AN INTELLIGIBILITY METRIC BASED ON A SIMPLE MODEL OF SPEECH COMMUNICATION

Steven Van Kuyk[1], W. Bastiaan Kleijn[1,2] and Richard C. Hendriks[2]

[1]Victoria University of Wellington, New Zealand
[2] Delft University of Technology, The Netherlands

ABSTRACT

Instrumental measures of speech intelligibility typically produce an index between 0 and 1 that is monotonically related to listening test scores. As such, these measures are dimensionless and do not represent physical quantities. In this paper, we propose a new instrumental intelligibility metric that describes speech intelligibility using bits per second. The proposed metric builds upon an existing intelligibility metric that was motivated by information theory. Our main contribution is that we use a statistical model of speech communication that accounts for noise inherent in the speech production process. Experiments show that the proposed metric performs at least as well as existing state-of-the-art intelligibility metrics.

Index Terms— Intelligibility, mutual information.

1. INTRODUCTION

When designing a speech-based communication system (e.g., a hearing aid or telecommunication device), it is important to understand how the system will affect speech intelligibility. Although listening tests provide the most reliable data, in many cases, instrumental measures of intelligibility are preferred as they provide a quicker and cheaper assessment.

Most instrumental intelligibility measures can be divided into two classes: those based on articulation index theory (e.g., the articulation index (AI) [1], speech intelligibility index (SII) [2], extended SII (eSII) [3], and coherence SII (CSII) [4]), and those based on spectral-temporal modulations (e.g., the speech transmission index (STI) [5], speech-based envelope power spectrum model (sEPSM) [6], short-time objective intelligibility measure (STOI) [7], and speech-based STI methods (sSTI) [8]). Additionally, there are some predictors that do not fall into either class (e.g., the glimpse proportion metric (GP) [9], and a measure based on the Dau auditory model (DAU) [10]). As a group, the above algorithms have been successful at predicting the intelligibility of speech subjected to speech-enhancement strategies such as spectral subtraction [11] and ideal time-frequency segregation (ITFS) [12] in a wide range of environments including additive noise, filtering, and reverberation. Though the algorithms are successful as a group, individually each algorithm tends to perform well for only a narrow subset of conditions. This is because each algorithm is heuristically motivated and designed for a specific purpose. Consequently, a unified intelligibility predictor is yet to emerge.

Recently, information theory (IT) has been proposed as a new paradigm for speech intelligibility prediction [13, 14, 15]. This is a natural approach to take given that the fundamental goal of speech communication is to transfer information from a talker to a listener.

There are two key advantages of IT. First, several studies have suggested that IT could provide a unified point of view. In [16] and [17] it was observed that the AI resembles the Shannon capacity of a memoryless Gaussian channel, and in [14] it was shown that STOI is related to the average amount of information shared between the temporal envelopes of clean and distorted speech signals. Second, IT provides a powerful theoretical framework. As an example, the concept of mutual information offers a generalized measure of dependency between two random variables that, unlike Pearson's correlation coefficient, can measure non-linear dependencies.

In the literature, two IT-based intelligibility metrics have been proposed [14, 15]. Both metrics are based on the hypothesis that intelligibility is monotonically related to the mutual information of the sub-band temporal envelope of a transmitted speech signal and the corresponding received speech signal. The main difference between the metrics is that [14] uses a lower bound on the information rate, whereas [15] measures mutual information using non-parametric techniques. Additionally, [14] estimates statistics over short-time segments, whereas [15] uses complete utterances.

In [17] a simple but effective model of speech communication was presented that includes variability inherent in the speech production process. It was argued that this variability, called 'production noise', causes the usefulness of a communication channel to saturate. In this paper, we extend the statistical model used in [14] to account for channel saturation. The outcome is a new intelligibility metric.

The remainder of this paper is organized as follows. In the following section we describe an information theoretical model of speech communication. Section 3 uses the model to develop an intelligibility predictor. Section 4 describes a procedure for measuring the statistics of production noise. Section 5 presents our evaluation, and finally, Section 6 concludes the work.

