Prediction of Intelligibility of Noisy and Time-Frequency Weighted Speech based on Mutual Information Between Amplitude Envelopes

Jensen, Jesper; Taal, C.H.

Published in:
14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)

Publication date:
2013

Document Version
Early version, also known as pre-print

Prediction of Intelligibility of Noisy and Time-Frequency Weighted Speech based on Mutual Information Between Amplitude Envelopes

Jesper Jensen[1] *and Cees H. Taal*[2]

[1]Oticon A/S, Smørum, and Aalborg University, Aalborg, Denmark.
[2]Leiden University Medical Center, Leiden, The Netherlands
jsj@oticon.dk and jje@es.aau.dk, and c.h.taal@lumc.nl

Abstract

This paper deals with the problem of predicting the average intelligibility of noisy and potentially processed speech signals, as observed by a group of normal hearing listeners. We propose a prediction model based on the hypothesis that intelligibility is monotonically related to the the amount of Shannon information the critical-band amplitude envelopes of the noisy/processed signal convey about the corresponding clean signal envelopes. The resulting intelligibility predictor turns out to be a simple function of the correlation between noisy/processed and clean amplitude envelopes. The proposed predictor performs well ($\rho > 0.95$) in predicting the intelligibility of speech signals contaminated by additive noise and potentially non-linearly processed using time-frequency weighting.
Index Terms: Intelligibility prediction. Mutual information. Auditory model. Time-frequency weighting. Single-channel noise reduction.

1. Introduction

Speech intelligibility prediction algorithms aim at predicting the average intelligibility of noisy and potentially processed speech signals, as judged by a group of listeners. Robust intelligibility predictors are of great practical importance, e.g., in guiding the development process of speech processing algorithms in early stages of the development phase, and have the potential to lead to a better understanding of human intelligibility capabilities.

Most existing intelligibility predictors are rooted in the *Articulation Index (AI)* [1], proposed first by French and Steinberg [2] and later refined by Kryter [3], or the *Speech Transmission Index (STI)* [4] proposed by Steeneken and Houtgast [5, 6].

The AI was originally intended for stationary noise situations. It has later been refined further and standardized as the *Speech Intelligibility Index (SII)* [7]. Several extensions of AI/SII exist, e.g., the *Extended SII* [8], which was developed to take into account fluctuating noise sources, and *Coherence SII (CSII)* which was proposed to better take into account non-linear distortions such as peak- and center-clipping [9].

The STI approach [5, 6] extends the type of degradations to convolutive noise sources. The STI has also seen extensions to take into account the effects of various non-linear processing operations, such as dynamic amplitude compression [10], and envelope clipping [11]. More recently, the class of *speech STI (sSTI)* methods [12] were proposed to improve performance for these processing types.

Generally speaking, existing AI and STI based intelligibility predictors are less suited for speech signals distorted by non-stationary noise sources and time-varying and non-linear processing, e.g., as in single-channel speech enhancement systems [13, 14]. Recently proposed intelligibility predictors, however, have shown promising results for this type of distortion, e.g., the *STOI* measure [15].

In this paper we focus on speech signals contaminated by additive noise, and potentially processed by a single-channel noise reduction type of system mentioned above. We use a standard signal processing model of the auditory periphery[1] consisting of a filter bank simulating the bandpass characteristics of the cochlea, and a full-wave rectification to simulate coarsely the hair cell transduction in the inner ear. The basic idea is to compare the resulting critical-band amplitude envelopes of the clean and noisy/processed signal. Specifically, we hypothesize that the intelligibility of the noisy/processed signal can be described by the amount of information the corresponding amplitude envelopes convey about the clean amplitude envelopes.

The proposed model shows similarities to STOI [15] both in terms of structure and performance. However, it avoids some of the heuristically motivated choices made in STOI and may thus be seen as a better motivated model. It also bears some similarities to the method described in [18], although the motivation for the proposed model is quite different: it arises as a consequence of describing speech information transmission in a simple model of the auditory periphery, whereas the method in [18] has a more heuristic foundation in that it replaces the linear correlation operation used in the STOI model with a generalization, namely Shannon mutual information (MI). Furthermore, the proposed model relies on lower bounds of MI, leading to simple equations, whereas the method in [18] tries to estimate MI, which is a harder problem.

