A WEIGHTED STOI INTELLIGIBILITY METRIC BASED ON MUTUAL INFORMATION

Leo Lightburn and Mike Brookes

Dept. of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

It is known that the information required for the intelligibility of a speech signal is distributed non-uniformly in time. In this paper we propose WSTOI, a modified version of STOI, a speech intelligibility metric. With WSTOI the contribution of each time-frequency cell is weighted by an estimate of its intelligibility content. This estimate is equal to the mutual information between two hypothetical signals at either end of a simplified model of human communication. Listening tests show that the modification improves the prediction accuracy of STOI at all performance levels on both long and short utterances. An improvement was observed across all tested noise types and suppression algorithms.

Index Terms— Intelligibility, intelligibility metric, intelligibility estimate, mutual information, speech entropy

1. INTRODUCTION

In situations where a measurement of the intelligibility of a degraded speech signal is required, but where listening tests are too impractical or time consuming, an algorithmic estimate of the intelligibility may provide an adequate substitute for its true value. A popular family of intelligibility metrics are based on a correlation-comparison between the spectral envelopes of clean and degraded versions of the speech. One such metric, the Short-Time Objective Intelligibility Measure (STOI) [22], has been shown to have a high correlation with the intelligibility scores of both unenhanced and enhanced noisy speech signals [12, 20].

More recently, metrics based on the mutual information between the spectral envelopes of the clean and degraded signal have been proposed [14, 23]. The metric proposed in [23], which estimates mutual information using a k-nearest neighbor estimator, achieved comparative results to STOI in one of the tested performance measures and marginally worse results in the other. The metric in [14], which is computed from the lower bounds of mutual information, achieved a performance approximately equal to that of STOI.

To obtain the final metric in [22, 14, 23], the intermediate intelligibility measures computed in different Time-Frequency (TF) regions are averaged using uniform weights. However, it is known that not all portions of a speech signal contain equal quantities of the information required for intelligibility. For example, multiple studies in which parts of a waveform corresponding to consonants and vowels are replaced with noise have observed that vowel phones appear to contribute more to speech intelligibility than consonants [7, 15, 9]. In [21] the authors investigated the link between the relative information carried by different sections of speech and the degree to which the signal in those sections changed as a function of time. Their findings were consistent with the sensitivity to change of human perceptual systems and also with the principle from information theory that the information a signal carries is related to its unpredictability. Encouraged by the results in [21], the authors of [5] compared the intelligibility prediction performance of two metrics after modifying them to exclude segments of speech containing little speech information. The authors found that the best performing segmentation schemes retained most segments corresponding to vowel-consonant transitions and excluded vowel-only or consonant-only segments.

In this paper we propose a modified version of STOI in which the contribution of each TF cell is weighted according to the estimated contribution of the cell to intelligibility. This estimate equals the mutual information between two versions of a hypothetical signal, representing the information bearing component of the clean speech envelope, at either end of a simplified model of human communication. The modification improves STOI by better accounting for the variation in information content of a speech signal with time and frequency. An added bonus is that, since "silent" frames contain little or no information and are therefore downweighted, it is no longer necessary to delete these frames before calculating the STOI metric. This is advantageous since STOI's deletion scheme is sensitive to high energy frames and can result in the concatenation of speech segments that are widely separated in time.

2. OVERVIEW OF STOI

We present here a brief overview of the STOI metric on which the work described in this paper is based; readers are referred to [22] for a more detailed description. The metric compares a clean speech sample with a degraded speech sample. The clean sample is first converted into the Short Time Fourier Transform (STFT) domain using 50%-overlapping Hanning analysis windows of length $25.6\,\mathrm{ms}$. STFT frames whose total energy is 40 dB or more below that of the frame with highest energy are deemed to be silent. These frames are deleted

from both the clean and degraded speech signals and are not used in calculating the STOI metric. The resultant complex-valued STFT coefficients, $X(k,m)$, are then combined into $J = 15$ third-octave bands by computing the TF cell amplitudes

