On Predicting the Difference in Intelligibility Before and After Single-Channel Noise Reduction

Cees H. Taal, Richard C. Hendriks, Richard Heusdens
Delft University of Technology, SIPlab,
2628 CD Delft, The Netherlands
Email: {c.h.taal, r.c.hendriks, r.heusdens}@tudelft.nl

Jesper Jensen
Oticon A/S
2765 Smørum, Denmark
Email: jsj@oticon.dk

***Abstract*—In general, single-channel noise-reduction algorithms do not improve the speech intelligibility for normal-hearing listeners. In order to understand this problem, a reliable objective intelligibility measure is of great interest. Such a measure could be used for analysis and/or optimization of noise-reduction algorithms. For this application 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. Five objective measures (STOI, CSTI, DAU, CSII and FWS) are evaluated in order to let them predict the intelligibility before and after noise reduction. The measures CSTI, DAU, CSII and FWS significantly overestimated the intelligibility of the noise-reduced speech. This was not the case with STOI, which is therefore a new potential candidate for analysis and/or optimization in the field of single-channel noise reduction.**

I. Introduction

Single-channel noise reduction is a common feature in many DSP-based speech-communication devices (e.g., in mobile phones, digital hearing aids) in order to recover an underlying clean speech signal from a noisy observed speech signal. It has been shown that single-channel noise-reduction algorithms can successfully improve the speech quality (i.e., pleasantness/naturalness of speech) [1]. However, a recent evaluation also showed that these algorithms, in general, do not improve the speech intelligibility for normal-hearing listeners [2]. 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. For these two applications (i.e., analysis and optimization) it is important that the objective measure can correctly predict 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) [3] incorrectly predicts a large intelligibility improvement due to noise reduction [4]. For many other objective intelligibility measures it is unknown if they can predict the effect on intelligibility due to noise-reduction.

Typically, evaluative studies (e.g., [5]) report figures of merit (e.g., correlation coefficient) from which it is difficult to conclude if an objective measure can correctly predict the difference in intelligibility before and after noise reduction. An (artificial) example of this problem is illustrated in Fig. 1, where a scatter-plot is shown 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 plot illustrates that the noise-reduced conditions, in general, are more to the right of the diagonal line 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 if an objective measure can predict the effect of noise reduction.

In this paper we present the results from an intelligibility listening experiment conducted for the evaluation of two different single-channel noise-reduction algorithms. Five objective measures are evaluated in order to let them predict the intelligibility scores from this listening test. Besides reporting several figures of merit, additional plots are given to reveal if these measures can correctly predict the difference in intelligibility before and after noise reduction.

II. 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) [6] which was developed under the assumption that speech and noise DFT coefficients are Gaussian, and B) an improved

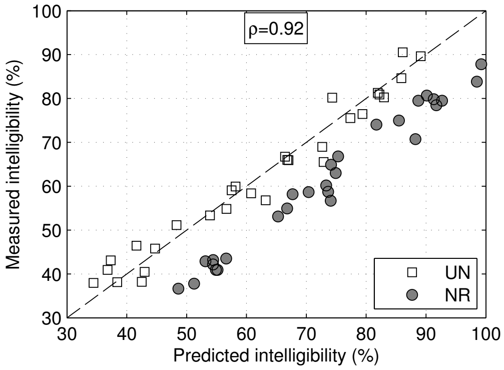

Fig. 1. Despite high correlation (ρ=0.92), an objective measure can report incorrectly higher intelligibility scores for noise-reduced (NR) signals compared to unprocessed (UN) noisy speech.

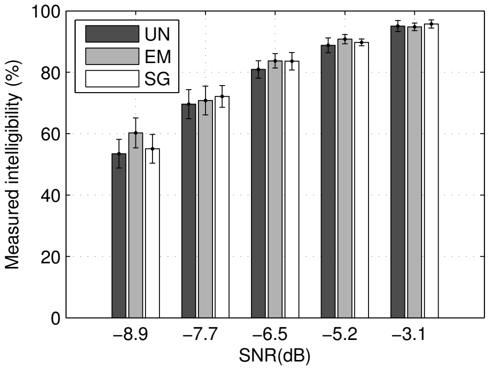

Fig. 2. Average intelligibility scores and standard errors for unprocessed (UN) speech-shaped noise degraded speech, and two noise-reduction schemes (EM, SG). See text for more details.

version by Erkelens *et al.* (SG) [7], 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 [6] with a smoothing factor of α=0.98. The noise PSD in EM and SG is estimated using Minimum Statistics [8] and the noise-tracker described in [9], 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 [7] for details.

