at n zero soon	blue	n	0
swab7a.wav	set white at b seven again	white	b	7
swihzn.wav	set white in h zero now	white	h	0
swiu2s.wav	set white in u two soon	white	u	2
swiu4a.wav	set white in u four again	white	u	4
swwc3p.wav	set white with c three please	white	c	3
swwc4a.wav	set white with c four again	white	c	4
swwi6s.wav	set white with i six soon	white	i	6
swwp1p.wav	set white with p one please	white	p	1
swwp2a.wav	set white with p two again	white	p	2
swwpzs.wav	set white with p zero soon	white	p	0
swwv3n.wav	set white with v three now	white	v	3
swwv5p.wav	set white with v five please	white	v	5

A.2 Training Database

audio file	utterance	colour	letter	number
bbax7n.wav	bin blue at x 7 now	blue	x	7
bbbr9a.wav	bin blue by r 9 again	blue	r	9
lgay9s.wav	lay green at y 9 soon	green	y	9
lgbmzs.wav	lay green by m 0 soon	green	m	0
prig9n.wav	place red in g 9 now	red	g	9
priu8s.wav	place red in u 8 soon	red	u	8
swbi2s.wav	set white by i 2 soon	white	i	2
swwv1p.wav	set white with v 1 please	white	v	1

B
Supplemental Figures

B.1 Intelligibility Listening Test

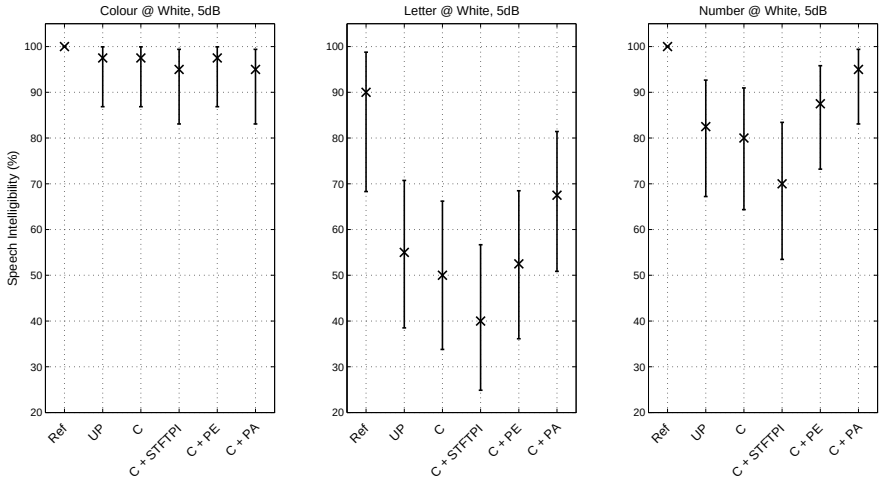

Figure B.1: Intelligibility scores showing the mean and 95% confidence interval separated by (left) colour, (middle) letter, and (right) number for white noise @ 5 dB SNR averaged over ten participants.

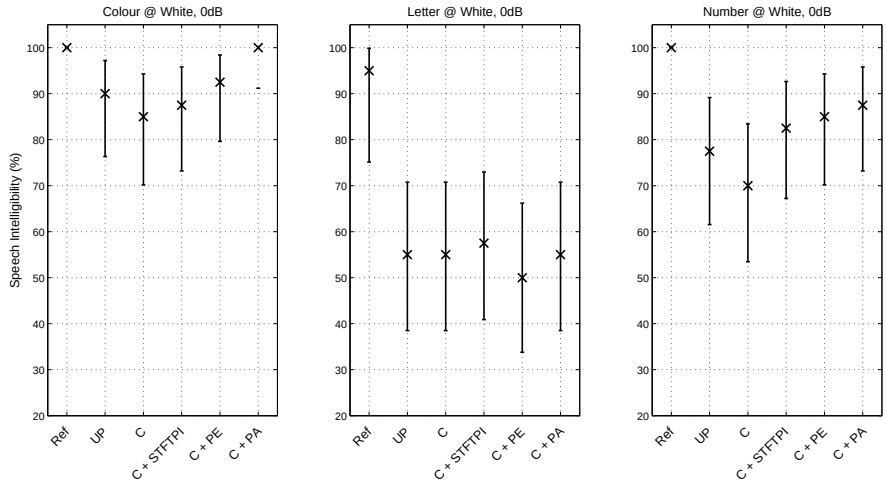

Figure B.2: Intelligibility scores showing the mean and 95% confidence interval separated by (left) colour, (middle) letter, and (right) number for white noise @ 0 dB SNR averaged over ten participants.

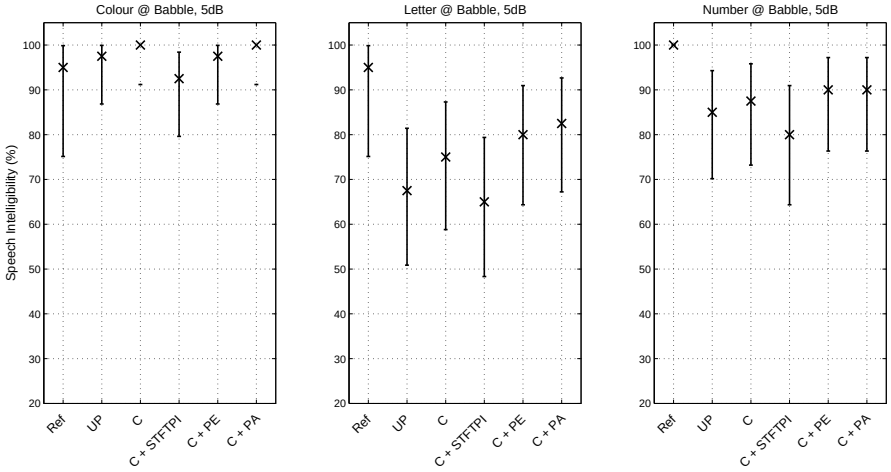

Figure B.3: Intelligibility scores showing the mean and 95% confidence interval separated by (left) colour, (middle) letter, and (right) number for babble noise @ 5 dB SNR averaged over ten participants.

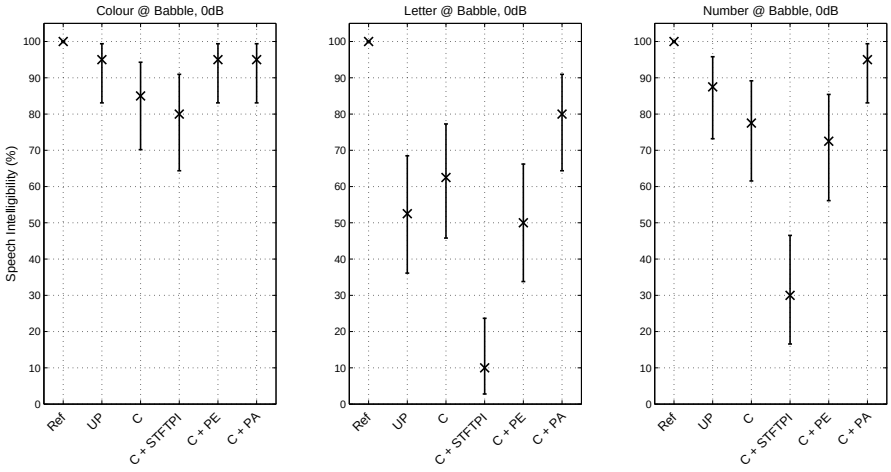

Figure B.4: Intelligibility scores showing the mean and 95% confidence interval separated by (left) colour, (middle) letter, and (right) number for babble noise @ 0 dB SNR averaged over ten participants.

