Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 65 (2014) 9–19

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

An improved speech transmission index for intelligibility prediction

Belinda Schwerin *, Kuldip Paliwal

Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Nathan, QLD 4111, Australia

Received 12 September 2012; received in revised form 11 April 2014; accepted 16 May 2014
Available online 2 June 2014

Abstract

The speech transmission index (STI) is a well known measure of intelligibility, most suited to the evaluation of speech intelligibility in rooms, with stimuli subjected to additive noise and reverberance. However, STI and its many variations do not effectively represent the intelligibility of stimuli containing non-linear distortions such as those resulting from processing by enhancement algorithms. In this paper, we revisit the STI approach and propose a variation which processes the modulation envelope in short-time segments, requiring only an assumption of quasi-stationarity (rather than the stationarity assumption of STI) of the modulation signal. Results presented in this work show that the proposed approach improves the measures correlation to subjective intelligibility scores compared to traditional STI for a range of noise types and subjected to different enhancement approaches. The approach is also shown to have higher correlation than other coherence, correlation and distance measures tested, but is unsuited to the evaluation of stimuli heavily distorted with (for example) masking based processing, where an alternative approach such as STOI is recommended.

Keywords: Speech transmission index; Modulation transfer function; Speech enhancement; Objective evaluation; Speech intelligibility; Short-time modulation spectrum

1. Introduction

The enhancement of speech corrupted by noise has received much focus in recent years, with many new technologies reliant on their ability to provide high quality or intelligible signals in a range of everyday environments. The intelligibility of speech in such environments is of particular importance, with many devices making use of voice activation or control interfaces as an alternative to traditional button or key based systems. Consequently, there has been much research into the development of improved methods for reducing noise in speech signals so as to improve the quality of the resulting sound and the intelligibility of speech.

The development of such methods relies on the ability to compare or evaluate the quality of speech processed using various approaches. Quality is typically characterised in terms of the level of distortion which is audible in the signal. Intelligibility, on the other hand, is characterised by the amount of speech that can be correctly recognised. While enhancement methods aim to improve one or both of these characteristics without introducing unwanted distortions, in practice, residual noise is often left in the processed stimuli, and other nonlinear types of distortion are often introduced by the processing method itself. Evaluation of the effectiveness of a proposed enhancement algorithm relies on comparison of stimuli processed using the proposed approach with that processed by other methods. For this purpose, listening tests are considered the most reliable indicator of performance. However, these are time consuming and costly, and therefore testing using objective metrics, which are comparatively fast and cost effective, are valuable in at least some stages of the development process. As a result, much attention to the

* Corresponding author.
E-mail address: belsch71@gmail.com (B. Schwerin).

http://dx.doi.org/10.1016/j.specom.2014.05.003
0167-6393/

development of effective objective indicators of quality and/or intelligibility are of particular interest.

Many of the metrics proposed for predicting speech intelligibility have been based on finding the distance between clean and corrupted speech, in either the time or spectral domain. Early efforts to predict intelligibility in this way, such as the articulation index (AI) (French and Steinberg, 1947) and the speech intelligibility index (SII) (ANSI, 1997), were based on the assumption that the intelligibility of speech is given by the sum of the contributions to intelligibility within individual frequency bands (French and Steinberg, 1947). In these approaches, a function of the signal-to-noise ratio (SNR) is used to represent the contributions within each frequency band. A weighted average across the bands was then used to calculate an intelligibility score, which was found to correlate well with the subjective intelligibility of stimuli corrupted with additive noise.

The speech transmission index (STI) (Steeneken and Houtgast, 1980) is a well known objective measure, widely used for assessing room acoustics. Instead of calculating the SNR in each band, it calculates the reduction in the modulation-depth as a function of modulation frequency via a modulation transfer function (MTF). The change in modulation-depth with reference to the original signal, is considered to reflect the intelligibility of speech (in a sound transmission system) (Houtgast and Steeneken, 1985). As a result, the STI measure has good correlation to subjective intelligibility scores for stimuli distorted by linear filtering, reverberation, and additive noise.

However, the STI measure suffers from the problem of being poorly correlated to the subjectively measured intelligibility of stimuli subjected to nonlinear processing (Goldsworthy and Greenberg, 2004). In speech enhancement, algorithms generally operate in the frequency domain, applying a suppression function to the noisy magnitude spectra in order to enhance speech. Additionally, short-time processing is typically used, with effects of processing varying from frame to frame. This nonlinear processing of speech signals makes the STI measure unsuited to the evaluation of stimuli subjected to speech enhancement.

The poor correlation of STI for evaluating the intelligibility of enhanced speech may in part be attributed to its assumption that the intensity envelope (referred to here as the modulation signal) is stationary, with STI performing Fourier analysis over the whole utterance. Not only is this assumption invalid, but recent work using the short-time modulation domain for speech enhancement has demonstrated that shorter modulation window durations of around 32 or 64 ms are more beneficial when using the modulation magnitude spectra for improving speech intelligibility, and 256–512 ms durations are more beneficial when processing the modulation spectra (*i.e.*, modulation magnitude and phase) (Paliwal et al., 2011). Therefore, in this work, we revisit the STI approach and propose a variation whereby we only assume quasi-stationarity of the modulation signal. We apply short-time processing to the modulation signals of degraded and clean speech and calculate the reduction in modulation depth (compared to the clean), then use this to calculate an intelligibility score. Results presented show that using the proposed variation, we achieve an improved correlation to subjective intelligibility scores for stimuli subjected to nonlinear processing.

The rest of the paper is organised as follows. In Section 2, we describe the proposed QSTI method of predicting speech intelligibility. Section 3 describes experimental procedures used to evaluate intelligibility measure performance. In Section 4 we begin with an evaluation of the parameters affecting the QSTI measure, then discuss results of intelligibility experiments comparing the correlation of the proposed and other well known objective measures with subjective intelligibility scores. Final conclusions are drawn in Section 5.

2. Method

STI is an objective measure that is widely used for assessing room-acoustics, having good correlation to subjective scores for stimuli distorted by reverberation, linear filtering, as well as additive noise. It measures the extent to which slow temporal intensity envelope modulations, which are important for speech intelligibility, are preserved in degraded listening environments (Payton and Braida, 1999). In the speech-based STI procedure, the original and degraded speech signals are passed separately through a bank of six (or seven) octave band filters with centre frequencies from 125 Hz to 4000 (or 8000) Hz. The envelope signal is then calculated and used to compute the transmission index for each band. The final STI score is found as a weighted sum of the transmission index for each band.