2. Mutual Information between Amplitude Envelopes

Fig. 1 shows a diagram of the proposed model. Let us focus first on the upper half of the figure where $S(n)$ denotes a random process modeling a clean speech input signal[2]. Band pass filtered signals $\tilde{S}(k,m)$ are obtained by applying the Discrete Fourier Transform to successive, overlapping analysis frames,

$$\tilde{S}(k,m) = \sum_{n=0}^{N-1} S(mD+n)w(n)e^{-j2\pi kn/N},$$

where k and m denote the frequency bin index and the frame index, respectively, D is the frame shift in samples, N is the

[1]This type of model is part of a typical automatic speech recognition front-end [16], and similar models have been used in speech enhancement systems [17] and for intelligibility prediction purposes [15].

[2]We use capital letters to denote random processes and variables and lower-case letters to denote the corresponding realizations.

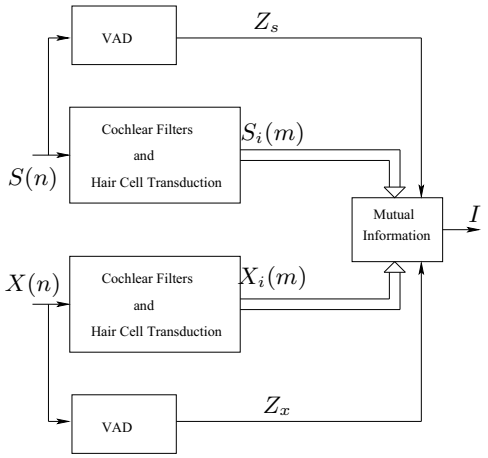

Figure 1: Intelligibility of $X(n)$ is estimated as the average information present in the amplitude envelopes $X_i(m)$ about the clean amplitude envelopes $S_i(m)$.

frame length in samples, and $w(n)$ is an analysis window.

A one-third octave band analysis is performed by grouping the DFT bins, resulting in clean critical-band amplitudes

$$S_i(m) = \sqrt{\sum_{k \in CB_i} |\tilde{S}(k,m)|^2}, \tag{1}$$

where CB_i is the frequency index set representing the ith one-third octave band, $i = 1, \ldots, L$. The upper Voice Activity Detection (VAD) block in Fig. 1 identifies signal frames of $S(n)$ with speech activity; the corresponding frame index set is denoted by Z_s. Finally, let M be the number of frames in a given speech sentence, and form random super vectors by stacking critical-band spectra

$$\mathcal{S} = [S_1(1)S_2(1)\cdots S_L(1)S_1(2)\cdots S_L(M)]^T.$$

The noisy and potentially processed signal $X(n)$, bandpass filtered signals $\tilde{X}(k,m)$, critical-band amplitudes $X_i(m)$, and corresponding super vector $\mathcal{X}$ are found in an analogous way.

We are interested in the average MI (to be defined exactly below) between clean and noisy/processed amplitude envelopes, i.e., $\frac{1}{L|Z_s|} I(\mathcal{S};\mathcal{X})$, where $|\cdot|$ denotes set cardinality, and $L|Z_s|$ estimates the number of clean speech-active critical-band amplitudes. Let us assume that the frame length N is large compared to the correlation time of the signals in question such that the entries in each super vector are statistically independent [19, 20]. Then $I(\mathcal{S};\mathcal{X})$ decomposes into a sum of MI $I(S_i(m); X_i(m))$ terms

$$\frac{1}{L|Z_s|} I(\mathcal{S};\mathcal{X}) = \frac{1}{L|Z_s|} \sum_{m \in Z_s \cap Z_x} \sum_{i=1}^{L} I(S_i(m); X_i(m)).$$

The equation follows because summation over the frame index set $m = Z_s \cap Z_x$, where both signals $S(n)$ and $X(n)$ are speech active, excludes $I(S_i(m); X_i(m))$ terms which are all zero.

For notational convenience, we skip the subband and frame index where possible, and replace $S_i(m)$ and $X_i(m)$ by S and X, respectively. The MI $I(S;X)$ between clean and noisy/processed critical-band amplitudes is given by [21]

$$I(S;X) = h(S) - h(S|X) \quad \text{[nats]}, \tag{2}$$

where the differential entropy of S is

$$h(S) = \int_S f_S(s) \ln f_S(s) ds,$$

and the conditional differential entropy $h(S|X)$ is

$$h(S|X) = \int_X \int_S f_{S,X}(s,x) \ln f_{S|x}(s|x) ds dx. \tag{3}$$

3. Upper Bound on $h(S|X)$

Estimating $I(S;X)$ in Eq. (2) from limited data with time-varying statistics is hard. Instead, we lower bound the MI $I(S;X)$; this turns out to require only second-order statistics of $f_{S,X}(s,x)$. To do so, we upper bound the conditional entropy $h(S|X)$ in Eq. (2), see [22] for another application of this procedure.