$$X_j(m) = \sqrt{\sum_{k=K_j}^{K_{j+1}-1} |X(k,m)|^2} \quad \text{for } j = 1, \ldots, J \quad (1)$$

where K_j is the lowest STFT frequency bin within frequency band j. The correlation between clean and degraded speech is performed on vectors of duration 384 ms. For each m, we therefore define the modulation vector

$$\mathbf{x}_{j,m} = [X_j(m-M+1), X_j(m-M+2), \ldots, X_j(m)]^T \quad (2)$$

comprising $M = 384/(0.5 \times 25.6) = 30$ consecutive TF cells within frequency band j. The degraded speech is similarly processed to obtain $Y(k,m)$, $Y_j(m)$ and $\mathbf{y}_{j,m}$. Before computing the correlation, the degraded speech amplitudes, $Y_j(m)$, are clipped to limit the impact of frames containing low speech energy. The clipped TF cell amplitudes, denoted by a tilde superscript, are determined as

$$\widetilde{Y}_j(m) = \min\left(Y_j(m),\ \lambda \frac{\|\mathbf{y}_{j,m}\|}{\|\mathbf{x}_{j,m}\|} X_j(m)\right) \quad (3)$$

where $\lambda = 6.623$ and $\|\ \|$ is the Euclidean norm. The corresponding modulation vectors are $\tilde{\mathbf{y}}_{j,m}$. The STOI contribution of the TF cell $(j,\ m)$ is then given by

$$d\left(\mathbf{x}_{j,m},\ \tilde{\mathbf{y}}_{j,m}\right) \triangleq \frac{\left(\mathbf{x}_{j,m} - \bar{x}_{j,m}\right)^T \tilde{\mathbf{y}}_{j,m}}{\|\mathbf{x}_{j,m} - \bar{x}_{j,m}\| \, \|\tilde{\mathbf{y}}_{j,m} - \bar{\tilde{y}}_{j,m}\|} \quad (4)$$

where $\bar{x}_{j,m}$ denotes the mean of vector $\mathbf{x}_{j,m}$. The overall STOI metric is found by averaging the contributions of TF cells over all bands, j, and all frames, m.

3. WEIGHTED-STOI

In this section we describe WSTOI, a modified version of STOI in which the contribution of each TF cell is weighted by an estimate of its intelligibility content. In WSTOI, it is no longer necessary to identify and delete "silent" frames since these frames contain little information in any case. In the block diagram of WSTOI shown in Fig. 1, the left panel is identical to STOI and calculates the STOI contribution, $d\left(\mathbf{x}_{j,m},\ \tilde{\mathbf{y}}_{j,m}\right)$, of each TF cell. The right panel determines the weight, $I_{j,m}$, to apply to each cell and the final metric in the lower block is a weighted sum of the contribution from each TF cell. To compute $I_{j,m}$ we estimate the mutual information between the unpredictable component of a hypothetical signal that we assume the speaker intended to produce, and the version of this signal which is perceived by a

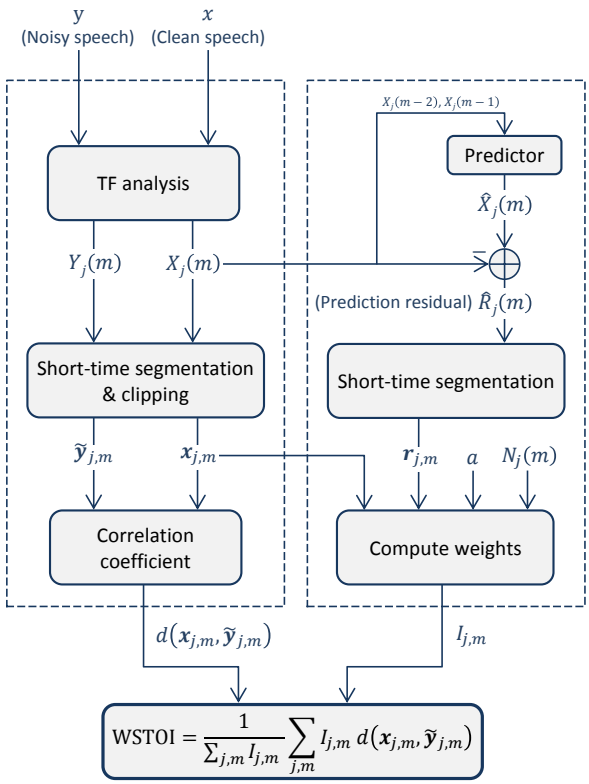

Fig. 1: Diagram of the modified version of STOI.

