JASA ARTICLE

Intelligibility prediction for speech mixed with white Gaussian noise at low signal-to-noise ratios

Simone Graetzer[a] and Carl Hopkins[b]
Acoustics Research Unit, School of Architecture, University of Liverpool, Liverpool L69 7ZN, United Kingdom

ABSTRACT:
The effect of additive white Gaussian noise and high-pass filtering on speech intelligibility at signal-to-noise ratios (SNRs) from -26 to 0 dB was evaluated using British English talkers and normal hearing listeners. SNRs below -10 dB were considered as they are relevant to speech security applications. Eight objective metrics were assessed: short-time objective intelligibility (STOI), a proposed variant termed STOI+, extended short-time objective intelligibility (ESTOI), normalised covariance metric (NCM), normalised subband envelope correlation metric (NSEC), two metrics derived from the coherence speech intelligibility index (CSII), and an envelope-based regression method speech transmission index (STI). For speech and noise mixtures associated with intelligibility scores ranging from 0% to 98%, STOI+ performed at least as well as other metrics and, under some conditions, better than STOI, ESTOI, STI, NSEC, $\mathrm{CSII}_{\mathrm{Mid}}$, and $\mathrm{CSII}_{\mathrm{High}}$. Both STOI+ and NCM were associated with relatively low prediction error and bias for intelligibility prediction at SNRs from -26 to 0 dB. STI performed least well in terms of correlation with intelligibility scores, prediction error, bias, and reliability. Logistic regression modeling demonstrated that high-pass filtering, which increases the proportion of high to low frequency energy, was detrimental to intelligibility for SNRs between -5 and -17 dB inclusive.

https://doi.org/10.1121/10.0003557

(Received 26 May 2020; revised 7 January 2021; accepted 28 January 2021; published online 25 February 2021)

[Editor: John H. L. Hansen] Pages: 1346–1362

I. INTRODUCTION

Speech communication can be impaired in adverse conditions such as those involving interfering noise, excessive reverberation, and distortion of the transmission channel. To estimate the magnitude of the impairment, the signals acquired before and after transmission or processing are compared, either by human listeners or by means of an algorithm. Such an algorithm needs to be effective across a range of signal-to-noise ratios (SNRs) and should take into account the non-stationarity of speech—and some maskers—such that human listeners can use speech information "present in the dips."[1]

In general, the literature considers objective methods to assess speech intelligibility that are relevant to the field of speech enhancement, where the aim is to obtain a high percentage of intelligible words with $SNR \geq -10$ dB using natural noise sources such as babble or cafeteria noise. However, in the field of speech security, where there is a need to assess the risk of only a few words being intelligible when overheard or covertly intercepted, typically, the aim is to identify percentage correct word scores that are <20%.[2] This tends to occur when $SNR < -10$ dB, and in this paper SNRs are considered down to -26 dB. For speech security situations where masking noise is required, a noise source such as road traffic or a nearby conversation is not reliable, as there is no control over the time-varying amplitude, and there is the risk of a substantial lull. For this reason, electronic or mechanical sources of stationary noise can be considered, and as an example of such a source, white Gaussian noise (WGN) is used in this paper (N.B. WGN can be more effective than speech-shaped noise in reducing the recognition of consonants).[3] In this paper, several speech intelligibility algorithms are considered, most of which use short-time methods to account for dip listening.

Various objective methods proposed for predicting speech intelligibility in additive noise are based on SNR estimates, such as the articulation index[4] (AI), the speech intelligibility index[5] (SII), and the speech transmission index[6] (STI). AI performs well for signals corrupted by additive, stationary noise[4] but is not able to account for the effects of reverberation, non-stationary noise, and nonlinear or time domain distortion (e.g., peak clipping or reverberation). According to ANSI S3.5,[5] SII can be used in cases of additive noise or linear filtering but not in cases of fluctuating maskers or nonlinear distortion such as dynamic envelope compression. Not only AI and SII, but also STI, are unsuitable in conditions involving (strongly) fluctuating maskers and nonlinear processing, such as spectral subtraction noise reduction methods (e.g., see Houtgast *et al.*[6]) Further, these metrics are not sensitive enough to distinguish between merely audible and intelligible speech signals at a very low SNR. Gover and Bradley[7] found that some words

[a] Present address: Acoustics Research Centre, School of Science, Engineering and Environment, University of Salford, Salford M5 4NT, United Kingdom.

[b] Electronic mail: carl.hopkins@liverpool.ac.uk, ORCID: 0000-0002-9716-0793.

from the Institute of Electrical and Electronics Engineers (IEEE) sentences[8] could be identified at values of AI and SII equal to 0, while all STI values below 0.3 are classified as indicating "bad" intelligibility.[9]

Since the introduction of SNR-based methods, research has focused more on correlation, covariance, and coherence methods. There has also been a movement toward using speech as a test signal (rather than, for example, modulated noise), which permits real-time intelligibility prediction. Speech-based SII and STI methods based on signal correlation/covariance include the normalised covariance metric[10,11] (NCM, also termed CSTI) and the coherence SII[12] (CSII), which is based on the SII but replaces the SNR with the signal-to-distortion ratio (SDR). Of the large number of measures considered by Ma *et al.*[13] for intelligibility prediction with signals created at 0 or 5 dB SNR, NCM and CSII with signal-dependent band importance weightings performed best.

The short-time objective intelligibility metric (STOI) was developed by Taal *et al.*[14] and is a correlation-based metric used to quantify the intelligibility benefits of time-frequency masking algorithms [e.g., ideal time-frequency segregation (ITFS)] and other nonlinear enhancement techniques. STOI values are converted to predicted speech intelligibility scores via a logistic (sigmoid) function.[14] Mean STOI scores have been used, in practice, as a standalone measure of the relative effectiveness of a speech enhancement algorithm (e.g., Kolbæk *et al.*[15] and Hsu *et al.*[16]). This requires that STOI can accurately and reliably predict intelligibility before and after noise reduction. It has been stated in publications that STOI varies between zero and one (e.g., Hsu *et al.*[16]). Taal *et al.*[14] claimed only that STOI had "a monotonic relation with speech intelligibility" and that the aim was "not necessarily to predict absolute intelligibility scores" (p. 2126); no claim was made that STOI should vary between zero and one. However, the use of a full range from zero to one can be advantageous for ease of interpretation, for example, when intelligibility scores are unavailable. In evaluating STOI for noisy signals, Taal *et al.*[17] found that for speech from the Dantale II corpus, which comprises only one female talker, when degraded by four noise types, STOI values close to 0.4 were associated with intelligibility scores of 0%. This indicates that a range of zero to one is not used. Other studies also show that STOI rarely falls below 0.3, even for signals associated with 0% intelligibility scores, and where SII, NCM, and CSII are zero (see, e.g., Tang *et al.*[18]). Taal *et al.*[14] found that for Dantale II speech degraded by speech-shaped noise (SSN), at SNRs above −10 dB, the magnitude of overestimation increased with increasing degradation for this noise type.