Let $E_X(\cdot)$ denote expectation with respect to the random variable X, and let $\mu_{S|x} = \int_y y f_{S|x}(y|x) dy$ and $\sigma^2_{S|x} = \int_y (y - \mu_{S|x})^2 f_{S|x}(y|x) dy$ denote the mean and variance, respectively, of the random variable distributed according to the conditional probability density function $f_{S|x}(s|x)$. Then, from Eq. (3) it follows that

$$\begin{aligned} h(S|X) &= -\int_x f_X(x) \int_s f_{S|x}(s|x) \ln f_{S|x}(s|x) ds dx \\ &\leq E_X\left(\frac{1}{2} \ln 2\pi e \sigma^2_{S|x}\right) \\ &\leq \frac{1}{2} \ln 2\pi e E_X\left(\sigma^2_{S|x}\right). \end{aligned} \tag{4}$$

The first inequality holds because the maximum entropy pdf for a random variable Y with a given variance σ^2_Y is the Gaussian pdf, which has a differential entropy of $h(Y) = \frac{1}{2} \ln 2\pi e \sigma^2_Y$. The second inequality follows from Jensen's inequality [21, Thm.2.6.2] and the fact that $\ln(\cdot)$ is concave.

How should we interpret $\sigma^2_{S|x}$? Recall that the conditional mean $\mu_{S|x}$ is equal to the minimum mean-square error (mmse) estimator $\hat{s}_{mmse}(x)$ of the clean random variable S upon observing the noisy and/or processed realization x [23]. Then,

$$\begin{aligned} \sigma^2_{S|x} &= \int_{y \geq 0} (y - \hat{s}_{mmse}(x))^2 f_{S|x}(y|x) dy \\ &\triangleq D_{mmse}(x). \end{aligned} \tag{5}$$

So, $\sigma^2_{S|x} \triangleq D_{mmse}(x)$ is simply the mean-square error (mse) resulting from estimating S upon observing x, using an mmse estimator. Let D_{mmse} denote $D_{mmse}(x)$ averaged across all realizations of the noisy/processed critical-band amplitude x,

$$D_{mmse} = \int_{x \geq 0} f_X(x) D_{mmse}(x) dx. \tag{6}$$

Inserting Eq. (5) in Eq. (4) and using Eq. (6), we arrive at

$$\begin{aligned} h_{mmse}(S|X) &\triangleq \frac{1}{2} \ln 2\pi e D_{mmse} \\ &\geq h(S|X). \end{aligned} \tag{7}$$

Evaluating the upper bound $h_{mmse}(S|X)$ via D_{mmse} in Eq. (7) may be difficult, since the pdf $f_{S,X}(s,x)$ is generally unknown. Instead we form a looser upper bound by replacing the mmse

estimator $\hat{s}_{mmse}(x) = E(S|x)$ with the *linear* mmse estimator $\hat{s}_{lmmse}(x)$. This results in an mse of

$$D_{lmmse}(x) = \int_{y\geq 0} (y - \hat{s}_{lmmse}(x))^2 f_{S|x}(y|x)dy \geq D_{mmse}(x),$$

with equality for jointly Gaussian (S, X). Furthermore

$$D_{lmmse} \triangleq \int_{x\geq 0} f_X(x) D_{lmmse}(x)dx \geq D_{mmse}.$$

It follows that a looser upper bound on $h(S|X)$ based on linear mmse estimators is given by

$$h_{lmmse}(S|X) \triangleq \frac{1}{2}\ln 2\pi e D_{lmmse} \geq h_{mmse}(S|X).$$