listener in an imagined scenario where the listener hears the clean speech signal at a comfortable listening level. To do this, we consider a simple model of communication between the speaker and listener.

3.1. STOI weights

The clean speech, $X_j(m)$, is expressed as $X_j(m) = S_j(m) + V_j(m)$ where $S_j(m)$ is the speaker's intended speech signal and $V_j(m)$ models the "production noise" proposed in [16]. The production noise models the natural variation in the human speech production process. We denote, by $\hat{S}_j(m)$, a linear prediction of $S_j(m)$ of order P, so that $\hat{S}_j(m) = \sum_{p=1}^{P} b_p S_j(m-p)$ where b_p are a set of frequency band-dependent prediction coefficients. The linear prediction residual, $R_j(m) = \hat{S}_j(m) - S_j(m)$, is the message signal at the input of our communication model. The output of our model, termed $W_j(m)$, is the sum of $R_j(m)$, $V_j(m)$ and a frequency band-dependent hypothetical internal ear noise, $N_j(m)$, which models the absolute threshold of human hearing. If we model $X_j(m)$, $R_j(m)$ and $N_j(m)$ as Gaussian random variables whose power is constant within each modulation vector, we can estimate the mutual information between the signals at the input and output of the model, in the modulation vector ending in $(j,\ m)$, as

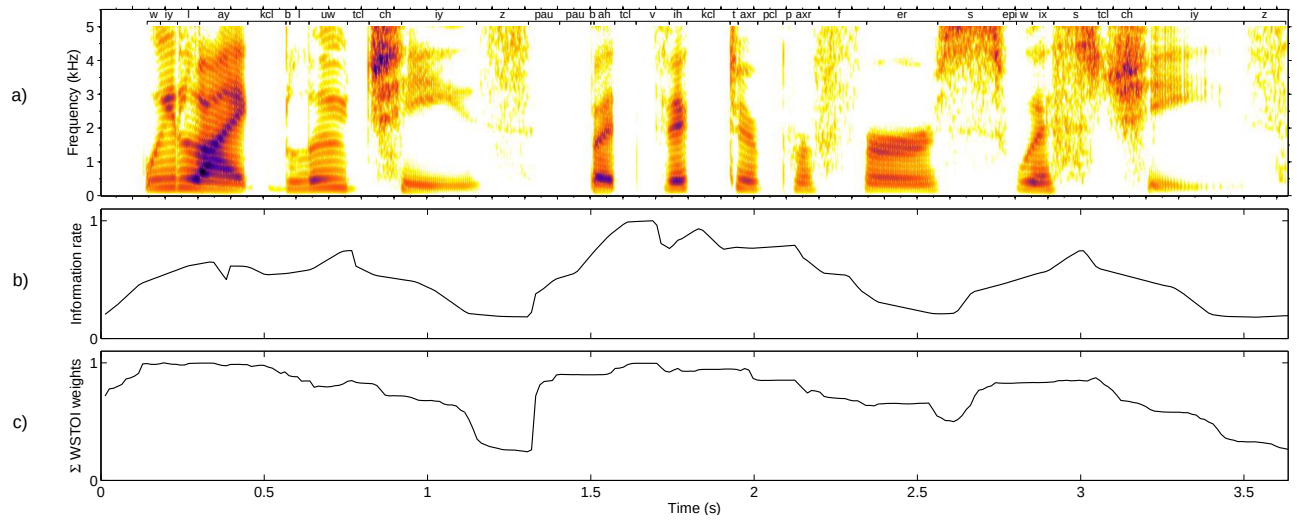

Fig. 2: a) Spectrogram of the utterance "We like blue cheese but Victor prefers Swiss cheese" with a phonetic transcription shown above. b) Speech information rate predicted by the phoneme-level trigram language model from Sec. 3.2. c) WSTOI weights, (5), summed over all frequency bands.

