# An Evaluation of Intrusive Instrumental Intelligibility Metrics

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

***Abstract*—Instrumental intelligibility metrics are commonly used as an alternative to listening tests. This paper evaluates 12 monaural intrusive intelligibility metrics: SII, HEGP, CSII, HASPI, NCM, QSTI, STOI, ESTOI, MIKNN, SIMI, SIIB, and sEPSM$^{\text{corr}}$. In addition, this paper investigates the ability of intelligibility metrics to generalize to new types of distortions and analyzes why the top performing metrics have high performance. The intelligibility data were obtained from 11 listening tests described in the literature. The stimuli included Dutch, Danish, and English speech that was distorted by additive noise, reverberation, competing talkers, preprocessing enhancement, and postprocessing enhancement. SIIB and HASPI had the highest performance achieving a correlation with listening test scores on average of $\rho = 0.92$ and $\rho = 0.89$, respectively. The high performance of SIIB may, in part, be the result of SIIBs developers having access to all the intelligibility data considered in the evaluation. The results show that intelligibility metrics tend to perform poorly on datasets that were not used during their development. By modifying the original implementations of SIIB and STOI, the advantage of reducing statistical dependencies between input features is demonstrated. Additionally, this paper presents a new version of SIIB called SIIB$^{\text{Gauss}}$, which has similar performance to SIIB and HASPI, but takes less time to compute by two orders of magnitude.**

***Index Terms*—Intelligibility prediction, instrumental measures, speech enhancement.**

## I. Introduction

When designing a speech-based communication system it is important to understand how the system will affect the intelligibility and quality of speech. Intelligibility is often defined as the proportion of words correctly identified by a listener [1], whereas speech quality refers to the pleasantness of the speech signal [2]. Many algorithms for predicting the intelligibility of a communication system have been proposed. This paper summarizes existing algorithms and evaluates their accuracy using data from formal listening tests.

Manuscript received January 25, 2018; revised May 7, 2018; accepted June 17, 2018. Date of publication July 16, 2018; date of current version August 8, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W. H. Khong. *(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 paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2856374

In [3], Shannon proposed that any communication system can be modelled by three components: a transmitter, a receiver, and a channel. In the context of speech communication, the transmitter is the vocal apparatus of the talker, the receiver is the auditory system of the listener, and the channel is the physical medium traversed by the speech signal. The channel may distort the speech signal and decrease the speech signal's intelligibility or quality. As an example, for telephone systems, the speech signal is sampled, quantized, and compressed prior to transmission. Additionally, environmental degradation such as additive noise and reverberation may be introduced at the far-end (i.e., at the talker) or the near-end (i.e., at the listener).

To combat environmental degradation, a variety of speech enhancement algorithms have been proposed (see [2] for an overview). There are two main approaches to speech enhancement: 1) the speech signal can be modified prior to degradation (e.g., optimal energy redistribution [4] and dynamic range compression [5]), or 2) the speech signal can be modified after degradation has been introduced (e.g., Wiener filters [6]). The former type of algorithm is referred to as a pre-processing algorithm and the latter as a post-processing algorithm.

A key component to the design of speech-based communication systems is an understanding of how they affect intelligibility. Although formal listening tests can provide valid data, such tests are time-consuming, laborious, and expensive. For this reason, quantities that are fast to compute and correlated with intelligibility are of interest. Such quantities are referred to as *instrumental intelligibility metrics*.

Rather than using human subjects, instrumental intelligibility metrics may rely on knowledge of the clean speech, distorted speech, and the communication channel. There are two types of intelligibility metrics: intrusive and non-intrusive. Intrusive intelligibility metrics require knowledge of the clean speech and either the channel or the distorted speech, whereas non-intrusive intelligibility metrics require only the distorted speech. Although non-intrusive metrics are more widely applicable, they tend to be less correlated with intelligibility than intrusive metrics [7], [8]. From here on, this paper focuses on intrusive intelligibility metrics.

One of the first intelligibility metrics was developed during the 1920's and is called the articulation index (AI) [9]. The AI is calculated by computing a weighted average of the signal-to-noise ratio (SNR) of several frequency bands. More recently, the AI has been refined to incorporate the results of new experiments and is now known as the speech intelligibility index (SII) [10].

Another intelligibility metric that was developed early on is the speech transmission index (STI) [11]. For this intelligibility metric, probe signals consisting of sinusoidally modulated Gaussian noise are transmitted through the communication system. The change in the modulation depth of the probe signals at the receiver is then measured and converted to an apparent SNR for each frequency band. Subsequently, the apparent SNRs are averaged similarly to the AI and SII.

Both the SII and STI have found widespread use by engineers and audiologists. However, the SII and STI have a number of limitations. First, both metrics are based on long-term statistics. This means that they do not accurately account for degradations caused by noise sources that fluctuate over time such as competing talkers and wind [12]. Second, neither metric can account for distortion introduced by enhancement algorithms [13], [14].

To overcome the limitations of the SII and STI, a number of intelligibility metrics have been proposed. Examples include the coherence SII (CSII) [15], the extended SII (ESII) [12], the quasi-stationary STI (QSTI) [16], the normalized covariance measure (NCM) [17], [18], the temporal fine-structure spectrum based index (TFSS) [19], the hearing-aid speech perception index (HASPI) [20], the Christiansen-Pedersen-Dau metric (CPD) [21], those based on the short-time objective intelligibility measure (STOI) (e.g., [22], [23]), those based on the speech-based envelope power spectrum model (sEPSM) (e.g., [24]–[26]), and those based on the glimpse proportion metric (GP) (e.g., [27]–[29]). Many of these metrics have not been extensively tested on data sets other than those used during their development. Additionally, the above metrics are often heuristically motivated, which suggests that they may not generalize well to new environments and enhancement strategies.

Recently, information theory has been proposed as a theoretically grounded approach to model speech communication. This is a natural direction to take given that the fundamental goal of speech communication is to transfer information from a talker to a listener. Information theory has been used to design state-of-the-art speech enhancement algorithms [30], [31] and intelligibility metrics [32]–[34]. Moreover, [35] used the information bottleneck principle [36] to argue that the structure of speech might be adapted to the coding capability of the mammalian auditory system (see also [37]).

Motivated by the fact that many intrusive intelligibility metrics have been recently proposed but have not been widely evaluated, this paper presents a study on the accuracy of 12 existing monaural intrusive intelligibility metrics. To assess the accuracy of each metric, the strength of the relationship between intelligibility and the metric is measured. The intelligibility data were obtained from 11 experiments described in the literature. The data include Dutch, Danish, and English speech that was degraded by additive noise, reverberation, and competing talkers, and subjected to pre-processing enhancement and post-processing enhancement.

The majority of the intelligibility metrics in this paper were developed with Germanic languages in mind, however, the studies in [38]–[41] have suggested that many intelligibility metrics can obtain good performance for Mandarin, Cantonese, and Korean.

In addition to evaluating the accuracy of pre-existing intelligibility metrics, this paper analyzes why the top performing metrics have high performance. Specifically, the effect of decorrelating input features, the effect of the auditory model, and the effect of using different distortion measures is investigated.

Previous evaluations of intrusive intelligibility metrics exist. For example [42], [43] evaluated the accuracy of intelligibility metrics for noise-reduced speech, and [44] evaluated the accuracy of intelligibility metrics for speech processed by ideal time-frequency segregation (ITFS). Those evaluations each considered a single type of degradation, whereas the evaluation in this paper considers data from many real-word scenarios.

Evaluations can also be found in publications that propose new intelligibility metrics, but in terms of the number of intelligibility metrics and the number of data sets, the scope of such evaluations is smaller than the present study. Two advantages of considering a broader scope are 1) it is easier to determine why some intelligibility metrics perform better than others, and 2) it is possible to investigate the ability of intelligibility metrics to generalize to new types of distortion.

The remainder of this paper is organized as followed. Section II describes the listening test data and Section III describes intelligibility metrics from the literature. Modified intelligibility metrics are proposed in Section IV. Performance criteria are described in Section V and results are presented in Section VI. Finally, Section VII concludes the paper.

## II. Listening Test Data

This paper considers the results of 11 intelligibility studies. From these studies, 13 data sets were created. In this section, each data set is described. Table I summarizes the data sets, while the accompanying references provide additional details. The naming convention for the data sets includes the first author of the publication that describes the data set in full, and an abbreviation that indicates the type of degradation or processing. The order that the data sets are presented in is such that similar data sets are grouped together.