The signals are taken from the Dantale II corpus [10] 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 [11]) 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 [12] where, in general, no noise-reduction scheme could improve the intelligibility of noisy speech.

III. Objective Intelligibility Measures

In total five different objective intelligibility measures are included in this evaluation. These are all a function of the clean and the modified speech signal (i.e., UN, EM, SG). We will only highlight the main aspects of each model and motivate why it is included in this evaluation. For further details about each objective measure the reader is referred to its corresponding reference.

The short-time objective intelligibility measure (STOI) [13] is developed by the authors and shows high correlation (ρ=0.95) with the speech intelligibility of ideal binary masked (IBM) noisy speech from [11]. Although IBM and conventional single-channel noise reduction algorithms are different techniques, they both apply some time-frequency varying gain function to the noisy speech. Therefore, we hypothesized in [13] that STOI could be a potential candidate for speech-intelligibility prediction of single-channel noise-reduced speech.

The perceptual model developed by Dau *et al.* (DAU) [14], acts as an artificial observer and is originally used for accurately predicting masking thresholds for various masking conditions. More recently it has also been shown that the model can be used as a good intelligibility predictor for IBM-processed speech [15]. Similarly as with STOI, it may therefore be a potential candidate for speech-intelligibility prediction of single-channel noise-reduced speech.

In [16], an evaluation is presented for various STI-based intelligibility measures. One STI-based measure included in this evaluative study is the normalized covariance-based STI (CSTI), which shows promising results with respect to spectral subtraction [16] (i.e., a conventional single-channel noise-reduction algorithm). It is of interest to see if these results also hold for the noise type and noise-reduction algorithms included in this evaluation.

Recently, a new version of the frequency weighted segmental SNR (FWS) was proposed in [12]. The measure has good performance with respect to speech-intelligibility prediction of single-channel noise-reduced speech in the evaluative study of [5]. However, it has not been evaluated yet in order to predict the difference in intelligibility before and after noise reduction.

Finally the coherence speech intelligibility index (CSII) [17] is included, which shows good correlation with various nonlinear distortions (e.g., peak clipping). An earlier study already reported that CSII predicts an incorrect intelligibility improvement due to noise reduction [18]. However, the model has not been evaluated with the combined noise type (SSN) and noise-reduction schemes (EM, SG) included in this paper. It is of interest to see if our results will be in line with [18] also under these conditions.

IV. General Procedure

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 [13].

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,

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

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. This logistic function is only fitted to the UN-conditions, which is then used to predict the absolute intelligibility scores for the noise-reduction conditions. In this manner the UN-conditions will be well predicted by all objective measures. Fig. 3 illustrates an example of this calibration process for DAU. The logistic function clearly fits the data points very well for the UN-conditions, which can now be used to let the model predict the absolute intelligibility scores for the other noise-reduced conditions.

The performance of all objective measures is evaluated with the RMS of the prediction error (RMSE),

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

where s refers to an intelligibility score obtained in processing condition i and S denotes the total number of processing conditions. In addition, the maximum absolute deviation (MAD) is included,

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

which reveals the worst-case prediction for each objective measure on speech intelligibility due to noise reduction. Since the UN-conditions were included in the calibration process they are *excluded* in the evaluation of Eqs. (2), (3).

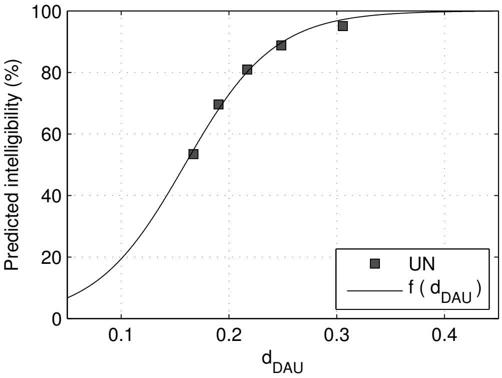

Fig. 3. Results of the calibration process for DAU. A logistic function is fitted to the intelligibility scores of the noisy unprocessed speech (UN), which is then used to predict the intelligibility scores for the other processing conditions.