$$I(\mathbf{r}_{j,m};\, \mathbf{w}_{j,m}) = 0.5 \times \log_2 \left(1 + \frac{\|\mathbf{r}_{j,m}\|^2}{\left(N_j(m) + a \, \|\mathbf{x}_{j,m}\|^2 \right)} \right), \tag{5}$$

where $I(\,)$ denotes the mutual information, $\mathbf{r}_{j,m}$ and $\mathbf{w}_{j,m}$ are defined in an analogous way to (2), and a is a fixed coefficient representing the ratio of the power of the production noise to that of the speech signal [16]. The value of a is determined from training data as described in Sec. 3.2. When forming $\mathbf{r}_{j,m}$ in (5) we approximate $R_j(m)$ as $\hat{R}_j(m) = \hat{X}_j(m) - X_j(m)$ since $S_j(m)$ is unavailable. Using (5) as weights, WSTOI is computed as a weighted average of (4) over all bands, j, and all frames, m.

3.2. Optimising weights with language model

The TF-dependent weight in (5) is a measure of the local information capacity of the communications channel. To determine the free parameter, a, in (5), we assume that the information content of the speech mirrors the channel capacity. Accordingly, the parameter a is chosen to maximize the correlation between (5) summed over all frequency bands, j, and the speech information as estimated from a phoneme-level trigram language model. The interpolated Kneser-Ney model from [6] was implemented; this is a modified version of the model from [17]. The optimisation was performed on TIMIT [10] with phoneme labels mapped to the reduced set from [18]. The correlation coefficients were computed over the length of each utterance. The output of the language model was the negative log conditional probability of the third phoneme given the previous two phonemes, divided by the duration of the phoneme. The language model output was smoothed with a moving average of window length $M = 30$, to replicate the smoothing effect of (5). The optimum was found to be $a = 2.2 \times 10^{-4}$.

Fig. 2 shows a) a spectrogram of the utterance "We like blue cheese but Victor prefers Swiss cheese", b) the smoothed output of the language model and c) the STOI weights from (5) summed over all frequency bands. The information rate estimated by the language model is high in time intervals containing many closely spaced phonemes. The summed weights are high in intervals with frequent changes in the speech spectrum. Since intervals with closely spaced phonemes coincide with intervals where the spectrum changes frequently, the summed weights mirror the information rate estimated by the language model.

4. EXPERIMENTAL PROCEDURE

The WSTOI algorithm was evaluated using the results of the intelligibility tests in [11]. Recordings of the IEEE sentences [19] spoken by a single male speaker combined with babble or car noise were played at one of five Signal-to-Noise Ratios (SNRs) to 60 listeners in either an unprocessed condition or after having been processed using one of three noise suppressors. The number of content words a listener was able to correctly identify in each sentence (between zero and five) was recorded. We define intelligibility as the % of content words correctly identified. The responses to a total of 200 sentences were recorded for each combination of noise type, SNR, noise suppressor and suppressor condition (On/Off). For car noise, $\mathrm{SNR} = -\{21, 18, 15, 12, 9\}$ dB, and for babble noise, $\mathrm{SNR} = -\{12, 9, 6, 3, 0\}$ dB. The suppressor algorithms were spectral subtraction (SS) [3, 4], minimum mean squared error log spectral estimation (MMSE) [8, 4] and subspace enhancement (SSA) [13]. STOI scores, d, were mapped