### A. JensenMOD

The first data set consists of speech degraded by noise with strong temporal modulations. In [23] phrases from the Dantale II corpus [51] were degraded by ten types of noise. Four of the noise types included Track 1, 4, 6, and 7 from the ICRA noise corpus [52]. The ICRA signals are synthetic signals with spectral and temporal properties similar to speech. Four of the noise types were constructed by multiplying speech-shaped noise (SSN) (i.e., Gaussian noise with a long-term power-spectrum that is similar to the power spectrum of clean speech) with $1 + \sin(2\pi f t + \phi)$ where $\phi$ is uniformly distributed between $\pm\pi$, $t$ is the sample index, and $f = 2$, 4, 8, or 16 Hz. The final two noise sources were machine-gun noise and destroyers-operation-room noise from the NOISEX corpus [53]. Six SNRs were chosen for each noise source so that some stimuli were unintelligible and others were perfectly intelligible. In total there are 10 noise sources $\times$ 6 SNRs $=$ 60 conditions. Stimuli were presented to 12 normal-hearing listeners. For each word in a

TABLE I
SUMMARY OF LISTENING TEST DATA SETS. $m$ IS THE NUMBER OF LISTENERS AND $n$ IS THE NUMBER OF LISTENING CONDITIONS

| Name | Degradation | Enhancement strategy | Bandwidth (kHz) | $m$ | $n$ |
|---|---|---|---|---|---|
| JensenMOD [23] | Modulated noise | None | 10.0 | 12 | 60 |
| SantosREV [45] | Noise & reverb | None | 8.0 | 10 | 17 |
| KjemsAN [46] | Noise | None | 7.7 | 15 | 40 |
| KjemsITFS [46] | Noise | Ideal time-frequency segregation. | 7.7 | 15 | 168 |
| TaalPOST [22] | Noise | Minimum mean-squared error estimate of the short-time spectral amplitude. | 8.7 | 15 | 15 |
| JensenPOST [47] | Noise | Minimum mean-squared error estimate of the short-time spectral amplitude. | 4.0 | 13 | 20 |
| HuPOST [48] | Noise | Spectral subtractive, sub-space, statistical model based, and Wiener-type algorithms. | 3.5 | 40 | 72 |
| HendriksPRE [49] | Noise & reverb | Optimal energy redistribution. | 8.0 | 8 | 20 |
| KleijnPRE [30] | Noise | Optimal energy redistribution. | 8.0 | 9 | 32 |
| CookePRE [50] | Noise & competing talker | Nine pre-processing enhancement algorithms. | 8.0 | 175 | 60 |
| KhademiJOINT [31] | Noise | MVDR beamformer, Wiener filter, & optimal energy redistribution. | 8.0 | 7 | 24 |
| DutchMRG | - | JensenPOST, HendriksPRE, KleijnPRE, and KhademiJOINT merged into a single data set. | - | - | - |
| DantaleMRG | - | KjemsAN, KjemsITFS, and TaalPOST merged into a single data set. | - | - | - |

given sentence, the listeners were shown ten candidate words from which they were instructed to select from. See [23] for more details.

### B. SantosREV

The second data set consists of speech corrupted by noise and reverberation. In [45], IEEE sentences [54] were degraded by three types of distortion: 1) additive noise, 2) reverberation, and 3) additive noise and reverberation. For the additive noise distortion, SSN and babble noise at SNRs of $-5$, 0, 5, and 10 dB were used. For the reverberant distortion, IEEE sentences were convolved with a room impulse response with T60 $=$ 0.3, 0.6, 0.8, 1, and 1.4 s. For the additive noise and reverberant distortion the sentences were convolved with room impulse responses with T60 $=$ 0.3 and 0.6 s and mixed with SSN at SNRs of 5 dB and 10 dB. In total there are 8 noise + 5 reverberant + 4 noise and reverberant $=$ 17 conditions. Stimuli were presented to ten normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. See [45] for more details.

Originally, the distorted stimuli in SantosREV were offset in time from the clean stimuli. However, time-alignment is a requirement for many intrusive intelligibility metrics. For this paper, the signals in SantosREV were aligned by finding the time-offset that maximised the cross-correlation of the clean and distorted stimuli. This resulted in significantly higher performance scores than those reported in [45].

### C. KjemsAN

The third data set consists of speech degraded by additive noise. In [46] phrases from the Dantale II corpus [51] were degraded by four types of noise: SSN, cafeteria noise, noise from a bottling factory hall, and car interior noise. The stimuli were presented to 15 normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. Based on the listening test results, Kjems *et al.* derived psychometric curves that relate intelligibility to SNR for each noise type.

For this paper, KjemsAN was created by adding the noise signals to the clean Dantale II sentences at ten SNRs. The SNRs were selected by sampling the psychometric curves at intervals of 10% intelligibility from 10% to 100%. In total there are 4 noise types $\times$ 10 SNRs $=$ 40 conditions.

### D. KjemsITFS

The fourth data set consists of speech subjected to ideal time-frequency segregation processing (ITFS) [55]. ITFS processing aims to eliminate the energy of a speech signal at particular time-frequency locations by multiplying the short-time Fourier transform of the speech signal with a binary gain function. Similarly to KjemsAN, the listening experiment was conducted by Kjems *et al.*, used phrases from the Dantale II corpus [51], involved 15 normal-hearing listeners, and used the same four types of noise. For each noise type, the noisy phrases 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. See [46] for more details.

### E. TaalPOST

The fifth data set consists of speech subjected to post-processing enhancement. In [22] phrases from the Dantale II corpus were degraded by SSN at SNRs of 8.9, 7.7, 6.5, 5.2, and 3.1 dB. The MMSE-STSA enhancement algorithm [56] and an improved version [57] were applied to the noisy phrases. In total there are 5 SNRs $\times$ (2 algorithms + 1 unprocessed) $=$ 15 conditions. Stimuli were presented to 15 normal-hearing listeners. The listeners were instructed to transcribe sentences without any

additional information, and the proportion of correctly identified words was recorded.

### F. JensenPOST

The sixth data set consists of speech subjected to post-processing enhancement. In [47] phrases from the Dutch version of the Hagerman test [58] were degraded by SSN at SNRs of $-8$, $-6$, $-4$, $-2$, and 0 dB and processed by three enhancement algorithms. The three algorithms compute a minimum mean-squared error estimate of the clean speech by multiplying the short-time spectral amplitude of the noisy speech with a gain function. In total there are 5 SNRs $\times$ (3 algorithms + 1 unprocessed) $=$ 20 conditions. Stimuli were presented to 13 normal-hearing listeners. For each word in a given sentence, the listeners were shown ten candidate words from which they were instructed to select from.

### G. HuPOST

The seventh data set consists of speech subjected to post-processing enhancement. In [48] IEEE sentences [54] were filtered by a simulated telephone channel, degraded by four noise types: babble, car, street, and train, at SNRs of 0 and 5 dB, and processed by eight enhancement algorithms encompassing spectral subtractive, sub-space, statistical model based and Wiener-type algorithms. In total there are 4 noise types $\times$ 2 SNRs $\times$ (8 algorithms + 1 unprocessed) $=$ 72 conditions. Stimuli were presented to 40 normal-hearing listeners where ten listeners were used for each of the four noise types. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. See [48] for more details.

### H. HendriksPRE

The eighth data set consists of speech subjected to pre-processing enhancement and degraded by reverberation and noise. In [49] phrases from the Dutch version of the Hagerman test [58] were processed by four enhancement algorithms, convolved with a room impulse response with a T60 time of 1 s, and then degraded by SSN at SNRs of $-2$, 0, 2, and 4 dB. Three of the enhancement algorithms optimally redistribute the energy of the clean speech according to a distortion criterion. The fourth algorithm uses steady-state suppression to reduce degradation caused by reverberation. In total there are 4 SNRs $\times$ (4 algorithms + 1 unprocessed) $=$ 20 conditions. Stimuli were presented to eight normal-hearing listeners. For each word in a given sentence, the listeners were shown ten candidate words from which they were instructed to select from. See [49] for more details.

### I. KleijnPRE

The ninth data set consists of speech subjected to pre-processing enhancement and degraded by noise. In [30] phrases from the Dutch version of the Hagerman test [58] were subjected to three pre-processing 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. Stimuli were presented to nine normal-hearing listeners. For each word in a given sentence, the listeners were shown ten candidate words from which they were instructed to select from. See [30] for more details.

### J. CookePRE

The tenth data set consists of speech subjected to pre-processing enhancement and degraded by noise. In [50] IEEE sentences [54] were processed by 19 pre-processing 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. Stimuli were presented to 175 normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. Short words (e.g., a, the, in, to) were not scored.

For this paper, a subset of the data in [50] was considered because the entire data set was not available. Ten of the IEEE sentences for each condition and nine of the enhancement algorithms were used. The algorithms are referred to in [50] as AdaptDRC, F0-shift, IWFEMD, on/offset, OptimalSII, RESSYSMOD, SBM, SEO, and SSS. In total there are 2 noise sources $\times$ 3 SNRs $\times$ (9 algorithms + 1 unprocessed) $=$ 60 conditions.

### K. KhademiJOINT

The eleventh data set consists of speech that has been jointly processed by far-end and near-end enhancement algorithms. In [31], four enhancement strategies were considered, all of which used an MVDR beamformer at the far-end. The first strategy used no near-end enhancement, the second used blind optimal energy redistribution at the near-end, the third used blind optimal energy redistribution at the near-end and an additional Wiener filter at the far-end, and the fourth used jointly optimal energy redistribution at the near-end. Three near-end SNRs ($-7.5$, 0, and 5 dB) and two far-end SNRs ($-10$ and 2.5 dB) were used. In total there are 4 enhancement strategies $\times$ 3 near-end SNRs $\times$ 2 far-end SNRs $=$ 24 conditions. For each condition phrases from the Dutch version of the Hagerman test [58] were presented to seven normal-hearing listeners. For each word in a given sentence, the listeners were shown ten candidate words from which they were instructed to select from. See [31] for more details.

