On Speech Intelligibility Estimation of Phase-Aware Single-Channel Speech Enhancement

Andreas Gaich and Pejman Mowlaee

Signal Processing and Speech Communication Lab
Graz University of Technology, Graz, Austria
gaich@student.tugraz.at pejman.mowlaee@tugraz.at

Abstract
To reduce time and costs in the development process of noise reduction algorithms, an objective intelligibility measure is crucial. Such a measure has to show high correlation with speech intelligibility determined by real listening experiments. In the past several measures were found that perform reliable in a particular scenario when only the spectral amplitude of a noisy signal is modified. Recent studies demonstrate the positive impact of a phase modification in a single-channel speech enhancement showing improved speech intelligibility while conventional methods relying on amplitude-only modification are known for reduced intelligibility. Further, another recent study shows that a distortion metric defined on the spectral phase outperforms state-of-the-art quality metrics when used in phase-aware speech enhancement. This raises two questions we account for in this work: First, to study the reliability of the existing intelligibility measures in predicting the performance of the phase-aware methods, and second to investigate candidates for new phase-aware instrumental metrics and evaluate their reliability in terms of intelligibility prediction. Our objective and subjective evaluations demonstrate that CSH-based and STOI as well as the proposed phase-aware metrics perform as reliable speech intelligibility estimators following the subjective results.

Index Terms: Phase estimation, speech intelligibility, phase-aware speech enhancement, subjective listening.

1. Introduction
Estimating the quality of a speech signal provided by a signal enhancement system is of high importance in the development step of a new methodology. Extracting desired signals out of some noisy observations is a commonplace scenario in digital speech transmission and in different applications including speech coding, enhancement and de-reverberation where the quality of service is a fundamental issue in the standardization and implementation. As speech communication systems including mobile phones are expected to reliably perform in different noisy environments, improving the speech intelligibility in a noisy environment has attracted a lot of interest recently in the literature (see [1] or [2, Ch. 11] for an overview). Several signal processing techniques have been proposed to improve intelligibility in noise, relying on a perceptual distortion measure [3], modifying features [4], re-distributing the energy of a speech signal within frequency bands in a way that some approximated speech intelligibility index is maximized [5] or through deriving an optimal linear filter [6].

The challenging part in the assessment of the speech quality is the rather complex process of the human perception of sound involving cognitive and auditory system whereby no unified framework or conclusive model is available to address every aspect in the auditory perception. Further, although many standards have been established by European Telecommunications Standards Institute (ETSI) and International Telecommunication Union (ITU) to estimate the speech quality, they are often recommended with some restrictions to certain conditions, motivating researchers to propose new intelligibility measures [7–9] or combinations of existing ones [10].

The speech intelligibility is a measure of whether a message is correctly interpreted by a human conversational partner [1]. The phase information was shown to have large impact on human listening and intelligibility as reported by Palwal [11]. In particular, recent advances in speech enhancement report on the possibility to improve the speech intelligibility by phase-only modification [12]. Further, the authors in [13] reported the strong dependency of the perception of intervocalic stops on the phase information, supporting the positive impact of an enhanced spectral phase for an improved speech intelligibility in speech enhancement as recently reported in [14–16].

From our recent study in [17], existing instrumental metrics perform not entirely reliable in terms of predicting the perceived quality of phase-enhanced speech signals. Consequently, in this paper, we aim to extend our investigation on the issue of estimating the speech intelligibility of an enhanced speech signal provided by a phase-aware speech enhancement system. The phase-awareness could occur at two levels: at signal reconstruction where the noisy phase can be replaced with an enhanced one [14, 18–20], or in a joint way where the estimated phase is used to find a phase-aware amplitude estimate [21, 22]. We aim to evaluate the correlation of the existing intelligibility measures with our subjective results collected for phase-aware speech enhancement scenario. Further, we consider new phase-aware instrumental metrics and evaluate their reliability in terms of predicting their improved intelligibility performance.

The rest of the paper is arranged as follows. Section II presents a review on the conventional speech intelligibility measures. Section III presents the problem definition for phase-aware intelligibility of a phase-aware speech enhancement system where in contrast to the conventional enhancement methods the spectral phase is also modified. Further, we present some new phase-aware metrics. Section IV presents the results and finally Section VI concludes on the work.

2. Conventional Intelligibility Measures
The existing intelligibility metrics rely on different concepts to approximate the speech intelligibility. The concepts followed in the literature are: weighted sum of SNRs at frequency

The work was supported by Austrian Science Fund (P28070-N33).
bands [23–25], incorporating some auditory model as the band-importance frequency weighting [26], relying on the mutual information between the reference and modified signal [8, 9], using some octave-level normalization and compression [7]. In the following we present a brief overview for each category.

