An Instrumental Intelligibility Metric Based on Information Theory

Van Kuyk, Steven; Kleijn, W. Bastiaan; Hendriks, Richard C.

DOI
10.1109/LSP.2017.2774250

Publication date
2018

Document Version
Final published version

Published in
IEEE Signal Processing Letters

An Instrumental Intelligibility Metric Based on Information Theory

Steven Van Kuyk, *Student Member, IEEE*, W. Bastiaan Kleijn, *Fellow, IEEE*, and Richard C. Hendriks, *Member, IEEE*

***Abstract*—We propose a monaural intrusive instrumental intelligibility metric called speech intelligibility in bits (SIIB). SIIB is an estimate of the amount of information shared between a talker and a listener in bits per second. Unlike existing information theoretic intelligibility metrics, SIIB accounts for talker variability and statistical dependencies between time-frequency units. Our evaluation shows that relative to state-of-the-art intelligibility metrics, SIIB is highly correlated with the intelligibility of speech that has been degraded by noise and processed by speech enhancement algorithms.**

***Index Terms*—Intelligibility, mutual information.**

I. Introduction

INTELLIGIBILITY is defined as the proportion of words correctly identified by a listener and is a natural measure for quantifying the effectiveness of speech-based communication systems [1]. Although listening tests can provide valid data, such tests are time-consuming to conduct. For this reason, instrumental intelligibility metrics that are correlated with intelligibility and quick to compute are often preferred.

We can distinguish two types of instrumental intelligibility metrics: intrusive and nonintrusive. Intrusive intelligibility metrics require knowledge of the clean speech and either the communication channel or degraded speech, whereas nonintrusive intelligibility metrics require only the degraded speech. In this letter, we develop a new intrusive intelligibility metric based on information theory [2].

Existing intrusive intelligibility metrics include the speech intelligibility index (SII) [3], the speech transmission index [4], the coherence SII (CSII) [5], the extended SII [6], the normalized covariance measure [7], [8], the hearing-aid speech perception index [9], the short-time objective intelligibility measure (STOI) [10], the extended STOI (ESTOI) [11], the speech-based envelope power spectrum model [12]–[14], and the glimpse proportion metric (GP) [15]–[17]. As a group, the above-mentioned algorithms have been successful at predicting speech intelligibility in a wide range of conditions including additive noise, filtering, reverberation, and nonlinear enhancement. However, each intelligibility metric tends to perform well for only a narrow subset of conditions. This is because the above-mentioned algorithms were heuristically motivated and were often designed with a specific type of distortion or dataset in mind.

Information theory provides a mathematical framework for modeling communication systems. Information theoretical concepts have previously been used in the analysis of linguistics [18], [19], speech production [20], and human hearing [21]. Additionally, state-of-the-art speech enhancement algorithms [20], [22] and intelligibility metrics [23]–[25] that are based on information theory have been developed.

Existing information theoretic intelligibility metrics, such as the mutual information k-nearest neighbor metric (MIKNN) [23], assume that speech can be described by a memoryless stochastic process and that the energy of a speech signal at one time-frequency location is statistically independent to the energy at all other time-frequency locations. In reality neither of these assumptions are valid, which leads to an overestimate of the information shared between a talker and a listener.

In this letter, we propose a conceptually simple intelligibility metric called SIIB. SIIB is a function of a clean acoustic signal produced by a talker and a degraded signal that is received by a listener. As described in Sections II and III, the acoustic signals are converted to a representation of speech based on a crude model of the human auditory system. A nonparametric estimate of the mutual information rate of the signals is then computed. Unlike existing metrics, SIIB partially accounts for time-frequency dependencies in the speech signals using the Karhunen–Loève transform (KLT) [26] and incorporates the theory developed in [20] to account for the effect that talker variability has on the information rate. In Sections IV and V, SIIB is evaluated by comparing its performance to STOI [10], ESTOI [11], and MIKNN [23] for speech degraded by noise and processed by enhancement algorithms.

II. Model of Speech Communication

In this section, we present a theoretical model of speech communication similar to that described in [20], [25], and [27]. The model considers the transmission of a message from a talker to a listener. Stochastic processes are denoted by $\{\cdot\}$, random variables are denoted by bold font, and their realizations are denoted by regular font.