### L. DutchMRG

The twelfth data set was created by merging JensenPOST, HendriksPRE, KleijnPRE, and KhademiJOINT. It is reasonable to merge these data sets because the associated listening tests all used phrases from the Dutch version of the Hagerman test [58] and were conducted using the same procedures by the Circuits and Systems Group at Delft University of Technology. Note,

TABLE II
PRE-EXISTING INTELLIGIBILITY METRICS CONSIDERED IN THIS STUDY

| Abbreviation | Description |
|---|---|
| SII | The speech intelligibility index [10]. |
| HEGP | The high-energy glimpse proportion metric [29]. |
| CSII-MID | The mid-level coherence SII [15]. |
| HASPI | The hearing-aid speech perception index [20]. |
| NCM-BIF | The normalized covariance measure with signal-dependent band-importance functions [42]. |
| QSTI | The quasi-stationary speech transmission index [16]. |
| STOI | The short-time objective intelligibility measure [22]. |
| ESTOI | The extended STOI measure [23]. |
| MIKNN | The k-nearest neighbour mutual information intelligibility measure [32]. |
| SIMI | Speech intelligibility prediction based on a mutual information lower bound [33]. |
| SIIB | Speech intelligibility in bits [34]. |
| sEPSM$^{\text{corr}}$ | The speech-based envelope power spectrum model with short-time correlation [26]. |

that the number of subjects differed for the four experiments. DutchMRG was included in the evaluation to test if the intelligibility metrics give consistent measurements for different enhancement strategies.

### M. DantaleMRG

The thirteenth data set was created by merging KjemsAN, KjemsITFS, and TaalPOST. It is reasonable to merge these data sets because the associated listening tests all used phrases from the Dantale II corpus. To prevent KjemsITFS from dominating the other data sets, 60 out of the 168 conditions from KjemsITFS were randomly selected, and all of the conditions for KjemsAN and TaalPOST were selected. Note that the listening tests were conducted by different laboratory groups. Similarly to DutchMRG, this data set was included to test if the intelligibility metrics give consistent measurements for different enhancement strategies. JensenMOD also used the Dantale II corpus, but was not included in DantaleMRG because the listening test for JensenMOD presented listeners with ten candidate words to select from, whereas the listening tests for KjemsAN, KjemsITFS, and TaalPOST did not.

## III. PRE-EXISTING INTELLIGIBILITY METRICS

Over the past decade a large number of intrusive intelligibility metrics have been proposed. In this section, 12 metrics from the literature, which are considered in this evaluation, are summarized. An overview of the metrics can be found in Table II. See the accompanying references for more detailed descriptions. Unless stated otherwise, all parameters were selected according to those recommended in the original publications.

### A. Speech Intelligibility Index

The speech intelligibility index (SII) [10] is based on the idea that intelligibility is related to audibility. To compute the SII, a bandpass filterbank is applied to the clean speech and the noise signal, and a weighted average of the long-term SNR of each frequency band is calculated. The weights define a band-importance function (BIF) that characterizes the relative importance of each frequency band. Prior to averaging, the SNR is clipped to be between $\pm 15$ dB and normalized to be between 0 and 1. This reflects the idea that below $-15$ dB the speech signal is inaudible and above 15 dB the intelligibility is at its maximum. The SII is known to perform well for speech degraded by stationary additive noise, but poorly for speech degraded by modulated noise sources [12].

In this paper, the SII was only evaluated using JensenMOD, KjemsAN, and CookePRE. For the remaining data sets, either the noise signal was not available, or noise was not the main cause of distortion. The implementation of the SII was obtained from the Acoustical Society of America (http://sii.to) and used the 1/3 octave band procedure with the BIF tabulated in [10, Table 3].

### B. High-Energy Glimpse Proportion Metric

The glimpse proportion metric (GP) is the initial stage of the glimpsing model of speech perception [27] and has been used as an intelligibility metric in various studies (e.g., [28], [29]). The GP is defined as the proportion of spectro-temporal regions where the clean speech has energy greater than the noise signal by a pre-defined threshold. The GP shares similarities with the SII in that both metrics assume that audibility is the determining factor of intelligibility. The difference is that the SII averages the long-term SNR of each frequency band, whereas the GP is the proportion of short-time frequency-local SNRs above a threshold.

In [29] a variation of the GP called the high-energy GP (HEGP) was shown to be more highly correlated with intelligibility than the original GP. The main difference between the metrics is that HEGP only uses spectro-temporal regions where the noisy speech has above average energy. Similarly to the SII, HEGP can only quantify distortion caused by additive noise signals. For this reason, HEGP was evaluated using KjemsAN, JensenMOD, and CookePRE only.

The implementation of HEGP used in this paper was obtained from its developers. Note that CookePRE is a subset of a data set that was used during the development of HEGP.

### C. Coherence Speech Intelligibility Index

The coherence speech intelligibility index (CSII) [15] is based on the SII, but replaces the SNR of each frequency band with a signal-to-distortion ratio (SDR). The SDR is estimated from the coherence function [59] of the clean and distorted speech signal. For the case of speech degraded by additive noise, the SDR and SNR are equivalent, making the CSII a generalization of the SII that can be applied to a wider range of distortions. In [15] it was found that the performance of the CSII could be improved by calculating the CSII separately for low, mid, and high-energy speech segments.

The implementation of the CSII used in this paper was obtained from [2] and is described in [42], where it is referred to as $\text{CSII}_{\text{mid}}$. Note that the implementation in [2] differs to

that originally proposed in [15] in that [2] averages the CSII over short-time segments. For this paper, the implementation in [2] was modified to make it more similar to that originally proposed (i.e., it does not use short-time segments) because we found that the original method had higher overall performance. In this paper the algorithm is referred to as CSII-MID.

### D. Hearing-Aid Speech Perception Index

The hearing-aid speech perception index (HASPI) [20] is based on an elaborate auditory model where the shape and bandwidth of the cochlear filters depend on the speech signal intensity and the outer hair-cell damage of the listener. Dynamic range compression is applied to the output of each cochlear filter in accordance with physiological measurements of compression in the cochlea and psychophysical estimates of compression in the human ear. Additionally, a time-alignment stage is included. The auditory model has two outputs: a sequence of short-time log-spectra, and a basilar membrane vibration signal for each frequency band.

From the outputs of the auditory model the cepstral correlation and auditory coherence are computed. To compute cepstral correlation, the log-spectra are converted to an approximation of Mel-frequency cepstral coefficients [60] by taking the inner product between the log-spectra and a set of cosine functions. Pearson's correlation coefficient between the cepstra of the clean and distorted speech is then computed for each cepstral dimension and the resulting coefficients are averaged.

The auditory coherence is computed by splitting the basilar membrane vibration signals into three sets that contain low, mid, and high-energy segments. For each set and each frequency band, short-time correlation coefficients between the clean vibration signals and the distorted vibration signals are computed and then averaged over the time dimension and the frequency dimension. This results in three auditory coherence terms corresponding to low, mid, and high energy segments.

HASPI is computed as a linear combination of the cepstral correlation and the three auditory coherence terms. The relative importance of each term depends on the type of distortion and thus is fitted to the intelligibility data. In this paper the weights of the cepstral correlation and auditory coherence terms were computed for each data set such that the mean squared error between the predicted and measured intelligibility scores was minimized. However, it was found that similar performance could be obtained simply by summing the cepstral correlation and high-energy auditory coherence. The implementation of HASPI used in this paper was obtained from its developers.

### E. Normalized Covariance Measure

The normalized covariance measure (NCM) [17], [18] is a variant of the STI that uses clean speech as the probe signal. To compute the NCM, a band-pass filterbank is applied to the clean and distorted speech signals, and the temporal envelope of the output of each filter is extracted. Subsequently, the normalized covariance (i.e., Pearson's correlation coefficient) between the clean and distorted envelopes is calculated and converted to an apparent SNR for each frequency band. Similarly to the SII, the apparent SNR is clipped before a weighted average over the frequency bands is computed.

In [42] it was found that the NCM is strongly correlated with intelligibility for speech subjected to post-processing enhancement. The correlation was particularly strong when new signal dependent BIFs were used. The implementation of the NCM used in this paper was obtained from [2] and is described in [42] where it is referred to as NCM $W_i^{(1)}, p = 1.5$. In this paper the algorithm is referred to as NCM-BIF. Note that HuPOST was used during the development of NCM-BIF.

### F. Quasi-Stationary Speech Transmission Index

The quasi-stationary speech transmission index (QSTI) was proposed in [16]. The QSTI is a variation of the STI that uses clean speech as the probe signal and averages the score over short-time segments. In [16] the QSTI was reported to be more strongly correlated with intelligibility than the traditional STI.

The implementation of the QSTI used in this paper was obtained from its developers webpage. Note that HuPOST, TaalPOST, and KjemsITFS were used during the development of QSTI.

### G. Short-Time Objective Intelligibility Measure

The short-time objective intelligibility measure (STOI) was proposed in [22] as an algorithm for predicting the intelligibility of time-frequency weighted noisy speech. To compute STOI, a simple model of the human auditory system is used to extract temporal envelopes of the clean speech and the distorted speech for various frequency bands. The temporal envelopes are segmented into short-time frames with a duration of 386 ms and a clipping procedure is used to ensure that the SDR of each frame is greater than $-15$ dB. STOI is calculated by computing Pearson's correlation coefficient between the clean and distorted envelopes for each short-time frame and each frequency band and then taking the mean.

The implementation of STOI used in this paper was obtained from its developer's webpage. Note that TaalPOST and KjemsITFS were used during the development of STOI.

### H. Extended Short-Time Objective Intelligibility Measure

The extended short-time objective intelligibility measure (ESTOI) was proposed in [23] to address the finding that STOI performs poorly for modulated noise sources (e.g., Gaussian noise that is amplitude modulated by a sinusoid). Rather than computing the correlation of the clean and distorted envelopes for short-time segments, ESTOI computes the correlation between clean and distorted spectra so that 'glimpses of clean speech' can be detected. Additionally, the clipping procedure in STOI was removed to make the new model more mathematically tractable.