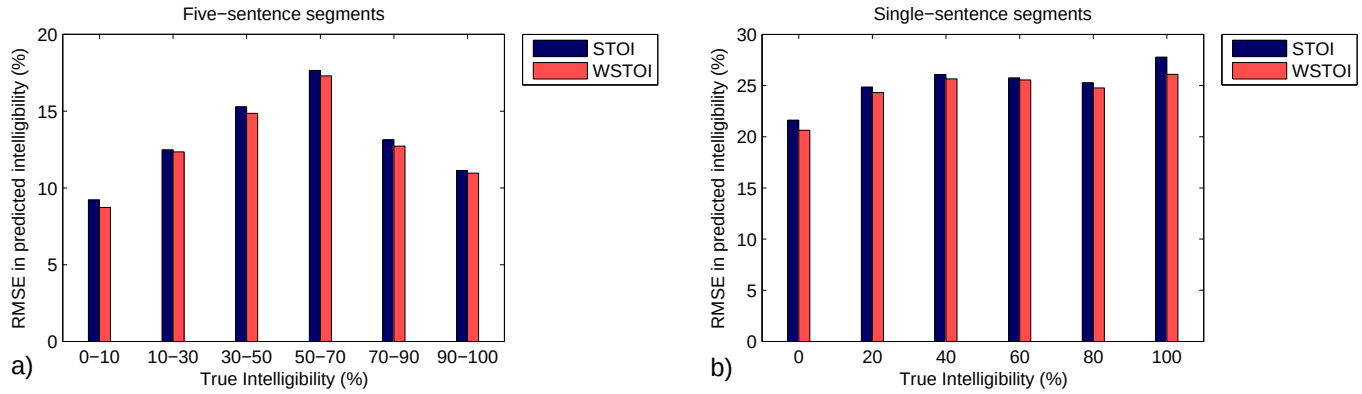

Fig. 3: Root mean square error in predicted intelligibility against intelligibility for STOI and WSTOI applied to a) five-sentence segments (25 content words) and b) single-sentence segments (5 content words).

to an intelligibility prediction using the logistic function from [22],

$$f(d) = \frac{100}{1 + \exp\left(cd + e\right)},$$

where c and e are free parameters which were fitted to the data using non-linear least squares optimization. Separate mappings were computed for STOI and WSTOI. The available data was split randomly, with half used to optimize the logistic mapping and the remaining half for algorithm evaluation. This process was repeated 1000 times using different splits, with the results from each repetition averaged to compute an overall set of results. The values $N_j(m)$ for $j = 1, \ldots, 15$ in (5) were obtained by integrating the reference internal noise spectrum levels from Table 3 of [2] over the width of each frequency band and then scaling the resulting values for each utterance so that the mean speech-to-internal-noise power ratio of the utterance during active speech periods matched the ratio of the speech and noise spectrum levels for a "normal" vocal effort. Active periods were identified using the procedure in [1]. The prediction order was $P = 3$.

5. RESULTS

Fig. 3a plots the root mean square error (RMSE) in predicted intelligibility against the true intelligibility, for STOI and WSTOI applied to five-sentence segments having the same noise type, SNR and suppressor condition. The histogram is grouped according to the true intelligibility of each segment. We see that both STOI and WSTOI were able to predict the true intelligibility with a root-mean-square error (RMSE) of between 8.7% and 17.7% and that WSTOI gave a lower RMSE in all bins. Fig. 3b shows the performance of STOI and WSTOI on single-sentence segments containing only five content words. Even with these short segments, both STOI and WSTOI were able to predict the intelligibility with an RMSE of between 20.6% and 27.8%. For every one of the 1000 splits the intelligibility prediction performance of WSTOI was significantly better than that of STOI with $p < 10^{-6}$ using a 1-sided sign test.

Fig. 4 shows the RMSE in predicted intelligibility for the algorithms applied to single-sentence segments, plotted for each suppressor and noise type. For every combination of suppressor and noise type WSTOI resulted in a lower RMSE than STOI.