STOI was defined by Taal *et al.*[14] to include a normalisation procedure to compensate for global level differences and a clipping procedure to put an upper bound on the sensitivity to severely degraded time-frequency (TF) units. In subsequent investigations or extensions of STOI, the clipping procedure has often been removed. For implementation with cochlear implants, Taal *et al.*[19] introduced a simplified version of STOI for which one of the simplifying steps was to remove the clipping procedure. However, no comparison of the approach with and without clipping was provided. Lightburn and Brookes[20] derived a binary mask for speech enhancement by maximising STOI, for which they also removed the clipping procedure on the basis that clipping was "very rare in the stochastic noise case" (p. 5079). Andersen *et al.*[21] modified STOI for use with binaural speech and removed the clipping procedure on the basis that this did not appear to significantly impair the prediction performance for Taal *et al.*[19] For modulated noise maskers, Jensen and Taal[22] developed the extended short-time objective intelligibility metric (ESTOI) to improve STOI performance for highly fluctuating or modulated noise sources and stated that it discards the clipping procedure. ESTOI is based on energy-normalised short-time spectrograms that are decomposed into orthogonal one-dimensional subspaces that are important for intelligibility. Kolbæk *et al.*[15] used a deep neural network to maximise an approximation to STOI for which the clipping procedure was not used on the basis that empirical observations from previous studies[19–22] indicated that omitting clipping tended not to affect the performance of STOI. These studies did not provide any comparison of results with and without clipping. Hence, in this paper, STOI is assessed alongside a proposed variant, STOI+, which does not use the normalisation and clipping proposed by Taal *et al.*,[14] to identify whether this variant would have a lower prediction error and metric bias, and better metric reliability, than the original STOI for low mixture SNRs and WGN. The justification for the proposed variant is discussed further in Sec. II C 1.

It is beneficial to test metrics on data sets other than those used in their development. For STOI, most evaluations have considered speech from a single speaker of Danish,[14] Dutch,[23] American English,[24] or Mandarin[25] (and therefore a single gender, although it differed between the languages). Van Kuyk *et al.*[26] found that amongst the speech intelligibility metrics they considered, including STOI, SII, and NCM with signal-dependent band importance functions, a form of CSII termed CSII_{Mid} and ESTOI tended to perform poorly when applied to data sets that were not used in their development. For Dantale II speech degraded by four types of noise, including SSN and car interior noise, STOI and speech intelligibility in bits (SIIB) obtained higher correlation coefficients than other metrics. STOI tends to outperform more commonly used objective metrics for ITFS-processed speech but performs less well for unprocessed noisy speech (at least for noise that is non-stationary) and less well for modified or synthetic speech.[27]

In speech security, there is usually a need to assess worst-case scenarios. One potential scenario is speech produced in the presence of background noise, which leads to a flattening of spectral tilt that can reliably increase speech intelligibility compared to speech produced in quiet (e.g., Lu and Cooke[28]). This is likely to be due to release from energetic masking at mid to high speech frequencies (1–4 kHz). Lu and Cooke[28] mixed speech with speech-shaped noise at $SNR = -9$ dB and used filtering to produce an artificial reduction in spectral tilt that led to an increase in intelligibility for native listeners, when compared to unmodified speech. For speech mixed with WGN, a high-pass filter (HPF) can improve speech intelligibility relative to unmodified speech by increasing the proportion of high to low frequency energy for signals presented at the same global SNR;[29,30] however, previous studies focused on $SNR \geq -10$ dB. Therefore, in this paper, the opportunity is taken to assess the effect of high-pass filtering over a wider range of SNRs down to −26 dB.

In the current study, speech signals are mixed with WGN at low mixture SNRs and presented to listeners with and without flattening of the spectral tilt. In total, eight invasive metrics are evaluated for the intelligibility prediction of noisy speech: STOI, STOI+, ESTOI, two forms of CSII ($\text{CSII}_{\text{High}}$ and CSII_{Mid}), NCM, the normalised subband envelope correlation metric[31] (NSEC), and a speech-based STI method[32] (hereafter termed STI). The main aim is to compare STOI with a variant, STOI+, for speech mixed with WGN at SNRs between −26 and 0 dB and to determine how these metrics compare with other well known measures, particularly in the context of speech security. This range of SNRs is used to give percentages of words correctly identified ranging from 0% up to almost 100% to evaluate metric behavior over the whole intelligibility score range. A secondary aim is to determine whether a HPF that decreases the spectral tilt without a strong attenuation of low frequencies ($f < 300$ Hz) improves the intelligibility of speech mixed with WGN at signal-to-noise ratios between −26 and 0 dB. To be able to make more defensible claims about British English speech in general and to provide more information about the intelligibility score-metric logistic function, which is advantageous for prediction, this study uses speech from 12 talkers (rather than the typical 1–3) with an equal gender split and 9 SNRs (rather than the typical 3–5).

Section II outlines the experimental procedures, including a brief discussion of how the proposed STOI variant, STOI+, differs from conventional STOI. Section III reports the effects of SNR, spectral tilt flattening, and talker gender on intelligibility scores and the performance of metrics in estimating those scores. In Sec. IV, the reasons for the variation in outcomes of spectral tilt flattening and the relative performance of STOI and STOI+ and the other metrics are discussed.

II. EXPERIMENTAL PROCEDURES

A. Speech signals

1. Speech recordings

Twelve talkers (six male, six female) between 21 and 47 yrs of age were recorded in an anechoic chamber using a 0.5 in. Brüel & Kjær (B&K) (Nærum, Denmark) type 4190 microphone at 1 m on axis, a B&K type 2669 preamplifier, and a B&K LAN-XI type 3050 front end with a B&K time data recorder. The sampling frequency for the recordings was 65.536 kHz. The talkers were native British English speakers with an accent similar to Received Pronunciation (Standard Southern English).

Talkers produced the IEEE sentences,[8] which form 72 word lists in total (where each list comprises ten sentences), in a pseudo-random order. Before the recording session, the talkers were asked to “speak normally as you would in everyday conversation” to elicit a normal vocal effort, where vocal effort is defined as the equivalent continuous A-weighted sound pressure level (SPL) of speech measured at a distance of 1 m in front of the mouth, i.e., on axis. If the talker hesitated or made an error, s/he repeated the sentence. These recordings are freely available for download in the ARU speech corpus.[33]

2. Signal processing

All speech recordings were initially filtered with a high-pass finite impulse response (FIR) filter using a Kaiser window method to remove energy below 60 Hz and low-pass filtered to attenuate energy above 9 kHz (predominantly electrical background noise). These signals are termed non-HP-filtered (where HP refers to high-pass).

In subsequent processing, a HPF was used to flatten the spectral tilt. The filter was designed to obtain desired amplitudes of zero and one at normalised frequencies between zero and one (Nyquist), with an approximately linear relationship between amplitude and normalised frequency. This was carried out with the MATLAB filter command *firpm* to give a 10th order optimal equiripple, linear-phase FIR filter using the Parks–McClellan algorithm (weights set to unity). To illustrate the effect of this filter, the long-term average speech spectra (calculated using MATLAB[34]) based on ten word lists are shown in Fig. 1, before and after the application of the filter. Talker fundamental frequencies were as low as approximately 70 Hz for males and 130 Hz for females; at and above these frequencies, one-third octave band levels of the speech were at least 10 dB above background noise.