2. THEORY

In the following, we present a theoretical framework for describing speech communication. We adopt the model presented in [17], which considers the transmission of a message from a talker to a listener and assumes a noisy speech production process.

2.1. Model of the communication chain

The model describes a speech signal $\mathcal{X}$ as a multi-dimensional ergodic stationary discrete-time random process. The process is composed of real scalar random variables $X(j,t)$ where j is the dimension index and t is the time index. A common representation of speech used for instrumental intelligibility measures, including this work, is the sub-band temporal envelope. Let $x(i)$ be a real-valued random process that represents the samples of an acoustic speech signal where i is the sample index. The sub-band temporal envelope

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

of $x(i)$ is

$$X(j,t) = \sqrt{\sum_k |h_j(k)|^2 |\tilde{x}(k,t)|^2}, \quad (1)$$

where j is the sub-band frequency index, t is the frame index, $h_j(k)$ is the transfer function of the jth filter in an auditory filterbank, and $\tilde{x}(k,t)$ is the Short-Time Fourier Transform (STFT) of $x(i)$ where k is the frequency bin index.

The transmission of speech from talker to listener is modelled by a Markov chain:

$$\mathcal{M} \rightarrow \mathcal{X} \rightarrow \mathcal{Y}, \quad (2)$$

where $\mathcal{X}$ is the clean speech produced by a talker and $\mathcal{Y}$ is the signal received by the listener. The model distinguishes between a hypothetical message $\mathcal{M}$ and a produced speech signal $\mathcal{X}$ to account for 'production noise' (e.g., inter-talker and intra-talker variability).

2.2. Information rate of the communication channel

The mutual information rate $I(\mathcal{M};\mathcal{Y})$ between $\mathcal{M}$ and $\mathcal{Y}$ describes the effectiveness of the communication channel. As a first approximation we assume that the time-frequency (TF) units that compose $\mathcal{M}$ are statistically independent, and likewise for $\mathcal{X}$ and $\mathcal{Y}$. For this case, the mutual information rate decomposes into a summation of mutual information terms

$$I(\mathcal{M};\mathcal{Y}) = \frac{1}{T}\sum_t \sum_j I(M(j,t);Y(j,t)), \quad (3)$$

where T is the sequence length. To enhance readability we drop the time index and frequency index where possible.

By exploiting Markov chain properties, for each TF unit we have the upper bound:

$$I(M;Y) \leq \min(I(M;X), I(X;Y)). \quad (4)$$

This equation shows that the communication model can be decomposed into two channels: the speech production channel, $\mathcal{M} \rightarrow \mathcal{X}$, and the environmental channel $\mathcal{X} \rightarrow \mathcal{Y}$. In the case of a distortionless environmental channel, $I(X;Y)$ is infinite and $I(M;Y)$ saturates at the information rate of the speech production channel.

2.2.1. Information rate of M and X

Let us consider the nature of speech production. It is obvious that multiple speech signals can be produced for a given linguistic message. This means that variability is inherent to the speech production channel. We name this variability 'production noise'.

The two main sources of variability can be attributed to learned speech habits (i.e., accents), and physiological differences between vocal-tracts [18]. Consequently, it is likely that the variability between speech signals manifests as differences in the vocal-tract characteristics of different talkers. The effect of different vocal tracts can be seen by modelling speech production as the convolution of a vocal-tract filter impulse response and an excitation signal [19]. In the time-frequency domain the convolution becomes a multiplication and we can write $\tilde{x}(k,t) = v(k,t)g(k,t)$, where $v(k,t)$ is the time-varying vocal-tract filter transfer function and $g(k,t)$ is the excitation. Under this speech production model, it is natural to assume that production noise has a multiplicative nature. Thus we define

$$P \triangleq \log(X) - \log(M), \quad (5)$$

where P is production noise, such that $X = e^P M$. Furthermore, we assume that P is zero mean and that P and M are statistically independent. The logarithm in (5) is applied so that P has an additive nature consistent with the model in [17]. Since the logarithm is an invertible function, we can transform the signals without affecting the information rate. That is,

$$I(M;X) = I(\log(M);\log(X)). \quad (6)$$

Our initial experiments have shown that $\log(X)$ and P are approximately Gaussian. Thus, the mutual information is given by [20]