B.2 Perceived Quality Evaluation

B.2.1 Mapped Scatter Plots

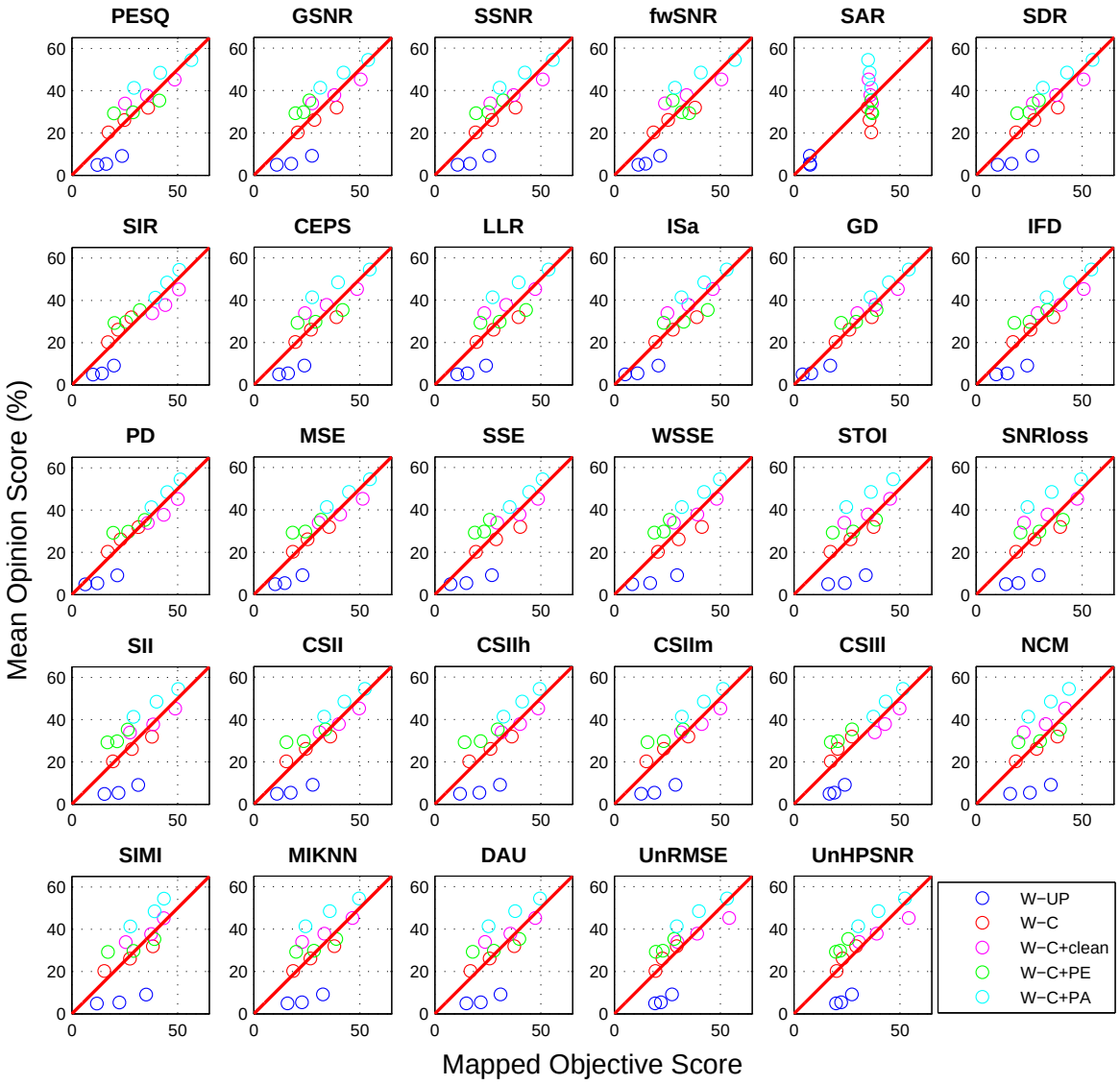

Figure B.5: Mapped scatter plots for all instrumental measures in a white noise scenario at 0, 5, 10 dB.

