Intelligibility Prediction of Single-Channel Noise-Reduced Speech

Cees Taal[1], *Richard Hendriks*[1], *Richard Heusdens*[1], *Jesper Jensen*[2]

[1]Delft University of Technology, SIPlab,
2628 CD Delft, The Netherlands
E-Mail: {c.h.taal, r.c.hendriks, r.heusdens}@tudelft.nl

[2]Oticon A/S,
2765 Smørum, Denmark
E-Mail: jsj@oticon.dk

Abstract

In general, single-channel noise-reduction algorithms do not improve the speech intelligibility for normal-hearing listeners. A reliable objective intelligibility measure is therefore of great interest. It could be used for analysis and/or optimization of noise-reduction algorithms. For these applications it is important that the objective measure can correctly predict the difference in intelligibility before and after noise reduction. Typically, existing studies do not evaluate objective measures for this property. Twelve objective measures are evaluated in order to let them predict the intelligibility before and after noise reduction. Best performance was obtained with a recently developed intelligibility predictor called STOI. Modest results were obtained with WSS and NSEC. The remaining measures significantly overestimated the intelligibility of the noise-reduced speech.

1 Introduction

In order to determine the perceptual consequences of a noise-reduction algorithm, the algorithm at hand can be evaluated by means of a listening test or an objective machine-driven quality assessment. Although a listening test can lead to a judgment as observed by the intended group of users, such tests are often costly and time consuming. Therefore, accurate and reliable objective evaluation methods are of interest since they might replace a listening test, at least in some stages of the algorithm development process. Although it is not straightforward to describe the overall quality of a speech processing system, people tend to divide the evaluation into the attributes speech quality, (i.e., pleasantness/naturalness of speech) and speech intelligibility. In this paper we focus on speech intelligibility.

It has been shown that single-channel noise-reduction algorithms can successfully improve the speech quality (i.e., pleasantness/naturalness of speech) [10]. However, a recent evaluation also showed that these algorithms, in general, do not improve the speech intelligibility for normal-hearing listeners [8]. Inventing a single-channel noise-reduction algorithm which improves speech intelligibility is currently one of the main challenges in this research field. In order to gain more knowledge in the field of intelligibility improvement of single-channel noisy speech, a reliable objective intelligibility measure (i.e., a distance measure which has high correlation with speech intelligibility) is of great interest. Such an objective measure could be used for the analysis of existing conventional noise-reduction algorithms and perhaps explain why there is no gain in intelligibility. In addition, new noise-reduction algorithms could be developed which optimize for such an objective measure.

To use a reliable intelligibility measure for noise reduction it is important that it correctly predicts the difference in intelligibility *before* and *after* noise reduction. For example, the predictions from such an objective measure applied to signals obtained from a conventional single-channel noise-reduction algorithm should be in line with the fact that there is no significant change in intelligibility (i.e., the predictions should also not change significantly due to noise reduction). Conversely, a significant improvement of the objective measure due to noise reduction should also imply an improvement in speech intelligibility. Unfortunately, a standard and well-known objective intelligibility measure like the speech transmission index (STI) [21] incorrectly predicts that single-channel noise reduction leads to a large intelligibility improvement [16]. For many other objective intelligibility measures it is unknown whether they can predict the effect on intelligibility due to noise-reduction.

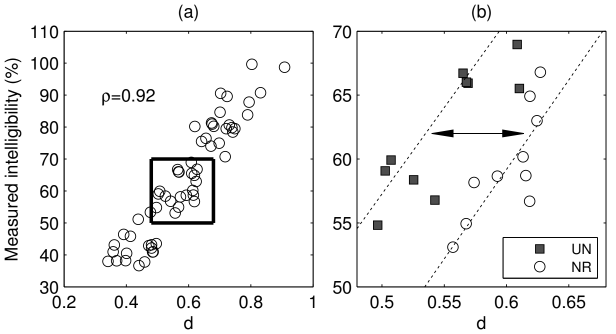

Figure 1: (a) Scatter plot between objective measure output *d* and the intelligibility scores from a listening test. The region within the rectangle is re-plotted in (b) where noisy unprocessed speech (UN) and noise reduced conditions (NR) are highlighted. Despite the high correlation ($\rho = 0.92$), Fig.(b) reveals that the objective measure overestimates the intelligibility of the noise-reduced conditions (denoted by the arrow).