While there are a number of variations on the STI method, they generally only differ in how the envelope signals and the transmission index are calculated. For the traditional STI (Steeneken and Houtgast, 1980), and variations such as that of Payton and Braida (1999) and Drullman et al. (1994), the envelope signals are calculated as follows. Each band filtered signal is squared, then low pass filtered with a 50 Hz cutoff frequency to extract the temporal intensity envelope of each signal. The modulation spectrum is then found by calculating the FFT of each intensity envelope for the entire signal. The resulting spectra of the original and degraded speech, in each band, can then be used to calculate the modulation transfer function, and subsequently the STI. In the variation by Drullman et al. (1994), the MTF is calculated by finding the ratio of the real part of the cross-power spectrum of the degraded and clean speech to the power spectrum of the clean signal. One-third octave band analysis over each of 14 intervals with centres ranging from 0.63 to 12.7 Hz (Goldsworthy and Greenberg, 2004) are then summed to produce 84 (or 98) MTF indices.

A number of variations on the STI method have been proposed in an effort to improve the correlation for stimuli subjected to other types of distortions, however they all

generally still suffer from the problem of being poorly correlated to the subjectively measured intelligibility of stimuli subjected to nonlinear processing (Goldsworthy and Greenberg, 2004). As mentioned earlier, this may in part be attributed to its assumption that the characteristics of the intensity envelope are stationary over the entire utterance. While the properties of noise and speech modulation spectrum are typically slow varying, they are not stationary. Therefore, in this work we propose a variation on the traditional speech-based STI approach in which the modulation signal in each band is processed using short-time Fourier analysis. The resulting method is now described.

2.1. QSTI intelligibility measure

The proposed quasi-stationary speech transmission index (denoted QSTI), is computed using a procedure shown in the form of a block diagram in Fig. 1. It uses degraded and clean speech to calculate a scalar score which we demonstrate to be monotonically related to speech intelligibility obtained subjectively from human listeners.[1] Both clean and degraded signals are assumed to be time-aligned and are resampled to 8000 Hz.

Similar to STI, the signals are initially filtered into 6 overlapping, octave spaced acoustic bands with centre frequencies at 125, 250, 500, 1000, 2000, and 4000 Hz. The intensity envelope for each band is then found by squaring the band-pass filtered signal and applying a low pass filter with cutoff frequency of 50 Hz. This envelope (or modulation signal) is then normalised to unit mean to account for the power in each signal.

The modulation signal for each band (denoted $X(n,b)$ for the clean signal, and $Y(n,b)$ for the degraded signal) are framed and processed using short-time Fourier analysis, to give the complex modulation spectra $\mathcal{X}(\eta,b,k)$ and $\mathcal{Y}(\eta,b,k)$ for each modulation frame and acoustic band. Thus, for the clean signal, the modulation spectra is given by

$$\mathcal{X}(\eta,b,k)=\sum_{\ell=0}^{N-1}X(\ell+\eta Z,b)w(\ell)e^{-j2\pi\ell k/N}, \tag{1}$$

where η is the modulation frame index, b is the acoustic band index, k is the modulation frequency index, N is the frame duration in samples, Z is the frame shift, and $w(\ell)$ is the analysis window function. A Hamming window was used for the analysis window. A frame duration (MFD) of 512 ms is used to adequately represent both magnitude and phase information in the signal, and a frame shift (MFS) of 1/8 of the MFD is used (see Section 4.1 for an evaluation of the effect of these parameters on the performance of QSTI).

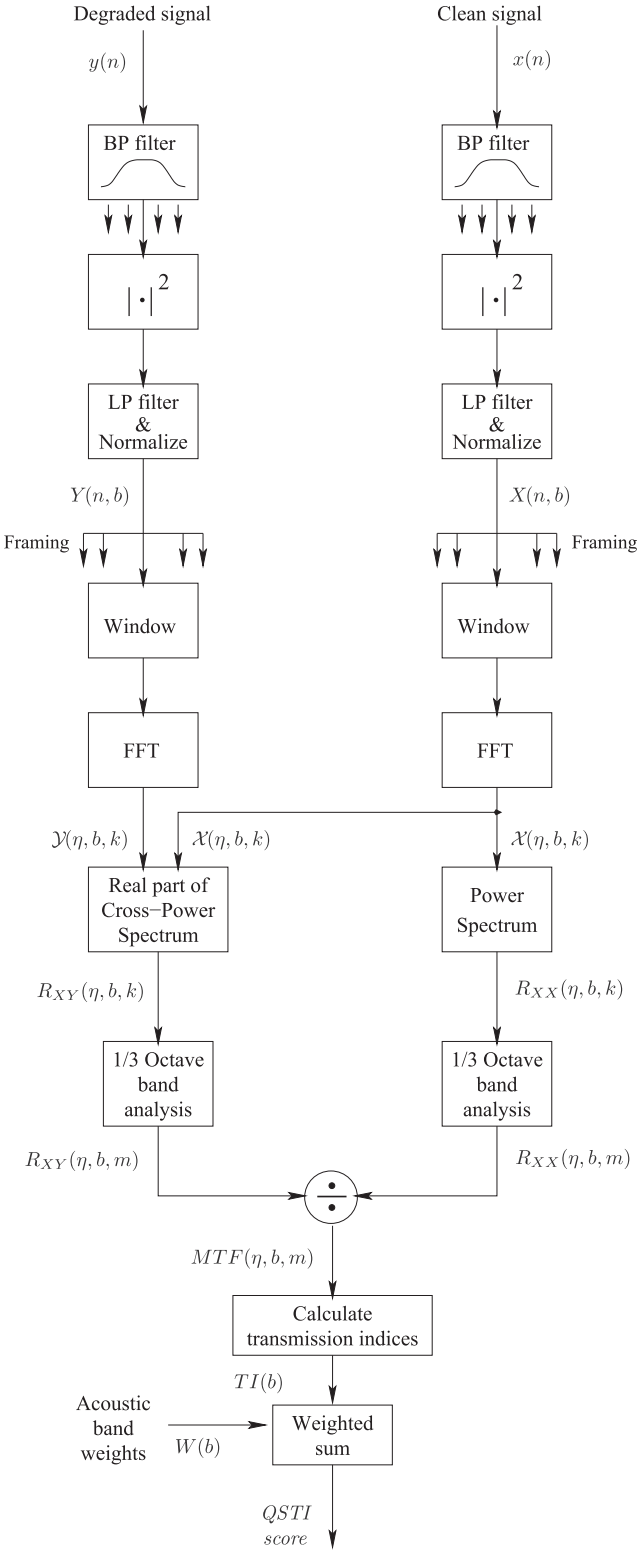

Fig. 1. Block diagram showing the method of calculating the QSTI objective intelligibility score.