For the illustrative example from Fig. 1 this general evaluation procedure would measure the distance of the NR-conditions from the dashed, diagonal line. Hence, the RMSE and MAD will reveal the ability of each objective measure to predict the effect on intelligibility due to noise reduction.

V. Results

The scatter plots of the predicted and measured intelligibility scores for all objective measures are shown in Fig. 4. A perfect prediction would imply that all points are fitted by the dashed diagonal line. Due to the calibration procedure, all objective measures almost perfectly predict the intelligibility scores for the unprocessed noisy conditions (UN). Remaining data points appearing to the right of the dashed diagonal line imply an incorrect predicted over-estimated intelligibility improvement due to 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 [11] for both unprocessed noisy and ideal binary masked noisy speech [13]. No parameters of STOI have been modified in this paper.

The measures CSTI, DAU, CSII and FWS all four significantly overestimated the intelligibility scores of the noise-reduced speech with RMSEs, 12.7%, 10.5%, 12.6% and 11.4%, respectively. 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 [3] (The conventional STI is different from the CSTI used in this paper) and CSII [17], it was already known from literature that the speech intelligibility is overestimated after noise reduction, [4] and [18], respectively.

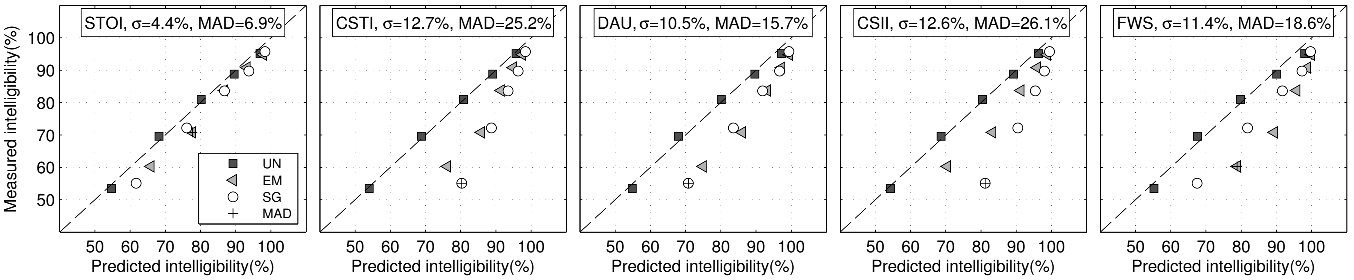

Fig. 4. Prediction results for all objective measures for unprocessed (UN) speech-shaped noise degraded speech followed by two noise-reduction algorithms (EM, SG). Measures are calibrated on (UN). RMS of the prediction error (σ) and the maximum absolute deviation (MAD) are indicated at the top of each plot. The condition responsible for the MAD is highlighted with '+'.

Our results show that this problem is also clearly present for CSTI, DAU and FWS 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 [18].

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 if the results from this paper also hold for other noise-reduction algorithms and/or noise-types more experiments are needed.

VI. Conclusions

In this paper five objective intelligibility measures (STOI, CSTI, DAU, CSII, FWS) are evaluated in order to predict the difference in intelligibility before and after single-channel noise reduction. Two noise-reduction algorithms are considered, applied to speech-shaped noise (SSN) degraded speech at various SNRs. From the results the following conclusions can be drawn:

- 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.
- The measures CSTI, DAU, CSTI and FWS all four 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.
- Good performance of an objective measure with respect to some figure of merit (e.g., correlation coefficient) is not sufficient to verify if an objective measure can correctly predict the effect on intelligibility due to noise-reduction.

Acknowledgment

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

References

[1] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[2] ——, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[3] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[4] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 124, no. 6, pp. 3937–3946, 2008.

[5] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, 2009.

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

[7] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," 2010, pp. 4266–4269.

[10] K. Wagener, J. L. Josvassen, and R. Ardenkjaer, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. of Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[11] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[12] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[13] 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*, 2010, pp. 4214 – 4217.

[14] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction," in *Proc. Interspeech*, 2009, pp. 1947–1950.

[16] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[17] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[18] 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. [Online]. Available: http://www.bsa-short-papers.org/Home/orals-2009