The implementation of ESTOI used in this paper was obtained from its developer's webpage. Note that JensenPOST, JensenMOD, KjemsITFS, and a data set similar to KjemsAN were used during the development of ESTOI.

### *I. K-Nearest Neighbour Mutual Information Intelligibility Measure*

The k-nearest neighbour (KNN) mutual information intelligibility measure (MIKNN) was proposed in [32] while investigating the use of information theoretical techniques for intelligibility prediction. MIKNN uses the same representation of speech as STOI, however, rather than using the short-time correlation coefficient to quantify distortion, MIKNN estimates the mutual information between the clean and distorted temporal envelopes using a non-parametric estimator based on k-nearest neighbours [61]. One advantage of mutual information is that unlike Pearson's correlation coefficient, mutual information can account for non-linear dependencies.

The implementation of MIKNN used in this paper was obtained from its developer's webpage. Note that TaalPOST and KjemsITFS were used during the development of MIKNN.

### *J. Speech Intelligibility Prediction Based on Mutual Information*

Similarly to MIKNN, the speech intelligibility prediction based on mutual information measure (SIMI) [33] is based on the hypothesis that intelligibility is related to the mutual information between the clean and distorted temporal envelopes. In contrast to MIKNN, SIMI estimates a lower bound on the mutual information by assuming a parametric statistical model. Another important difference between SIMI and MIKNN is that SIMI operates on short-time segments of 250 ms, whereas MIKNN uses whole utterances. In [33] SIMI was used to justify some of the heuristic design decisions of STOI.

The implementation of SIMI used in this paper was obtained from its developer's webpage. Note that JensenPOST, KjemsITFS, and a data set similar to KjemsAN were used during the development of SIMI.

### *K. Speech Intelligibility in Bits*

Speech intelligibility in bits (SIIB) is an information theoretic intelligibility metric that was recently proposed in [34]. Similar to MIKNN, a non-parametric mutual information estimator [61] is used to estimate the information shared between a clean and distorted speech signal.

There are three main differences between SIIB and MIKNN. First, SIIB uses the Karhunen-Loève transform (KLT) [62] to reduce statistical dependencies between spectro-temporal regions, and thus reduces overestimation of the information rate.

Second, SIIB accounts for 'production noise', which incorporates differences in pronunciation between talkers. Importantly, production noise causes the information rate of the communication channel to saturate [30].

Third, SIIB uses an auditory model that more accurately accounts for the frequency masking [63] and temporal masking [64] of the human auditory system. To account for frequency masking, the temporal envelopes are extracted using an equivalent rectangular bandwidth (ERB) gammatone filterbank [65]. To account for temporal masking, the forward masking function suggested in [66] is used. Additionally, logarithmic compression is applied to the envelopes.

The end result of SIIB is an estimate of the information shared between a talker and a listener in bits per second. Note that all of the data sets considered in this paper were used during the development of SIIB.

### *L. Speech-Based Envelope Power Spectrum Model With Short-Time Correlation*

The speech-based envelope power spectrum model forms the basis of three intelligibility metrics: sEPSM [24], mr-sEPSM [25], and sEPSM$^{\text{corr}}$ [26]. All of the sEPSM metrics use the Hilbert transform and a gammatone filterbank to extract temporal envelopes for different frequency bands. A second bandpass filterbank called a modulation filterbank is then applied to each envelope signal. This results in a multi-dimensional representation that includes a time, frequency, and modulation dimension. Within this multi-dimensional domain, sEPSM and mr-sEPSM quantify distortion using a SNR metric, whereas sEPSM$^{\text{corr}}$ quantifies distortion using short-time correlation coefficients similarly to STOI. In this paper only the most recent metric is considered: sEPSM$^{\text{corr}}$.

Note that the output of sEPSM$^{\text{corr}}$ increases as the duration of the stimulus increases. This is a consequence of the 'multiple looks' strategy that sEPSM$^{\text{corr}}$ uses to integrate information over the time dimension. For this reason, when comparing results from multiple data sets (i.e., for the merged data sets), it is important that the duration of the stimuli is held constant. In this paper, when evaluating sEPSM$^{\text{corr}}$, all stimuli were truncated to have a duration of 20 seconds.

The implementation of sEPSM$^{\text{corr}}$ used in this paper was obtained from its developers. Note that KjemsITFS was used during the development of sEPSM$^{\text{corr}}$.

## IV. Modified Intelligibility Metrics

One of the goals of this paper is to investigate why some intelligibility metrics have higher performance than others. In this section we modify existing intelligibility metrics so that effective strategies can be identified.

### *A. Investigating the Effect of Decorrelating Input Features*

The majority of the intelligibility metrics in the previous section quantify distortion by comparing time and/or frequency local features. SIIB and HASPI are exceptions to this. SIIB decorrelates log-spectra over the time and frequency dimension using the KLT, and HASPI decorrelates log-spectra over the frequency dimension using a cosine expansion similar to the type-1 discrete cosine transform (DCT) [67]. Recall that for stationary signals the DCT asymptotically approximates the KLT.

To investigate the effect of decorrelating input features, SIIB and STOI were modified to produce two intelligibility metrics denoted SIIB$^{\text{noKLT}}$ and STOI$^{\text{KLT}}$. To compute SIIB$^{\text{noKLT}}$, the implementation of SIIB described in [34] was used, but the KLT was not applied. To compute STOI$^{\text{KLT}}$ three changes are made to the original STOI implementation [22]:

1) Instead of using temporal envelopes to represent speech signals, log-temporal envelopes are used. To prevent singularities, a small amount of uniformly distributed noise is added to the envelopes before applying the logarithm.
2) The KLT is used to decorrelate the log-temporal envelopes over the frequency dimension. To do so, the eigenvectors of the covariance matrix of the clean log-temporal envelopes are computed.
3) Short-time correlation coefficients for the eigenchannels are computed and then averaged to produce a final value. The short-time segmentation approach in [22] is used, but the clipping procedure is not.

By comparing the performance of STOI with $\text{STOI}^{\text{KLT}}$, and SIIB with $\text{SIIB}^{\text{noKLT}}$ the effect of decorrelating input features can be investigated.

### B. Investigating the Effect of the Auditory Model

The auditory model that is used to extract features could have a significant impact on performance. To investigate this effect, the auditory model used for $\text{STOI}^{\text{KLT}}$ (i.e., STOIs auditory model) was replaced with the auditory model used by SIIB. The differences between the auditory models are: 1) SIIB uses an ERB gammatone filterbank, whereas STOI uses a 1/3 octave band rectangular filterbank, 2) SIIB considers frequencies up to 8 kHz, whereas STOI considers frequencies up to 5 kHz, and 3) SIIB includes a forward temporal masking function, whereas STOI does not. The resulting intelligibility metric is denoted $\text{STOI}^{\text{KLT}}_{\text{gamma}}$.

### C. Investigating the Effect of Mutual Information Estimation

The majority of the intelligibility metrics in the previous section rely on the correlation coefficient to quantify distortion. On the other hand, SIIB and MIKNN use a non-parametric mutual information estimator. Recall that if the clean and degraded signals are jointly Gaussian, then the mutual information is a function of the correlation coefficient only. In [33] this observation was used to justify the use of the correlation coefficient. However, a direct comparison between the performance obtained using a non-parametric mutual information estimator and the performance obtained using the capacity of a Gaussian channel has not been made.

To investigate the effect of mutual information estimation, SIIB was modified to produce a simpler metric called $\text{SIIB}^{\text{Gauss}}$. The original SIIB algorithm [34] quantifies distortion using a KNN mutual information estimator, whereas $\text{SIIB}^{\text{Gauss}}$ uses the information capacity of a Gaussian channel. Concretely,

$$\text{SIIB}^{\text{Gauss}} = -\frac{F}{2K}\sum_{j}\log_2(1-r^2\rho_j^2), \tag{1}$$

where $F$ is the frame rate, $K = 15$ is the number of stacked log-spectra, $r = 0.75$ is the production noise correlation coefficient, $j$ is the eigenchannel index, and $\rho_j$ is the correlation coefficient between the $j$th clean eigenchannel and the $j$th distorted eigenchannel. The values for $F$, $K$ and $r$ are the same as those in [34].

## V. Performance Criteria

The key requirement of an intelligibility metric is that it has a strong monotonic increasing relationship with intelligibility. This paper uses two performance criteria to quantify the strength of the relationship: Kendall's tau coefficient, $\tau$, and Pearson's correlation coefficient, $\rho$. Both performance criteria are discussed below.

In the following, $p_c$ is the intelligibility in terms of percentage of words correctly identified for condition $c$ in a particular data set and $d(x_c, y_c)$ is the corresponding score computed by an intelligibility metric. The clean signal $x_c$ is formed by concatenating all available clean sentences for condition $c$ and likewise for the distorted signal $y_c$.

### A. Kendall's Tau Coefficient

Kendall's tau coefficient [68], $\tau$, measures the ordinal association between two quantities and ranges between $-1$ and 1. If $\tau = -1$ then $p_c$ and $d(x_c, y_c)$ have a monotonic decreasing relationship, if $\tau = 1$ they have a monotonic increasing relationship, and if they are statistically independent then $\tau = 0$.