To create the noisy speech signals and present these signals to listeners with a Nyquist frequency of 12 kHz, first, WGN was generated with a sampling frequency of 24 kHz, and the speech signals were downsampled to the same sampling frequency. Second, the active speech levels of all speech signals (non-HP-filtered and HP-filtered) were equalised using the procedure in ITU-T P.56.[35] Finally, these speech signals were mixed with a pseudo-randomly selected segment of the WGN at nine SNRs ranging from −26 to 0 dB. The additive WGN was gated on and off 1 s before and after the speech signal.

B. Listening tests

Forty-eight untrained listeners (24 male, 24 female) aged between 19 and 49 yrs ($\mu = 27.8$ yrs, $\sigma = 8.2$ yrs) took part in the experiment. No listeners had been exposed previously to the speech material. All listeners used British English as a first language and reported having a good

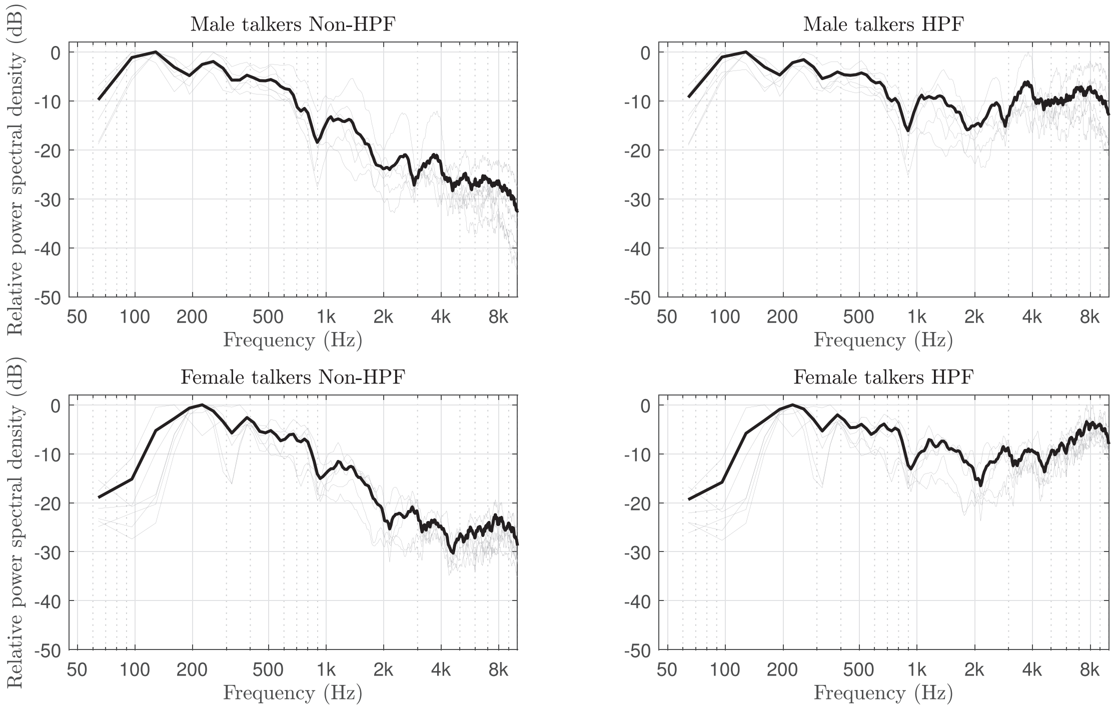

FIG. 1. Long-term average speech spectra from ten word lists per talker gender and filter condition before (left) and after (right) the application of the HPF. The six talkers are shown in gray, with the average of those talkers shown as a thick black line. Note that individual talkers vary by up to 24 dB across the frequency range, and whilst the HPF flattens the spectra, this variation remains with or without the HPF.

spelling ability. Their hearing thresholds were tested according to ISO 8253–1[36] and did not exceed 20 dB hearing level (HL) between 125 and 8 kHz. The tests were conducted in a sound-attenuated booth. The background noise at the entrance to the ear canal during testing was estimated to be 22 dB L_{Aeq} using the B&K type 4100 head-and-torso simulator (HATS) wearing circumaural headphones [Beyerdynamic (Heilbronn, Germany) DT770 Pro] connected to the PC. Diotic presentation of the stimuli used a playback system comprising the same headphones connected to a PC running MATLAB code with a custom GUI. The audio output of the system was calibrated using the HATS with type 4189 microphones in each ear canal. Subjects chose their preferred listening level as 70 or 75 dB L_{Aeq}. Twenty-eight listeners chose a playback level of 70 dB L_{Aeq}, while 20 chose a level of 75 dB L_{Aeq}. In the familiarisation stage, listeners heard one clean sentence and four noisy sentences at SNRs equal to 0, −5, −8, or −11 dB. Sentences were selected at random. Listeners heard at least one sentence from each of the four talkers assigned to that listener. These sentences were later presented in the full test, as the experimental design required 720 unique sentences.

Two female and two male talkers were randomly allocated to each of the 48 listeners in such a way that each talker was allocated to eight female and eight male listeners. For each talker, one word list was used per SNR and filter (HPF, non-HPF) combination. Signals were presented in a randomised order. Each listener participated in a total of 72 listening conditions (4 talkers × 9 SNRs × 2 filter conditions).

Listeners were asked to identify as many words as possible in each sentence. They had approximately 15 s after the sentence had played to enter the words they heard into the GUI text box and were able to correct their spelling during this time. Listeners were allowed to pause the test at any time and were offered a break of up to 5 min after every ≈30 min. Tests were completed in approximately 2 h including breaks. The ability to pause the test at any time and the randomised presentation order that ranged from "easier" sentences (e.g., 0 dB SNR) to "harder" sentences (e.g., −26 dB SNR) was intended to reduce the likelihood of any fatigue.

Listener responses were scored according to the number of words identified correctly. Scores were expressed as the percentage of words identified correctly in each word list, which comprised ten sentences. After Robinson *et al.*,[2] homophones and some alternative spellings were allowed, according to the following rules: (a) ignore punctuation such as apostrophes, (b) allow homophones, (c) allow either American English or British English spelling, and (d) allow certain misspellings. Regarding (d), words were judged to be correct when two words were identified as one when permitted in modern British English, e.g., "should not" could be given as "shouldn't"; "cannot" was identified as "can't"; some regular plurals were provided in singular form and vice versa, e.g., "desk" could be given as "desks"; some regular verbs conjugated with "-s" or "-ed" were missing the suffix, e.g., "asks" could be "ask," "baulked" could be "baulk"; nouns with a possessive "'s" suffix were missing the suffix, e.g., "pirate's" could be given as "pirate"; an initial "a-" was missing and the result was a word, e.g., "account" could be given as "count"; an initial "h" was inserted if the result was a word, e.g., "air" could be given as "hair," and "man" was identified as "men" and vice versa. While scoring was automated, results were carefully monitored by the authors. These rules were appropriate in the security context, where the interest is in identifying as few as one or two words and identifying the root of the word may be sufficient for a breach. After Robinson *et al.*,[2] the words "a" and "the" were considered to have negligible information content and were therefore removed from the analysis. The article "an" occurs very rarely and so was not removed.