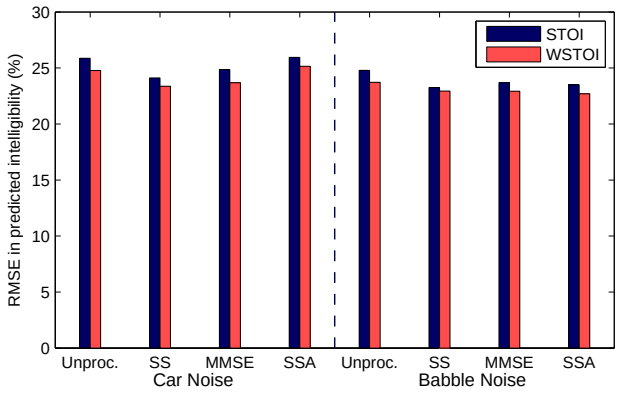

Fig. 4: Root mean square error in predicted intelligibility for STOI and WSTOI applied to single-sentence segments, plotted for each suppressor and noise type.

6. CONCLUSION

We have presented WSTOI, a modified version of STOI in which the contribution of each TF cell is weighted by an estimate of its intelligibility content. The proposed method improves STOI's performance in active speech frames by weighting TF cells containing important speech information more heavily than cells containing less important information. Listening tests showed that the modification improved the prediction accuracy of STOI at all performance levels on both long and short utterances. An improvement was observed across all tested noise types and suppression algorithms.

7. REFERENCES

[1] Objective measurement of active speech level. Recommendation P.56, International Telecommunications Union (ITU-T), March 1993.

[2] ANSI. Methods for the calculation of the speech intelligibility index. ANSI Standard S3.5–1997 (R2007), American National Standards Institute, 1997.

[3] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 208–211, 1979.

[4] M. Brookes. VOICEBOX: A speech processing toolbox for MATLAB. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 2011.

[5] F. Chen and P. C. Loizou. Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise. *J. Acoust. Soc. Am.*, 131(5):4104–4113, May 2012.

[6] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393, 1999.

[7] Ronald A. Cole, Yonghong Yan, Brian Mak, Mark Fanty, and Troy Bailey. The contribution of consonants versus vowels to word recognition in fluent speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 853–856, 1996.

[8] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 33(2):443–445, 1985.

[9] D. Fogerty and D. Kewley-Port. Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *J. Acoust. Soc. Am.*, 126(2):847–857, August 2009.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. TIMIT acoustic-phonetic continuous speech corpus. Corpus LDC93S1, Linguistic Data Consortium, Philadelphia, 1993.

[11] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale. Effects of noise suppression on intelligibility: dependency on signal-to-noise ratios. *J. Acoust. Soc. Am.*, 131(1):531–539, 2012.

[12] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale. Effects of noise suppression on intelligibility. II: An attempt to validate physical metrics. *J. Acoust. Soc. Am.*, 135(1):439–450, January 2014.

[13] Y. Hu and P. C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech Audio Process.*, 11(4):334–341, July 2003.

[14] J. Jensen and C. H. Taal. Speech intelligibility prediction based on mutual information. *IEEE Trans. Audio, Speech, Lang. Process.*, 22:430–440, 2014.

[15] Diane Kewley-Port, T. Zachary Burkle, and Jae Hee Lee. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.*, 122(4):2365–2375, October 2007.

[16] Bastiaan W. Kleijn and R. C. Hendriks. A simple model of speech communication and its application to intelligibility enhancement. *IEEE Signal Process. Lett.*, 22(3):303–307, March 2015.

[17] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, May 1995.

[18] Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Trans. Acoust., Speech, Signal Process.*, 37(11):1641–1648, November 1989.

[19] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.*, 17(3):225–246, 1969.

[20] Belinda Schwerin and Kuldip Paliwal. An improved speech transmission index for intelligibility prediction. *Speech Communication*, 65:9–19, 2014.

[21] Christian E. Stilp and Keith R. Kluender. Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proc. National Academy of Sciences*, 107(27):12387–12392, 2010.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2125–2136, September 2011.

[23] Jalal Taghia, Rainer Martin, and Richard C. Hendriks. On mutual information as a measure of speech intelligibility. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 65–68, 2012.