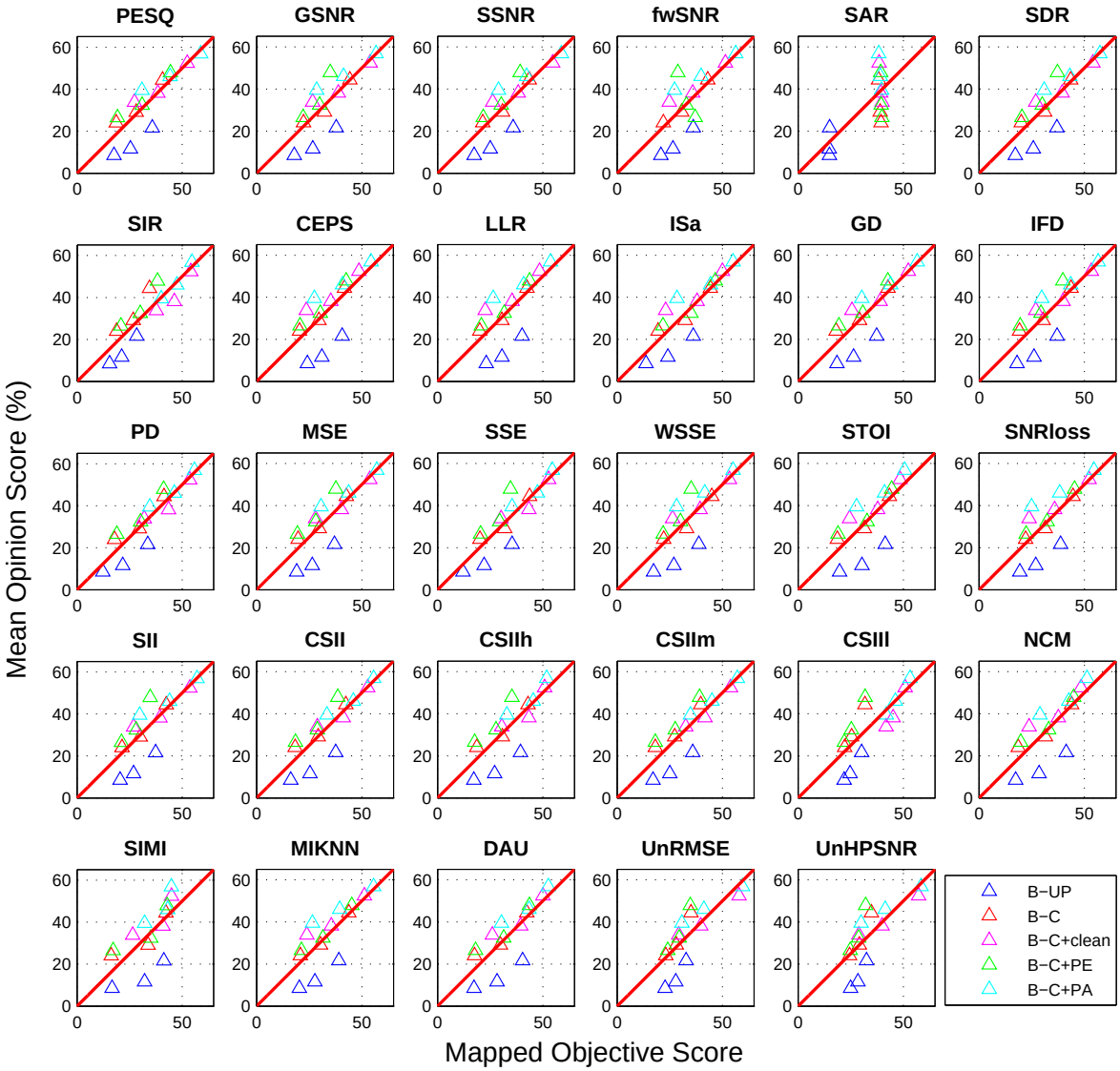

Figure B.6: Mapped scatter plots for all instrumental measures in a babble noise scenario at 0, 5, 10 dB.

B.2.2 Unmapped Scatter Plots

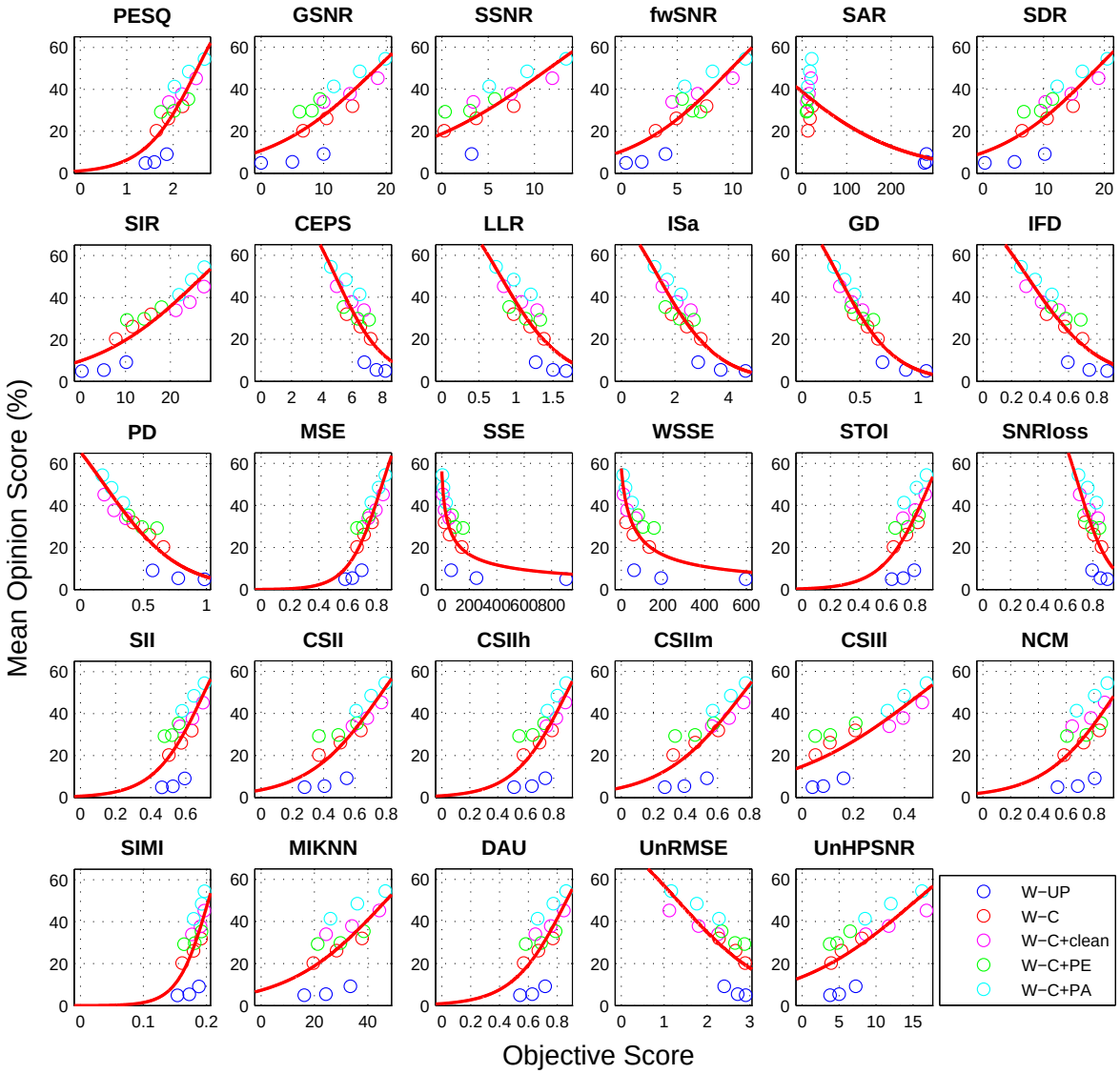

Figure B.7: Scatter plots for all instrumental measures together with the fitted mapping function in a white noise scenario at 0, 5, 10 dB.