### B. Pearson's Correlation Coefficient

Pearson's correlation coefficient, $\rho$, is defined as the normalized covariance between two quantities. To use $\rho$ effectively, the relationship between the quantities must be linear. For this reason, a monotonic function $f$ is applied to $d(x_c, y_c)$ to linearize the relationship before computing $\rho$. The function $f$ can be thought of as a mapping from the metric to predicted intelligibility scores, but more generally it is simply a tool for quantifying the strength of the relationship between $d(x_c, y_c)$ and $p_c$.

In the literature $f$ is commonly assumed to be a logistic function, e.g., [15], [22], [69]:

$$f(d(x_c, y_c)) = \frac{100}{1 + e^{a\,(d(x_c, y_c) - b)}}, \tag{2}$$

where $b$ is the midpoint and $a$ is the slope at the midpoint. These parameters are fitted to the data to minimize the mean squared error between $p_c$ and $f(d(x_c, y_c))$.

In the literature $\rho$ is sometimes also computed without applying a mapping function. However, we believe that such a measure is misleading because without $f$, a metric with a strong non-linear relationship between $p_c$ and $d(x_c, y_c)$ will have a small value for $\rho$, but could also have a monotonic increasing relationship with intelligibility.

Note that $p_c$ depends on the experimental procedures used to measure intelligibility, but that $d(x_c, y_c)$ does not. For example, the intelligibility of a given stimulus can be increased by changing an open listening test to a closed listening test.[1] It follows that the relationships between intelligibility and intelligibility metrics also depend on experimental procedures. For this reason, $f$ is fit individually to each data set. Finally, negative values of $\rho$ and $\tau$ are set to zero.

[1] In a closed listening test, subjects are given a list of possible speech sounds, e.g., phones or words, and are asked to identify the sounds that they heard. In an open listening test, no list is provided, which makes the test more difficult.

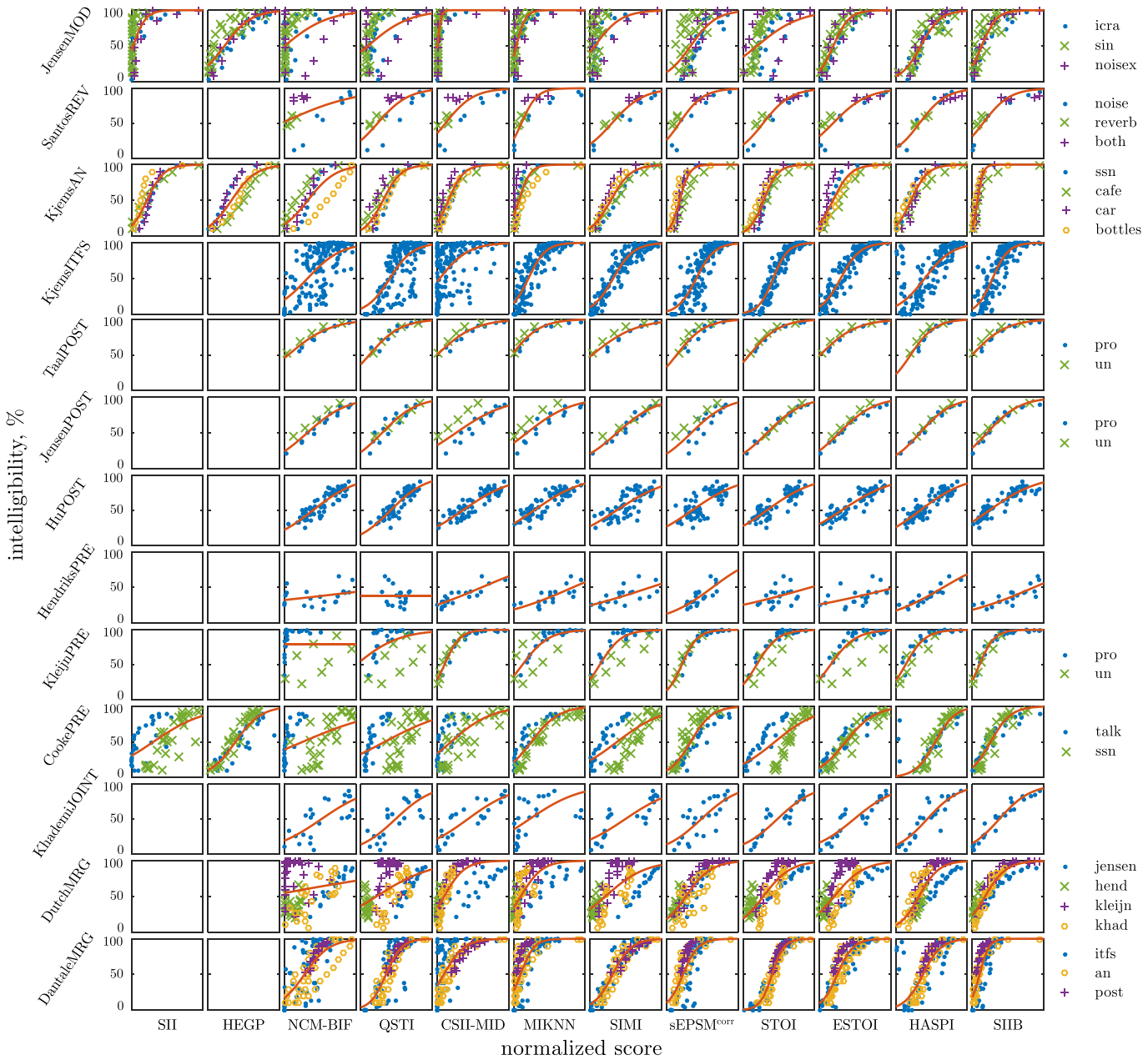



Fig. 1. Scatter plots for all data sets and pre-existing intelligibility metrics. The vertical axis is the 'ground-truth' intelligibility in terms of the percentage of words correctly identified during listening tests, and the horizontal axis is the score computed by an intelligibility metric. The horizontal axis of each plot has been normalized to be between 0 and 1. Each data point corresponds to a processing condition. The mapping function in (2) is also shown.

## VI. Results

Scatter plots for all data sets described in Section II and all pre-existing intelligibility metrics described in Section III are displayed in Figure 1. Each row of plots corresponds to a data set and each column of plots corresponds to an intelligibility metric. The vertical axis of each scatter plot is the 'ground-truth' intelligibility in terms of the percentage of words correctly identified during listening tests, and the horizontal axis is the score computed by an intelligibility metric. To facilitate an easy visual comparison, the horizontal axis of each scatter plot is normalized to be between 0 and 1. Each point on a scatter plot corresponds to a condition in the respective data set. The function in (2) that was used to linearize the relationship between the intelligibility scores and the metric for each data set is also shown. For an ideal intelligibility metric, all points would fall exactly on top of the fitted curve.

The labels 'icra', 'sin', 'noisex', 'noise', 'reverb', 'both', 'ssn', 'cafe', 'car', 'bottles', 'talk', and 'ssn' in Figure 1 indicate the type of environmental degradation in the data set. The labels 'pro' and 'un' indicate whether a stimulus was processed by an enhancement algorithm or was unprocessed. The labels 'jensen', 'hend', 'kleijn', 'khad', 'itfs', 'an', and 'post' refer to individual data sets within the merged data sets.

TABLE III
PERFORMANCE IN TERMS OF KENDALL'S TAU COEFFICIENT, $\tau$, FOR ALL DATA SETS AND INTELLIGIBILITY METRICS. THE INTELLIGIBILITY METRICS ARE LISTED IN ORDER OF MEAN PERFORMANCE AND ARE GROUPED BY PRE-EXISTING METRICS (LEFT) AND MODIFIED METRICS (RIGHT)