The fractional change in modulation depth is then calculated for each modulation frequency as the real part

[1] Subjective speech intelligibility is used throughout this work to refer to the intelligibility of speech stimuli as determined by experiments in which human listeners identify words in speech stimuli. Subjective intelligibility scores and listener recognition scores are used synonymously to refer to the results of these listening experiments.

of the cross-power spectrum of the clean and noisy modulation signals,

$$R_{XY}(\eta,b,k) = Re\{\mathcal{X}(\eta,b,k).\mathcal{Y}^*(\eta,b,k)\}, \quad (2)$$

where $\mathcal{Y}^*(\eta,b,k)$ is the complex conjugate of $\mathcal{Y}(\eta,b,k)$. The power spectrum of the original signal is calculated for each modulation frequency as

$$R_{XX}(\eta,b,k) = \mathcal{X}(\eta,b,k).\mathcal{X}^*(\eta,b,k), \quad (3)$$

where $\mathcal{X}^*(\eta,b,k)$ is the complex conjugate of $\mathcal{X}(\eta,b,k)$.

Third octave band analysis is next applied to both the cross-power spectrum and clean power spectrum, with band centre frequencies ranging from 0.63 Hz to 16 Hz, to give $R_{XY}(\eta,b,m)$ and $R_{XX}(\eta,b,m)$. This corresponds to use of 15 one-third octave spaced, rectangular shaped, overlapping bands. Spectra values are averaged across each of these bands. Given the resolution of the spectrum, this acts to remove frequencies less than 1 Hz and greater than around 18 Hz.

The MTF for each acoustic and modulation band and modulation frame is then calculated as the ratio of the clean power spectrum and cross-power spectrum,

$$\mathrm{MTF}(\eta,b,m) = \frac{R_{XX}(\eta,b,m)}{R_{XY}(\eta,b,m)}, \quad (4)$$

where m is the modulation band index. Note that the calculation of the MTF in Eq. (4) differs from the original STI where MTF was calculated as $\mathrm{MTF}(b,m) = R_{XY}(b,m)/R_{XX}(b,m)$. This difference enables QSTI to account for negative spectral distortions, where information in the spectrum has been lost due to over-suppression by an enhancement algorithm. That is, for frames where envelope information is over-suppressed in the degraded envelope compared to the clean, a small MTF (close to 0) and therefore intelligibility score results. When the envelope of the clean and degraded stimuli have similar shape (well correlated), a higher MTF value results.

The MTFs of each modulation band and frame are then used to calculate a transmission index (TI) for each acoustic band as follows. Mean MTFs are first found by averaging MTFs across modulation bands for each modulation frame and acoustic band,

$$\mathrm{uMTF}(\eta,b) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{MTF}(\eta,b,m), \quad (5)$$

and then limiting them to values between 0 and less than 1 by a clipping operation.

The signal-to-noise ratio (SNR) is then calculated as

$$\mathrm{SNR}(\eta,b) = 10\ \log\left(\frac{\mathrm{uMTF}(\eta,b)}{1-\mathrm{uMTF}(\eta,b)}\right), \quad (6)$$

and limited to values between -15 and $+15$ dB by a clipping operation. Finally, SNRs are averaged across modulation frames, and mapped to values between 0 and 1, to give a TI for each acoustic band as follows,

$$\mathrm{TI}(b) = \frac{\frac{1}{N}\sum_{\eta=1}^{N}\mathrm{SNR}(\eta,b)+15}{30}. \quad (7)$$

The QSTI score is finally calculated as a linearly weighted sum of TIs as follows:

$$\mathrm{QSTI} = \frac{\sum_{b=1}^{B}\mathrm{TI}(b).W(b)}{\sum_{b=1}^{B}W(b)}, \quad (8)$$

where $W(b)$, $b = 1,2,\ldots,B$, are the weights assigned to individual acoustic bands and are given as $W(b) = \{1,1,1,1,0.75,0.425\}$. These weights were determined empirically, through experiments using the testing corpus (see Section 3), as those resulting in predictions with the highest correlation to subjective intelligibility scores.

Eq. (8) gives the final QSTI intelligibility score with values between 0 and 1.

3. Experimental procedure

In the remainder of this work we evaluate the proposed QSTI method described in Section 2.1. Experiments conducted evaluate the effectiveness of this method at predicting speech intelligibility of enhanced stimuli with reference to the clean signal,[2] by calculating the correlation of predicted intelligibility with human listener recognition scores. The proposed method is then compared with the correlation of well known methods in the literature. Methods included in comparisons are listed with their abbreviations in Table 1.

The STI method is implemented using the MTF calculation method of Drullman et al. (1994) (STI-D) as described in (Goldsworthy and Greenberg, 2004) and briefly reviewed in Section 2. The SII method (ANSI, 1997) is implemented with use of programs publicly available at url: www.sii.to/html/programs.html. For correlation and coherence based measures, we include ESC, NCM_w, STOI, MSC, and CSII. The ESC method (Ma and Loizou, 2011) is based on calculating the correlation between enhanced and clean stimuli along frequency then averaging across frames. This method is noted to better account for non-linear distortions. The NCM_w method (Ma et al., 2009) is based on the normalised covariance metric of Holube and Kollmeier (1996), which calculates an intelligibility score from the weighted sum of transmission index (TI) values determined from the covariance between the clean and degraded envelope signals in each frequency band (Goldsworthy and Greenberg, 2004). The variation of Ma et al. (2009) applies weights to TI values which are signal and frequency dependent to achieve improved correlation for stimuli subjected to nonlinear processing. The STOI approach (Taal et al., 2011) is based on calculating the correlation along short-time segments, using clipping instead of limited SNR ranges to limit the maximum

[2] Use of the clean signal as a reference is in contrast to non-intrusive intelligibility measures, which rate the intelligibility of stimuli without knowledge of the original or clean speech signal.

Table 1
Objective measures evaluated and their abbreviations.

Abbrev.	Objective measure
STI-D	Speech transmission index (Drullman et al., 1994)
ESC	Excitation spectra correlation method (Ma and Loizou, 2011)
NCM_w	Normalised covariance measure with modified weights (Ma et al., 2009)
STOI	Short-time objective intelligibility method (Taal et al., 2011)
SII	Speech intelligibility index (ANSI, 1997)
CSII	Coherence speech intelligibility index (Kates and Arehart, 2005)
MSC	Magnitude-squared coherence method (Carter et al., 1973)
LLR	Log-likelihood ratio (Quackenbush et al., 1988)
fwSegSNR	Frequency weighted segmental SNR (Tribolet et al., 1978)
PESQ	Perceptual evaluation of speech quality (ITU-T P.862, 2001)

allowed distortions. Its use of short-time segments enables it to better account for non-stationary distortions.