In general, evaluative studies (e.g., [17]) report figures of merit (e.g., correlation coefficient) from which it is difficult to conclude whether an objective measure can correctly predict the effect of noise reduction on intelligibility. An (artificial) example of this problem is illustrated in Fig. 1, where the left plot illustrates a scatter-plot between the predicted and measured intelligibility scores for different noisy and noise-reduced conditions (e.g., different SNRs and noise types). For this example a correlation coefficient of $\rho = 0.92$ is obtained, which is generally considered as good performance. However, the right plot reveals that the noise-reduced conditions, in general, are higher with respect to the intelligibility scores compared to the noisy, unprocessed conditions. This implies that the measure overestimates the intelligibility of the noise-reduced speech. Hence, next to conventional figures of merit, additional information (e.g., a plot like Fig. 1) is needed to determine whether an objective measure can predict the effect of noise reduction.

In this paper, twelve objective speech-quality and speech-intelligibility measures are evaluated in order to let them predict the intelligibility scores from noisy unprocessed and noise-reduced speech. Additional plots are given to reveal if these measures can correctly predict the difference in intelligibility before and after noise reduction.

This research is supported by the Oticon Foundation and the Dutch Technology Foundation STW.

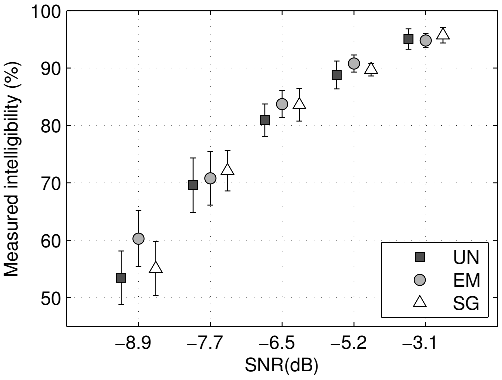

Figure 2: Average intelligibility scores (% correct words) and standard errors for unprocessed (UN) speech-shaped noise degraded speech, and two noise-reduction schemes (EM, SG). See text for more details.

2 Listening Experiment

A listening experiment is conducted to evaluate the intelligibility of unprocessed (UN) noisy speech followed by two different single-channel noise-reduction algorithms. That is, A) the standard MMSE-STSA algorithm by Ephraim-Malah (EM) [3] which was developed under the assumption that speech and noise DFT coefficients are Gaussian distributed, and B) an improved version by Erkelens *et al.* (SG) [4], which assumes the speech and noise DFT coefficients to be super-Gaussian and Gaussian distributed, respectively. For both algorithms, the a priori SNR is estimated with the decision directed approach [3] with a smoothing factor of α=0.98. The noise PSD in EM and SG is estimated using Minimum Statistics [18] and the noise-tracker described in [6], respectively. Maximum attenuation is limited to 10 dB in both algorithms. In SG, the parameters describing the assumed super-Gaussian density of the speech DFT coefficients are γ=1 and ν=0.6, see [4] for details.

The signals are taken from the Dantale II corpus [24] and are degraded by additive speech-shaped noise (SSN) at a sample rate of 20 kHz. Five different SNRs are considered (-8.9 dB, -7.7 dB, -6.5 dB, -5.2 dB and -3.1 dB), which were chosen such that the psychometric function of clean speech degraded by SSN (derived from earlier experiments, see [13]) was sampled approximately between 50% and 100% intelligibility.

Fifteen normal-hearing Danish-speaking listeners were asked to judge the intelligibility of the noisy signals and the two enhanced versions. The signals were presented diotically through head-phones (Sennheiser HD 280 pro) at a sound pressure level of approximately 65 dB SPL (A-weighted). The three processing conditions (i.e., UN, EM and SG) and 5 SNR values make up 3*5=15 conditions. For each of the 15 conditions, each listener is presented with 10 five-word sentences. The average score for all users and for one condition was consequently obtained by the average percentage of correct words.

The results from the listening experiment are shown in Fig. 2. As can be observed, both noise-reduction algorithms have a very small effect on the speech intelligibility compared to the intelligibility of the noisy unprocessed speech. No statistical significant intelligibility improvements were measured due to either of both noise-reduction algorithms. This result is in line with the results from [11] where, in general, no noise-reduction scheme could improve the intelligibility of noisy speech.

