A SHORT-TIME OBJECTIVE INTELLIGIBILITY MEASURE FOR TIME-FREQUENCY WEIGHTED NOISY SPEECH

Cees H. Taal, Richard C. Hendriks, Richard Heusdens

Delft University of Technology,
Signal Information & Processing Lab,
2628 CD Delft, The Netherlands
{c.h.taal, r.c.hendriks, r.heusdens}@tudelft.nl

Jesper Jensen

Oticon A/S
2765 Smørum, Denmark
jsj@oticon.dk

ABSTRACT

Existing objective speech-intelligibility measures are suitable for several types of degradation, however, it turns out that they are less appropriate for methods where noisy speech is processed by a time-frequency (TF) weighting, e.g., noise reduction and speech separation. In this paper, we present an objective intelligibility measure, which shows high correlation (rho=0.95) with the intelligibility of both noisy, and TF-weighted noisy speech. The proposed method shows significantly better performance than three other, more sophisticated, objective measures. Furthermore, it is based on an intermediate intelligibility measure for short-time (approximately 400 ms) TF-regions, and uses a simple DFT-based TF-decomposition. In addition, a free Matlab implementation is provided.

Index Terms— intelligibility prediction, speech enhancement, noisy speech.

1. INTRODUCTION

Speech processing systems, such as a speech-enhancement scheme or an intelligibility improvement algorithm in a hearing aid, often introduce degradations and modifications to clean or noisy speech signals. To determine the effect of these methods on the speech-intelligibility, the algorithm at hand can be evaluated by means of subjective listening tests and/or an objective intelligibility measure (OIM). Accurate and reliable objective evaluation methods are of interest, since they might replace costly and time consuming subjective tests, at least in some stages of the algorithm development process.

One of the first OIMs was developed at AT&T Bell Labs by French and Steinberg in 1947 [1], currently known as the articulation index (AI) [2]. AI evolved to the speech-intelligibility index (SII), and has been standardized in 1997 under ANSI S3.5-1997 [3]. Later, the speech transmission index (STI) [4] was proposed, which, in contrast to AI, is also able to predict the intelligibility of various simple nonlinear degradations, e.g. clipping. The majority of recent published models are still based on the fundamentals of AI, e.g. [5, 6] and STI (see [7] for an overview of STI-based measures).

Although the just mentioned OIMs are suitable for several types of degradation (e.g., additive noise, reverberation, filtering, clipping), it turns out that they are less appropriate for methods where noisy speech is processed by a time-frequency (TF) weighting. This includes single-microphone speech-enhancement algorithms, e.g., [8], but also speech separation techniques like ideal time frequency segregation (ITFS) [9], where typically a binary TF-weighting is used. For example, STI and various STI-based measures predict an intelligibility improvement when spectral subtraction is applied [7, 10]. This is not in line with the results of listening experiments in literature, where it is reported that general single-microphone speech-enhancement algorithms are not able to improve the intelligibility of noisy speech [10, 11]. Furthermore, OIMs like the coherence SII [5] and a covariance-based STI procedure, [12, 7], both show low correlation with the intelligibility of ITFS-processed speech [13, 14]. Only recently, two different OIMs are proposed which indicate promising results for ITFS-processed speech [14, 15].

To analyze the effect of certain signal degradations on the speech-intelligibility in more detail, the OIM must be of a simple structure, i.e., transparent. However, some OIMs are based on a large amount of parameters which are extensively trained for a certain dataset. This makes these measures less transparent, and therefore less appropriate for these evaluative purposes. Moreover, OIMs are often a function of long-term statistics of entire speech signals, and do not use an intermediate measure for local short-time TF-regions. With these measures it is difficult to see the effect of a time-frequency localized signal-degradation on the speech intelligibility.

In this paper we present an OIM, which shows high correlation with the intelligibility of both noisy and ITFS-processed noisy speech. It has a relatively simple structure, and is based on an intermediate measure for short-time (≈400 ms) TF-regions. In addition, a simple DFT-based TF-decomposition is used. Due to these properties, it is a transparent model which is suitable for evaluative purposes.