The MSC (Carter et al., 1973) and CSII (Kates and Arehart, 2005) methods are both coherence based techniques. The MSC method is based on the magnitude-squared coherence between signals, and represents the fraction of power that is linearly related between clean and degraded signals along frequency. In using the cross-spectral density of overlapped short-time frames, MSC also accounts for in-phase and out-of-phase spectra. CSII makes use of the complement of the degraded speech spectra to calculate the output power that is unrelated, that is, the non-linear distortion and noise.

Techniques based on spectral distance measurement have also been included in comparisons. Of this type are the frequency weighted segmental SNR method (Tribolet et al., 1978), the log-likelihood ratio (Quackenbush et al., 1988), as well as the PESQ quality measure (ITU-T P.862, 2001). While these methods were intended for objectively evaluating speech quality, they also work well as an intelligibility metric.

In order to evaluate the above described methods, the corpus and the listener recognition scores from the sentence intelligibility evaluation study reported in (Hu and Loizou, 2007) have been used in this work. In the cited study, the recordings available in (Loizou, 2007), consisting of all the sentences from the IEEE sentence database (Rothauser, 1969)[3] recited by a male speaker, were down-sampled to 8 kHz and additively corrupted with 4 real-world recorded noises from the AURORA database (babble, car, street, and train) (Pearce and Hirsch, 2000) at SNRs of 0 and 5 dB. Then, 8 noise-suppression algorithms were applied to produce a total of 72 treatments, including unprocessed noisy stimuli. Using these sentences as the corpus, the study in (Hu and Loizou, 2007) conducted subjective intelligibility measuring experiments involving 40 native American English speaking participants. Each listener assessed a total of 18 different treatments, each one consisting of 20 sentences, ensuring no subject listened to the same sentence twice. Finally, mean subjective intelligibility scores were found for each treatment type from the percentage of words correctly identified (where all words were considered in the scoring).

In this work, two experimentation sections are included. In the first, parameter values are evaluated, and in the second, the QSTI measure is compared to other objective intelligibility measures. In the first case, 16 of the 72 treatment types of the above corpus are used. Those selected represent a spread of the SNRs, noise types and enhancement method types applied.[4] This subset of the corpus is referred to as the tuning set, and are used for experiments presented in Section 4.1. The remaining 56 treatment types of the corpus are used for comparative experiments presented in Section 4.2, and are referred to as the testing set. In this way, separate stimuli are used for tuning of parameters and testing experiments.

For experiments presented in Section 4, each objective intelligibility measure tested is applied to each of the sentences of the tuning set (for experiments of Section 4.1) or testing set (for experiments of Section 4.2) from the above corpus. A mean (unmapped) score for each treatment type is then obtained by averaging objective scores for each sentence.

To enable secondary evaluation of the above described methods, the corpus and listener recognition scores from the intelligibility evaluation study employed in (Taal et al., 2010) to evaluate STOI, has also been utilised. This is a much smaller corpus consisting of 10 utterances (each consisting of 5 Danish words) corrupted with speech-shaped noise added at SNRs of −8.9, −7.7, −6.5, −5.2 and −3.1 dB, and processed using two different noise suppression algorithms (namely AME (Ephraim and Malah, 1984), and the estimation approach of Erkelens et al. (2007)). This gives a total of 15 different treatment types (including noisy). Fifteen native Danish-speaking listeners participated in the recognition experiment, listening to each of the sentences of the corpus, to give mean recognition scores for each treatment type.

In order to ensure the monotonic relationship between objective scores and subjective recognition scores, a mapping function is also applied. For this purpose, and as done by many works in the literature (*e.g.*, Ma and Loizou, 2011; Taal et al., 2011; Kates and Arehart, 2005; Boldt and Ellis, 2009; Christiansen et al., 2010), we use a logistic function such as

$$l(x) = \frac{1}{1 + e^{a+bx}}, \tag{9}$$

where x is the objective score, and parameters a and b are known as regression coefficients. Values for these

[3] IEEE database contains phonetically balanced sentences with low word-context predictability.

[4] The tuning set includes 2 treatment types for each SNR and noise type combination, and different enhancement method types such that each enhancement type is included only twice. This results in a total of 16 different treatments types for the tuning corpus.

coefficients are computed through a logistic regression (Balakrishnan, 1992), as those which best fit the objective scores to the subjective intelligibility scores. The logistic function always takes on values between zero and one, and can be interpreted as a way of describing the relationship between one or more independent variables and a probability, in our case the probability of correctly identifying words (*i.e.*, subjective intelligibility score).

The leave-one-out cross-validation procedure is applied to parameter fitting and logistic function mapping in order to ensure use of mutually exclusive training and testing sets, and thereby prevent over-fitting. In this procedure, regression parameters are determined using the entire data set excepting one treatment. Intelligibility predictions for the excluded treatment are then evaluated via the above logistic function with these determined parameters, providing a mapped score. The procedure is repeated for every treatment in the tuning/testing set from the corpus.

To assess the performance of each technique as a predictor of corrupted speech intelligibility, Pearson's correlation coefficient, r, between subjective and objective scores is used. Correlation is calculated both for the unmapped intelligibility scores, as well as the scores mapped using the above described logistic function. In addition, the standard deviation of the prediction error σ_e is calculated as $\sigma_e = \sigma_d\sqrt{1-r^2}$, where σ_d is the standard deviation of the speech intelligibility scores for a given treatment type. Absolute values of r nearer to one and smaller values of σ_e indicate a better speech intelligibility prediction. Additionally, maximum absolute prediction errors (MAE) are calculated to give further indication of measure performance. Here, MAE is the maximum error in the predicted intelligibility for treatments in the corpus, calculated as

$$\mathrm{MAE} = \mathrm{Max}_i(x_i - y_i). \tag{10}$$

4. Results

4.1. Evaluation of QSTI and its parameters

QSTI has a number of parameters which effect the performance of the method, including the modulation frame duration (MFD) and modulation frame shift (MFS).