3 Objective Measures

In Table 1 all evaluated objective measures are presented in combination with their corresponding abbreviations. This includes several conventional speech-quality measures (SSNR, LLR, IS, CEP, WSS, FWS, PESQ), a sophisticated auditory based model (DAU) which correlates well with intelligibility [23] and several measures specifically meant for intelligibility prediction (CSII, CSTI, STOI). Note that some of these measures are designed for speech-quality prediction rather than intelligibility. However, since they are typically used to evaluate noise-reduction algorithms it is of interest to study whether to which extent such measures can predict the effect of noise reduction on speech intelligibility.

Abbr.	Objective Measure
SSNR	Segmental SNR [19]
LLR	Log-Likelihood Ratio [19]
IS	Itakura-Saito [19]
CEP	Cepstral Distance Measure [19]
WSS	Weighted-Spectral Slope Metric [14]
FWS	Normalized Frequency Weighted SSNR [9]
PESQ	PESQ [20]
DAU	Dau auditory model [1]
CSII	Coherence SII [12]
CSTI	Covariance based STI [5]
STOI	Short-time Objective Intelligibility Measure [22]

Table 1: *The evaluated objective measures with their corresponding abbreviations.*

All models are a function of the clean speech and the modified speech signal (e.g., noisy or noise-reduced speech). Matlab implementations of LLR, IS, CEP, WSS, FWS, and PESQ are based on the software included in [15].

To evaluate the objective measures, 30 five-word sentences are used from the corpus, where for each sentence the corresponding modified sentence (e.g., UN, EM, SG) is obtained. The clean speech sentences and the modified speech sentence are then concatenated separately, resulting in one clean and one modified speech signal with a length approximately equal to 90 seconds. Before evaluation, noise-only regions (i.e., regions where no speech is present) are removed as described in [22].

4 Evaluation Procedure

Typically, objective measures do not directly predict an absolute intelligibility score but instead some monotonic relation is present between the objective scores and the results from the listening experiment. A mapping is needed in order to obtain an absolute intelligibility score between 0% and 100%. The logistic function is used for this,

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

where a and b are free parameters, which are fitted to the intelligibility scores with a nonlinear least squares procedure, and d denotes the objective score for one particular objective measure. The mapping is only fitted to the UN-conditions, which is then used to predict the intelligibility scores for the noise-reduction conditions. Fig. 3 illustrates an example of this calibration process. The logistic function clearly fits the data points very well for the UN-conditions. Since the mapping is now reliable for the noisy speech, the prediction error for the noise-reduction case (NR) reveals if the objective measure can predict the effect on the speech intelligibility due to the applied noise reduction.

The performance of all objective measures is evaluated with the RMS of the prediction error (RMSE) and the mean absolute deviation (MAD). Since the UN-conditions were included in the calibration process they are *excluded* in the evaluation of these merits. The RMSE is defined as,

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

where s refers to an intelligibility score obtained for the i^{th} NR-condition and S denotes the total number of processing condi-

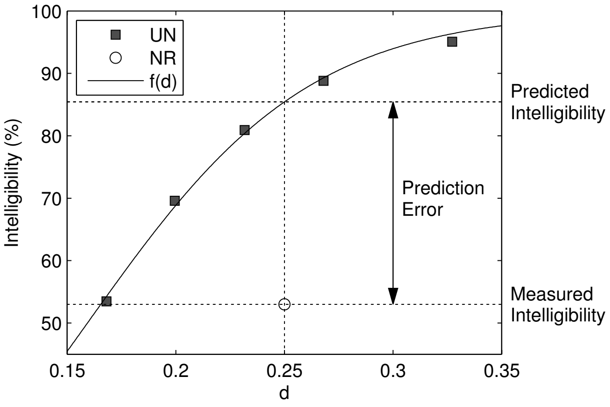

Figure 3: The function f(d) is fitted to the intelligibility scores of the noisy unprocessed speech (UN), which is then used to predict the intelligibility score of speech followed by noise-reduction (NR). In this example, the intelligibility of the noise-reduced speech condition is largely overestimated.

tions. The MAD is given by,

$$MAD = \max_{i} \left(\left| s_i - f\left(d_i\right) \right| \right), \tag{3}$$

and reveals the worst-case prediction for each objective measure on speech intelligibility due to noise reduction.

5 Results

The figure of merits are shown in Fig. 4 and the scatter plots of the predicted and measured intelligibility scores for all objective measures are shown in Fig. 5. A perfect prediction would imply that all points are fitted by the logistic function. If the fitted monotonic mapping is *increasing*, conditions on the right of the fitting imply an incorrect overestimation of speech intelligibility (e.g., SSNR). Conditions appearing on the left of a monotonic *decreasing* function are also linked to an incorrect overestimation of speech intelligibility (e.g., IS). Hence, Fig. 5 reveals that almost all objective measures significantly over-estimate the intelligibility of the noise-reduced speech compared to the unprocessed noisy speech.