The quantity D_{lmmse} is a function of second-order statistics, rather than the joint pdf $f_{S,X}(s, x)$: let $\mu_S = E_S(S)$ and $\mu_X = E_X(X)$ denote expected values of S and X, respectively, and let $r_{SX} = E_{S,X}(SX)$, $\sigma_S^2 = E_S(S^2) - \mu_S^2$ and $\sigma_X^2 = E_X(X^2) - \mu_X^2$ denote the cross-correlation between S and X, the variance of S, and the variance of X, respectively. Finally, let $\rho_{SX} = \frac{r_{SX} - \mu_S\mu_X}{\sqrt{\sigma_S^2\sigma_X^2}}$ denote the linear correlation coefficient between S and X. It can be shown that the linear mmse (lmmse) is given by

$$D_{lmmse} = \sigma_S^2\left(1 - \rho_{SX}^2\right). \quad (8)$$

With the derived upper bounds on $h(S|X)$ we have the following lower bound on $I(S; X)$,

$$I_{lmmse}(S; X) \leq I(S; X), \quad (9)$$

where

$$I_{lmmse}(S; X) \triangleq \max\left\{h(S) - h_{lmmse}(S|X), 0\right\}. \quad (10)$$

4. Differential Entropy $h(S)$

The MI lower bound in Eq. (9) depends on the differential entropy $h(S)$ of the clean speech critical-band amplitudes. To derive an expression for $h(S)$, note that when the frame length N is large compared to the correlation time of the clean signal $S(n)$, then the real and imaginary parts of the DFT coefficients $\tilde{S}(k, m)$ can be considered independent zero-mean Gaussian variables [19, 20]. Assuming further that the DFT coefficients within the same critical band $\tilde{S}(k, m),\ k \in CB_i$ are identically distributed (i.e., the power spectral density is constant), then $S_i(m)$ given in Eq. (1) follows a (scaled) chi-distribution with $k' = 2|CB_i|$ degrees of freedom.

It can be shown that the differential entropy of the corresponding critical-band amplitude is

$$h(S) = h(Z) - \frac{1}{2}\ln\sigma_Z^2 + \frac{1}{2}\ln\sigma_S^2. \quad (11)$$

Thus, the differential entropy $h(S)$ is a simple function of the variance σ_S^2 of the critical-band amplitudes, because the two first terms in Eq. (11) are functions only of the number of degrees of freedom k' and can therefore be computed offline.

Inserting Eq. (11) in Eq. (10), we find

$$I_{lmmse}(S; X) = \max\left(c' + \frac{1}{2}\ln(1 - \rho_{SX}^2)^{-1}, 0\right), \quad (12)$$

where $c' = h(Z) - \frac{1}{2}\ln\sigma_Z^2 - \frac{1}{2}\ln 2\pi e$ is a signal-independent constant (but which depends on k').

5. Implementation

Signals are resampled to a sampling frequency of 10 kHz, divided into frames of length $N = 256$ samples, and a Hann analysis window $w(n)$ is applied. The frame shift is $D = N/2 = 128$ samples, and a DFT order of $N = 256$ is used. DFT coefficients are grouped into $L = 15$ third-order octave bands, with a center frequency of the lowest band of 150 Hz, and the center frequency of the highest band set to approximately 4.3 kHz, see [15]. The VADs in Fig. 1 identify signal frames with energy no less than Δ_E dB of the signal frame with maximum energy.

Let $\bar{S}_i(m)$ and $\bar{X}_i(m)$ denote critical-band amplitudes with frame indices $m \in Z_s \cap Z_x$. The first- and second-order moments needed to evaluate $I(\bar{S}_i(m), \bar{X}_i(m))$ via Eqs. (12) and Eq. (8) are estimated using first-order recursive smoothing, i.e.,

$$\hat{r}_{S_iX_i}(m+1) = \alpha\hat{r}_{S_iX_i}(m) + (1-\alpha)\bar{S}_i(m+1)\bar{X}_i(m+1),$$

and similarly for the other moments.

Let $\hat{I}(S_i(m); X_i(m))$ denote the estimate of $I_{lmmse}(S_i(m); X_i(m))$ obtained by replacing expected values by recursively estimated moments. The average per sentence MI is finally computed as

$$\tilde{I}(\mathcal{S}; \mathcal{X}) = \frac{1}{L|Z_s|}\times \sum_{m\in Z_x\cap Z_s}\sum_{i=1}^{L}\min\left(\hat{I}(S_i(m); X_i(m)), I_{max}\right),$$

where the upper limit I_{max} takes into account that at a sufficiently high SNR, generally speaking, a signal is perfectly intelligible, and increasing the SNR further cannot increase intelligibility.

The values of the three parameters, α, Δ_E, and I_{max} are summarized in Table 1. The value of $\alpha = 0.95$ corresponds to a time constant of 250 ms. This is in the same range as for STOI [15] where statistics were computed across time spans of roughly 400 ms. The value of $I_{max} = 0.2$ nats was determined heuristically. Intelligibility prediction performance appears to be insensitive to the exact values of any of these parameters.