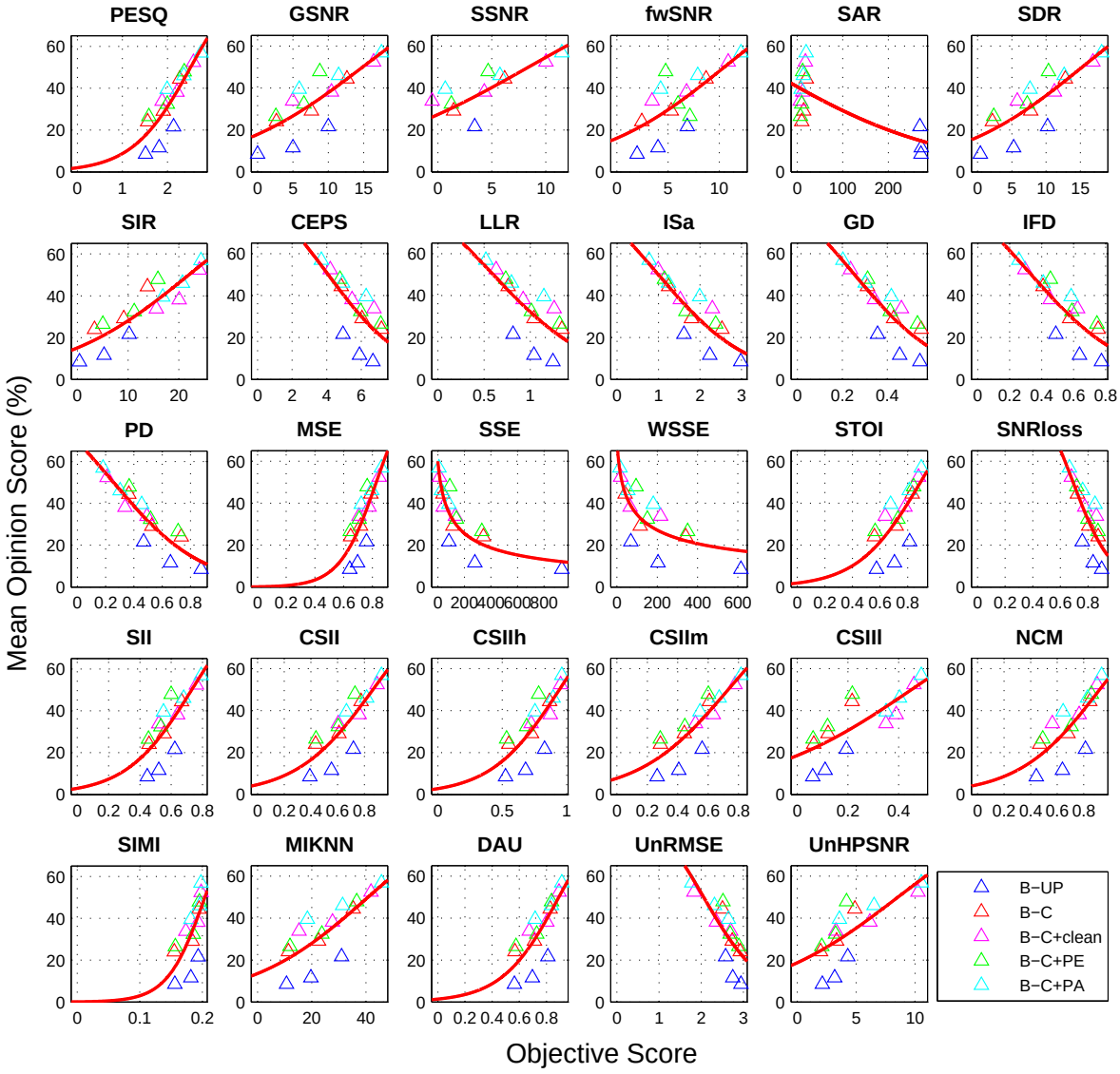

Figure B.8: Scatter plots for all instrumental measures together with the fitted mapping function in a babble noise scenario at 0, 5, 10 dB.

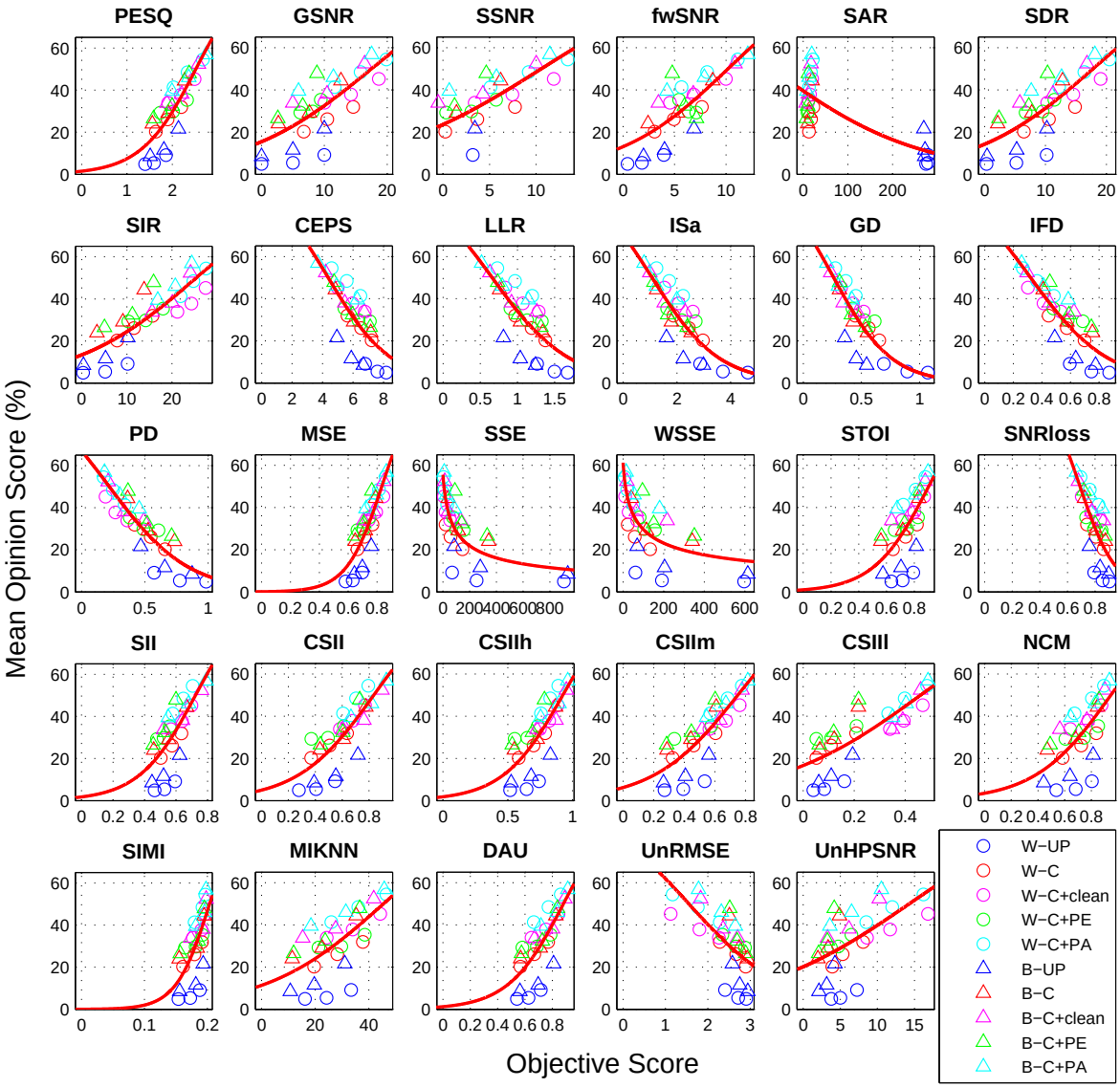

Figure B.9: Scatter plots for all instrumental measures together with the fitted mapping function in a combined white and babble noise scenario at 0, 5, 10 dB.