2. A SHORT-TIME OBJECTIVE INTELLIGIBILITY MEASURE

The proposed method is a function of the clean and processed speech, denoted by x and y, respectively. The model is designed for a sample-rate of 10000 Hz, in order to cover the relevant frequency range for speech-intelligibility. Any signals at other sample-rates should be resampled. Furthermore, it is assumed that the clean and the processed signal are both time-aligned.

First, a TF-representation is obtained by segmenting both signals into 50% overlapping, Hanning-windowed frames with a length of 256 samples, where each frame is zero-padded up to 512 samples and Fourier transformed. Then, an one-third octave band analysis is performed by grouping DFT-bins. In total 15 one-third octave bands are used, where the lowest center frequency is set equal to 150 Hz. Let $\hat{x}(k, m)$ denote the k^{th} DFT-bin of the m^{th} frame of the clean speech. The norm of the j^{th} one-third octave band, referred to as a TF-unit, is then defined as,

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

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

where k_1 and k_2 denote the one-third octave band edges, which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly, and will be denoted by $Y_j\left(m\right)$.

The intermediate intelligibility measure for one TF-unit, say $d_j\left(m\right)$, depends on a region of N consecutive TF-units from both $X_j(n)$ and $Y_j(n)$, where $n\in\mathcal{M}$ and $\mathcal{M}=\{\left(m-N+1\right),\left(m-N+2\right),...,m-1,m\}$. First, a local normalization procedure is applied, by scaling all the TF-units from $Y_j(n)$ with a factor $\alpha = \left(\sum_n X_j\left(n\right)^2/\sum_n Y_j\left(n\right)^2\right)^{1/2}$, such that its energy equals the clean speech energy, within that TF-region. Then, $\alpha Y_j(n)$ is clipped in order to lower bound the signal-to-distortion ratio (SDR), which we define as,

$$SDR_j\left(n\right)=10\log_{10}\left(\frac{X_j\left(n\right)^2}{\left(\alpha Y_j\left(n\right)-X_j\left(n\right)\right)^2}\right). \tag{2}$$

Hence,

$$Y'=\max\left(\min\left(\alpha Y,X+10^{-\beta/20}X\right),X-10^{-\beta/20}X\right), \tag{3}$$

where Y' represents the normalized and clipped TF-unit and β denotes the lower SDR bound. The frame and one-third octave band indices are omitted for notational convenience. The intermediate intelligibility measure is defined as an estimate of the linear correlation coefficient between the clean and modified processed TF-units,

$$d_j(m)=\frac{\sum\limits_n\left(X_j(n)-\frac{1}{N}\sum\limits_l X_j(l)\right)\left(Y_j'(n)-\frac{1}{N}\sum\limits_l Y_j'(l)\right)}{\sqrt{\sum\limits_n\left(X_j(n)-\frac{1}{N}\sum\limits_l X_j(l)\right)^2\sum\limits_n\left(Y_j'(n)-\frac{1}{N}\sum\limits_l Y_j'(l)\right)^2}}, \tag{4}$$

where $l\in\mathcal{M}$. Finally, the eventual OIM is simply given by the average of the intermediate intelligibility measure over all bands and frames,

$$d=\frac{1}{JM}\sum_{j,m}d_j\left(m\right), \tag{5}$$

where M represents the total number of frames and J the number of one-third octave bands.

In our experiments, we used different values of $N\in$[20, 30, 40, 50, 60] and $\beta\in$[-∞ , -30, -20, -15, -10] [1]. Maximum correlation is obtained with β=-15 and N=30, which means that the intermediate measure depends on speech information from the last $\approx$400 ms.