$$I(\log(M);\log(X)) = -\frac{1}{2}\log(1-\rho_p^2), \quad (7)$$

where ρ_p is the correlation coefficient between $\log(M)$ and $\log(X)$. We call ρ_p the 'speech production correlation coefficient'. Using (5) it is easy to show that

$$\rho_p^2 = \frac{\mathrm{E}[(\log X)^2] - \mathrm{E}[\log X]^2 - \mathrm{E}[P^2]}{\mathrm{E}[(\log X)^2] - \mathrm{E}[\log(X)]^2}. \quad (8)$$

We assume that the speech production correlation coefficient is an inherent property of speech communication and does not depend on a specific realization of $\mathcal{X}$. We also assume that ρ_p does not depend on the time index, but may depend on the frequency index. In Section 4 we estimate $\rho_p(j)$ using an appropriate speech corpus.

2.2.2. Information rate of X and Y

In [14] an algorithm was developed for estimating a lower bound on $I(X;Y)$. It was argued that X follows a chi distribution with d degrees of freedom. The resulting lower bound is given by:

$$\begin{aligned} I(X;Y) \leq{}& \log\Gamma(d/2) + \frac{1}{2}\Big(d - \log 2 - (d-1)\psi(d/2)\Big) \\ &- \frac{1}{2}\log 2\pi e(d - 2\frac{\Gamma^2((d+1)/2)}{\Gamma^2(d/2)}) \\ &- \frac{1}{2}\log(1-\rho_{XY}^2) \\ \triangleq{}& I_{\mathrm{low}}(X;Y), \end{aligned} \quad (9)$$

where $\Gamma(.)$ and $\psi(.)$ denote the gamma and the digamma function, respectively, and ρ_{XY} is the correlation coefficient between X and Y given by

$$\rho_{XY} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\sqrt{(\mathrm{E}[X^2] - \mathrm{E}[X]^2)(\mathrm{E}[Y^2] - \mathrm{E}[Y]^2)}}. \quad (10)$$

3. PROPOSED INTELLIGIBILITY METRIC

In this section we use the theory from Section 2 to develop a speech intelligibility metric. The metric is a function of a clean acoustic speech signal $x(i)$ that is produced by a talker, and a distorted acoustic speech signal $y(i)$ that is received by a listener. The output is an upper bound on the amount of information shared between the talker and listener in bits per second. The basic structure of the proposed intelligibility metric shows strong similarities to [7, 14, 15, 21].

3.1. Implementation

First we derive an internal representation based on a simplified model of the auditory system. To this end, $x(i)$ and $y(i)$ are resampled to a sampling frequency of 10 kHz. Subsequently the signals are transformed to the STFT domain using a 512-point Hann window with 50% overlap. This results in a frame rate of $R \approx 39$

frames/second, which is sufficient for capturing the spectral modulations necessary for speech intelligibility [22, 23]. The sub-band temporal envelopes of $x(i)$ and $y(i)$ are then calculated according to (1). We use an auditory filterbank that consists of $J = 25$ gammatone filters with center frequencies linearly spaced between 100 Hz and 4500 Hz on the ERB-rate scale [24].

The proposed intelligibility metric is computed by evaluating

$$\mathrm{I} = \frac{R}{T} \sum_{t} \sum_{j} \min(\hat{I}_{MX}(j), I_{\mathrm{low}}(X(j,t); Y(j,t))), \qquad (11)$$

where $\hat{I}_{MX}(j)$ is an estimate of the information rate of the speech production channel, which does not depend on a specific realization of $X(j,t)$ or $Y(j,t)$, and I is in the units of bits per second.

The moments in (10) needed to compute $I_{\mathrm{low}}(X(j,t); Y(j,t))$ are estimated using a causal moving-average filter. For example, $\mathrm{E}[X(j,t)]$ is estimated according to

$$\hat{\mu}_X(j,t) = \frac{1}{\alpha} \sum_{\tau=t-\alpha+1}^{t} X(j,\tau). \qquad (12)$$

We use $\alpha = 30$ which corresponds to an analysis window of 768 ms. Although a larger value of α could be used to reduce the variance and bias of $\hat{\mu}_X(j,t)$, $\alpha > 30$ would also limit the ability of the algorithm to account for non-stationary distortions.