| | SII | HEGP | NCM-BIF | QSTI | CSII-MID | MIKNN | SIMI | sEPSM$^{corr}$ | STOI | ESTOI | HASPI | SIIB | SIIB$^{noKLT}$ | STOI$^{KLT}$ | STOI$^{KLT}_{gamma}$ | SIIB$^{Gauss}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| JensenMOD | 0.52 | 0.71 | 0.41 | 0.34 | 0.57 | 0.55 | 0.34 | 0.51 | 0.38 | 0.75* | 0.75 | 0.74* | 0.59 | 0.72 | 0.71 | 0.74 |
| SantosREV | – | – | 0.38 | 0.61 | 0.57 | 0.70 | 0.72 | 0.72 | 0.82 | 0.79 | 0.85 | 0.82* | 0.82 | 0.79 | 0.79 | 0.80 |
| KjemsAN | 0.76 | 0.75 | 0.65 | 0.78 | 0.80 | 0.65 | 0.81* | 0.74 | 0.81 | 0.74* | 0.79 | 0.82* | 0.74 | 0.74 | 0.76 | 0.84 |
| KjemsITFS | – | – | 0.48 | 0.51* | 0.41 | 0.71* | 0.80* | 0.70* | 0.82* | 0.81* | 0.66 | 0.73* | 0.69 | 0.73 | 0.74 | 0.73 |
| TaalPOST | – | – | 0.85 | 0.87* | 0.81 | 0.83* | 0.81 | 0.79 | 0.92* | 0.96 | 0.83 | 0.87* | 0.87 | 0.79 | 0.79 | 0.87 |
| JensenPOST | – | – | 0.81 | 0.80 | 0.60 | 0.68 | 0.92* | 0.66 | 0.89 | 0.83* | 0.95 | 0.92* | 0.65 | 0.82 | 0.82 | 0.94 |
| HuPOST | – | – | 0.67* | 0.68* | 0.63 | 0.64 | 0.55 | 0.44 | 0.59 | 0.69 | 0.61 | 0.74* | 0.39 | 0.72 | 0.70 | 0.73 |
| HendriksPRE | – | – | 0.30 | 0.00 | 0.69 | 0.56 | 0.52 | 0.59 | 0.26 | 0.43 | 0.78 | 0.66* | 0.72 | 0.53 | 0.62 | 0.60 |
| KleijnPRE | – | – | 0.13 | 0.20 | 0.86 | 0.71 | 0.57 | 0.88 | 0.70 | 0.58 | 0.79 | 0.86* | 0.77 | 0.78 | 0.88 | 0.86 |
| CookePRE | 0.44 | 0.72* | 0.38 | 0.38 | 0.46 | 0.72 | 0.52 | 0.71 | 0.56 | 0.77 | 0.75 | 0.76* | 0.71 | 0.87 | 0.84 | 0.77 |
| KhademiJOINT | – | – | 0.50 | 0.51 | 0.71 | 0.53 | 0.74 | 0.60 | 0.79 | 0.80 | 0.77 | 0.89* | 0.90 | 0.82 | 0.87 | 0.90 |
| DutchMRG | – | – | 0.13 | 0.29 | 0.57 | 0.54 | 0.44 | 0.68 | 0.59 | 0.46 | 0.64 | 0.75* | 0.58 | 0.54 | 0.67 | 0.74 |
| DantaleMRG | – | – | 0.54 | 0.64 | 0.53 | 0.61 | 0.80 | 0.66 | 0.83 | 0.75 | 0.67 | 0.68* | 0.58 | 0.70 | 0.73 | 0.71 |
| Mean | 0.57 | 0.73 | 0.48 | 0.51 | 0.63 | 0.65 | 0.66 | 0.67 | 0.69 | 0.72 | 0.76 | 0.79 | 0.69 | 0.73 | 0.76 | 0.79 |
| $CI_{low}$ | 0.50 | 0.68 | 0.43 | 0.46 | 0.59 | 0.61 | 0.61 | 0.63 | 0.65 | 0.68 | 0.72 | 0.75 | 0.66 | 0.70 | 0.73 | 0.76 |
| $CI_{high}$ | 0.64 | 0.77 | 0.52 | 0.55 | 0.67 | 0.69 | 0.69 | 0.70 | 0.73 | 0.75 | 0.78 | 0.81 | 0.72 | 0.76 | 0.79 | 0.81 |

TABLE IV
PERFORMANCE IN TERMS OF PEARSON'S CORRELATION COEFFICIENT, $\rho$, FOR ALL DATA SETS AND INTELLIGIBILITY METRICS. THE INTELLIGIBILITY METRICS ARE LISTED IN ORDER OF MEAN PERFORMANCE AND ARE GROUPED BY PRE-EXISTING METRICS (LEFT) AND MODIFIED METRICS (RIGHT)

| | SII | HEGP | NCM-BIF | QSTI | CSII-MID | MIKNN | SIMI | sEPSM$^{corr}$ | STOI | ESTOI | HASPI | SIIB | SIIB$^{noKLT}$ | STOI$^{KLT}$ | STOI$^{KLT}_{gamma}$ | SIIB$^{Gauss}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| JensenMOD | 0.65 | 0.88 | 0.45 | 0.43 | 0.65 | 0.72 | 0.51 | 0.68 | 0.47 | 0.92* | 0.92 | 0.89* | 0.78 | 0.90 | 0.88 | 0.89 |
| SantosREV | – | – | 0.46 | 0.76 | 0.72 | 0.90 | 0.94 | 0.87 | 0.94 | 0.91 | 0.97 | 0.93* | 0.98 | 0.93 | 0.95 | 0.93 |
| KjemsAN | 0.89 | 0.89 | 0.80 | 0.90 | 0.92 | 0.78 | 0.93* | 0.87 | 0.93 | 0.87* | 0.93 | 0.94* | 0.88 | 0.88 | 0.89 | 0.94 |
| KjemsITFS | – | – | 0.67 | 0.72* | 0.49 | 0.88* | 0.95* | 0.84* | 0.96* | 0.95* | 0.78 | 0.89* | 0.83 | 0.89 | 0.91 | 0.89 |
| TaalPOST | – | – | 0.95 | 0.95* | 0.93 | 0.95* | 0.92 | 0.90 | 0.98* | 0.97 | 0.95 | 0.96* | 0.96 | 0.92 | 0.92 | 0.96 |
| JensenPOST | – | – | 0.95 | 0.93 | 0.78 | 0.86 | 0.97* | 0.80 | 0.99 | 0.97* | 0.99 | 0.98* | 0.77 | 0.95 | 0.96 | 0.98 |
| HuPOST | – | – | 0.89* | 0.89* | 0.89 | 0.88 | 0.77 | 0.73 | 0.87 | 0.90 | 0.88 | 0.92* | 0.65 | 0.91 | 0.92 | 0.92 |
| HendriksPRE | – | – | 0.29 | 0.00 | 0.86 | 0.76 | 0.66 | 0.78 | 0.35 | 0.47 | 0.92 | 0.82* | 0.91 | 0.65 | 0.77 | 0.73 |
| KleijnPRE | – | – | 0.00 | 0.34 | 0.98 | 0.82 | 0.87 | 0.98 | 0.92 | 0.81 | 0.94 | 0.97* | 0.97 | 0.91 | 0.99 | 0.98 |
| CookePRE | 0.62 | 0.90* | 0.47 | 0.49 | 0.65 | 0.90 | 0.69 | 0.89 | 0.70 | 0.94 | 0.86 | 0.94* | 0.90 | 0.96 | 0.97 | 0.95 |
| KhademiJOINT | – | – | 0.74 | 0.80 | 0.87 | 0.53 | 0.84 | 0.75 | 0.90 | 0.90 | 0.87 | 0.96* | 0.96 | 0.91 | 0.97 | 0.95 |
| DutchMRG | – | – | 0.19 | 0.49 | 0.74 | 0.72 | 0.65 | 0.85 | 0.82 | 0.69 | 0.81 | 0.92* | 0.77 | 0.75 | 0.87 | 0.91 |
| DantaleMRG | – | – | 0.72 | 0.81 | 0.68 | 0.76 | 0.94 | 0.78 | 0.96 | 0.90 | 0.77 | 0.82* | 0.72 | 0.86 | 0.89 | 0.85 |
| Mean | 0.72 | 0.89 | 0.58 | 0.66 | 0.78 | 0.80 | 0.82 | 0.82 | 0.83 | 0.86 | 0.89 | 0.92 | 0.85 | 0.88 | 0.91 | 0.92 |
| $CI_{low}$ | 0.65 | 0.85 | 0.53 | 0.61 | 0.74 | 0.76 | 0.79 | 0.79 | 0.80 | 0.82 | 0.86 | 0.90 | 0.83 | 0.86 | 0.89 | 0.89 |
| $CI_{high}$ | 0.78 | 0.92 | 0.63 | 0.70 | 0.81 | 0.83 | 0.85 | 0.85 | 0.86 | 0.89 | 0.91 | 0.93 | 0.87 | 0.90 | 0.93 | 0.93 |

Table III displays Kendall's tau coefficient for all data sets and intelligibility metrics and, similarly, Table IV displays Pearson's correlation coefficient. In both tables, an asterisk is used to indicate when a data set was used during the development of an intelligibility metric. For the remainder of the paper, 'unseen' refers to a data set that was not used during development, and 'seen' refers to a data set that was used during development. The mean performance of each intelligibility metric and a confidence interval, $[\mathrm{CI}_{\mathrm{low}}, \mathrm{CI}_{\mathrm{high}}]$, with 95% coverage of the mean performance is also included. The confidence intervals were calculated using the non-parametric $\mathrm{BC}_{\mathrm{a}}$ bootstrap approach [70]. To do so, 5000 bootstrap sample sequences of $p_c$ and $d(x_c, y_c)$ were generated for each data set and intelligibility metric. The sample distribution of the mean performance of each intelligibility metric was then estimated from the bootstrap sample sequences.

From here on, subscripts are used to indicate performance criteria for particular intelligibility metrics. For example, $\rho_{\mathrm{SIIB}}$, refers to the correlation coefficient that SIIB achieved on some data set.

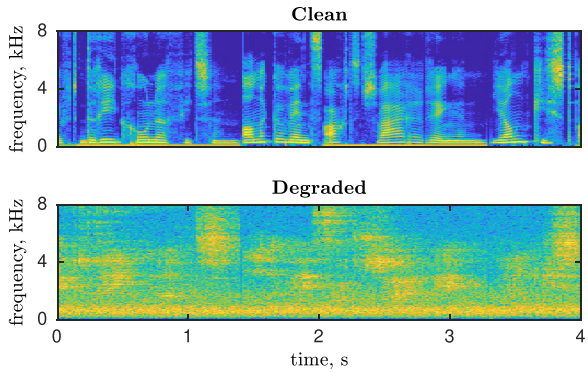



Fig. 2. An example of a clean and degraded stimulus from HendriksPRE. The severe reverberant distortion 'blurs' the time-alignment between the stimuli.

### A. Remarks for the Preexisting Metrics

It is clear that out of the pre-existing metrics SIIB and HASPI have the highest performance overall, on average achieving $\tau_{\mathrm{SIIB}} = 0.79$ and $\rho_{\mathrm{SIIB}} = 0.92$, and $\tau_{\mathrm{HASPI}} = 0.76$ and $\rho_{\mathrm{HASPI}} = 0.89$. This performance is followed closely by ESTOI, which has an average score of $\tau_{\mathrm{ESTOI}} = 0.72$ and $\rho_{\mathrm{ESTOI}} = 0.86$. HEGP has high performance for data sets distorted by additive noise achieving an average score of $\tau_{\mathrm{HEGP}} = 0.73$ and $\rho_{\mathrm{HEGP}} = 0.89$, but its usefulness is limited to situations where noise is the main source of degradation and where the noise signal is available.