6. Simulation Results

We evaluate the proposed intelligibility predictor using noisy speech signals processed with different time-frequency weighting strategies.

6.1. Signals and Processing Conditions

6.1.1. Additive Noise

The first set of signals is from the study described by Kjems et. al. in [24]. Speech signals from the Dantale II sentence test [25] are contaminated by four different additive noise sources: stationary speech-shaped noise, car cabin noise, bottle hall noise, and two-talker babble noise. While the first two noise sources are essentially stationary, the last two are highly non-stationary. Noisy test signals were generated with SNRs from -20 dB to 5 dB in steps of 2.5 dB, and the intelligibility was evaluated

Parameter	α	Δ_E [dB]	I_{max} [nats]
Value	0.95	30	0.2

Table 1: Parameter values in proposed model.

by 15 normal-hearing subjects. The test included a total of 44 different test conditions. For details we refer to [24].

6.1.2. Ideal Binary Mask Signals

In a second experiment, Kjems [24] processed the noisy signals from above using the technique of ideal time-frequency segregation (ITFS) [26], and measured the intelligibility of the resulting signals for different versions of the ITFS processing. For each of the four noise types, three different SNRs were used: two SNRs were selected corresponding to the 20% and 50% speech reception threshold[3], while the third SNR was fixed at -60 dB. As above, 15 subjects participated. The test encompassed a total of 168 test conditions.

6.2. Per Sentence Mutual Information vs Intelligibility

In order to estimate *absolute* intelligibility, e.g. the proportion of correctly identified words in an intelligibility test, a mapping is needed between the outcome of the intelligibility predictor, and the true underlying intelligibility. This mapping is a function of many factors, including the noise type, the test type, the processing applied to the noisy signal, the redundancy of the speech material, and, obviously, the intelligibility predictor. As in [27, 15], we use a mapping which is a logistic function of the outcome $\tilde{I}$ of the intelligibility predictor

$$f(\tilde{I}) = \frac{1}{1+\exp(a\tilde{I}+b)},$$

where $a, b \in \Re$ are test specific model parameters, which are estimated to fit the intelligibility data.

For numerical performance evaluation of intelligibility predictors, we use the linear correlation coefficient ρ between average intelligibility scores obtained through listening tests, and the outcomes of the intelligibility predictors, and the root mean-square prediction error σ [15]. Let $\tilde{I}_k$ denote the intelligibility prediction for the kth processing condition, and let SI_k denote the average across listeners in the corresponding intelligibility test. Furthermore, let $\mu_{f(\tilde{I})}$, and μ_{SI} denote the averages of $f(\tilde{I}_k)$ and SI_k, respectively, across listening test conditions, and let K denote the number of test conditions. The linear correlation coefficient is then defined as

$$\rho = \frac{\sum_k \left(f(\tilde{I}_k) - \mu_{f(\tilde{I})}\right)(SI_k - \mu_{SI})}{\sqrt{\sum_k \left(f(\tilde{I}_k) - \mu_{f(\tilde{I})}\right)^2 \sum_k (SI_k - \mu_{SI})^2}},$$

and the root mean-square prediction error σ is defined as

$$\sigma = \sqrt{\frac{1}{K}\sum_k \left(f(\tilde{I}_k) - SI_k\right)^2}.$$

Finally, cross-validated values $\bar{\rho}$ and $\bar{\sigma}$ are computed using n-fold ($n = 4$) cross-validation. Specifically, for each data set, the set is randomly divided into $n = 4$ equal size subsets, the free parameters a, b in the logistic function are fitted to the $n-1$ subsets, after which ρ and σ are computed based on the remaining subset. This procedure is repeated for each subset, and the averages, $\bar{\rho}$ and $\bar{\sigma}$, of the resulting ρ and σ values are computed.

[3] The x % SRT is defined as the SNR at which the average listener correctly identifies x percent of the test words.

Method Name	Remarks
STOI [15]	The short-time objective intelligibility measure (STOI).
CSII-MID [9]	The mid-level coherence SII.
CSII-BIF [28]	The coherence SII with signal-dependent band importance functions (named CSII_{mid}, $W_4, p = 1$ in [28]).
STI-NCM [12]	The normalized covariance speech transmission index.
NSEC [29]	The normalized subband envelope correlation method.