Lastly, an energy-based voice activity detection algorithm with a 40 dB threshold is applied to locate silent frames in $x(i)$ and $y(i)$. For the silent frames, we set $I_{\mathrm{low}}(X(j,t); Y(j,t)) = 0$ before applying (11). This is reasonable because no information is transferred when either $x(i)$ or $y(i)$ is silent.

4. PRODUCTION NOISE ESTIMATION

By definition it is impossible to generate a production noise signal separately from a speech signal. This means that production noise cannot be observed directly. However, it is possible to estimate production noise using an ensemble of time-aligned speech signals.

Consider an ensemble of N acoustic speech signals where each speech signal is produced by a different talker and where each signal is composed of the same sequence of speech sounds where the duration of each speech sound is the same for each talker. This means that each speech signal contains the same linguistic information at each unit of time. Consequently, $\mathcal{M}$ is the same for all signals in the ensemble. Production noise can then be estimated by considering the variability of each TF unit over the ensemble.

More concretely, (1) is applied to all signals in the ensemble to recover their internal representations. We denote the resulting data $X_n(j,t)$, where n is the talker index. Using (5) and the fact that P is zero mean, the production noise for each TF unit of each talker is estimated according to

$$\hat{P}_n(j,t) = \log X_n(j,t) - \frac{1}{N-1} \sum_{l,\, l \neq n} \log X_l(j,t). \qquad (13)$$

The idea behind (13) is that because $\mathcal{M}$ is the same for each speech signal, $\log M(j,t)$ can be estimated by taking the expectation over the ensemble. The expectation can then be subtracted from $\log X_n(j,t)$ to obtain $P_n(j,t)$. In practice, the sample mean is used as an estimator of the expectation. Removing the n'th observation from the sample mean results in an unbiased estimator of $P_n(j,t)$.

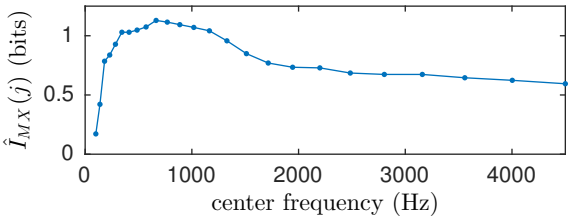

Fig. 1. The information rate of the speech production channel plotted against the center frequencies of the J sub-bands.

Next, for each j we define the following $1 \times TN$ vectors:

$$\begin{aligned} \tilde{X}_j(t') = [&\log X_1(j,1), \ldots, \log X_1(j,T), \ldots, \\ &\log X_N(j,1), \ldots, \log X_N(j,T)] \end{aligned} \qquad (14)$$

and

$$\tilde{P}_j(t') = [\hat{P}_1(j,1), \ldots, \hat{P}_1(j,T), \ldots, \hat{P}_N(j,1), \ldots, \hat{P}_N(j,T)], \qquad (15)$$

which are obtained by stacking the TF units from each talker. The moments in (8) are then estimated according to

$$\begin{aligned} \hat{\sigma}^2_{\log X}(j) &= \frac{1}{TN-1} \sum_{t'} \left(\tilde{X}_j(t') - \frac{1}{TN} \sum_{s} \tilde{X}_j(s) \right)^2 \\ \hat{\sigma}^2_{P}(j) &= \frac{N-1}{N} \frac{1}{TN} \sum_{t'} \left(\tilde{P}_j(t') \right)^2 \end{aligned} \qquad (16)$$

where $\sigma^2_{\log X} = \mathrm{E}[(\log X)^2] - \mathrm{E}[(\log X)]^2$, $\sigma^2_P = \mathrm{E}[P^2]$, and $\hat{}$ denotes their estimates. Note that the $(N-1)/N$ factor in the second equation is a bias reduction factor. The bias exists because σ^2_P is estimated using $\hat{P}_n(j,t)$ rather than the true production noise $P_n(j,t)$.