All listening tests received prior approval from the University of Liverpool Committee on Research Ethics.

C. Implementation of metrics

In this section, metrics are introduced that consider a clean signal, x, and a degraded or processed signal, y, where m and j are used to denote frame and frequency band, respectively, and n denotes the short-time region of the signal. Furthermore, M, J, and N denote the total number of frames, number of bands, and number of frames within a region, respectively. Metric indices were averaged over the ten sentences within each IEEE word list.

1. STOI and STOI+

STOI is based on the correlation between the envelopes of clean and degraded speech signals (10 kHz sampling rate) decomposed into regions that are approximately 386 ms (30 samples) in length. As described by Taal *et al.*,[14] the output of STOI, d, takes values $-1 < d \leq 1$ but is in practice nonnegative and has a monotonic relationship with speech intelligibility scores. Signals x and y are divided into Hanning windowed frames with 50% overlap, and where the energy of each x frame is more than 40 dB below the maximum clean speech energy, both the x frame and the corresponding y frame are discarded. Subsequently, a short-time discrete Fourier analysis is undertaken, where the frequency bins are grouped into 15 one-third octave bands with centre frequencies from 150 to 3800 Hz. Within each frequency band and region, the degraded signal energy is normalised and clipped. Normalisation is performed to compensate for global level differences, which are assumed to have a limited effect on intelligibility.[14] As mentioned, clipping is performed to limit the sensitivity of the model toward severely degraded or noise-only time-frequency units—according to Taal *et al.*[14]—and place a lower bound on the SDR. This was determined by Taal and colleagues to be optimal for their noisy and ITFS-processed speech corpus on the basis of results for the Dantale II corpus, which used one female Danish talker. Subsequently, the correlations between signals in each band and each region are calculated, and the correlation coefficients are averaged to obtain d. In this paper, STOI was calculated using publicly available code from Taal *et al.*[14]

After short-time Fourier transformation of x and y, short-time (386 ms) temporal envelopes in each band and frame are denoted $X_{j,m}$ and $Y_{j,m}$, where each short-time region has a length $N = 30$. A short-time region of the clean speech signal can be represented in vector notation as $X_{j,m} = [X_j(m-N+1), X_j(m-N+2), \ldots, X_j(m)]^T$.

The normalisation factor, α, is calculated for each region and band as shown in Eq. (1),

$$\alpha_{j,m}(n) = \sqrt{\frac{\sum_{m-N+1}^{m} X_{j,m}(n)^2}{\sum_{m-N+1}^{m} Y_{j,m}(n)^2}}. \tag{1}$$

$Y_{j,m}(n)$ is multiplied by α to obtain normalised $Y'_{j,m}(n)$, which can be represented as $X_{j,m}/Y_{j,m}Y_{j,m}(n)$, where $\|\cdot\|$ indicates the $l2$ norm. $Y'_{j,m}(n)$ is then clipped to obtain $\overline{Y}_{j,m}(n)$ using Eq. (2), where $\beta = -15$ dB,

$$\overline{Y}_{j,m}(n) = \min\left(Y'_{j,m}(n),\ \left(1 + 10^{-\frac{\beta}{20}}\right)X_{j,m}(n)\right). \tag{2}$$

This means that any $Y'_{j,m}$ that comprises values close to zero in any band, j, will result in $\overline{Y}_{j,m}(n) = \left(1 + 10^{-\beta/20}\right)X_{j,m}(n)$. Taal *et al.*[14] state that clipping is performed to place a lower bound on the SDR at -15 dB, where SDR is defined as

$$SDR_{j,m}(n) = 10\log_{10}\left(\frac{X_{j,m}(n)^2}{\left(Y'_{j,m}(n) - X_{j,m}(n)\right)^2}\right). \tag{3}$$

The correlation between the signals in each frame and band is given by

$$d_{j,m} = \left(\frac{\left(X_{j,m} - \mu_{X_{j,m}}\right)^T \left(\overline{Y}_{j,m} - \mu_{\overline{Y}_{j,m}}\right)}{\|X_{j,m} - \mu_{X_{j,m}}\|\,\|\overline{Y}_{j,m} - \mu_{\overline{Y}_{j,m}}\|}\right), \tag{4}$$

where $\mu_{X_{j,m}}(\cdot)$ and $\mu_{Y_{j,m}}(\cdot)$ are sample averages of the vectors $X_{j,m}$ and $\overline{Y}_{j,m}$. When clipping is not performed, normalisation has no effect on the correlation coefficients.

STOI+ was calculated as in the case of STOI but without normalisation and clipping. The effect of the normalisation and clipping procedure at the level of the 386 ms region is illustrated in Fig. 2 for global SNRs of 0 and -20 dB. For $SNR = 0$ dB, there is only a small increase in the intermediate correlation coefficient after clipping. However, for $SNR = -20$ dB, the intermediate correlation coefficient changes from 0.02 before clipping, indicating no correlation, to 0.54 after clipping, indicating a moderate positive correlation. Given such findings, one motivation of this paper is to assess whether removing the normalisation and clipping procedure reduces the prediction error for additive WGN and low SNRs.

For STOI and STOI+, correlation coefficients were averaged over all J bands and M frames for all possible 386 ms regions to obtain the final value, d, as given by