3. SUBJECTIVE DATA

The subjective data is obtained from a listening experiment conducted by Kjems *et al.* [16], where noisy speech signals are ITFS-processed. ITFS is a technique which can improve the intelligibility of noisy speech significantly by applying a binary modulation pattern in a TF-representation. This binary modulation pattern has a value equal to one, when the SNR within a certain TF-component exceeds a user-defined local criterion (LC), and is commonly referred to as the ideal binary mask (IBM). It is ideal in the sense that the noise is needed separately from the clean speech. A mathematical description for the IBM is given as follows,

$$IBM\left(t,f\right)=\begin{cases}1 & \text{if } T\left(t,f\right)-M\left(t,f\right)>LC\\ 0 & \text{otherwise}\end{cases}, \tag{6}$$

where $T\left(t,f\right)$ and $M\left(t,f\right)$ denote the signal power in dBs, at time t and frequency f, for the target (clean speech) and the masker (noise only), respectively. An alternative way of calculating the IBM is also included in [16] which is only based on the clean speech. This so-called 'Target Binary Mask' (TBM) is obtained by comparing the clean speech power with the power of a signal with the long-term spectrum of the clean speech, within a TF-component. Therefore, the noise itself is not needed in order to determine the binary mask. For more details about the used TF-decomposition and reconstruction of the eventual speech signals, the reader is referred to [16].

The test signals are taken from the Dantale II corpus [17], where each excerpt consists of five words, all spoken by the same Danish female speaker. These sentences are degraded by four different types of additive noise: speech shaped noise (SSN), cafeteria noise, noise from a bottling factory hall and car interior noise at three different SNRs: 20% and 50% speech reception threshold (SRT) and an SNR of -60dB, which represents essentially pure noise. Eight different LC-values are chosen, including an unprocessed condition where only the noisy speech is presented, i.e., LC=-∞.

For the listening experiment, 15 normal-hearing native Danish speaking subjects participated, where the correctly recognized words are recorded by an operator without providing any form of feedback. Each subject listened to two five-word sentences for each condition. The average score for all users for one condition is then obtained by the average percentage of correct words. In total, this gives us (4$*$IBM + 3$*$TBM)$*$(3$*$SNR)$*$(8$*$LC)=168 conditions to be tested in the listening experiment. Only three TBM conditions are included since the TBM equals the IBM for the case that SSN is used, by definition.

4. MODEL EVALUATION

In order to evaluate the model, we compare the proposed method with three other OIMs which will be described in the next section, followed by a description of the evaluation procedure.

4.1. Reference Objective Measures

The first model is a normalized covariance-based STI-procedure [12, 7] (CSTI), which determines the correlation coefficient between the band intensity envelopes of the processed and clean speech. This correlation coefficient is then translated to an SNR, such that it can be plugged into the original STI-procedure [4]. Compared to other STI-based measures, CSTI showed good results with several types of nonlinear signal degradations, e.g., clipping and spectral subtraction [7]. Implementation details can be found in [7].

The second model is a sophisticated perceptual model (DAU) developed by Dau *et al.* [18], which can be used as an artificial observer for accurately predicting masking thresholds for various maskers. Its final distance measure is calculated by means of a linear correlation coefficient between the internal spectro-temporal representations, of the clean and processed speech, on short, 50% overlapping segments, as was proposed in [14]. The final outcome is then based on the average of these intermediate correlation coefficients. The model shows high correlation with the intelligibility of ITFS-processed speech, e.g., [13].

[1] The SDR-based clipping in Eq. (3) is general for all β. However, since $\beta<0$ in our experiments, the second argument of the max-operator is always negative, and can therefore be discarded.

The last model is a simple objective measure based on the normalized subband envelope correlation (NSEC) [15]. This model shows very good results with respect to the same IBM-conditions used in this paper [15]. First, a TF-decomposition is performed, after which the temporal envelope is frequency normalized, compressed, and followed by a DC-removal. The eventual outcome is then calculated by means of the correlation between all TF-points of the clean and processed speech.

4.2. General Procedure

Out of all the 168 ITFS-processing conditions, 75 conditions have a subjective intelligibility score above 80%. In order to prevent clustering for these high scores, which may bias the objective intelligibility prediction results, 41 randomly picked conditions, with a score above 80%, are discarded. As a consequence, the scores of the remaining subset are approximately uniformly distributed between 0%-100%.