For our experiment, we used data from the CHAINS speech corpus [25]. This corpus includes easy reading material spoken by $N = 36$ talkers consisting of 18 females, and 18 males. A dynamic time warping algorithm [26] was applied to all signals to ensure that $\mathcal{M}$ was constant over the ensemble. Figure 1 plots our measurement of $\hat{I}_{MX}(j) = -\frac{1}{2}\log(1 - \rho_p(j)^2)$, which was obtained using (8) and (16). Note that $\hat{I}_{MX}(j)$ can be interpreted as a band-importance function [17].

5. EVALUATION

We now present our evaluation of the proposed intelligibility metric. This includes a description of two listening tests and the corresponding results.

5.1. Experimental procedures

The performance of the proposed intelligibility metric (I) was evaluated by considering two listening experiments. The first experiment involved ITFS processed speech (ITFSS), and the second involved speech corrupted by additive noise (ANS).

Three competing intelligibility predictors were also evaluated as reference: STOI [7], the lower-bound speech intelligibility mutual information metric (SIMI) [14], and the k-nearest neighbour mutual information metric (MI-KNN) [15]. For all intelligibility metrics, the implementation was provided by the original authors with the parameters described in the accompanying papers.

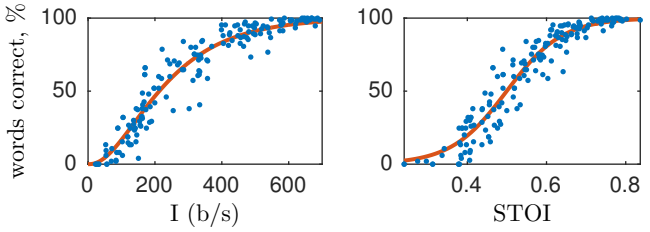

Fig. 2. Scatter plots between listening test scores for ITFSS against the proposed measurement of information (left) and STOI (right).

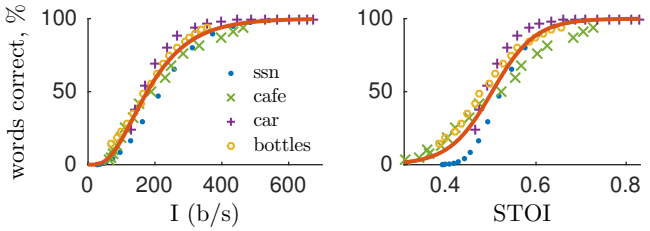

Fig. 3. Scatter plots between listening test scores for ANS against the proposed measurement of information (left) and STOI (right).

5.1.1. Listening test data

The intelligibility data for the ITFSS experiment was obtained from a listening test conducted by Kjems *et al.* [27]. The speech signals were selected from the Dantale II corpus [28] and degraded by four types of additive noise: speech-shaped noise, cafeteria noise, noise from a bottling factory hall, and car interior noise, at three different speech-to-noise ratios (SNR): 20% and 50% of the speech reception threshold and -60 dB. Two types of binary mask were applied to the degraded signals: an ideal binary mask, and a target binary mask. Fifteen normal-hearing listeners participated in the listening test. For a full description of the listening test conditions and the binary mask parameters, see [27].

The intelligibility data for the ANS experiment was also obtained from [27]. Using the same speech signals and noise sources as the ITFSS experiment, we generated noisy speech signals with SNRs ranging from -20 dB to 0 dB in steps of 1.25 dB. The corresponding listening test intelligibility scores were obtained by sampling the psychometric functions derived in [27] for each noise source at the appropriate SNR values. See [14], which used the same technique.

5.1.2. Performance measures

The performance of the intelligibility metrics was evaluated using two figures of merit: the correlation coefficient, ρ, and Kendall's tau coefficient [29], τ. To use ρ effectively, the relationship between instrumental intelligibility scores and listening test scores needs to be linear (see [7]). For the proposed intelligibility metric the relationship was linearized using the function [30]

$$s = 100\big(1 - 10^{-a\mathrm{I}}\big)^b, \tag{17}$$

where s is the predicted listening test score (percentage of words correct), and a and b are free parameters. The free parameters are functions of the speech corpus and reflect the difficulty of the speech material. A non-linear least squares procedure was applied to estimate a and b from the data. For the other intelligibility metrics, denoted d, the mapping function of the original authors was used:

$$s = \frac{100}{1 + e^{ad+b}}. \tag{18}$$

As pointed out in [15], linearizing the relationship can decrease the transparency of the results. Thus, we also compute τ, which does not require a linear relationship.