B.3 Intelligibility Evaluation

B.3.1 Mapped Scatter Plots

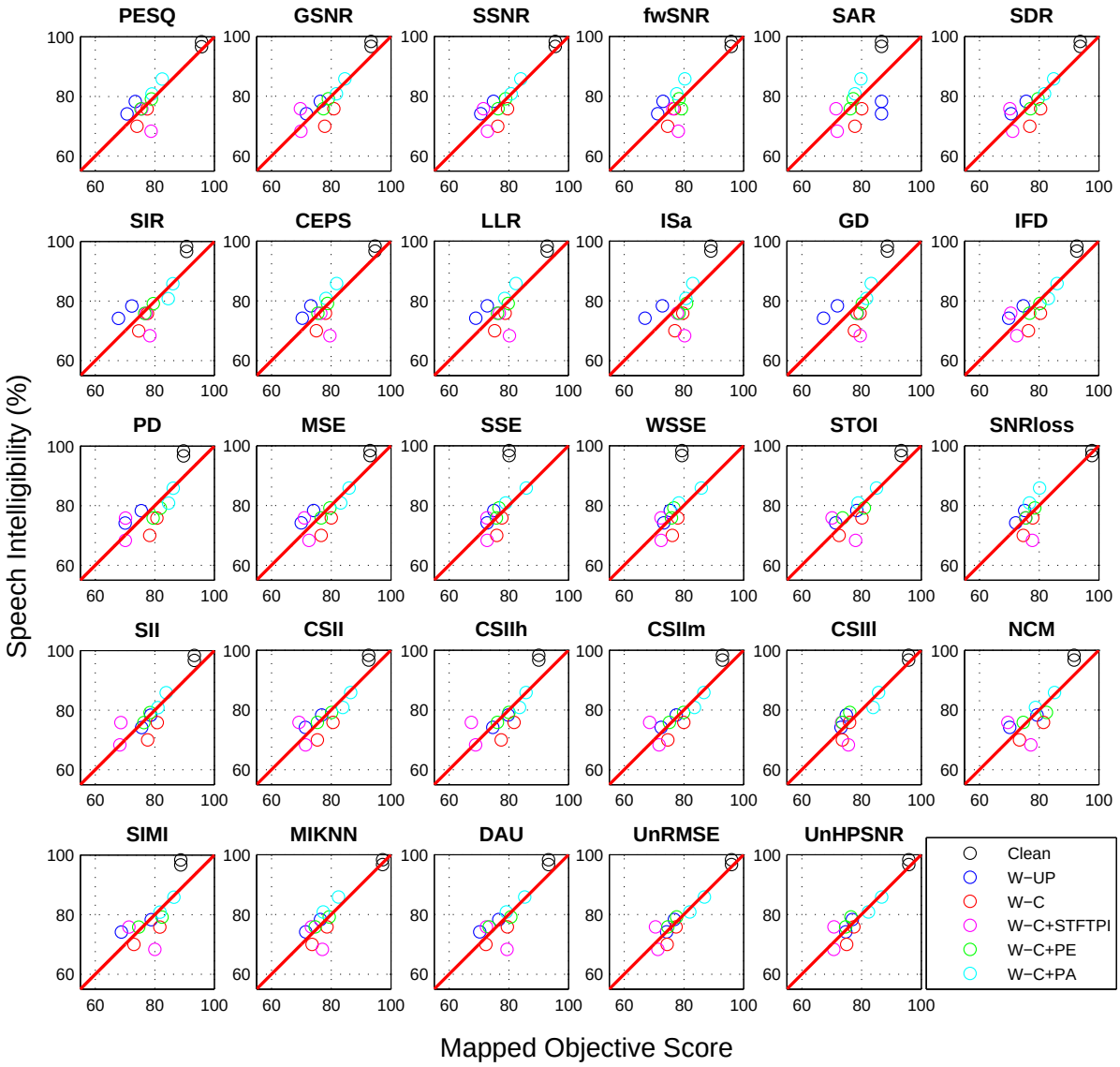

Figure B.10: Mapped scatter plots for all instrumental measures in a white noise scenario at 0, 5, 10 dB.

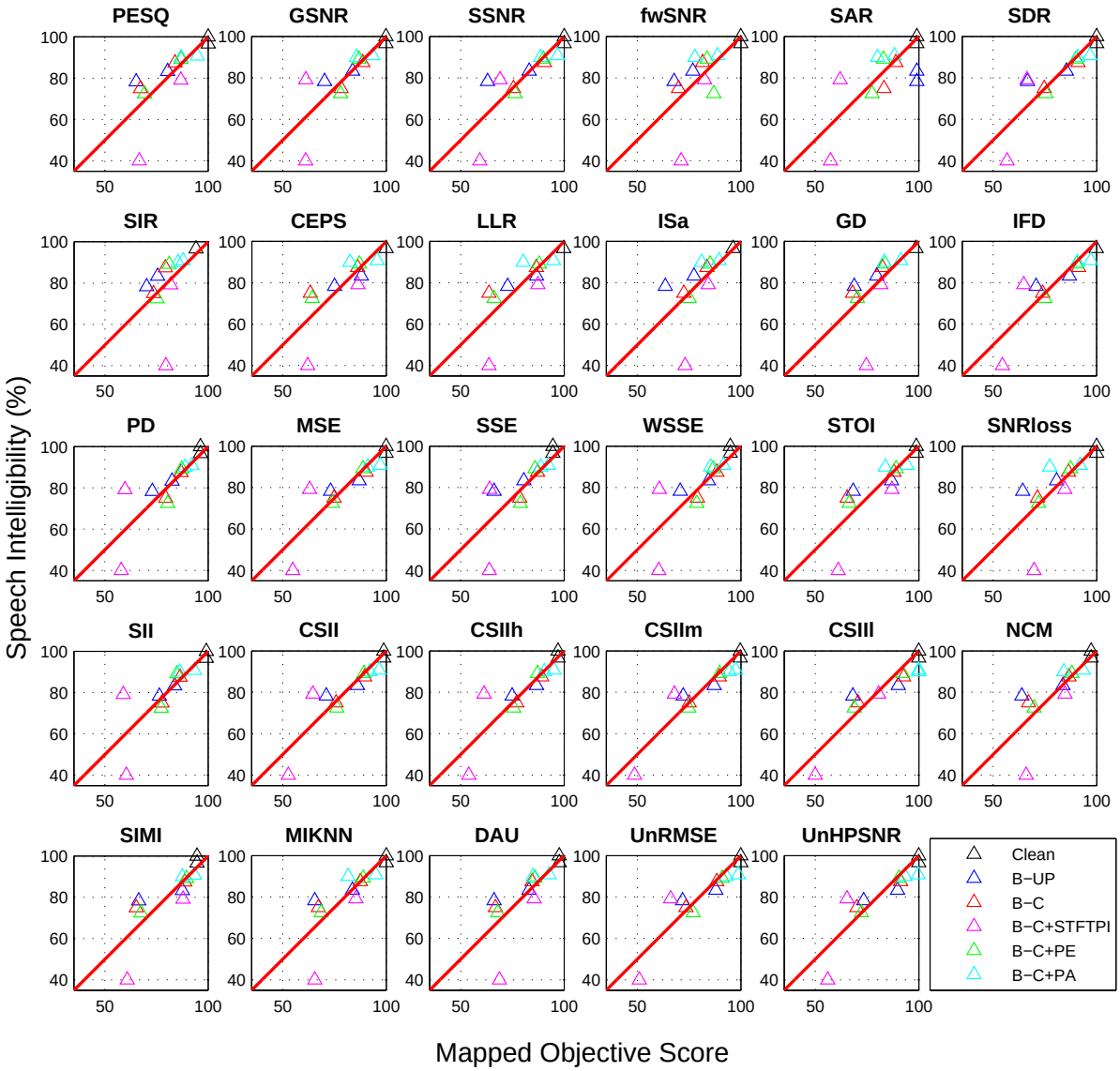

Figure B.11: Mapped scatter plots for all instrumental measures in a babble noise scenario at 0, 5, 10 dB.