The first group estimates intelligibility via calculating the speech audibility at frequency bands, expressed as SNR. Those frequency bands possessing positive SNR contribute to speech intelligibility and consequently the overall speech intelligibility could be predicted as the weighted sum of SNRs defined across all frequency bands [2, Ch. 13]. The frequency bands are assigned following the band-importance function $W_k$ with $k$ as the frequency band index. The band-importance function is designed as such to reflect some auditory perception characteristics. Finally, an intelligibility score is provided as the weighted sum of the SNRs in frequency band $k$ denoted by $\text{SNR}_k$ weighted according to some psychoacoustic principle:

$$\text{SI} = \sum_{k=1}^{K} W_k \text{SNR}_k,$$  \hspace{1cm} (1)

where $K$ is the total number of frequency bands with $k \in [1, K]$ and $W_k$ denotes the band-importance function normalized to fulfill $\sum_{k=1}^{K} W_k = 1$, and SI is the objective speech intelligibility (SI) score. Equation (1) has been used to define several intelligibility measures including: speech intelligibility index (SII) [23], SNRloss, coherence SII (CSII) [25], normalized covariance measure (NCM), which all differ to one another in the way they calculate SNR and the selected weighting function. In the following, we present an overview on each measure.

As the first measure, SII incorporated two aspects: the hearing threshold level as well as equivalent masking spectrum in the form of level distortion factors that are taken into account in the calculation of an effective band SNR (also called audibility function). The band-importance function $I_k$ was specified for each individual band $k$ by a table based on fitting to a database [23]. The equivalent speech and noise spectrum levels denoted by $E_k$ and $D_k$ measured in decibels (termed as high presentation level) were used to approximate $\text{SNR}_k$:

$$\text{SII} = \sum_{k=1}^{K=18} I_k A_k(\text{SNR}_k),$$  \hspace{1cm} (2)

where $A_k(\text{SNR}_k)$ is the weighting function or band-audibility function, as a function of the calculated frequency band SNR. The SII metric ranges between 0 to 1, where a value below 0.45 and above 0.75 corresponds to a system with low and high intelligibility performance, respectively. SII has been used to optimize the near end listening enhancement [24].

Articulation index (AI) has been first introduced in [27], as a predictor of intelligibility. In [28], intelligibility was defined as the correct recognition of words which later Allen continued to find a relationship between the automatic speech recognition errors and the articulation index using Fletcher’s model (importance of articulation tests was reported by Fletcher [27]). The AI as a common index of speech audibility helps to emphasize the importance of a particular frequency relative to the contribution of incoming speech. It is calculated using the intensities of speech and noise received by the ear, both as a function of frequency. In [29], AI was used to define the intelligibility measure SII, calculated as a weighted sum of AI contributions weighted according to a band-importance function (similar to (1)).

Alternatively, CSII is similar to SII measure and differs by replacing the SNR term with the $\text{SNR}_{\text{est}}$ called signal-to-distortion ratio, calculated at critical bands filtered by a bank of rounded-exponential (Ro-ex) filters (for details see Table 1 in [23]) providing approximations to the auditory filters.

In [30], NCM was proposed as an intelligibility measure which relied on calculating the covariance between the input and output envelope signals in the frequency bands. The approximated SNR provided as such was limited between -15 to 15 decibels. A linear mapping scheme is used to compute the transmission index (TI) defined as a linear mapping of the SNR values at frequency band $k$ given by: $\text{TI}_k = \max(\text{SNR}_k, 0)$. The NCM index is then given by averaging all bands as follow:

$$\text{NCM} = \frac{\sum_{k=1}^{K} W_k \text{TI}_k}{\sum_{k=1}^{K} W_k}.$$  \hspace{1cm} (3)

The authors in [31] proposed the $\text{SNR}_{\text{loss}}$ metric calculated at critical-band spectral representation between the clean and noise-suppressed signals to predict the speech intelligibility loss introduced by a speech enhancement algorithm. The $\text{SNR}_{\text{loss}}(k, l)$ at frame $l$ and frequency band $k$ is defined as:

$$\text{SNR}_{\text{loss}}(k, l) = SNR_X(k, l) - SNR_{X}(k, l)$$ \hspace{1cm} (4)

where $X$ and $X$ are the STFT magnitudes of the input reference signal and the enhanced signal.

In [26], Dau proposed to incorporate his psychoacoustically validated model of auditory processing first presented in [32] for intelligibility prediction. The auditory model measured the similarity between the modified signal and the reference one at some internal representation level. At frames of length 20 ms, averaged cross-correlation coefficients are calculated and classified to three low, mid and high levels. The predicted intelligibility by the DAI metric is finally given by applying a logistic function on a weighted sum of these three level scores.