Table 1. The performance of the proposed intelligibility metric (I) and three reference metrics. Left: ITFSS. Right: ANS.

	I	MI-KNN	SIMI	STOI	I	MI-KNN	SIMI	STOI
ρ	0.96	0.88	0.95	0.96	0.99	0.87	0.97	0.97
τ	0.82	0.72	0.82	0.83	0.91	0.70	0.86	0.83

5.2. Results

Figure 2 and Figure 3 plot the listening test scores against the proposed intelligibility metric and STOI for each experiment. The mapping functions (17) and (18) are also included. The plots show that there is a strong relationship between listening test scores and the proposed intelligibility metric. Observe that with the ANS experiment, there are four distinguishable relationships for STOI. This shows that STOI scores are sensitive to the noise source. This is also true for the proposed metric, but to a lesser degree.

The performance results of the proposed intelligibility metric and the reference metrics are summarized in Table 1. We see that the proposed metric has good performance in all categories.

Note that like the proposed metric, SIMI also includes a saturation threshold. However, the threshold in SIMI was heuristically motivated whereas in our work channel saturation is a natural consequence of production noise. A further difference is that the threshold in SIMI is the same for each sub-band channel, but saturation for the proposed metric depends on the sub-band j.

6. DISCUSSION & CONCLUSION

We have developed a new intelligibility metric motivated by a simple model of speech communication. The model consists of a speech production channel and an environmental channel. We measured the saturation level of the speech production channel and used this knowledge to estimate the amount of information shared between a talker and a listener. Our evaluation showed that the proposed measure of information has a strong monotonic relationship with listening test scores.

A limitation of this work is the assumption that the frequency sub-bands are statistically independent. In practice we have found that this is not the case. Our preliminary experiments suggest that accounting for frequency dependencies could reduce the given information rate by as much as a factor of 4. This would give a maximum information rate of approximately 150 b/s, which is closer to the linguistic information rate of speech than previous measures (see [31]).

In Section 2.2.1 we assumed that $\log(X)$ follows a Gaussian distribution and in Section 2.2.2 we assumed that X follows a chi distribution. We adopted this parametric approach for the sake of mathematical tractability. In practice, neither distribution fits the data perfectly, but this does not pose a problem when the performance of the proposed intelligibility metric is considered.

7. REFERENCES

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 17, no. 1, pp. 103–103, 1945.

[2] "Methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, vol. 19, pp. 90–119, 1997.

[3] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 3988–3997, 2006.

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

[5] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[6] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Amer.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[7] 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, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

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

[9] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[10] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7, pp. 678–692, 2010.

[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE/ Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

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

[13] W. B. Kleijn, J. B Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 43–54, 2015.

[14] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, 2014.

[15] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, 2014.

[16] J. Allen, "The articulation index is a Shannon channel capacity," in *Auditory Signal Processing*, pp. 313–319. Springer, 2005.

[17] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, 2015.

[18] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Proc. Interspeech*, 2001, pp. 1377–1380.

[19] J. L. Flanagan, *Speech analysis synthesis and perception*, vol. 3, Springer Science & Business Media, 2013.

[20] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[21] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016.

[22] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[23] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, pp. e1000302, 2009.

[24] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep.*, vol. 35, pp. 8, 1993.

[25] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in *Proc. of SPECOM*, 2006, vol. 6, pp. 431–435.

[26] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[27] 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. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[28] R. Ardenkjœr, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiol.*, vol. 42, pp. 10–17, 2003.

[29] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[30] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[31] R. M. Fano, "The information theory point of view in speech communication," *J. Acoust. Soc. Amer.*, vol. 22, no. 6, pp. 691–696, 1950.