B.3.2 Unmapped Scatter Plots

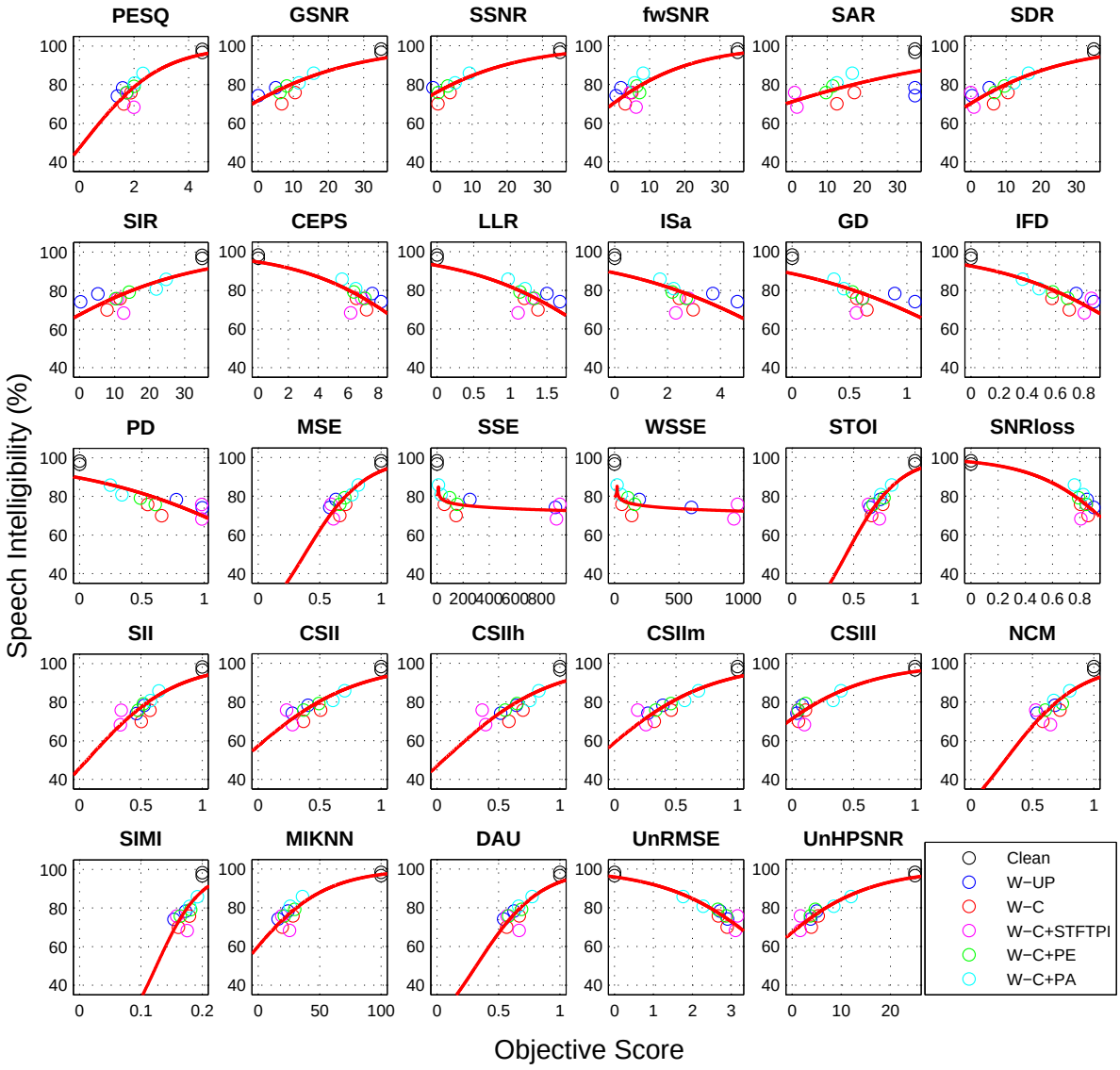

Figure B.12: Scatter plots for all instrumental measures together with the fitted mapping function in a white noise scenario at 0, 5 dB.