The choice for MFD may be rationalised from previous work utilising short-time modulation domain processing. In (Paliwal et al., 2011), the contributions of the modulation magnitude and phase spectra towards intelligibility were investigated, and found that modulation phase spectra has more speech information and makes a greater contribution to intelligibility when using larger MFD (such as 512 ms or 1024 ms) than when shorter MFD was used. It was also shown that when processing the modulation magnitude spectra, much shorter MFD should be used to improve intelligibility. Here, we utilise the complex modulation spectrum (with both magnitude and phase components), and therefore we have chosen an in-between duration of 512 ms. To affirm this rationalisation we compare the correlation coefficients of the QSTI method using MFD values ranging from 32 ms to 1024 ms. In each case, the MFS was kept to 16 ms (half the shortest MFD tested), and the Fourier analysis length was set to give a FFT resolution of 0.9765 Hz (i.e., an FFT length of 8192). The resulting correlations between objectively measured intelligibility scores (both mapped and unmapped) for the tuning set from the sentence corpus and the recognition scores from real listeners are shown in Table 2. As can be seen, there is a much higher correlation to listener intelligibility scores using a MFD of 512 ms than using either higher or lower durations. The improved performance of the QSTI measure using an MFD of 512 ms is therefore consistent with previous findings.

For enhancement methods which process the short-time modulation spectrum, MFSs ranging between just a few samples and up to half a frame are used, depending on the modification method applied, the stationarity of the noise corrupting stimuli, and the importance of calculation time. To evaluate the effect of MFS on the performance of the QSTI measure, we calculated intelligibility scores for the tuning set of the sentence corpus using the QSTI measure and MFS values ranging from 16 to 256 ms. In each case, MFD was held at 512 ms and the FFT resolution at 0.9765 Hz. Resulting correlation coefficients between unmapped and mapped objective intelligibility scores and listener recognition scores are shown in Table 3. Results show there is only a small difference in the performance of the QSTI measure with change in MFS. Best results were for one eighth of a frame shift (64 ms), but given the small difference in performance, a larger MFS could be used to reduce calculation time with only a small drop in performance.

In addition to these parameters, the varying amounts of silence in stimuli being processed can often adversely affect or alter the performance of an objective measure. Therefore some approaches, such as STOI, incorporate silence removal as part of their algorithm. We now investigate the effect of silence removal on the performance of the proposed QSTI measure. For this purpose, we determine and compare the correlation of the QSTI measure with listener recognition scores for QSTI, both with and without silence removal.

Silence removal is a pre-processing procedure which removes frames from the clean and degraded signals that

Table 2
Correlation coefficients (r) and standard deviation of the prediction errors (σ_e) between the predictions from QSTI and the recognition scores from real listeners, when QSTI is applied using each of the MFDs shown.

MFD (ms)	Unmapped r	Mapped r	Mapped σ_e
32	0.495	0.273	0.186
64	0.508	0.286	0.186
128	0.646	0.488	0.167
256	0.817	0.751	0.124
512	0.929	0.905	0.080
768	0.889	0.873	0.093
1024	0.811	0.808	0.112

Table 3
Correlation coefficients (r) and standard deviation of the prediction errors (σ_e) between the predictions from QSTI and the recognition scores from real listeners, when QSTI is applied to the corpus tuning set using each of the MFSs shown.

MFS (ms)	Unmapped r	Mapped r	Mapped σ_e
16	0.929	0.905	0.080
32	0.931	0.908	0.079
64	0.933	0.912	0.077
128	0.927	0.902	0.082
256	0.913	0.886	0.087

are identified as being silence, before continuing with the QSTI calculation procedure as described in Section 2.1. Here, 32 ms frames with 50% overlap and the Hanning window function are used. Silence frames are identified by considering the energy of each frame in the clean signal. Frames where the energy of the clean signal is lower than the silence threshold are identified as silence, and removed from both the clean and degraded stimuli. The silence threshold is set as the maximum clean frame energy minus the silence threshold range (θ_R), where θ_R is a parameter of the silence removal procedure. Silence removal is therefore more aggressive for smaller values of θ_R, and less aggressive for larger values of θ_R.

Table 4 gives correlation coefficients and standard deviation of prediction errors between QSTI predictions and listener recognition scores without silence removal and with silence removal for the indicated silence threshold range θ_R. Results show that QSTI has best performance for a silence threshold range of 40 dB. Use of more aggressive silence removal (i.e., reduced θ_R), is shown to result in reduced correlation. This is due to the removal of lower energy speech regions, which correspond to speech entries and exits and lower energy consonants, which are important to the intelligibility of speech. QSTI is shown to still perform well without any silence removal, due to its use of sufficiently long modulation frame durations.

Based on the results above, remaining experiments give results for QSTI implemented with a 512 ms MFD, 64 ms MFS, and silence removal with a threshold of 40 dB below maximum.

Table 4
Correlation coefficients (r) and standard deviation of the prediction errors (σ_e) between the predictions from QSTI and the recognition scores from real listeners, when applied to the tuning set corpus using the indicated level of silence removal. Values for θ_R shown are the silence threshold range, and indicate the SNR below maximum to be used as the lower limit for identifying frames containing speech.

θ_R	Unmapped r	Mapped r	Mapped σ_e
–	0.917	0.890	0.085
40 dB	0.933	0.912	0.077
30 dB	0.912	0.883	0.088
20 dB	0.879	0.852	0.099

Table 5
Correlation coefficients (r) and standard deviation of the prediction errors (σ_e) between the predictions from each method and the recognition scores from real listeners, when applied to the testing set corpus.

Measure	Unmapped r	Mapped r	Mapped σ_e
STI-D	0.497	0.450	0.154
fwSegSNR	0.820	0.811	0.101
ESC	0.827	0.814	0.100
NCM_w	0.891	0.883	0.081
STOI	0.854	0.842	0.093
SII	0.655	0.616	0.136
CSII	0.871	0.859	0.088
MSC	0.822	0.814	0.100
LLR	−0.640	0.607	0.137
PESQ	0.800	0.790	0.106
QSTI	0.938	0.939	0.059

4.2. Comparison with other objective methods

In this work we have proposed an STI-based objective measure of intelligibility which uses short-time Fourier analysis of the modulation signal in different frequency bands to characterise speech. In this section, we now compare the performance of this proposed method with that of the other objective measures listed in Table 1.

Table 5 summarises the correlation coefficients (r) and the standard deviation of the prediction errors (σ_e) between the objective scores of each method and the recognition scores from real listeners across the testing set of the sentence corpus. Results show QSTI preformed very well having a correlation of $r = 0.94$, with low prediction error standard deviation of $\sigma_e = 0.06$. NCM_w and CSII, both non-linear measures were also shown to perform well for this sentence corpus, having the next highest correlation of $r = 0.88$ and $r = 0.86$, respectively.