Manuscript received July 23, 2017; revised October 13, 2017; accepted November 5, 2017. Date of publication November 16, 2017; date of current version December 8, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James E. Fowler. *(Corresponding author: Steven Van Kuyk.)*

S. Van Kuyk is with the Victoria University of Wellington, Wellington 6012, New Zealand (e-mail: steven.van.kuyk@ecs.vuw.ac.nz).

W. B. Kleijn is with the Victoria University of Wellington, Wellington 6012, New Zealand, and also with the Delft University of Technology, Delft 2628 CD, The Netherlands (e-mail: bastiaan.kleijn@ecs.vuw.ac.nz).

R. C. Hendriks is with the Delft University of Technology, Delft 2628 CD, The Netherlands (e-mail: r.c.hendriks@tudelft.nl).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2774250

A. Communication Channel

A message $\{\mathbf{M}_t\}$, speech signal $\{\mathbf{X}_t\}$, and degraded speech signal $\{\mathbf{Y}_t\}$ are represented by ergodic stationary discrete-time vector-valued random processes, where $t \in \mathbb{Z}$ is the time index. The message can be thought of as a sequence of latent variables that represent, for example, a sequence of sentences, phonemes, or neural states. The talker encodes the message into a speech signal according to a conditional probability distribution $p_{\{\mathbf{X}_t\}|\{\mathbf{M}_t\}}(\{X_t\}|\{M_t\})$. In this way, the variability of different talkers encoding the same message into a speech signal is incorporated into the model.

The speech signal is transmitted to a listener through a communication channel that may distort the signal. Examples of distortion include noise, reverberation, speech coding algorithms, and speech enhancement algorithms. Overall, the communication process is described by a Markov chain:

$$\{\mathbf{M}_t\} \rightarrow \{\mathbf{X}_t\} \rightarrow \{\mathbf{Y}_t\}. \tag{1}$$

We call $\{\mathbf{M}_t\} \rightarrow \{\mathbf{X}_t\}$ the speech production channel and $\{\mathbf{X}_t\} \rightarrow \{\mathbf{Y}_t\}$ the environmental channel.

The representation of speech used in this letter is based on a crude model of the human auditory system and was motivated using information theoretic arguments in [21] and [27]. Let $\{\mathbf{x}_i\}$ be a real-valued random process that represents the samples of an acoustic speech signal, where i is the sample index, and let $\{\hat{\mathbf{x}}_t\}$ be the short-time Fourier transform (STFT) of $\{\mathbf{x}_i\}$, where t is the frame index. We define $\mathbf{X}_t$ as an $\mathbb{R}^J$-valued random variable that represents auditory log-spectra given by

$$\mathbf{X}_t = \ln G|(\hat{\mathbf{x}}_t)|^2 \tag{2}$$

where $G \in \mathbb{R}^{J\times N}$ is a matrix that represents an auditory filterbank, and the logarithm and squared magnitude operators are applied element wise. To account for temporal masking in the auditory system, the masking function described in [28] is applied to $\mathbf{X}_t$. The degraded speech $\mathbf{Y}_t$ is defined similarly.

B. Information Rate of the Communication Channel

The proposed intelligibility metric is based on the hypothesis that intelligibility is a function of the mutual information rate between the message and the degraded speech. Let $\mathbf{M}^K = [(\mathbf{M}_1)^T, (\mathbf{M}_2)^T, \ldots, (\mathbf{M}_K)^T]^T$, where T denotes the transpose, be a vector obtained by stacking K consecutive message vectors and similarly for $\mathbf{Y}^K$. The mutual information rate is defined by

$$I(\{\mathbf{M}_t\};\{\mathbf{Y}_t\}) = \lim_{K\rightarrow\infty} \frac{1}{K} I(\mathbf{M}^K;\mathbf{Y}^K) \tag{3}$$

where $I(\mathbf{M}^K;\mathbf{Y}^K)$ is the mutual information between $\mathbf{M}^K$ and $\mathbf{Y}^K$ given by

$$I(\mathbf{M}^K;\mathbf{Y}^K) = \int_{\mathbf{M}^K,\mathbf{Y}^K} p(M^K,Y^K)\log_2 \frac{p(M^K,Y^K)}{p(M^K)p(Y^K)} dM^K dY^K. \tag{4}$$