For each subset-condition, 30 five-word sentences are randomly chosen from the corpus and concatenated. The clean and processed signal are then segmented into 50% overlapping, Hanning-windowed frames with a length of 256 samples where the maximum energy frame of the clean speech is determined. Finally, both signals are reconstructed, excluding all the frames where the clean speech energy is lower than 40 dB with respect to the maximum clean speech energy frame. With this procedure, time-frames with no significant speech energy (mainly silence regions), and therefore no contribution to the intelligibility, will not be included.

To compare the results between the objective measures and the subjective intelligibility scores directly, a mapping is needed in order to account for a nonlinear relation between the objective and subjective values. For the proposed method, and the CSTI a logistic function is applied,

$$f\left(d\right) = \frac{100}{1+\exp\left(ad+b\right)}, \tag{7}$$

while for DAU and NSEC a better fit was observed with the following function,

$$f\left(d\right) = \frac{100}{1+\left(ad+b\right)^{c}}, \tag{8}$$

where a, b and c in (7) and (8) are free parameters, which are fitted to the subjective data with a nonlinear least squares procedure, and d denotes the objective outcome. Due to better results with the latter proposed mapping for NSEC, the original logistic mapping which was proposed in [15] is omitted.

The performance of all the objective measures is evaluated by means of the root of the mean squared prediction error (RMSE),

$$\sigma = \sqrt{\frac{1}{S}\sum_{i}\left(s_i - f\left(d_i\right)\right)^2}, \tag{9}$$

where s refers to the subjective score, S denotes the total number of conditions in the subset, and i runs over all subset-conditions. In addition, the correlation coefficient between the subjective and objective data is calculated,

$$\rho = \frac{\sum_{i}\left(s_i-\mu_s\right)\left(f\left(d_i\right)-\mu_{f(d)}\right)}{\sqrt{\sum_{i}\left(s_i-\mu_s\right)^2\sum_{i}\left(f\left(d_i\right)-\mu_{f(d)}\right)^2}}, \tag{10}$$

where $\mu_{f(d)}$ and μ_s denote the average values of the objective and subjective data, respectively.

Model	a	b	c
PROP	-13.1903	6.5192	-
DAU	-2.8892	2.1710	2.4187
NSEC	-3.1805	2.8792	1.9055
CSTI	-5.5795	1.6113	-

Table 1. *Used values for the free parameters of the nonlinear mappings, for each OIM.*

5. RESULTS AND DISCUSSION

The results of the proposed method (PROP), together with the results of the three reference OIMs are shown in Fig. 1. Each plot shows the objective versus the subjective data, together with the nonlinear mapping. The unprocessed noisy speech conditions, i.e. $LC = -\infty$, are denoted by the crosses and the remaining ITFS-processed speech is represented by the dots. The two figures of merit are presented on top of each plot. Table 1 denotes the obtained values for the free parameters of the nonlinear mappings.

The proposed method shows the best results for both figures of merit, with σ=10.2% and ρ=0.95. After the proposed method, both the complex DAU-model and NSEC indicate reasonable results, however, the plots clearly show that these models are only reliable for the conditions where the intelligibility score is relatively high. This behavior is not present with the proposed method, where the mapping shows a better fit with the data over the complete intelligibility range. The lowest performance, with respect to the two figures of merit, is observed with the CSTI, which makes it a less reliable intelligibility estimator for these ITFS-processed speech signals.

The crosses in the scatter-plots reveal that all three reference objective measures significantly underestimate the intelligibility scores for most unprocessed noisy speech conditions, compared to the ITFS-processed signals. Lower bounding the SDR per TF-unit to -15dB, with the proposed method, was found to be of critical importance in order to avoid this bias. This makes the proposed method both reliable for noisy and ITFS-processed speech. Note, that the normalization of the processed speech with α in Eq. (3), before the clipping, must be applied, despite the fact that the correlation coefficient in Eq. (4) itself is independent on any linear scaling of the clean and processed signal. Omitting this stage might result in clipping all TF-units, due to large global energy differences between the clean and processed speech for certain ITFS-conditions (e.g., noisy mixtures at -60dB).