The significant improvement in the performance of the QSTI method compared to the STI method of, for example, Drullman et al. (1994) (STI-D), is firstly attributed to the use of short-time processing of the modulation signal. This is in contrast to the assumption of stationarity made in traditional STI approaches, which process the long-time modulation signal. The benefit of short-time processing is also demonstrated by the improvement of CSII compared to MSC. Secondly, the use of 6 acoustic bands, with weights emphasising lower frequencies and de-emphasising higher ones, was found to further improve correlation. Finally, like previous STI, the third octave band analysis, which removes frequencies lower than 1 Hz and greater than around 18 Hz, was found to have a considerable affect on the resulting correlation.[5] Based on the modulation spectral resolution of QSTI (approximately 1 Hz), the effect

[5] The effect of different weights, and number of acoustics and third octave bands used by QSTI on the resulting correlation was investigated using tuning experiments similar to those described in Section 4.1. Bands and weights used were those found to result in the highest correlation to subjective intelligibility scores.

was to emphasise the 2–4 Hz frequency range, where most of the important speech information is contained.

Intelligibility scores predicted by QSTI verses human listener recognition scores, before and after mapping, are plotted in Figs. 2 and 3, respectively. QSTI used mean mapping function parameters of $a = 13.33$ and $b = -16.82$ where the mapping function is given by Eq. 9. For comparison, a similar plot of predicted intelligibility versus human listener recognition scores (after mapping) have been included for STI-D in Fig. 4, and NCM_w and CSII (which recorded the next highest correlations) in Figs. 5 and 6, respectively.

Fig. 3 shows mapped predicted values using QSTI extend over a good range, with values ranging from around 0.2 to nearly 0.9. This is preferred over smaller ranges, such as that shown by STI-D in Fig. 4 (only around 0.4 to 0.75), as it gives better separation and distinction between scores for stimuli that have similar intelligibility.

The density of predictions verses actual scores about the 1–1 line is also shown to be greater for QSTI. This indicates QSTI to have greater prediction accuracy, corresponding to the higher correlation and lower standard deviation of prediction errors (σ_e) reported in Table 5. NCM_w and CSII are also quite dense about this line, consistent with their high correlation and low σ_e. STI-D, on the other hand, has very spread predictions versus actual scores, indicating much poorer accuracy in predictions, consistent with the lower correlation and higher σ_e recorded.

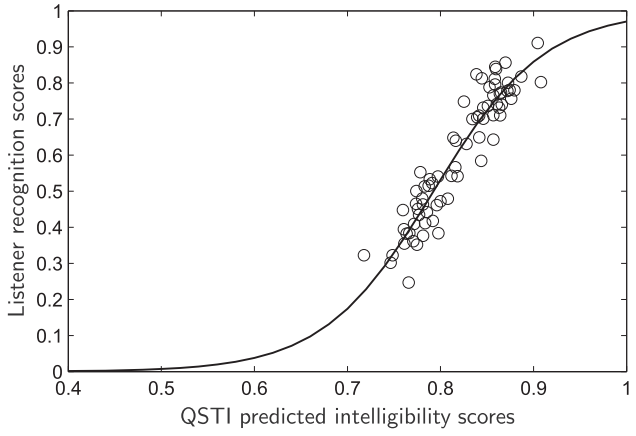

Fig. 2. Intelligibility score predicted (without mapping) by QSTI intelligibility measure versus human intelligibility score.

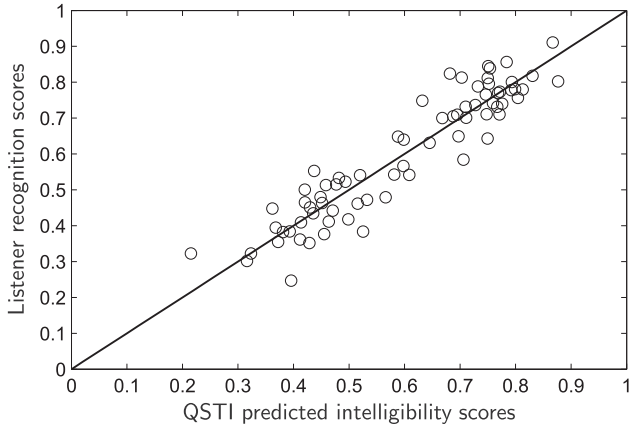

Fig. 3. Intelligibility score predicted (after mapping) by QSTI intelligibility measure versus human intelligibility score.

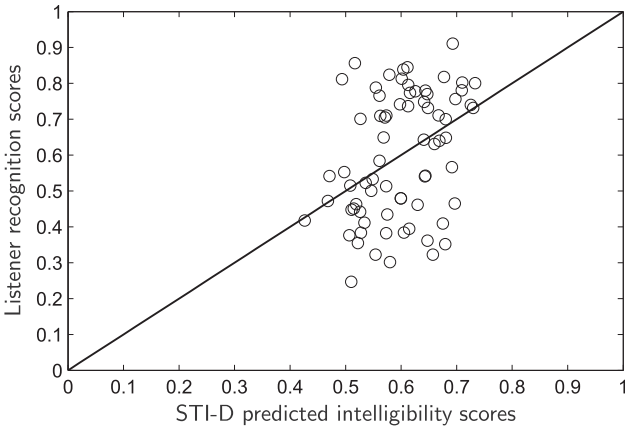

Fig. 4. Intelligibility score predicted (after mapping) by STI-D intelligibility measure versus human intelligibility score.

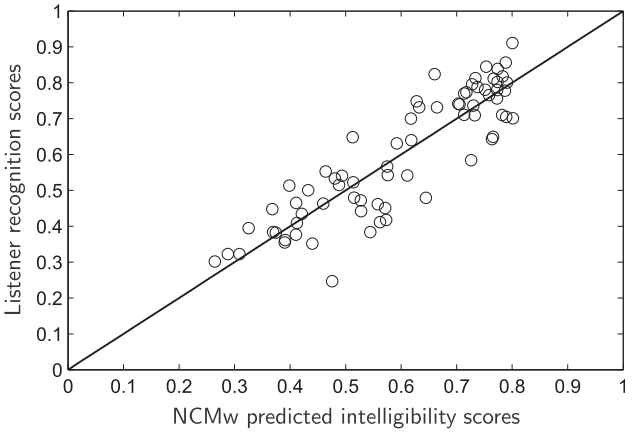

Fig. 5. Intelligibility score predicted (after mapping) by NCM_w intelligibility measure versus human intelligibility score.