To estimate (3), realizations of $\mathbf{M}_t$ and $\mathbf{Y}_t$ are needed. Estimating a realization of $\mathbf{M}_t$ requires a chorus of speech signals (see [27]). In typical applications of intelligibility prediction, such a chorus is not available; so, instead we use an upper bound on (3). By applying the data processing inequality twice, we have [29]

$$I(\{\mathbf{M}_t\};\{\mathbf{Y}_t\}) \leq \min\big(I(\{\mathbf{M}_t\};\{\mathbf{X}_t\}), I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\})\big). \tag{5}$$

In the case of a distortionless environmental channel, $I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\})$ is unbounded from above, and $I(\{\mathbf{M}_t\};\{\mathbf{Y}_t\})$ saturates at the information rate of the speech production channel [20]. This maximum information rate is determined by the variability in pronunciation between different talkers. The following sections describe how $I(\{\mathbf{M}_t\};\{\mathbf{X}_t\})$ and $I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\})$ can be calculated.

C. Information Rate of the Environmental Channel

The mutual information rate of the environmental channel is given by

$$I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\}) = \lim_{K\rightarrow\infty} \frac{1}{K} I(\mathbf{X}^K;\mathbf{Y}^K). \tag{6}$$

Estimating the mutual information between vectors of high dimensionality is a challenging task [30], particularly when the vector elements have strong statistical dependencies [31]. For this reason, we introduce an invertible transform $f(\cdot)$ that aims to remove the dependencies between the vector elements.

Let $\tilde{\mathbf{X}}^K = f(\mathbf{X}^K)$ and $\tilde{\mathbf{Y}}^K = f(\mathbf{Y}^K)$. In the following, we assume that the elements of $\tilde{\mathbf{X}}^K$ can be approximated as statistically independent, and likewise for $\tilde{\mathbf{Y}}^K$. Then (6) can be decomposed into a summation:

$$\begin{aligned} I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\}) &= \lim_{K\rightarrow\infty} \frac{1}{K} I(\mathbf{X}^K;\mathbf{Y}^K) \\ &= \lim_{K\rightarrow\infty} \frac{1}{K} I(\tilde{\mathbf{X}}^K;\tilde{\mathbf{Y}}^K) \\ &= \lim_{K\rightarrow\infty} \frac{1}{K} \sum_{j=1}^{KJ} I(\tilde{\mathbf{X}}_j^K;\tilde{\mathbf{Y}}_j^K) \end{aligned} \tag{7}$$

where j denotes the element index in the vector.

Finding an invertible $f(\cdot)$ that simultaneously removes the dependencies in both $\mathbf{X}^K$ and $\mathbf{Y}^K$ is difficult. Early speech recognition systems used the discrete cosine transform (DCT), which results in Mel-frequency cepstral coefficients [32]. It can be shown that the DCT approximates the KLT for stationary signals [33]. The KLT is the transformation that we use here and it is given by

$$\tilde{\mathbf{X}}^K = U(\mathbf{X}^K - \mathrm{E}[\mathbf{X}^K]) \tag{8}$$

and

$$\tilde{\mathbf{Y}}^K = U(\mathbf{Y}^K - \mathrm{E}[\mathbf{Y}^K]) \tag{9}$$

where U is a matrix with rows equal to the unit-magnitude eigenvectors of the covariance matrix of $\mathbf{X}^K$, and $\mathrm{E}[\cdot]$ is the expected value operator. The KLT ensures that the elements of $\tilde{\mathbf{X}}^K$ are statistically uncorrelated, and if $\mathbf{X}^K$ is Gaussian, which is a reasonable approximation, then the elements are also statistically independent.