An unwanted over-estimation of speech intelligibility due to noise reduction is less present with STOI and NSEC. Moreover, WSS seems to correctly predict the effect of noise-reduction only for the SG conditions. IS and SSNR both show the worst performance and both report that almost all speech is fully intelligible after noise reduction.

The best performance is obtained with STOI, which had the lowest RMSE of 4.4%. Even for the worst-case MAD only an intelligibility improvement of 6.9% is reported. These overestimated improvements are very low and fall within a similar range as the standard errors from the estimated mean intelligibility scores from the listening experiment (See Fig. 2). Note, that STOI has already high correlation (ρ=0.95) for a different large dataset [13] (See also [22]). No parameters of STOI have been modified in this paper.

Most of the remaining measures significantly overestimated the intelligibility scores of both EM and SG. Hence, one should take into account this overestimation when using these measures for analysis and optimization in the field of single-channel noise reduction. For the conventional STI [21] (The conventional STI is different from the CSTI used in this paper) and CSII [12], it is already known from literature that the speech intelligibility is overestimated after noise reduction, [2] and [7], respectively. Our results show that this problem is also clearly present for other measures when used for the evaluation of SSN-degraded speech followed by noise-reduction from EM or SG. The observation that CSII also overestimated the speech-intelligibility after noise-reduction is in line with the results from [7].

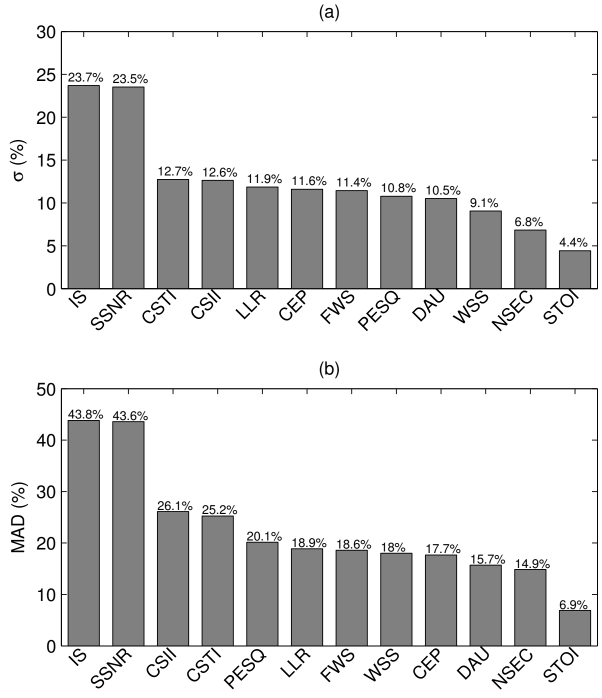

Figure 4: Prediction results for all objective measures for unprocessed speech-shaped noise degraded speech followed by two noise-reduction algorithms.(a) shows the RMS of the prediction error (σ) and (b) the maximum absolute deviation (MAD). Measures on the right perform better.

Speech corrupted by SSN and the noise-reduction algorithms EM and SG are relatively basic conditions in the field of single-channel noise reduction. Reliable objective intelligibility measures for these particular conditions are already of great interest. However, in order to verify whether the results from this paper also hold for other noise-reduction algorithms and/or noise-types more experiments are needed.

6 Conclusions

In this paper twelve objective measures are evaluated in order to predict the difference in intelligibility before and after single-channel noise reduction. Two noise-reduction algorithms are considered and applied to speech-shaped noise degraded speech at various SNRs. Out of all measures, STOI correctly predicted no large changes in intelligibility due to noise-reduction. This makes STOI a new potential candidate for analysis and/or optimization in the field of single-channel noise reduction. Moderate results were obtained with NSEC and WSS. The remaining measures largely overestimated the intelligibility scores of the noise-reduced speech compared to the noisy unprocessed speech. One should take into account this overestimation when using one of these measures for analysis and/or optimization in the field of single-channel noise reduction.

References

[1] T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the ”effective” signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622, 1996.

[2] F. Dubbelboer and T. Houtgast. The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *J. Acoust. Soc. Am.*, 124(6):3937–3946, 2008.