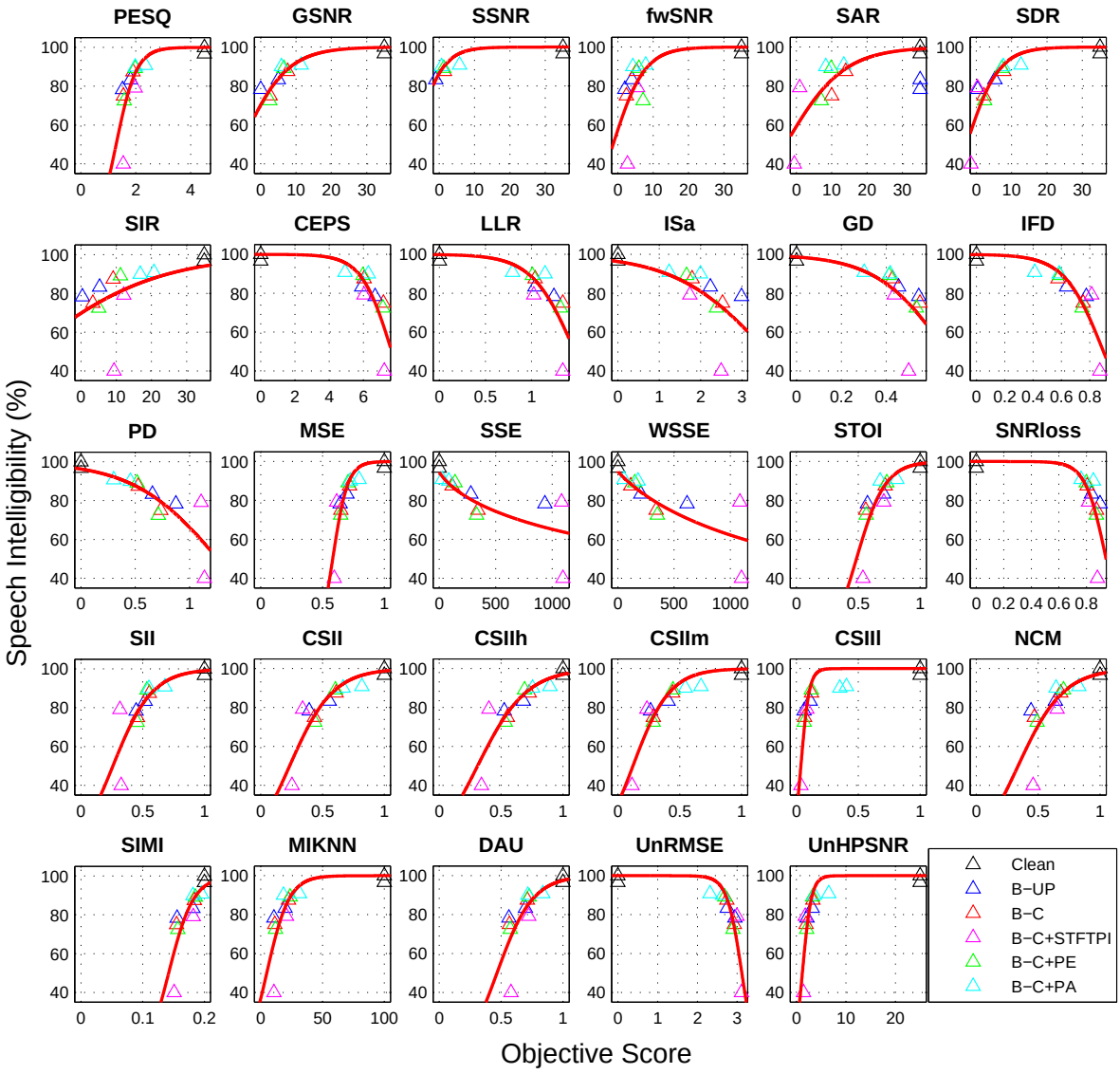

Figure B.13: Scatter plots for all instrumental measures together with the fitted mapping function in a babble noise scenario at 0, 5 dB.

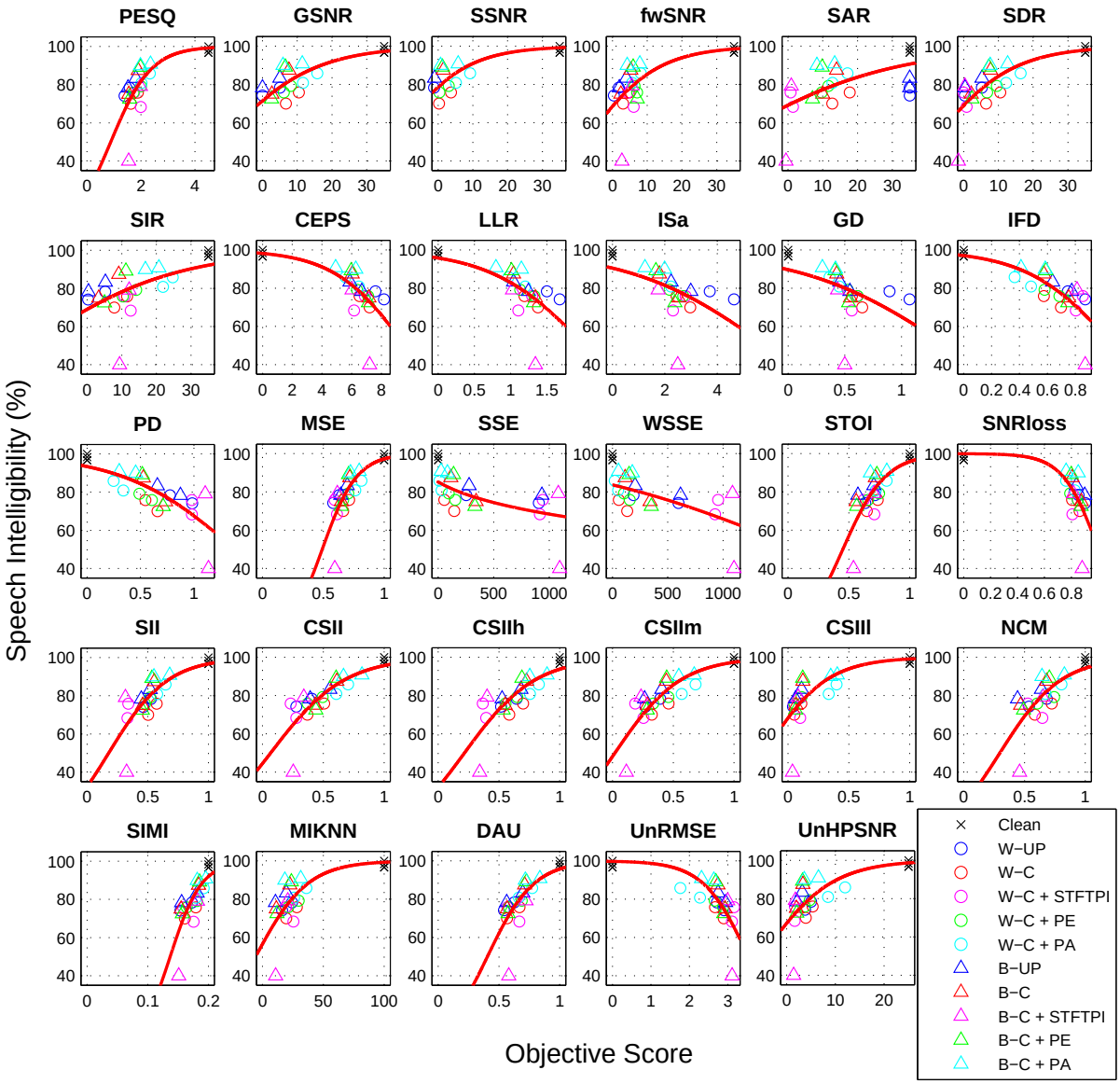

Figure B.14: Scatter plots for all instrumental measures together with the fitted mapping function in a combined white and babble noise scenario at 0, 5 dB.

Bibliography

[1] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[2] P. Mowlaee, M. Watanabe, and R. Saeidi, "Show & tell: Phase-aware single-channel speech enhancement," in *Proc. INTERSPEECH*, 2013, pp. 1–4.

[3] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[4] ——, "On phase importance in parameter estimation in single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2013, pp. 7462–7466.

[5] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 129 –132, Feb 2013.

[6] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent Wiener filtering: generalized time-frequency masking respecting spectrogram consistency," in *9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 89–96.

[7] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421 –424, May 2010.

[8] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[9] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in *The International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[10] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE signal processing letters*, vol. 20, no. 3, pp. 217 – 220, 2013.

[11] T. Kleinschmidt, S. Sridharan, and M. Mason, "The use of phase in complex spectrum subtraction for robust speech recognition," *Computer Speech and Language*, vol. 25, no. 3, pp. 585 – 600, 2011.

[12] R. Schluter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 133–136.

[13] P. Aarabi, *Phase-Based Speech Processing*. World Scientific Publishing, 2006.

[14] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4581–4584.

[15] B. J. Shannon and K. K. Paliwal, "Role of phase estimation in speech enhancement," in *Proc. Interspeech*, 2006.

[16] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, 2003, pp. 2117–2120.

[17] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2013.

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

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

[20] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 598–602, May. 2015.