The KLT does not guarantee the same properties for $\tilde{\mathbf{Y}}^K$ unless $\mathbf{Y}^K$ is also Gaussian and has a covariance matrix equal to a scalar multiple of the covariance matrix of $\mathbf{X}^K$. In practice the environmental channel can result in non-Gaussian $\mathbf{Y}^K$ or can

introduce statistical dependencies in $\mathbf{Y}^K$ that are not present in $\mathbf{X}^K$. An example of the latter is a reverberant channel. In this case, the statistical dependencies in the source are accounted for by the KLT, but the statistical dependencies in the received signal are not accounted for. The consequence is that (7) underestimates the mutual information rate. Although the KLT does not meet all of the requirements for $f(\cdot)$, we found that it improves performance.

D. Information Rate of the Speech Production Channel

Approximating $\{\mathbf{M}_t\}$ and $\{\mathbf{X}_t\}$ as Gaussian, the information rate of the speech production channel is

$$I(\{\mathbf{M}_t\};\{\mathbf{X}_t\}) = \lim_{K\to\infty} \frac{1}{K}\sum_{j=1}^{KJ} I(\tilde{\mathbf{M}}_j^K;\tilde{\mathbf{X}}_j^K)$$

$$= \lim_{K\to\infty} -\frac{1}{K}\sum_{j=1}^{KJ} \frac{1}{2}\log_2(1-r_j^2) \qquad (10)$$

where $\tilde{\mathbf{M}}^K$ is defined similarly to $\tilde{\mathbf{X}}^K$, and r_j is called the production noise correlation coefficient. The production noise correlation coefficient describes the efficiency of encoding a message into a speech signal according to $p_{\{\mathbf{X}_t\}|\{\mathbf{M}_t\}}(\{X_t\}|\{M_t\})$. Based on the measurements in [25] and [27], this letter uses $r_j = 0.75$ for all j.

III. Proposed Intelligibility Metric

The proposed intelligibility metric combines (7), (10), and (5) to give an estimate of the amount of information shared between $\{\mathbf{M}_t\}$ and $\{\mathbf{Y}_t\}$ in bits per second. It is given by

$$\text{SIIB} = \frac{F}{K}\sum_{j=1}^{KJ} \min\left(-\frac{1}{2}\log_2(1-r_j^2),\ I(\tilde{\mathbf{X}}_j^K;\tilde{\mathbf{Y}}_j^K)\right) \qquad (11)$$

where F is the frame rate in Hz.

We now describe our implementation. An estimate of $I(\tilde{\mathbf{X}}_j^K;\tilde{\mathbf{Y}}_j^K)$ is computed by applying a k-nearest neighbor mutual information estimator [34] to observed sample sequences $\tilde{X}_{j,t}^K$ and $\tilde{Y}_{j,t}^K$. To obtain $\tilde{X}_{j,t}^K$ and $\tilde{Y}_{j,t}^K$, a clean acoustic speech signal and a degraded signal are resampled to a sampling rate of 16 kHz. An energy-based voice activity detector with a 40-dB threshold is applied to remove silent segments. Subsequently, the signals are transformed to the STFT domain using a 400-point Hann window with 50% overlap. This gives a frame rate of $F = 80$ Hz, which is sufficient for capturing the spectral modulations required for high intelligibility [35].

A gammatone filterbank [36] that includes $J = 28$ filters linearly spaced on the ERB-rate scale [37] between 100 and 6500 Hz is used to obtain X_t and Y_t according to (2). A sequence of stacked vectors for the clean speech is then formed by stacking $K = 15$ consecutive vectors:

$$X_t^K = [(X_{t-K+1})^T, (X_{t-K+2})^T, \ldots, (X_t)^T]^T \qquad (12)$$

and similarly for Y_t^K. Setting $K = 15$ means that dependencies spanning 187.5 ms are considered. For comparison, the mean duration of a phoneme is 80 ms [38]. The sample covariance matrix of X_t^K is computed and the KLT in (8) and (9) is applied to obtain $\tilde{X}_t^K$ and $\tilde{Y}_t^K$.

IV. Evaluation Procedures

This section describes the procedures used to evaluate SIIB. The evaluation considered four intelligibility datasets and used two performance measures to quantify the strength of the relationship between SIIB and intelligibility.