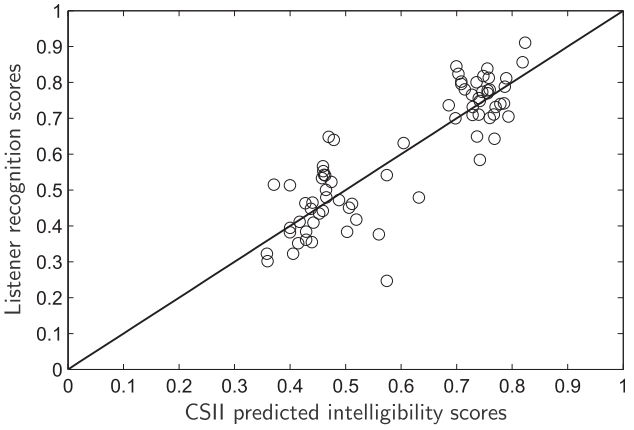

Fig. 6. Intelligibility score predicted (after mapping) by CSII intelligibility measure versus human intelligibility score.

A notable difference in performance between NCM_w and CSII, compared with QSTI, can be seen in the outliers, where predictors have considerably over or under estimated intelligibility. These are shown to have a higher error for NCM_w and CSII, resulting in a larger maximum absolute prediction error for NCM_w and CSII than for QSTI, as shown by MAE results shown in Table 6.

The significance of the improvement by QSTI, in terms of absolute prediction error, was also evaluated. QSTI was found to give a statistically significant improvement in predication compared to STOI ($p = 0.0013$), CSII ($p = 0.0069$) and NCM_w ($p < 0.001$).

Fig. 7 shows a comparison of correlation coefficients obtained for each technique after restricting the test corpus by noise type (babble, car, street, and train). While the performance of STI-D varied considerably between noise types, with noises such as babble and car being much less correlated, QSTI performed well and more consistently across all noises investigated. This is again attributed to the use of shorter modulation frame durations in its calculation of the MTF, making it better able to account for the changing properties of the degraded stimuli. NCM_w is also shown to perform well across noise types. The coherence methods (CSII and MSC), which overall performed very well, had a greater difference in performance between noise types, especially for street noise. FwSegSNR and ESC, on the other hand, while not performing quite as well overall, were generally consistent across the different noise types.

Table 6
Maximum absolute error (MAE) of scores from each intelligibility measure compared with human intelligibility scores.

Measure	MAE
STI-D	0.3395
fwSegSNR	0.2468
ESC	0.2685
NCM_w	0.2291
STOI	0.2563
SII	0.3594
CSII	0.3276
MSC	0.2404
LLR	0.3263
PESQ	0.2668
QSTI	0.1489

In the above results, we have made use of a corpus based on noise corrupted English speech sentences which have been processed with noise suppression algorithms, and evaluated by subjective intelligibility experiments. To further evaluate the performance of the proposed QSTI measure compared to other measures in the literature we consider two other smaller corpora.

The first corpus is also based on evaluating stimuli processed by speech enhancement algorithms. This corpus and the corresponding listener recognition scores have been reported in a study conducted by Taal et al. (2010). The corpus contains 10 utterances, each consisting of 5 Danish words, and corrupted with speech-shaped noise added at SNRs of −8.9, −7.7, −6.5, −5.2, and −3.1 dB. Utterances are then processed using two different noise suppression algorithms, namely MMSE magnitude estimation (Ephraim and Malah, 1984), and the estimation approach of Erkelens et al. (2007). This results in 15 different treatment types, including noisy. Fifteen native Danish-speaking listeners participated in the recognition experiment, resulting in mean recognition scores being calculated for each treatment type. Sentences for each treatment type are concatenated into a singular long speech signal, which is then evaluated by each measure.

Each of the objective measures utilised in previous experiments have been again applied to this corpus, without alteration, and the Pearson's correlation was calculated.

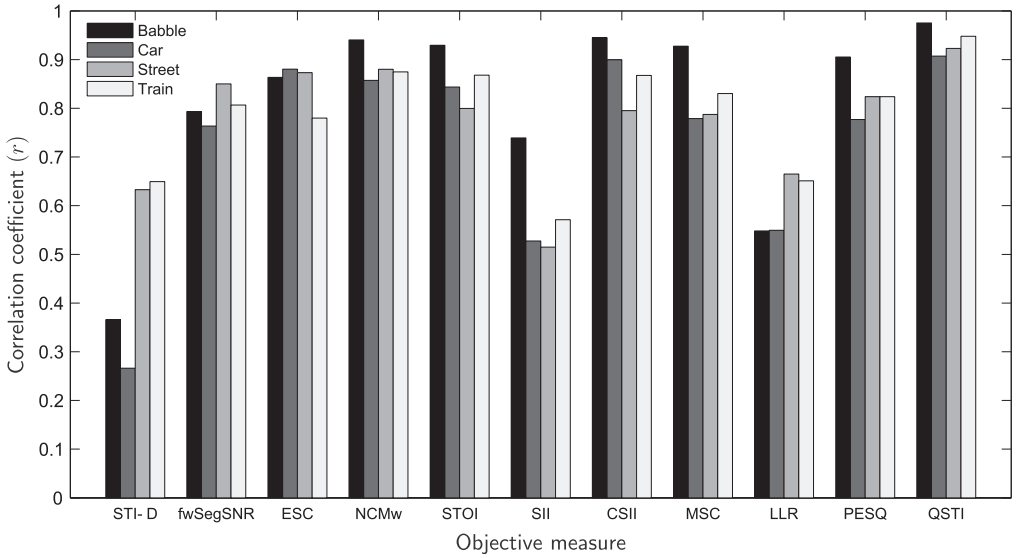

Fig. 7. Correlation of objective intelligibility scores to subjective intelligibility where noisy stimuli were corrupted with the indicated noise type. Correlations shown are for unmapped scores and include all treatment types of the indicated noise type in the full sentence corpus.

Table 7
Correlation coefficients (r) and standard deviation of the prediction errors (σ_e) between the predictions from each method and the recognition scores from real listeners averaged across the enhanced speech corpus of Taal et al. (2010).

Measure	Unmapped r	Mapped r	Mapped σ_e
STI-D	0.252	0.499	0.173
fwSegSNR	0.819	0.873	0.072
ESC	0.861	0.853	0.076
NCM_w	0.929	0.911	0.060
STOI	0.928	0.971	0.033
SII	0.478	0.194	0.147
CSII	0.880	0.952	0.045
MSC	0.792	0.801	0.088
LLR	−0.855	0.880	0.069
PESQ	0.803	0.790	0.090
QSTI	0.935	0.960	0.040

Table 8
Correlation coefficients (r) and standard deviation of the prediction errors (σ_e) between the predictions from each method and the recognition scores from real listeners averaged across the ITFS-processed speech corpus of Taal et al. (2011).