Despite the similarity between ITFS, and conventional speech-enhancement algorithms, it is not guaranteed that the proposed OIM can predict the intelligibility of such algorithms. However, preliminary experiments indicate that the proposed method does not report a significant intelligibility change for a well-known speech-enhancement scheme [8]. This observation is more in line with literature [11], than the results of conventional OIMs, which predict a significant intelligibility improvement [7, 10]. Currently, the model is evaluated more extensively with respect to these types of enhanced signals.

6. CONCLUSIONS

A simple objective intelligibility measure (OIM) is presented, which shows high correlation (ρ=0.95) with the intelligibility of time-frequency (TF) weighted noisy speech. The method shows significantly better performance than three other, more sophisticated, reference OIMs. Furthermore, it turned out that the model is also re-

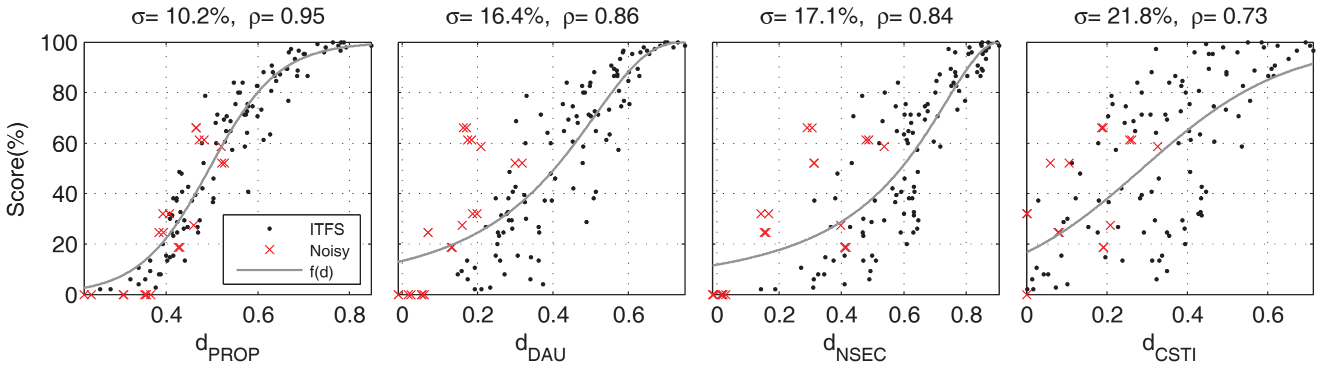

Fig. 1. *Results for the proposed method (left plot), together with the results for the three reference OIMs. The unprocessed noisy speech conditions are denoted by the crosses, and the remaining ITFS-processed conditions are represented by the dots. The gray line denotes the mapping used to translate the objective output to an intelligibility score. On top of each plot, the RMSE (σ) and the correlation coefficient (ρ), between the subjective and objective intelligibility scores, are given.*

liable for the noisy unprocessed speech. This was not the case for the reference OIMs, which all underestimated the intelligibility of the unprocessed noisy speech, compared to the TF-weighted noisy speech. The model is based on an intermediate intelligibility measure for short-time (≈400 ms) TF-regions and uses a simple DFT-based TF-decomposition. These properties make the model transparent, and therefore suitable for more detailed analysis of the effects of TF-weighting on noisy speech. A free Matlab implementation is provided at `http://siplab.tudelft.nl/users/cees-taal/`.

7. REFERENCES

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[2] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[3] ANSI, "Methods for calculation of the speech intelligibility index (American national standards institute, New York)," *ANSI S3.5-1997*, 1997.

[4] 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.

[5] 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.

[6] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[7] 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.

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

[9] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[10] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method - comparison between observed scores and scores predicted from STI," *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[11] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. ICASSP*, 2007, vol. 4, pp. IV–561–IV–564.

[12] Koch. R., *Gehörgerechte Schallanalyse zur Vorhersage und Verbesserung der Sprachverständlichkeit*, Ph.D. thesis, Universität Göttingen, 1992.

[13] 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.

[14] C. Christiansen, "Speech intelligibility prediction of linear and nonlinear processed speech in noise," M.S. thesis, Technical University of Denmark, 2008.

[15] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proc. EUSIPCO*, 2009, pp. 1849–1853.

[16] 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.

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

[18] 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. 3615–22, 1996.