The top performance rating of SIIB may be criticized on the grounds that SIIB has been 'over-designed' for the data sets in this evaluation. Although the parameters of SIIB were not intentionally optimized for the data sets in this paper, the developers of SIIB were the only researchers with access to all the data sets and thus had greater opportunity to redesign their algorithm when weaknesses were exposed during SIIBs development.

Many of the intelligibility metrics performed poorly on HendriksPRE. This is likely due to the large T60 time of the room impulse response that causes severe reverberant distortion. As shown in Fig. 2, the large T60 time somewhat 'blurs' the time-alignment of clean and degraded temporal envelopes. Many intrusive intelligibility metrics require that the clean and degraded signals are strictly time-aligned, and thus are over-sensitive to temporal blurring. Out of all the intelligibility metrics in this evaluation, HASPI achieved the highest performance for HendriksPRE ($\tau_{\mathrm{HASPI}} = 0.78$, $\rho_{\mathrm{HASPI}} = 0.92$) and is also the only intelligibility metric that included time-alignment processing.

Recall that HASPI is computed as a linear combination of four terms: the cepstral correlation, and three auditory coherence terms. The weights in the linear combination were optimized for each data set to maximize performance. None of the other intelligibility metrics modify their parameters based on the data, suggesting that the high performance of HASPI may be attributed to overfitting. To test this hypothesis, HASPI was computed simply by summing the cepstral correlation term and the high-energy auditory coherence term with equal weight. Doing so reduced the mean performance of HASPI to $\tau_{\mathrm{HASPI}} = 0.73$ and $\rho_{\mathrm{HASPI}} = 0.88$, which is still very high. Thus, the high performance of HASPI is unlikely the result of overfitting.

Another criteria that can be used to evaluate performance is whether a metric gives consistent predictions across classes of distortions. For example, CookePRE has two distinct classes: stimuli degraded by a competing talker, and stimuli degraded by SSN. Metrics may give consistent intelligibility predictions within a class, but could give inconsistent predictions between classes. An example of this can be seen in the scatter plot corresponding to STOI and DutchMRG. STOI gives consistent predictions for JensenPOST, KleijnPRE, and KhademiJOINT, but when the data sets are merged together we see distinct clusters corresponding to each data set. This means that for a given clean stimulus, a STOI score of 0.5 for noise-reduced speech and a STOI score of 0.5 for pre-processed speech could correspond to different intelligibility scores.

### B. Investigating the Performance in Terms of Generalization

Considering only entries in Table III and Table IV that have an asterisk, the mean performance of all such entries for all pre-existing metrics and data sets is $\tau = 0.78$ and $\rho = 0.92$. Considering only entries that do not have an asterisk, the mean performance for all pre-existing metrics and data sets is $\tau = 0.62$ and $\rho = 0.76$. This result demonstrates that, in general, intelligibility metrics have high performance for seen data sets, and poor performance for unseen data sets.

To further investigate the performance of intelligibility metrics in terms of their ability to generalize, Table V displays the mean performance for unseen data sets and seen data sets for each pre-existing intelligibility metric. HASPI has the highest performance for unseen data sets achieving $\tau_{\mathrm{HASPI}}^{\mathrm{unseen}} = 0.76$ and $\rho_{\mathrm{HASPI}}^{\mathrm{unseen}} = 0.89$. HEGP also has high performance for unseen data sets, however, recall that HEGP was evaluated exclusively on data sets with additive noise degradation.

STOI and SIMI both have outstanding performance for seen data sets ($\tau_{\mathrm{STOI}}^{\mathrm{seen}} = 0.87$, $\rho_{\mathrm{STOI}}^{\mathrm{seen}} = 0.97$, and $\tau_{\mathrm{SIMI}}^{\mathrm{seen}} = 0.84$, $\rho_{\mathrm{SIMI}}^{\mathrm{seen}} = 0.95$), but poor performance for unseen data sets ($\tau_{\mathrm{STOI}}^{\mathrm{unseen}} = 0.66$, $\rho_{\mathrm{STOI}}^{\mathrm{unseen}} = 0.80$, and $\tau_{\mathrm{SIMI}}^{\mathrm{unseen}} = 0.60$, $\rho_{\mathrm{SIMI}}^{\mathrm{unseen}} = 0.78$). This is because STOI and SIMI were specifically designed for speech processed by ITFS and noise-reduction algorithms, whereas the data sets in this evaluation include degradation caused by reverberation and modulated noise sources. Similarly, NCM-BIF was designed specifically for speech processed by noise-reduction algorithms. Observe that in Fig. 1 NCM-BIF has good performance for the data sets with noise-reduction: HuPOST, JensenPOST, and TaalPOST, but poor performance for the remaining data sets. These results show the danger of using intelligibility metrics outside of their intended domain.

In light of the above paragraphs, to ensure that future intelligibility metrics generalize to new data sets and give consistent predictions between classes, it may be more beneficial to gather data points with different types of degradation than to collect many data points for a single type of degradation. This notion is consistent with the high performance of HASPI, which considered six types of degradation during development:

TABLE V
MEAN PERFORMANCE OF PRE-EXISTING INTELLIGIBILITY METRICS FOR 'SEEN' AND 'UNSEEN' DATA SETS

| | SII | HEGP | NCM-BIF | QSTI | CSII-MID | MIKNN | SIMI | $\text{sEPSM}^{\text{corr}}$ | STOI | ESTOI | HASPI | SIIB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean $\tau^{\text{seen}}$ | – | 0.72 | 0.67 | 0.69 | – | 0.77 | 0.84 | 0.70 | 0.87 | 0.78 | – | 0.79 |
| mean $\tau^{\text{unseen}}$ | 0.57 | 0.73 | 0.46 | 0.45 | 0.63 | 0.63 | 0.60 | 0.66 | 0.66 | 0.69 | 0.76 | – |
| mean $\rho^{\text{seen}}$ | – | 0.90 | 0.89 | 0.86 | – | 0.92 | 0.95 | 0.84 | 0.97 | 0.93 | – | 0.92 |
| mean $\rho^{\text{unseen}}$ | 0.72 | 0.88 | 0.56 | 0.60 | 0.78 | 0.78 | 0.78 | 0.82 | 0.80 | 0.84 | 0.89 | – |

additive noise, envelope-clipping, ITFS processing, frequency-compression, noise reduction, and vocoded-speech.

### C. Remarks for the Modified Intelligibility Metrics

In general, removing the KLT from SIIB significantly reduced performance (on average $\tau_{\text{SIIB}^{\text{no\_KLT}}} = 0.69$ and $\rho_{\text{SIIB}^{\text{no\_KLT}}} = 0.85$). Furthermore, introducing the KLT to STOI improved performance (on average $\tau_{\text{STOI}^{\text{KLT}}} = 0.73$ and $\rho_{\text{STOI}^{\text{KLT}}} = 0.88$). The increase in overall performance for $\text{STOI}^{\text{KLT}}$ is mainly due to large increases in performance for JensenMOD, HendriksPRE, and CookePRE. Note that $\text{STOI}^{\text{KLT}}$ performs worse than STOI for KjemsITFS and TaalPOST, however, these are the same data sets that were used to tune the parameters of STOI during STOIs development.

The five intelligibility metrics with the highest performance: SIIB, $\text{SIIB}^{\text{Gauss}}$, $\text{STOI}^{\text{KLT}}_{\text{gamma}}$, HASPI, and $\text{STOI}^{\text{KLT}}$ are also the only metrics that decorrelate log-spectra. This outcome clearly demonstrates the advantage that can be obtained by reducing the statistical dependencies between input features.

Recall that ESTOI was proposed as an extension to STOI that can 'listen to glimpses of clean speech'. Interestingly, for the data sets that contain modulated noise, $\text{STOI}^{\text{KLT}}$ has similar performance to ESTOI (for JensenMOD, $\tau_{\text{STOI}^{\text{KLT}}} = 0.72$, $\rho_{\text{STOI}^{\text{KLT}}} = 0.90$, and for CookePRE, $\tau_{\text{STOI}^{\text{KLT}}} = 0.87$, $\rho_{\text{STOI}^{\text{KLT}}} = 0.96$). SIIB and $\text{SIIB}^{\text{Gauss}}$, which are based on long-term statistics, also have good performance for JensenMOD and CookePRE. Such results contest the idea that short-time segmentation is necessary for predicting the intelligibility of modulated noise sources.

On average $\text{STOI}^{\text{KLT}}_{\text{gamma}}$ achieved $\tau_{\text{STOI}^{\text{KLT}}_{\text{gamma}}} = 0.76$ and $\rho_{\text{STOI}^{\text{KLT}}_{\text{gamma}}} = 0.91$. Thus, by introducing the KLT to STOI and using a more realistic auditory model, performance competitive with SIIB could be obtained. This means that for some representations of speech signals, the correlation coefficient and the KNN mutual information estimator can quantify distortion equally well. A partial explanation for this result can be found by considering the high performance of $\text{SIIB}^{\text{Gauss}}$ ($\rho_{\text{SIIB}^{\text{Gauss}}} = 0.92$ and $\tau_{\text{SIIB}^{\text{Gauss}}} = 0.79$), which suggests that the Gaussian communication channel is a reasonable approximation of the true communication channel for many real-word distortions.