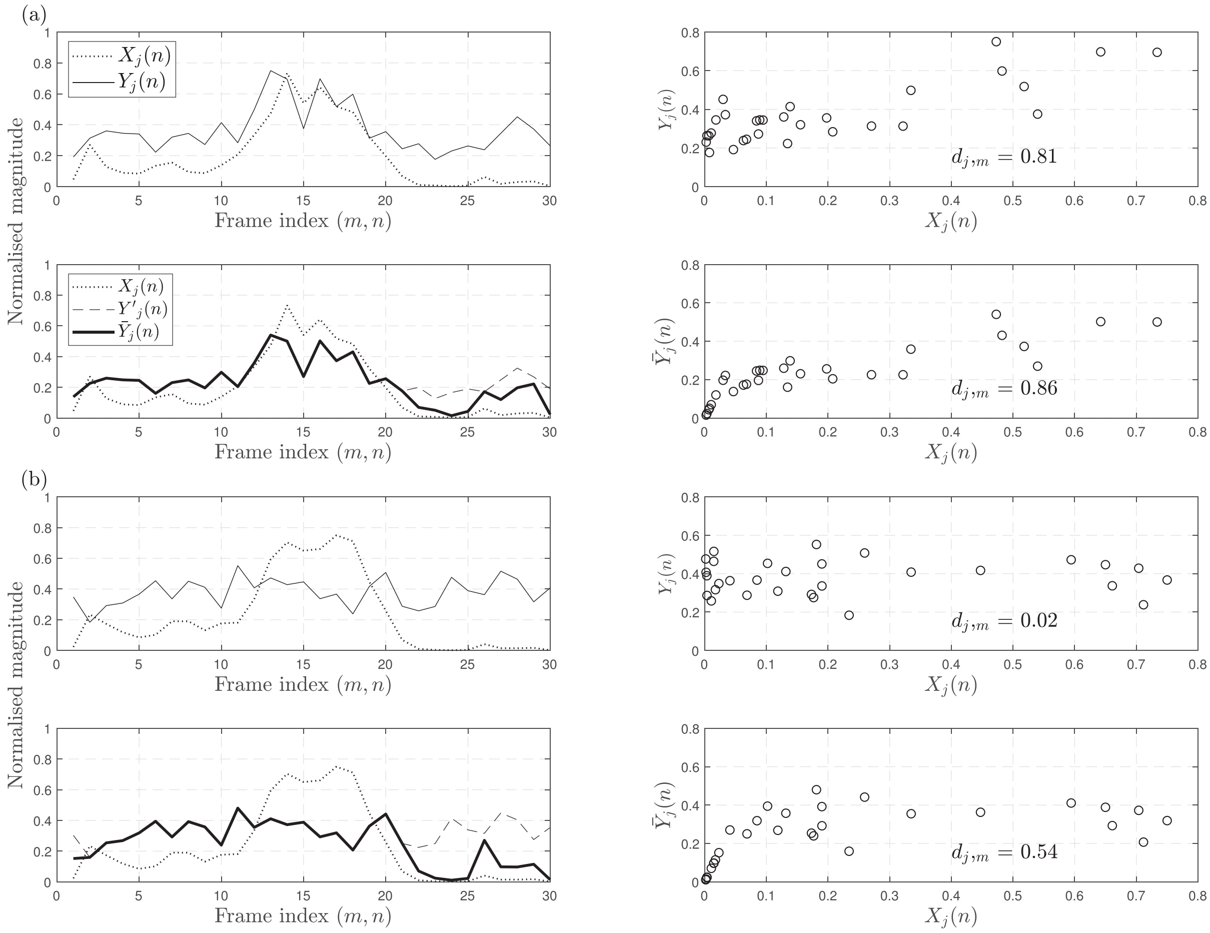

FIG. 2. Examples showing the effect of the normalisation and clipping procedures in STOI on a clean speech vector, $X_j(n)$, together with a normalised, $Y'_j(n)$, and clipped, $\bar{Y}_j(n)$, degraded speech vector in one frequency band, j: (a) $SNR = 0$ dB and (b) $SNR = -20$ dB. The intermediate correlation coefficient, $d_{j,m}$ is reported before and after normalisation and clipping for both SNRs.

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m}. \tag{5}$$

As the relationship between STOI-based measures and intelligibility scores is monotonic, as mentioned, and in order to predict absolute intelligibility, STOI-based values were converted to mapped values via a logistic function. This linearises the relationship between STOI-based measures and intelligibility scores and therefore allows the reporting of linear correlation coefficients and the determination of the prediction error distribution. The logistic function maps the variable d (representing STOI or STOI+) with the free parameters, a (slope) and b (centre), as follows:

$$f(d) = \frac{100}{1 + \exp(ad + b)}. \tag{6}$$

Free parameter values a and b were derived from the data under each filter and gender condition using a non-linear least squares procedure with starting values derived from Taal *et al.*[14] In all cases in this paper, mapping was performed by means of the *lsqcurvefit* function in MATLAB.

2. ESTOI

Jensen and Taal[22] proposed ESTOI as a measure to improve on STOI in the case of highly modulated noise sources, but also to work well under other noise conditions. Like STOI, ESTOI operates within a 384 ms analysis region on amplitude envelopes of clean and degraded signals, but as mentioned, it does not use the clipping procedure. Publicly available code was used in this study.[22] Signals are passed through a one-third octave filterbank, and temporal envelopes are extracted in each frequency band. The resulting row- and column-normalised short-time envelope spectrograms are decomposed into orthogonal one-dimensional subspaces, which are assigned intelligibility scores. Intermediate intelligibility scores derived from these subspace intelligibility scores are averaged to obtain the final intelligibility index, d. ESTOI is mapped using the logistic function given in Eq. (6). For details of the procedure, see Jensen and Taal.[22]

3. NCM

NCM was calculated using publicly available code.[37] This measure is based on apparent SNRs within frequency bands that are calculated on the basis of the squared normalised covariance—hence, correlation—between the envelopes of x and y. The covariance in each frequency band is used to derive an apparent or modulation signal-to-noise ratio (aSNR), which is treated in the manner of SNR values in the STI method to derive a final, band-weighted value, $0 \leq NCM \leq 1$.

Signals x and y are bandpass filtered into 20 frequency bands with centre frequencies ranging from 335 to 6910 Hz with eighth-order Butterworth filters. The signal envelopes are extracted with the Hilbert transform and smoothed by low-pass filtering and downsampling to 32 Hz to limit envelope modulation frequencies to ≤16 Hz. In each frequency band, j, the aSNR of the entire envelope is calculated using

$$aSNR_j = 10\log_{10}\left(\frac{{r_j}^2}{1-{r_j}^2}\right), \tag{7}$$

where r_j is the normalised covariance between x_j and y_j. The remaining calculations are consistent with the standard STI procedure. The aSNR is clipped to values ±15 dB to obtain the transmission indices. Using (interpolated) standard ANSI S3.5[5] weighting for short passages, the sum of the weighted values is divided by the sum of the weights to obtain the final NCM value of between 0 and 1. Logistic mapping was performed after Taal *et al.*[14] using Eq. (6).

4. NSEC

Boldt and Ellis[31] developed NSEC based on the correlation of the envelopes of the original speech and the degraded speech after time-frequency decomposition, equalisation of energy in frequency bands, amplitude compression, and Direct Current (DC) component removal. In this implementation, the energy envelopes are derived with a 16 channel gammatone filterbank with centre frequencies from 80 Hz to 8 kHz, equally spaced on the equivalent rectangular bandwidth (ERB) scale, and with a window length of 0.08 s with a 50% overlap.

With STOI, the irrelevance to intelligibility of high energy regions of y where x is low in energy is accounted for by removing these regions before calculating the correlation. In the case of NSEC, the same issue is addressed by normalisation, by dividing by the Frobenius norm of x and y [see Eq. (2) in Boldt and Ellis[31]]. Hence, NSEC is bounded between zero and one. The original mapping function proposed by Boldt and Ellis is given as

$$f(x) = \frac{1}{1+\exp((b-x)/a)}. \tag{8}$$

However, Taal *et al.*[17] obtained better performance with the following equation, which was applied in this paper:

$$f(x) = \frac{100}{1+(ax+b)^c}. \tag{9}$$

For details of the NSEC algorithm, see Boldt and Ellis.[31]

5. CSII

CSII was originally developed for predicting the speech intelligibility of peak- or centre-clipping distortions, such as those associated with hearing aids.[12] CSII assesses the coherence of the clean and degraded/processed signals on the basis of the magnitude squared coherence function. In later work, CSII was separated into three, separate indices, $\mathrm{CSII_{High}}$, $\mathrm{CSII_{Mid}}$, and $\mathrm{CSII_{Low}}$, based on the root mean square (rms) level of the signal envelope.[38] The $\mathrm{CSII_{High}}$ index is associated with segments at or above the overall rms level of the signal, the $\mathrm{CSII_{Mid}}$ index is associated with segments at or up to 10 dB below the same level, and the $\mathrm{CSII_{Low}}$ index is associated with segments from 10 to 30 dB below the level. Each Hanning windowed frame of the signal envelopes is assigned to one of the three amplitude regions. $\mathrm{CSII_{Low}}$ and $\mathrm{CSII_{Mid}}$ are combined linearly and transformed with a simple logistic function to derive a fourth measure, termed *I3*. In this paper, the short-time CSII implementation developed by Loizou[37] was used, in which CSII was averaged over short-time segments of 30 ms in length with a 25% window skip rate. In addition, the critical band weighting function of NCM and CSII was set to ANSI S3.5 weighting, as the masker is stationary.