A. Intelligibility Datasets

1) JensenSCNR: The first dataset consists of speech subjected to single channel noise reduction. In [39], phrases from the Dutch version of the Hagerman test [40], [41] were degraded by speech-shaped noise (SSN) at SNRs of $-8, -6, -4, -2$, and 0 dB and processed by three noise reduction algorithms. The three algorithms compute a minimum mean-squared error estimate of the clean speech by multiplying the short-time spectral magnitude of the degraded speech with a gain function. In total there are 5 SNRs $\times$ (3 algorithms + 1 unprocessed) $= 20$ conditions. The stimuli were presented to 13 normal-hearing subjects for identification.

2) KleijnPRE: The second dataset consists of speech subjected to preprocessing enhancement and degraded by noise. In [20], phrases from the Dutch version of the Hagerman test were subjected to three preprocessing enhancement algorithms and then degraded either by SSN at SNRs of $-15, -12, -9$, and -6 dB, or car noise at SNRs of $-23, -20, -17$, and -14 dB. The three enhancement algorithms optimally redistribute the energy of the clean speech according to a distortion criterion. In total there are 2 noise types $\times$ 4 SNRs $\times$ (3 algorithms $+$ 1 unprocessed) $= 32$ conditions. The stimuli were presented to nine normal-hearing listeners for identification.

3) CookePRE: The third dataset also consists of speech subjected to preprocessing enhancement. In [42], Harvard sentences [43] were processed by 19 preprocessing enhancement algorithms and degraded either by SSN at SNRs of $1, -4$, and -9 dB or by speech from a competing talker at SNRs of $-7, -14$, and -21 dB. The stimuli were presented to 175 normal-hearing listeners for identification. For this letter, a subset of the data in [42] was considered because the entire dataset was not available. Ten of the Harvard sentences and nine of the enhancement algorithms were used. The algorithms are referred to in [42] as AdaptDRC, F0-shift, IWFEMD, on/offset, OptimalSII, RESSYSMOD, SBM, SEO, and SSS. In total there are 2 noise types $\times$ 3 SNRs $\times$ (9 algorithms $+$ 1 unprocessed) $= 60$ conditions.

4) KjemsITFS: The fourth dataset consists of speech subjected to ideal time-frequency segregation processing (ITFS). In [44], phrases from the Dantale II corpus [45] were degraded by four types of noise: SSN, cafeteria noise, noise from a bottling factory, and car noise. For each noise type, the degraded signals were processed by two types of ITFS called an ideal binary mask and a target binary mask. Three SNRs were used (-60 dB, and SNRs corresponding to 20% and 50% intelligibility) and eight variants of each ITFS algorithm were considered. In total there are 168 conditions. The stimuli were presented to 15 normal-hearing subjects for identification.

B. Performance Measures

The most important characteristic of an intelligibility metric is that it has a strong monotonic increasing relationship with

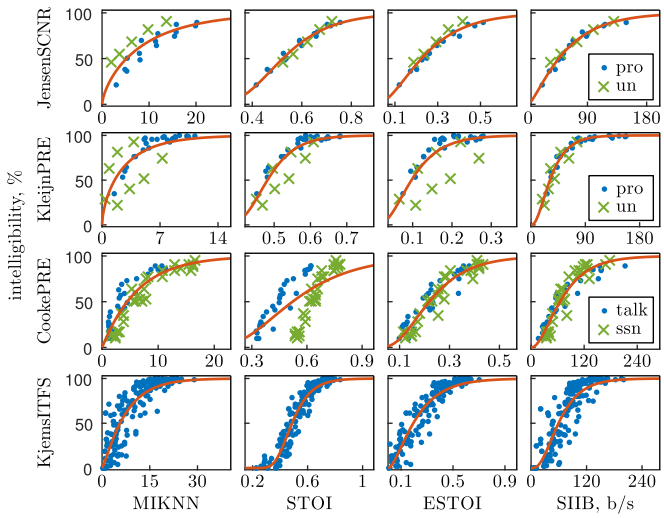

Fig. 1. Scatter plots of listening test scores (percentage of words correct) against scores computed by intelligibility metrics. For an ideal metric, all points would lie on the fitted curves. Some stimuli involved speech processed by enhancement algorithms (pro) and other stimuli were unprocessed (un). The type of noise in CookePRE is indicated by "talk" and "ssn."