Mutual information between the message transmitted by talker and the message interpreted by the listener has been widely used as a natural measure to assess the intelligibility [1]. In particular, mutual information calculated in the frequency bands was shown to be a function of the SNR defined in frequency bands. Recently, the authors in [8] proposed the mutual information k-nearest neighbor (MIKNN) objective intelligibility measure relying on estimating the mutual information between the processed speech and the corresponding clean speech signal at temporal envelope. Finally, the authors in [9] proposed speech intelligibility prediction based on mutual information (SIM) as a function of the mean square error assuming that the intelligibility is monotonically related to the mutual information between the amplitude envelope of the clean and the processed signal at critical bands. Their results confirmed the reduced intelligibility obtained by amplitude-only modification of critical band amplitude.

Most of the intelligibility models rely on the global statistics captured at a sentence-level. In contrast, in [7] Taal proposed a short-time objective intelligibility measure (STOI) relying on shorter time segments (386 ms), shown to provide a higher correlation with the listening test results.

### 3. Motivation for a Phase-Aware Metric

#### 3.1. Phase and Speech Intelligibility

Alsteris and Paliwal in [11] reported that spectral phase information is as important as the STFT magnitude spectrum for...
speech intelligibility, however, it is not clear whether either of the amplitude and phase or both contribute to speech intelligibility in a complementary (or independent) fashion. To answer the question, the authors performed a detailed analysis of confusion matrices for consonant identification obtained from their experiments with no consistent pattern in their observation.

Here, we present a proof-of-concept example in order to investigate the importance of the clean phase spectrum solely on the speech intelligibility. We further include the phase estimation method proposed in [19] relying on a harmonic model-based approach to reconstruct the STFT phase spectrum across time and frequency. Results are shown in Figure 1 where plots of noisy and clean signals are also included for comparison. As noisy phase presents no distinct pattern, here we consider spectrogram, group delay and phase variance representations. The noisy signal is produced by contaminating a female utterance saying “bin blue at p four soon” with white noise added at 10 (dB) signal-to-noise ratio. The utterance consists of stop consonants and fricatives, reported important for speech intelligibility [33].

Looking at the STOI scores in Figure 1 it is observed that the phase enhancement method STFTPI [19] degrades the speech intelligibility by 10%, with respect to the unprocessed noisy input speech. This counter example shows that phase enhancement leads to a degradation in the intelligibility performance predicted by the existing intelligibility measures (similar degraded performance was found using the other instrumental intelligibility metrics). Our observation motivates to investigate for other phase-aware intelligibility measures as well as to find out the correlation of the existing intelligibility measures with some subjective listening results.

3.2. Proposed Phase-Aware Intelligibility Measures

As our list of phase-aware intelligibility measure candidates, we consider our recent proposals in [17] consisting of three metrics: 1) group-delay (GD), 2) instantaneous frequency deviation (IFD), and 3) phase deviation (PD) [34]. We propose two new metrics as UnHPSNR and UnRMSE, defined in the following to measure the phase estimation error in the unwrapped domain.

As our first proposed metric, we propose unwrapped harmonic phase SNR (UnHPSNR) defined as follow:

$$\text{UnHPSNR} = \frac{1}{\sum_{h,l} X^2(h,l)} \left( \sum_{h,l} X^2(h,l) - \sum_{h,l} X^2(h,l) \left( 1 - \cos(\psi(h,l) - \hat{\psi}(h,l)) \right) \right),$$ (5)

measured in decibels where we define $$X(h,l)$$ as the spectral amplitude sampled at harmonic $$h$$ and frame index $$l$$, respectively, and $$\psi(h,l)$$ and $$\hat{\psi}(h,l)$$ are the corresponding unwrapped harmonic phase spectra. The unwrapped phase spectra are provided after subtracting the linear phase part, following the pitch-synchronous framing with phase decomposition principle proposed in [35].

As our second metric we propose unwrapped root mean square estimation error (UnRMSE) defined as:

$$\text{UnRMSE} = \sqrt{\frac{1}{\sum_{h,l} X^2(h,l)} \left( \sum_{h,l} X^2(h,l) \right) \left( \sum_{h,l} X^2(h,l) \left( \psi(h,l) - \hat{\psi}(h,l) \right)^2 \right)},$$ (6)

measured in (dB). The choice of phase variance without the mean was motivated by the recent findings on the importance of phase variance for voice quality assessment [36] and the negligible impact of the phase mean [35]. The weighting by the spectral amplitude emphasizes on the phase variance error at harmonics that are arguably perceptually more important.

It is important to note that the proposed UnHPSNR and UnRMSE measures concentrate on quantifying the estimation error occurred in the unwrapped harmonic phase, introduced by the phase modification procedure. Therefore, in the design of both metrics, we assume that the harmonic amplitude $$X(h,l)$$ is known considered as the weighting. The clean phase attains the infinity value for both UnHPSNR and UnRMSE measures.