Preliminary testing indicated that $\mathrm{CSII_{Low}}$ performed poorly and $\mathrm{CSII_{I3}}$ performed no better than $\mathrm{CSII_{Mid}}$ and so were not considered further in this paper. The best fitting nonlinear function was found for $\mathrm{CSII_{High}}$ and $\mathrm{CSII_{Mid}}$ measures from the following set: the original function used for STOI, as shown in Eq. (6); the second function provided by Taal *et al.*,[39] as shown in Eq. (10); and a linear fit,

$$f(x) = \frac{100}{1+(ax+b)^c}. \tag{10}$$

The prediction error indicated that Eq. (6) tended to perform as well as or better than these alternatives. Hence, the same logistic model was fit to $\mathrm{CSII_{High}}$ and $\mathrm{CSII_{Mid}}$ as to STOI, STOI+, ESTOI, and NCM.

6. Speech-based STI

The envelope regression-based approach to the speech-based STI developed by Payton and Shrestha[32] and derived from earlier work by Ludvigsen *et al.*[10] and Goldsworthy and Greenberg[11] and implemented in the AARAE toolbox for MATLAB (Cabrera *et al.*[40]) were used in this paper. Signals x and y are filtered by a bank of six sixth-order Butterworth octave band filters with centre frequencies from 125 Hz to 4 kHz. To extract the 8 kHz band, a sixth-order Butterworth HPF is used with a cutoff frequency of 6 kHz. For each frequency band, j, the intensity envelopes of x and y are extracted and downsampled to reduce the computation time. For each octave band, a modulation metric is calculated on the basis of a comparison of the intensity envelopes

https://doi.org/10.1121/10.0003557

with a rectangular window length set to 1 s and a 75% overlap and where the output, MOD_j, is normalised by the term μ_{xj}/μ_{yj}. When using such a window length (which is adequate for stationary noise), STI derived by this method approaches the values derived from the "true" STI and the long-term STI method derived using the magnitude cross-power spectrum.[32] The aSNR is calculated as in Eq. (7) but replacing the term r_j^2 with MOD_j. Subsequently, as in NCM, the aSNR is clipped to values ±15 dB to obtain the transmission indices. Finally, the overall STI value is calculated as a weighted sum of these transmission indices, where the weights and redundancy correction factors are as specified in IEC 60268–16.[41] For the intelligibility scores presented in this paper, there was no clear improvement in correlations between predicted and measured scores when using the 90th percentile rather than the mean STI results, so the mean results are reported in this paper (cf. Opsata *et al.*[42]) However, their environments differed in that they were reverberant, with low background noise.

D. Evaluation procedures

Objective measures were compared on the basis of summary statistics such as minimum and maximum value, correlation coefficients, estimates of the prediction error, and estimates of metric bias and reliability. The distribution of metric values relative to intelligibility scores was also considered.

The figures of merit included Pearson's product-moment (ρ) and Kendall's tau (τ) correlations between the metrics and intelligibility scores and the standard deviation of the prediction error (σ_e). A significant difference in metric performance can be expressed in terms of non-overlapping confidence intervals for ρ. After Ma *et al.*,[13] the standard deviation of the prediction error was calculated using

$$\sigma_e = \sigma_d\sqrt{1-\rho^2}, \tag{11}$$

where σ_d is the standard deviation of the intelligibility scores in a given condition. Figures of merit, ρ and σ_e, were applied to the mapped objective scores (with the exception of STI), while τ is rank based and therefore independent of the mapping.

Metric bias and reliability were calculated after Hilkhuysen *et al.*[43] To compute metric bias, b, both per SNR and across SNRs, the measured scores, v, were subtracted from the corresponding predicted scores, w. Similarly, the mean bias, $\overline{b}$, was calculated using

$$\overline{b} = \frac{1}{C}\cdot\sum_{i=1}^{C}(w_i - v_i), \tag{12}$$

where C is the number of measured scores.

Predicted scores were mapped metric values for all metrics other than STI, and unmapped metric values for STI, multiplied by 100 if a fraction. In boxplots of the prediction bias for each metric, the interquartile range, indicated by the length of the box, and the length of the box whiskers, which extend to approximately $\pm 2.7\sigma$ for a normal distribution, indicate the reliability of the predictions, with smaller boxes and shorter whiskers indicating higher reliability. The position of the box plus whiskers indicates overall prediction bias, with positions above the zero line indicating metrics that overpredict intelligibility and positions below the zero line indicating underprediction.

Logistic regression models were fitted via the *glm* function in R software[44] (version 3.5.1) to the word recognition scores expressed as the number of words correctly identified ("successes") and the number of words incorrectly identified ("failures") and with talker gender, and SNR and filter condition and their interaction, as fixed effects. The resulting logistic regression model can be described as follows:

$$logit(p) = \beta_0 + \beta_1 SNR + \beta_2 Filter + \beta_3 Gender + \beta_4 SNR\cdot Filter + e, \tag{13}$$

where p is a probability, SNR is treated as a discrete variable, *Filter* indicates filter condition (non-HPF = 0, HPF = 1), and *Gender* indicates talker gender (male = 0, female = 1). The reference levels were $SNR = -17$ dB (justified by the results in Sec. III A), non-HPF, and male. As nested model comparisons using likelihood ratio tests indicated that there was an interaction of SNR and filter and therefore to provide statistical information about the effects of the filter at each SNR, it was necessary to limit the number of SNR levels to be included in the model (due to complexity of interpretation and limited space). As median intelligibility scores at $SNR < -17$ dB were close to zero, only SNR levels equal to or greater than −17 dB were included. The Tukey method was used to conduct *post hoc* pairwise tests of SNR and filter. Adjusted p values were calculated using the Bonferroni method. Random effects were not incorporated into the model for reasons of interpretability (i.e., so that the coefficients did not have an interpretation conditional on the random effects). Note that the reduced range of SNRs from −17 to 0 dB is used only in the logistic regression model, unless stated otherwise.