TABLE I
PERFORMANCE OF INTELLIGIBILITY METRICS IN TERMS OF KENDALL'S TAU COEFFICIENT, τ

	MIKNN	STOI	ESTOI	SIIB
JensenSCNR	0.68	0.89	0.83	0.92
KleijnPRE	0.71	0.70	0.58	0.86
CookePRE	0.72	0.56	0.77	0.76
KjemsITFS	0.71	0.82	0.81	0.73
Mean	**0.71**	**0.75**	**0.75**	**0.82**

intelligibility. This letter uses two performance measures to quantify the strength of the relationship: Kendall's tau coefficient [46] τ and Pearson's correlation coefficient ρ. To use ρ effectively, the relationship between the metric d and intelligibility w must be linear. For this reason, a monotonic function $g(\cdot)$ is applied to d to linearize the relationship:

$$g(d) = 100(1 - e^{-ad})^b \tag{13}$$

where $a, b > 0$ are free parameters that are fit to each dataset to minimize the mean-squared error between w and $g(d)$ over all conditions. These free parameters are affected by the speech corpus, apparatus, and experimental procedures used during the listening test. Pearson's correlation coefficient between w and $g(d)$ is then computed.

V. RESULTS

The performance of SIIB is compared to three state-of-the-art intelligibility metrics: STOI [10], ESTOI [11], and MIKNN [23]. Fig. 1 shows scatter plots for each dataset and each intelligibility metric. The vertical axis shows the intelligibility w, and the horizontal axis shows the score computed by an intelligibility metric d. Each point represents a different condition in the dataset. The function in (13) that is used to linearize the relationship is also shown. Table I displays τ for each dataset and metric and, similarly, Table II displays ρ.

TABLE II
PERFORMANCE OF INTELLIGIBILITY METRICS IN TERMS OF PEARSON'S CORRELATION COEFFICIENT, ρ

	MIKNN	STOI	ESTOI	SIIB
JensenSCNR	0.86	0.99	0.98	0.99
KleijnPRE	0.80	0.91	0.81	0.98
CookePRE	0.90	0.69	0.95	0.95
KjemsITFS	0.88	0.96	0.95	0.88
Mean	**0.86**	**0.89**	**0.92**	**0.95**

The row of scatter plots corresponding to KleijnPRE shows that all of the reference metrics struggle to predict the effect that optimal energy redistribution has on intelligibility. In contrast SIIB is strongly correlated with intelligibility for this dataset ($\tau = 0.86$ and $\rho = 0.98$).

For CookePRE all of the metrics have reasonable performance except for STOI. This is in agreement with [11], which showed that STOI performs poorly for speech degraded by modulated noise sources such as interfering talkers. An assumption sometimes made by the speech processing community is that in order to predict intelligibility for modulated noise sources, statistics have to be averaged over short-time segments to capture the affect of "listening for glimpses of clean speech" [15]. It is then surprising that SIIB performs well on this dataset ($\tau = 0.76$ and $\rho = 0.95$) because SIIB is based on global statistics only.

Compared to the reference metrics, SIIB has excellent performance for JensenSCNR, KleijnPRE, and CookePRE, but poorer performance for KjemsITFS ($\tau = 0.73$ and $\rho = 0.88$). In [47], 17 intelligibility metrics were evaluated using KjemsITFS and only five metrics achieved $\rho \geq 0.85$. SIIB may not perform as well on KjemsITFS because ITFS processing generates some stimuli with distortions that are not normally encountered in nature. For these stimuli, it is plausible that humans are poor decoders. SIIB may correctly estimate the mutual information rate, but humans may be unable to efficiently use all of the information. This hypothesis could be tested by extensively training listeners to decode ITFS processed speech before conducting a listening test.

Notice that for maximum intelligibility, SIIB estimates an information rate of about 150 b/s. This is higher than estimates based on linguistic models of speech communication, where the information rate is 50–100 b/s [48]–[50]. This overestimate is likely the consequence of approximating $\mathbf{X}^K$ as Gaussian. Since $\mathbf{X}^K$ is only approximately Gaussian, the KLT does not remove all statistical dependencies. Accounting for the remaining dependencies would give a lower information rate.