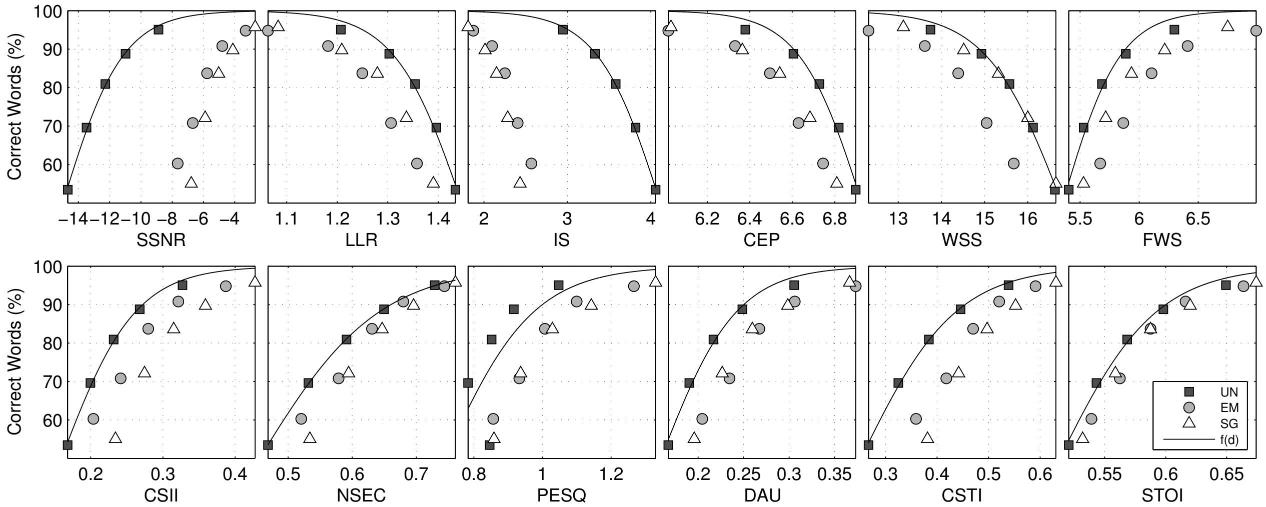

Figure 5: Prediction results for all objective measures for unprocessed (UN) speech-shaped noise degraded speech followed by two noise-reduction algorithms (EM, SG). Measures are calibrated on (UN). A perfect prediction would imply that all points are fitted by $f(d)$.

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

[4] J.S. Erkelens, R.C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. Audio Speech Lang. Process.*, 15(6):1741–1752, 2007.

[5] R. L. Goldsworthy and J. E. Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Am.*, 116(6):3679–3689, 2004.

[6] R. C. Hendriks, R. Heusdens, and J. Jensen. MMSE based noise PSD tracking with low complexity. In *Proc. ICASSP*, pages 4266–4269, 2010.

[7] G. Hilkhuysen. Understanding the intelligibility of speech after noise reduction : a comparison of predictive models. In *British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness*, 2009.

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

[9] Y. Hu and P. C. Loizou. A comparative intelligibility study of speech enhancement algorithms. In *Proc. ICASSP*, volume 4, pages IV–561–IV–564, 2007.

[10] Y. Hu and P. C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7-8):588–601, 2007.

[11] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.*, 16(1):229–238, 2008.

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

[13] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am.*, 126(3):1415–1426, 2009.

[14] D. Klatt. Prediction of perceived phonetic distance from critical-band spectra: A first step. In *Proc. ICASSP*, volume 7, 1982.

[15] P. C. Loizou. *Speech enhancement: theory and practice*. CRC, Boca Raton, FL, 2007.

[16] C. Ludvigsen, C. Elberling, and G. Keidser. Evaluation of a noise reduction method - comparison between observed scores and scores predicted from STI. *Scandinavian Audiology. Supplement.*, 38:50–55, 1993.

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

[18] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9(5):504–512, 2001.

[19] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[20] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends. Perceptual evaluation of speech quality (PESQ): the new ITU standard for end-to-end speech quality assessment part I-time-delay compensation. *J. of the Audio Engineering Society*, 50(10):755–764, 2002.

[21] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67(1):318–326, 1980.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. ICASSP*, pages 4214 – 4217, 2010.

[23] 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*, pages 1947–1950, 2009.

[24] K. Wagener, J. L. Josvassen, and R. Ardenkjaer. Design, optimization and evaluation of a Danish sentence test in noise. *Int. J. of Audiology*, 42(1):10–17, 2003.