Finally, recall that $\text{SIIB}^{\text{Gauss}} = -\frac{F}{2K}\sum_j \log_2(1 - r^2\rho_j^2)$. Since $r$ and $\rho_j$ are between $-1$ and 1, the product of their squares is likely to be small, particularly for challenging listening environments. Using the approximation $\log_2(1+a) \approx a/\ln(2)$ for small $a$, we have that $\text{SIIB}^{\text{Gauss}} \approx \frac{F}{2K\ln(2)} r^2 \sum_j \rho_j^2$. This approximation strongly resembles the distortion measure used by $\text{STOI}^{\text{KLT}}$ and $\text{STOI}^{\text{KLT}}_{\text{gamma}}$, which can be written as $\sum_j \sum_t \rho_{j,t}$, where $t$ is the short-time segment index.

## VII. CONCLUSION

In this paper, the accuracy of 12 intelligibility metrics from the literature was evaluated using the results of 11 listening tests. The stimuli included pre-processing enhancement, post-processing enhancement, and environmental distortions such as noise and reverberation. In order to analyze why the top performing metrics have high performance, four new intelligibility metrics were proposed. The main conclusions are as follows.

1) Out of the pre-existing metrics, SIIB and HASPI had the highest overall performance.
2) Many intrusive metrics struggle with severe reverberant distortion. This may be because they are over-sensitive to the time-alignment of clean and distorted temporal envelopes.
3) In general, intelligibility metrics perform more poorly on unseen data sets than on seen data sets. For this reason, caution should be taken when using intelligibility metrics outside of their intended domain.
4) For unseen data sets, HASPI had the highest performance. This suggests that HASPI is appropriate for situations where many types of potentially new speech material and distortions are likely. Additionally, unlike the other metrics, HASPI has built-in time-alignment processing and can account for hearing impairments.
5) The five intelligibility metrics with the highest overall performance are also the only metrics that decorrelate log-spectra. On average, introducing the KLT to STOI improved performance and removing the KLT from SIIB reduced performance. These results demonstrate the advantage of removing statistical dependencies between input features.
6) The high performance of $\text{SIIB}^{\text{Gauss}}$ suggests that the Gaussian communication channel is a reasonable approximation of the true communication channel for many real-world distortions. Additionally, $\text{SIIB}^{\text{Gauss}}$ has performance similar to SIIB, but takes less time to compute by two orders of magnitude.[2]
7) It was shown that $\text{STOI}^{\text{KLT}}$ and $\text{STOI}^{\text{KLT}}_{\text{gamma}}$ can be interpreted as approximations of $\text{SIIB}^{\text{Gauss}}$.

[2]MATLAB implementations of $\text{SIIB}^{\text{Gauss}}$ and SIIB are available at www.stevenvankuyk.com/MATLAB_code

## ACKNOWLEDGMENT

The authors would like to thank the following researchers for providing intelligibility data and MATLAB implementations of their intelligibility metrics: A. Andersen, F. Chen, M. Cooke, J. Jensen, J. Kates, H. Relano-Iborra, J. Santos, and Y. Tang. The authors would also like to thank U. Kjems, K. Paliwal, J. Taghia, and C. Taal, for making their materials publicly available. Finally, the authors would like to thank the three anonymous reviewers for their insightful comments.

## REFERENCES

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

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

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

[4] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 858–872, 2014.

[5] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 635–638.

[6] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, vol. 7. Cambridge, MA, USA: MIT Press, 1949.

[7] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[8] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5085–5089.

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

[10] American National Standards Institute, *American National Standard Methods for Calculation of the Speech Intelligibility Index*, ser. S3.5, New York, NY, USA, 1997.

[11] 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.

[12] 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.

[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.*, 1993, pp. 50–55.

[14] P. C. Loizou and J. Ma, "Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms," *J. Acoust. Soc. Amer.*, vol. 130, no. 2, pp. 986–995, 2011.

[15] 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.

[16] B. Schwerin and K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Commun.*, vol. 65, pp. 9–19, 2014.

[17] R. Koch, "Auditory sound analysis for the prediction and improvement of speech intelligibility," Ph.D. dissertation, Univ. Göttingen, Göttingen, Germany, 1992.

[18] 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.

[19] F. Chen, L. L. N. Wong, and Y. Hu, "A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech," *Speech Commun.*, vol. 55, no. 10, pp. 1011–1020, 2013.

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

[21] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Commun.*, vol. 52, no. 7, pp. 678–692, 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.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

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

[30] 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.

[31] S. Khademi, R. C. 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.

[32] 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.

[33] 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.

[34] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[35] 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.

[36] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control Comput.*, 1999, pp. 368–377.

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

[38] I.-K. Jin, "Development of the speech intelligibility index (SII) for Korean," Ph.D. dissertation, Dept. Speech, Lang., Hearing Sci., Univ. Colorado Boulder, Boulder, CO, USA, 2014.

[39] L. L. N. Wong, A. H. S. Ho, E. W. W. Chua, and S. D. Soli, "Development of the Cantonese speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 121, no. 4, pp. 2350–2361, 2007.

[40] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in Mandarin," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4465–4468.

[41] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *J. Acoust. Soc. Amer.*, vol. 129, no. 5, pp. 3281–3290, 2011.

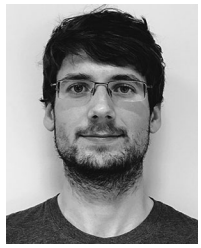
[42] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[43] 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. IEEE. Int. Workshop Acoust. Speech Enhancement*, 2010.

[44] 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.

[45] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. IEEE Int. Workshop Acoust. Speech Enhancement*, 2014, pp. 55–59.

[46] 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.
[47] 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.
[48] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.
[49] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 851–862, May 2015.
[50] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: The Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.
[51] 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.
[52] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.
[53] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.
[54] E. H. Rothauser *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 3, pp. 225–246, Sep. 1969.
[55] 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.
[56] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[57] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
[58] R. Houben *et al.*, "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," *Int. J. Audiol.*, vol. 53, no. 10, pp. 760–763, 2014.
[59] G. Carter, C. Knapp, and A. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 4, pp. 337–344, Aug. 1973.
[60] 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.
[61] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, 2004, Art. no. 066138.
[62] K. Karhunen, *Über lineare Methoden in der Wahrscheinlichkeitsrechnung (On Linear Methods in Probability and Statistics)*. Helsinki, Finland: Universitat Helsinki, 1947.
[63] R. Wegel and C. Lane, "The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear," *Phys. Rev.*, vol. 23, no. 2, pp. 266–285, 1924.
[64] A. J. Oxenham, "Forward masking: Adaptation or integration?" *J. Acoust. Soc. Amer.*, vol. 109, no. 2, pp. 732–741, 2001.
[65] M. Slaney, "An efficient implementaion of the Patterson-Holdsworth auditory filter bank," Apple Computer, Inc., Cupertino, CA, USA, Tech. Rep. 35, Apple Comput. Tech. Rep. #35, 1993.
[66] 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.
[67] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. San Diego, CA, USA: Academic, 1990.
[68] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.
[69] S. Gordon-Salant and P. J. Fitzgibbons, "Comparing recognition of distorted speech using an equivalent signal-to-noise ratio index," *J. Speech, Lang., Hear. Res.*, vol. 38, no. 3, pp. 706–713, 1995.
[70] B. Efron, "Better bootstrap confidence intervals," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 171–185, 1987.



**Steven Van Kuyk** received the B.E. (Hons.) degree in 2015 in electronic and computer systems engineering from the Victoria University of Wellington, Wellington, New Zealand, where he is currently working toward the Ph.D. degree. In 2014, he was with Fisher & Paykel Healthcare, Auckland, New Zealand, and in 2017, he was with Apple, Inc., Cupertino, CA, USA. His research interests include signal processing, information theory, and machine learning, for applications in audio processing.

**W. Bastiaan Kleijn** (F'99) received the M.S.E.E. degree from the Stanford University, Stanford, CA, USA, the M.Sc. degree in physics and the Ph.D. degree in soil science from the University of California, Riverside, CA, USA, and the Ph.D. degree in electrical engineering from the Delft University of Technology (DUT), Delft, The Netherlands. He is a Professor with the Victoria University of Wellington, Wellington, New Zealand, and a Professor (part-time) with the DUT. From 1996 to 2010, he was a Professor and the Head of the Sound and Image Processing Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. He was a founder of the Global IP Solutions, a company that provided the enabling audio technology to Skype. It was acquired by the Google in 2010. He has served on a number of editorial boards including those of the IEEE TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, *Signal Processing*, the IEEE SIGNAL PROCESSING LETTERS, and the *IEEE Signal Processing Magazine*. He was the Technical Chair of the ICASSP 1999, EUSIPCO 2010, and two IEEE workshops.

**Richard Christian Hendriks** was born in Schiedam, The Netherlands. He received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively. He is currently an Associate Professor with the Circuits and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. His main research interests include biomedical signal processing and audio and speech processing, including speech enhancement, speech intelligibility improvement, and intelligibility modeling. In March 2010, he received the prestigious VENI grant for his proposal Intelligibility Enhancement for Speech Communication Systems. He is a recipient of several best paper awards, among which the IEEE Signal Processing Society Best Paper Award in 2016. He is an Associate Editor for the IEEE/ACM TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the *EURASIP Journal on Advances in Signal Processing*.