[21] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *International Workshop on Acoustic Signal Enhancement; Proceedings of IWAENC*, 2012, pp. 1–4.

[22] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215 – 1229, 2006.

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

[24] ——, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[25] A. P. Stark and K. K. Paliwal, "Group-delay-deviation based spectral analysis of speech," in *Proc. INTERSPEECH*, 2009, pp. 1083–1086.

[26] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727 – 736, 2006.

[27] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529 – 541, May 1981.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Acoustical Society of America Journal*, vol. 120, p. 2421, 2006.

[29] S. Saito and F. Itakura, "The theoretical consideration of statistically optimum methods for speech spectral density," Electrical Communication Laborotory, NTT, Tokyo, Tech. Rep. Report No. 3107, Dec 1966.

[30] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," Tokyo, Tech. Rep. Reports of 6th Int. Cong. Acoust., 1968, ed. by Y. Kohasi.

[31] ——, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53, no. 1, pp. 36–43, 1970.

[32] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4199–4208, Aug 2014.

[33] S. P. Patil and J. N. Gowdy, "Exploiting the baseband phase structure of the voiced speech for speech enhancement," in *ICASSP*, May 2014, pp. 6092–6096.

[34] K. K. Paliwal, K. K. Wojcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[35] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *Signal Processing Magazine, IEEE*, vol. 32, no. 2, pp. 55–66, March 2015.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, 2001, pp. 749–752.

[37] ANSI S3.5, *American National Standard Methods for the Calculation of the Articulation Index*. American National Standard, 1969.

[38] T. Houtgast and J. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acta Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

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

[40] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, Jun 2004.

[41] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*, ser. T-Labs Series in Telecommunication Services, 1. Dordrecht: Springer, 2011.

[42] "The impact of voice processing on modern telecommunications," *Speech Communication*, vol. 17, no. 3–4, pp. 217 – 226, 1995.

[43] J. R. J. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, ser. An IEEE Press classic reissue. Wiley, 2000.

[44] J. M. Tribolet, P. Noll, B. McDermott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in *ICASSP.*, vol. 3, Apr 1978, pp. 586–590.

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

[46] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, Oct 1976.

[47] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 242–248, Feb 1988.

[48] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *IN PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SPEECH AND LANGUAGE PROCESSING*, 1998, pp. 2819–2822.

[49] K. D. Kryter, "Methods for the calculation and use of the articulation index," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1689–1697, 1962.

[50] C. Févotte, R. Gribonval, and E. Vincent, "Bss eval toolbox user guide," *IRISA Tech. Rep. 1706*, 2005.

[51] L. Rabiner and R. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

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

[53] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2012, pp. 69–72.

[54] K. Kondo, *Subjective Quality Measurement of Speech*. Berlin Heidelberg: Springer-Verlag, 2012.

[55] P. Mowlaee, R. Saeidi, and Y. Stylanou, "Phase importance in speech processing applications," in *in Proceedings of the 15th International Conference on Spoken Language Processing*, 2014.

[56] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice Hall, 1975.

[57] S. Kay, "A fast and accurate single frequency estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1987–1990, Dec 1989.

[58] L. D. Alsteris and K. K. Paliwal, "Asr on speech reconstructed from short-time fourier phase spectra," in *Proc. INTERSPEECH*, 2004, pp. 1–4.

[59] F. Charpentier, "Pitch detection using the short-term phase spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, 1986, pp. 113–116.

[60] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, May 1995, pp. 784–787.

[61] A. P. Stark and K. K. Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. INTERSPEECH*, 2008, pp. 22–26.

[62] P. Mowlaee and R. Saeidi, "Time-frequency constraint for phase estimation in single-channel speech enhancement," *The International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[63] P. Vary, "Noise suppression by spectral magnitude estimation mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387 – 400, 1985.

[64] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 245–256, Jul 2002.

[65] A. Gaich and P. Mowlaee, "On speech quality estimation of phase-aware single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015.

[66] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and snr information," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 9, Sep 2015.

[67] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*, ser. Fundamentals of Statistical Signal Processing. Prentice-Hall PTR, 1993.

[68] H. Pobloth and W. B. Kleijn, "Squared error as a measure of perceived phase distortion," *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 1081–1094, 2003.

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

[70] ANSI S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index*. American National Standard, 1997.

[71] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[72] J. Ma and P. C. Loizou, "SNR Loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, no. 3, pp. 340–354, Mar 2011.

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

[74] 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," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 3013–3027, 2011.

[75] W. Kleijn, J. Crespo, R. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, March 2015.

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

[77] J. Jensen and C. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb 2014.

[78] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *Journal of The Acoustical Society of America*, vol. 102, pp. 2892–2905, 1997.

[79] Y. Hu and P. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–561–IV–564.

[80] L. Li, H. Jialong, and P. Günther, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, Sept. 1997.

[81] E. Jokinen, M. Takanen, H. Pulakka, and P. Alku, "Enhancement of speech intelligibility in near-end noise conditions with phase modification," in *Proc. INTERSPEECH*, 2014, pp. 1643–1647.

[82] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP, Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 38, 2014.

[83] S. Gonzales and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, Feb 2014.

[84] C. E. Osgood, "The nature and measurement of meaning," *Psychological Bulletin*, vol. 49, no. 3, pp. 197 – 237, May 1952.

[85] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX–92 Study on the Effect of Additive Noise on Automatic Speech Recognition," *Technical Report, DRA Speech Research Unit*, 1992.

[86] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466 – 475, Sept. 2003.

[87] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J Acoust Soc Am*, vol. 111, pp. 1917–1930, 2002.

[88] N. Evans, N. Hastings, and B. Peacock, *von Mises Distribution.* New York: Wiley, 2000.

[89] "ITU-R BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems," 2001.

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

[91] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech and Language*, vol. 28, no. 4, pp. 858 – 872, 2014.

[92] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, Jan 2011.

[93] 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*, 2009, pp. 1947–1950.

[94] A. Gaich and P. Mowlaee, "On speech intelligibility estimation of phase-aware single-channel speech enhancement," in *Proc. INTERSPEECH*, 2015, forthcoming 2015.

[95] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *Selected Areas in Communications, IEEE Journal on*, vol. 6, no. 2, pp. 242–248, Feb 1988.

[96] M. Hollier, M. Hawksford, and D. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *Vision, Image and Signal Processing, IEEE Proceedings* -, vol. 141, no. 3, pp. 203–208, Jun 1994.

[97] S. H. Parthasarathi, R. Padmanabhan, and H. A. Murthy, "Robustness of group delay representations for noisy speech signals," *International Journal of Speech Technology*, vol. 14, no. 4, pp. 361–368, 2011.

[98] B. Fox, A. T. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation." in *ICA*, ser. Lecture Notes in Computer Science, vol. 4666. Springer, 2007, pp. 454–461.

[99] S. Greenberg, *a multi-tier theoretical framework for understanding spoken language.* Mahwah, 2005, pp. 411–433.

[100] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[101] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *Proc. INTERSPEECH*, Singapore, September 2014.