Measure	Unmapped r	Mapped r	Mapped σ_e
STI-D	0.2853	0.2853	0.3201
fwSegSNR	0.6112	0.6024	0.2647
ESC	0.4544	0.4211	0.3008
NCM_w	0.6636	0.6756	0.2442
STOI	0.9386	0.9622	0.0903
SII	0.4630	0.4332	0.2989
CSII	0.1646	0.4699	0.3054
MSC	0.4354	0.4139	0.3019
LLR	−0.0957	0.1281	0.3291
PESQ	0.3544	0.3499	0.3111
QSTI	0.6714	0.6768	0.2441

Correlations between predictions by each measure and mean listener recognition scores are shown in Table 7. STOI is shown to have a slightly higher correlation of 0.97, with QSTI having then next highest with a correlation of 0.96. CSII also performed very well. These results affirm that QSTI provides a large improvement on the prediction accuracy of STI-D. Secondly, the good performance of QSTI on this second corpus supports the reliability of the good performance indicated for the larger sentence corpus.

So far we have focused on the evaluation of noisy speech processed using well known noise suppression algorithms. However, there are many different types of distortions. One such distortion occurs as a result of applying some sort of time–frequency varying gain function, such as that resulting from speech separation techniques where binary time–frequency weighting is used, e.g. ideal time frequency segregation. Ideal time–frequency segregation (ITFS) applies a binary modulation pattern in a time–frequency representation. The binary mask is derived according to some criterion and then a gain is applied to the speech representation before reconstruction is derived from this mask. The resulting processed speech can improve intelligibility, but also heavily distorts speech signal. Further details on stimuli of this corpus can be found in (Taal et al., 2011).

Using a similar procedure to that used in evaluating other corpora, the correlations between predictions by each objective measure and mean listener recognition scores were calculated and are given in Table 8.[6] In these final set of results we see that QSTI does not perform well for this type of masking based processing, with STOI as the only method which scores highly. It is also noted that QSTI and NCM_w are the next best, with other methods performing very poorly. Consequently, for this type of distortion, the STOI is clearly the better method to use to objectively evaluate speech intelligibility.

[6] Results shown for the above two experiments are marginally different from that reported by Taal et al. (2011) due to a difference in the mapping function used. This difference does not, however, effect the overall interpretation of results.

The processing of heavily distorted speech will be further investigated by future work.

5. Conclusion

In this paper we have proposed an improved STI-based intelligibility measure which calculates the MTF from the short-time modulation spectra of clean and degraded signals. The use of short-time modulation processing results in improved correlation to subjectively determined intelligibility scores compared to the traditional STI, for degraded speech corrupted with a range of noise types and processed with different enhancement methods. Results show that the proposed QSTI measure also has higher correlation with human listener intelligibility than the coherence, correlation and distance based objective measures tested. The modulation frame duration was shown to be an important parameter of the approach, and predominantly responsible for its improved performance. Finally, while the QSTI measure proposed is shown to work well for speech stimuli processed using typical noise suppression algorithms, for the evaluation of stimuli heavily distorted with (for example) masking based processing, an alternative approach such as STOI is recommended.

Acknowledgements

The authors wish to thank Yi Hu and the late Philipos Loizou for providing access to their English speech intelligibility corpus (Hu and Loizou, 2007), and Cees Taal, Richard Hendriks, Richard Heusdens, and Jesper Jensen for providing access to their Danish speech intelligibility corpora (Taal et al., 2010, 2011).

References

ANSI, 1997. Methods for Calculation of the Speech Intelligibility Index (ANSI S3.5-1997). American National Standards Institute.

Balakrishnan, N., 1992. Handbook of the Logistic Distribution. Marcel Dekker Inc., NewYork, USA.

Boldt, J., Ellis, D., 2009. A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. In: Proc. EUPSIPCO 2009. Glasgow, Scotland, pp. 1849–1853.

Christiansen, C., Pedersen, M.S., Dau, T., 2010. Prediction of speech intelligibility based on an auditory preprocessing model. Speech Commun. 52 (7–8), 678–692.

Carter, G., Knapp, C., Nuttall, A., 1973. Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. IEEE Trans. Audio Electroacoust. 21, 337–344.

Drullman, R., Festen, J., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95 (5), 2670–2680.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. ASSP 32 (6), 1109–1121.

Erkelens, J., Hendriks, R., Heusdens, R., Jensen, J., 2007. Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors. IEEE Trans. Audio, Speech, Lang. Process. 15 (6), 1741–1752.

French, N., Steinberg, J., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1), 90–119.

Goldsworthy, R., Greenberg, J., 2004. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J. Acoust. Soc. Am. 116 (6), 3679–3689.

Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100 (3), 1703–1716.

Houtgast, T., Steeneken, H., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77 (3), 1069–1077.

Hu, Y., Loizou, P., 2007. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 122 (3), 1777–1786.

ITU-T P.862, 2001. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs. ITU-T Recommendation P.862.

Kates, J.M., Arehart, K.H., 2005. Coherence and the speech intelligibility index. J. Acoust. Soc. Am. 117 (4), 2224–2237.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. Taylor and Francis, Boca Raton, FL.

Ma, J., Hu, Y., Loizou, P., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125 (5), 3387–3405.

Ma, J., Loizou, P., 2011. SNR loss: a new objective measure for predicting the intelligibility of noise-suppressed speech. Speech Commun. 53 (3), 340–354.

Paliwal, K., Schwerin, B., Wójcicki, K., 2011. Role of modulation magnitude and phase spectrum towards speech intelligibility. Speech Commun. 53 (3), 327–339.

Payton, K., Braida, L., 1999. A method to determine the speech transmission index from speech waveforms. J. Acoust. Soc. Am. 106 (6), 3637–3648.

Pearce, D., Hirsch, H., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. of International Conference on Spoken Language Processing (ICSLP). Beijing, China, pp. 29–32.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs, NJ, USA.

Rothauser, E., 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17 (3), 225–246.

Steeneken, H., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. 67 (1), 318–326.

Taal, C., Hendriks, R., Heusdens, R., Jensen, J., 2010. Intelligibility prediction of single-channel noise-reduced speech. In: Proc. ITG-Fachtagung Sprachkommunikation. Bochum, Germany.

Taal, C., Hendriks, R., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio, Speech, Lang. Process. 19 (7), 2125–2136.

Tribolet, J., Noll, P., McDermott, B., Crochiere, R., 1978. A study of complexity and quality of speech waveform coders. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. pp. 586–590.