4. RESULTS

4.1. Experimental Data

Throughout objective and subjective evaluations, here we aim to address the following questions: First, to address how much the existing metrics correlate with subjective speech intelligibility results for phase-aware speech enhancement, and secondly to find out whether the newly proposed phase-aware metrics provide a reasonable prediction of the intelligibility performance provided by the phase-aware speech enhancement methods. The test material consisted of 50 sentences from the GRID corpus [37] including male and female speakers downsamplled to 8 kHz. Each sentences have a duration of approximately two seconds. The clean speech signals are mixed with white and babble noise taken from NOISEX-92 [38] at SNRs of 0 and 5 dB. We included four enhancement scenarios: Conventional (C) [MMSE-LSA [39], Conventional + STFT Phase Improvement (C + STFTPI) [19], Conventional + Phase Enhanced (C + PE) [14] and Phase-Aware (PA) [21]. Together with the unprocessed speech signals (UP) 1000 speech files were used in the analysis. Scores of the instrumental metrics were obtained by averaging over all utterances for each method and SNR.

4.2. Subjective Intelligibility Test

A panel of 12 normal-hearing listeners participated in the subjective listening test, which was held in a quiet room at the Graz University of Technology. GRID sentences were presented to the listeners over AKG K 601 High-End Stereo Headphones. We followed the principle and standard described in [40] to conduct a subjective listening test for speech intelligibility performance evaluation. A graphical user interface was designed to collect the inputs from the participants left in the analysis. The participants were instructed to choose for the right color, letter and number at each presented utterance selected from the GRID corpus. The test was organized in four blocks according to noise type and decreasing SNR. Within each block for each method
four randomly selected utterances were presented to the participants, where the order of the methods itself also was randomized. To check the reliability of the participants, clean reference utterances were included. Participants were taken out of the follow up statistical analysis of the listening test outcome once their intelligibility score for the clean utterances showed a value lower than the noisy utterances (as clean signal is supposed to have the highest performance). This led to 10 participants in the test. A training session was provided.

Figure 2 shows the intelligibility scores and 95% confidence intervals differentiated in terms of noise type and SNR. An overall improvement in intelligibility is more pronounced in babble rather than in white noise except for C + PE compared to C in white than babble noise verified with the results we reported in [17] (where for C + PE against C better quality improvement was shown in white noise case). The iterative amplitude-phase estimation method is labeled as PA which outperforms all the other methods in every scenario. For babble 0 dB case this improvement is significant. 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 5 dB case. Finally, the method C degrades the intelligibility of the noisy input, confirming the previous observations made in [3, 41].

4.3. Performance Evaluation

The Pearson’s correlation coefficient ($\rho$), the normalized root-mean-square error ($\sigma$) and Kendall’s Tau ($\tau$) were used to evaluate the prediction power of the presented instrumental metrics. To account for the non-linear relationship between objective and subjective scores a logistic function was applied.

Figure 3 shows the results separated by noise type and a third graph combining both noise types. The proposed measure UnRMSE 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 the fact that 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. Both metrics performed better for white noise than babble which is a result of the inaccuracy in $f_0$ estimation required for phase decomposition in the babble noise scenario. CSII-based measures exhibited reliable prediction in both noise scenarios. For babble noise and the overall scenario CSIIm was the top performing metric. This is supported by an observation of Kates [25] that the mid-level CSII contains much information on envelope transients. These transients are mostly effected by phase-enhancement which as a consequence led CSIIm as a reliable intelligibility predictor for phase-enhanced speech. The correlation results combining both noise types were in general at a lower level. This is due to two facts: first, no metric was capable to predict the intelligibility loss by STFTPI at 0 (dB) for babble noise as seen in Figure 1 and second, an overestimation of intelligibility occurred in white noise, in particular for the PA outcome predicted by UnRMSE, leading to an intermediate correlation score.

5. Acknowledgment

The authors would like to thank Prof. Torsten Dau for sharing the DAU model implementation [26], and Prof. Rainer Martin and Jalal Taghia for the Matlab codes in [8], and Prof. Peter Vary for fruitful discussions. We also thank Dr. Anton Schlesinger for his helpful discussions.

6. Conclusion

In this paper we addressed two questions; the reliability of the existing instrumental metrics when used to estimate the intelligibility of a phase-aware single-channel enhanced speech signal where both spectral amplitude and phase of the noisy input are modified, and to find out whether a new proposed phase-aware intelligibility measure could outperform the existing ones in terms of having a reliable prediction of the subjective intelligibility results. Throughout statistical analysis, we quantified the correlation between the intelligibility metrics and the listening results. Our results showed that the proposed phase-aware metrics present a comparable reliable prediction compared to the top-performing ones as CSII and STOI measures.

7. References