VI. CONCLUSION

In this letter, we proposed an intrusive instrumental intelligibility metric called SIIB. SIIB is based on the hypothesis that intelligibility is related to the amount of information shared between a clean and degraded speech signal in bits per second. Compared to existing metrics, SIIB is conceptually simple, theoretically motivated, and has high performance. According to Occam's razor, these properties suggest that SIIB might generalize well to new datasets. A MATLAB implementation is available at https://stevenvankuyk.com/matlab_code/.

REFERENCES

[1] J. B. Allen, "Articulation and intelligibility," *Synthesis Lectures Speech Audio Processing*, vol. 1, no. 1, pp. 1–124, 2005.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[3] *American National Standard Methods for Calculation of the Speech Intelligibility Index*, ANSI/ASA S3.5–1997 (R2012), 2012.

[4] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[5] 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.

[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. Amer.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[7] R. Koch, *Auditory Sound Analysis for the Prediction and Improvement of Speech Intelligibility*. Goettingen, Germany: Univ. of Goettingen, 1992.

[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] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index," *Speech Commun.*, vol. 65, pp. 75–93, 2014.

[10] 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.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[11] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[12] 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.

[13] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 134, no. 1, pp. 436–446, 2013.

[14] H. Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, and T. Dau, "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. 2670–2679, 2016.

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

[16] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Commun.*, vol. 49, no. 5, pp. 402–417, 2007.

[17] Y. Tang and M. Cooke, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions," in *Proc. Interspeech*, 2016, pp. 2488–2492.

[18] C. E. Shannon, "Prediction and entropy of printed english," *Bell Labs Tech. J.*, vol. 30, no. 1, pp. 50–64, 1951.

[19] F. Pellegrino, C. Coupé, and E. Marsico, "Across-language perspective on speech information rate," *Language*, vol. 87, no. 3, pp. 539–558, 2011.

[20] 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, Mar. 2015.

[21] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[22] S. Khademi, R. Hendriks, and W. B. Kleijn, "Intelligibility enhancement based on mutual information," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

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

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

[25] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An intelligibility metric based on a simple model of speech communication," in *Proc. IEEE. Int. Workshop Acoust. Speech Enhancement (IWAENC)*, 2016, pp. 1–5.

[26] K. Karhunen, *Über Lineare Methoden in der Wahrscheinlichkeitsrechnung*. Helsinki, Finland: Universitat Helsinki, 1947.

[27] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "On the information rate of speech communication," in *Proc. IEEE. Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 5625–5629.

[28] 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.

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2012.

[30] G. Doquire and M. Verleysen, "A comparison of multivariate mutual information estimators for feature selection," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2012, pp. 176–185.

[31] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2015, pp. 277–286.

[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[33] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. San Diego, CA, USA: Academic, 1990.

[34] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, 2004, Art. no. 066138.

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

[36] M. Slaney, "An efficient implementaion of the Patterson-Holdsworth auditory filter bank," Apple Comput. Tech. Rep. 35, Cupertino, CA, USA, 1993.

[37] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1, pp. 103–138, 1990.

[38] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *J. Acoust. Soc. Amer.*, vol. 83, no. 4, pp. 1553–1573, 1988.

[39] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.

[40] R. Houben, J. Koopman, H. Luts, K. C. Wagener, A. Van Wieringen, H. Verschuure, and W. A. Dreschler, "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," *Int. J. Audiol.*, vol. 53, no. 10, pp. 760–763, 2014.

[41] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scand. Audiol.*, vol. 11, no. 2, pp. 79–87, 1982.

[42] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3553–3556.

[43] E. H. Rothauser *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[44] 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.

[45] 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.

[46] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.

[47] 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," *J. Acoust. Soc. Amer.*, vol. 130, no. 5, pp. 3013–3027, 2011.

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

[49] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. New York, NY, USA: Springer, 1972.

[50] J. Villasenor, Y. Han, D. Wen, E. Gonzalez, J. Chen, and J. Wen, "The information rate of modern speech and its implications for language evolution," in *Proc. Int. Conf. Evol. Lang.*, 2012, pp. 376–383.