III. RESULTS

A. Intelligibility scores

Intelligibility scores computed as percentages of words correctly identified per wordlist for a given talker and listener combination are shown in Fig. 3. These scores extend from 0 to 98% to allow investigation of the relationship between each metric and intelligibility score over the full range of scores in Sec. III B. For SNRs between −26 and −8 dB, the median scores are ≤20%, which is the region of particular interest for speech security. The 50% speech reception threshold (SRT) is −4.1 dB for male talkers and −4.7 dB for female talkers in the non-HPF condition and is −3.8 dB for male talkers and −3.3 dB for female talkers in the HPF condition. In the non-HPF condition, the maximum percentage of words correctly identified is 4.5% (three words) at −20 dB SNR and 11% (eight words) at −17 dB SNR. Even at SNRs of −26 and −23 dB, words were identified in the non-HPF condition: 1.6%, or one word.

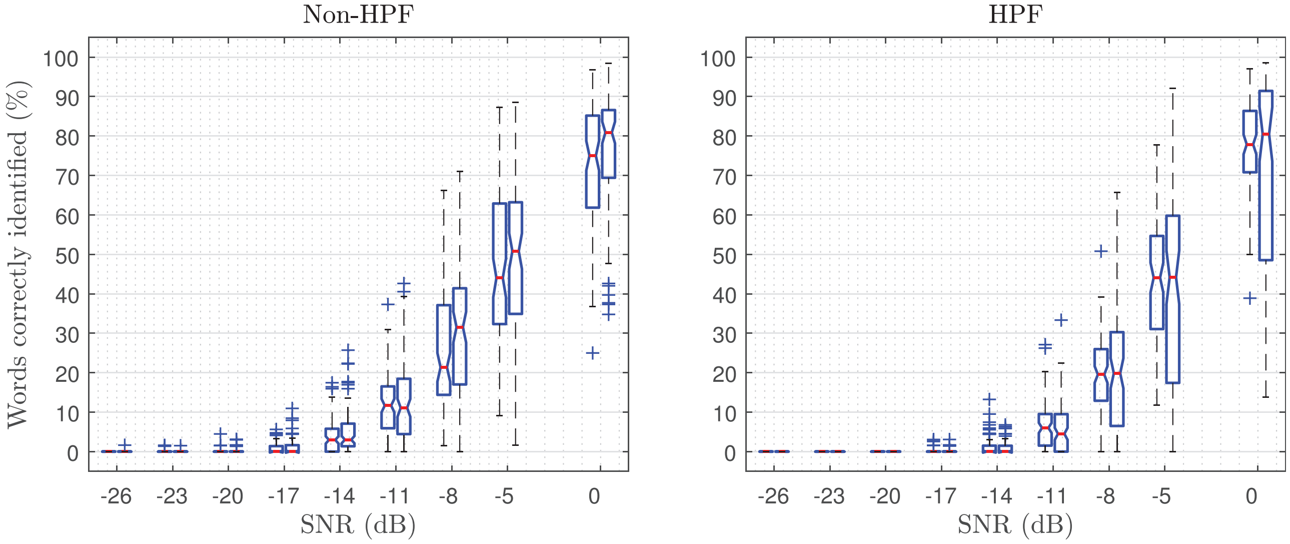

FIG. 3. (Color online) Boxplots of words correctly identified by SNR and filter condition. At each SNR, the left- and right-side box and whisker correspond to male and female talkers, respectively. At SNRs below −17 dB, at least one word was identifiable in the non-HPF condition but not in the HPF condition. At SNRs between −8 and 0 dB, the whiskers (± 2.7σ assuming a normal distribution) cover a range of words correctly identified of at least 40% in both filter conditions.

A logistic regression model is fitted for WGN mixed with non-HP-filtered and HP-filtered speech with effects of SNR, filter, and talker gender and the interaction of SNR and filter (see Table I). Model coefficients (described in Table I as *estimates*) are log odds. The p values indicate the probability of obtaining the observed effect (or larger) under a null hypothesis. The model output indicates that $SNR = -17$ dB is associated with reduced log odds of identifying a word correctly compared with higher SNRs, as would be predicted. At $SNR = -17$ dB, the log odds are approximately −2.05 when speech is HP-filtered relative to non-HP-filtered, i.e., the odds of identifying a word correctly decrease by about 87%. The log odds are 0.05 when the speech is produced by a female vs a male talker, i.e., the odds of identifying a word correctly increase by about 5%. The approximate R^2 derived from the full model deviance and the null model deviance is 0.80, or 80%.

A likelihood ratio test of nested models with and without the interaction of SNR and the filter HPF condition was significant ($p < 0.0001$). To evaluate the interaction, *post hoc* Tukey tests were run with p values adjusted for the number of comparisons. In this context, the concern is whether at a given SNR there is an effect of the HPF. At all SNRs considered in the model except 0 dB, the log odds of identifying a word correctly are lower in the HPF condition than in the non-HPF condition, with the log odds decreasing as the SNR is lowered. The result for $SNR = -17$ dB has already been reported. At $SNR = -14$ dB, the log odds decreased by 1.48 [standard error $(SE) = 0.10$, $z = -15.32$, $p < 0.0001$)]; at $SNR = -11$ dB, the log odds decreased by 0.78 ($SE = 0.05$, $z = -17.11$, $p < 0.0001$); at $SNR = -8$ dB, the log odds decreased by 0.45 ($SE = 0.03$, $z = -15.45$, $p < 0.0001$); and at $SNR = -5$ dB, the log odds decreased by 0.24 ($SE = 0.03$, $z = -9.70$, $p < 0.0001$). At $SNR = 0$ dB, there is no difference between filter conditions ($p = 1$). In sum, the HPF does not improve the intelligibility of speech mixed with WGN at $-17 \leq SNR \leq 0$ dB.

TABLE I. Logistic regression model output for WGN mixed with non-HP-filtered and HP-filtered. The interaction of SNR and filter is discussed in Sec. III A.

	Estimate	SE	z	p
(Intercept)	−4.66	0.09	−51.06	<0.0001
−14 dB SNR	1.58	0.10	15.71	<0.0001
−11 dB SNR	2.69	0.09	28.41	<0.0001
−8 dB SNR	3.73	0.09	40.09	<0.0001
−5 dB SNR	4.54	0.09	49.01	<0.0001
0 dB SNR	5.72	0.09	61.39	<0.0001
Filter HPF	−2.05	0.27	−7.71	<0.0001
Gender female	0.05	0.01	3.10	<0.001

B. Objective intelligibility metric results

In Fig. 4, the relationship between each metric and intelligibility score is shown per talker gender for the non-HPF filter and HPF filter conditions. With the exception of STI, the fitted lines derive from the logistic functions described in Sec. II C. The values for the free parameters a and b—and c for NSEC—are provided in the Appendix. A linear fit is assumed for STI as indicated for sentence material in ISO 9921.[45] For the purposes of illustration, the fitted lines extend to zero and one for all metrics except STI. The prediction bounds provide the interval with a 95% level of confidence for a single intelligibility score given a single metric value. Note that when the slope of the fitted line is relatively steep, as in the case of STOI and $CSII_{Mid}$, the bounds associated with predicting an intelligibility score from a single metric value may be relatively wide.