Table 2: Intelligibility predictors for comparison.

.

6.3. Comparison to Other Intelligibility Predictors

We compare the performance of the proposed intelligibility predictor, which will be abbreviated *SIMI* (Speech Intelligibility prediction based on Mutual Information), to several methods from the literature, see Table 2.

Table 3 summarizes the intelligibility prediction performance in terms of $\bar{\rho}$ and $\bar{\sigma}$. For additive noise all intelligibility predictors work well ($\bar{\rho} > 0.93$). For the ITFS processed signals, *SIMI* and *STOI* work well, resulting in $\bar{\rho} > 0.95$ and $\bar{\sigma} < 9.0$, while NSEC shows reasonable performance. The remaining methods fail in this situation. The results of Table 3 are in general agreement with the results reported in [30, 15].

Test	Add.Noise		ITFS Proc.	
	$\bar{\rho}$	$\bar{\sigma}$	$\bar{\rho}$	$\bar{\sigma}$
SIMI	0.975	8.95	0.957	8.49
STOI	0.969	9.45	0.966	8.20
CSII-MID	0.943	12.72	0.352	27.45
CSII-BIF	0.978	7.95	0.517	25.73
STI-NCM	0.934	13.41	0.613	20.56
NSEC	0.951	11.48	0.834	14.59

Table 3: Performance of intelligibility predictors in terms of cross-validated linear correlation coefficient $\bar{\rho}$, and root mean-square prediction error $\bar{\sigma}$.

7. Conclusion

We propose that intelligibility could be monotonically related to the mutual information between amplitude envelopes of the clean and the noisy/processed speech signal. We derive lower bounds on the mutual information which turn out to be simple functions of the linear correlation between these amplitude envelopes. Interestingly, the use of linear correlation coefficients is not a heuristically motivated choice, but follows as a consequence of the assumed auditory model (and the hypothesis that intelligibility is related to mutual information); this is in contrast to existing intelligibility predictors, e.g. STOI [15], where the use of linear correlation is less well motivated. Simulation experiments with the proposed method show that it is able to reliably estimate the average intelligibility of speech signals contaminated by stationary and non-stationary noise sources as well noisy speech processed with the ideal time-frequency segregation (ITFS) technique [26]. Future research includes evaluation of the proposed intelligibility predictor for other types of signal distortions.

8. References

[1] "ANSI S3.5, American National Standard Methods for the Calculation of the Articulation Index," American National Standards Institute, New York, 1969.

[2] 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.

[3] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, pp. 1689–1697, 1962.

[4] "IEC60268-16, Sound System Equipment – Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index," International Electrotechnical Commission, Geneva, 2003.

[5] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, pp. 355–367, 1971.

[6] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, pp. 318–326, 1980.

[7] "ANSI S3.5, Methods for the Calculation of the Speech Intelligibility Index," American National Standards Institute, New York, 1995.

[8] 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.

[9] 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.

[10] V. Hohmann and B. Kollmeier, "The effect of multichannel dynamic compression on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, pp. 1191–1195, 1995.

[11] R. Drullmann, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 585–592, January 1995.

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

[13] 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.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[15] ——, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio., Speech, Language Processing*, vol. 19, no. 7, pp. 2125–2136, September 2011.

[16] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[17] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2001, pp. 167–170.

[18] J. Taghia, R. Martin, and R. C. Hendriks, "On mutual information as a measure of speech intelligibility," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2012, pp. 65–68.

[19] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Philadelphia: SIAM, 2001.

[20] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech, Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[21] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley Series in Communications, 1991.

[22] W. Bialek, F. DeWeese, and D. Warland, "Bits and brains: Information flow in the nervous system," *Physica A*, vol. 200, no. 1-4, pp. 581–593, 1993.

[23] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall International, Inc., 1992.

[24] U. Kjems *et al.*, "Role of mask pattern in intelligibility of ideal binary-masked nosy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, September 2009.

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

[26] D. Brungart *et al.*, "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, December 2006.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An Evaluation of Objective Quality Measures for Speech Itelligibility Prediction," in *Proc. Interspeech*. Brighton, UK: ISCA, September 6-10 2009, pp. 1947–1950.

[28] 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.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[29] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proc. 17th European Signal Processing Conference (EUSIPCO)*, 2009, pp. 1849–1853.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *Journal of the Acoustical Society of America*, vol. 130, pp. 3013–3027, 2011.