Descriptive statistics on the different metric values are given in Table II to accompany the scatterplots (Figs. 5–8) of the metrics by intelligibility scores. In these plots, the fitted lines represent the best nonlinear least squares fit given the logistic functions described in Sec. II C, with the exception of

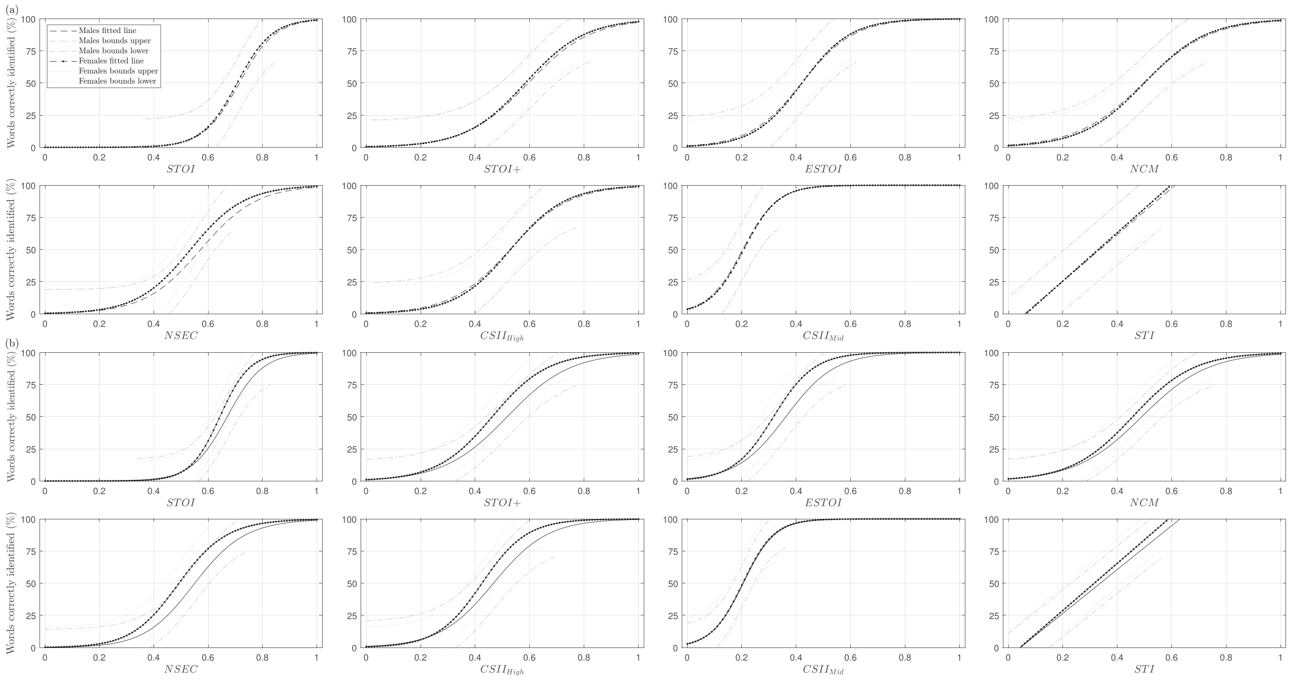

FIG. 4. Relationship between metrics and measured intelligibility scores in the (a) non-HPF and (b) HPF conditions per talker gender. These are shown with 95% prediction bounds, which, apart from STI, vary across the range of metric values. For fitted lines that have intelligibility scores close to 0%, the upper prediction bound tends to be higher in the non-HPF condition.

STI. Figure 5 shows the scatterplots for STOI and STOI+. Although intelligibility scores extend from 0 to 98%, STOI and STI cover a range of 0.52 and 0.56, respectively. This is not problematic if mapping functions are always used between the metric and the words correctly identified. However, for some indicators, such as STI, there is an expectation that a simple intelligibility rating (e.g., “bad,” “fair,” “excellent”) can be assigned to values between zero and one. STOI has the highest minimum value of 0.34, whereas the lowest value for all other metrics is zero, or close to zero. In contrast, STOI+ has the largest range (0–0.83) of all metrics considered. Accordingly, STOI+ is associated with shallower slopes and a lower sigmoid centre than STOI. The slope is similar or slightly steeper for HP-filtered than non-HP-filtered speech. ESTOI and NCM results are shown in Fig. 6, with NCM displaying a clear discontinuity for female speech in the region of intelligibility scores of 75%. ESTOI starts at zero and covers a range of 0.62. Both NCM and NSEC (shown in Fig. 9) metrics have a range from 0 up to ≈0.75, which is similar to that of STOI+ and $\text{CSII}_{\text{High}}$.

TABLE II. Summary of metric statistics for the non-HPF and HPF conditions.

Metric	Minimum	Median	Mean	Maximum	Interquartile range	Range
STOI	0.34	0.50	0.54	0.87	0.18	0.52
STOI+	0	0.20	0.27	0.83	0.36	0.83
ESTOI	0	0.13	0.18	0.62	0.25	0.62
NCM	0	0.14	0.21	0.76	0.33	0.76
NSEC	0	0.24	0.28	0.75	0.37	0.75
$\text{CSII}_{\text{High}}$	0	0.21	0.25	0.77	0.30	0.77
CSII_{Mid}	0	0.04	0.09	0.36	0.12	0.36
STI	0.01	0.10	0.15	0.56	0.20	0.56

Comparing $\text{CSII}_{\text{High}}$ and CSII_{Mid} in Fig. 7, the former covers a wider range of values and therefore is associated with shallower slope values. $\text{CSII}_{\text{High}}$ varies from 0 to 0.77, while CSII_{Mid} only covers a range from 0 to 0.36. CSII_{Mid} has a discontinuity in the data for values from 0.21 to 0.22; this is most evident for non-HP-filtered speech. STI, shown in Fig. 8, extends to only 0.56, which corresponds to a 100% sentence score and an intelligibility rating of “fair” for the original STI method (see ISO 9921[45]).

Figures of merit are reported in Table III for each metric per talker gender and filter condition. All correlation tests were significant at $p < 0.001$. For non-HP-filtered male speech, the 95% confidence intervals for ρ overlap for STOI, STOI+, NCM, and NSEC for male talkers non-HP-filtered speech, while NSEC has a higher ρ than ESTOI, $\text{CSII}_{\text{High}}$, CSII_{Mid}, and STI. STOI+ and NCM also outperform STI. For HPF male speech, NSEC has a higher ρ than STOI and ESTOI, while NSEC, STOI+, and NCM have a higher ρ than $\text{CSII}_{\text{High}}$ and STI. However, ρ is less useful in identifying differences in the other situations. For non-HP-filtered speech, the highest Kendall’s τ occurs with NCM and NSEC for male talkers and STOI+, NCM, and $\text{CSII}_{\text{High}}$ for female talkers. The lowest prediction error occurs with NSEC for male talkers and NCM for female talkers. For the HPF condition, the highest Kendall’s τ value occurs with NCM, NSEC, and CSII_{Mid} for male talkers and NCM for female talkers. The lowest prediction error for male talkers occurs with NSEC and for female talkers with STOI+. Across all conditions, STOI+ is associated with a lower prediction error than STOI, and in all conditions except female non-HPF, STOI+ is associated with a lower prediction error than ESTOI.

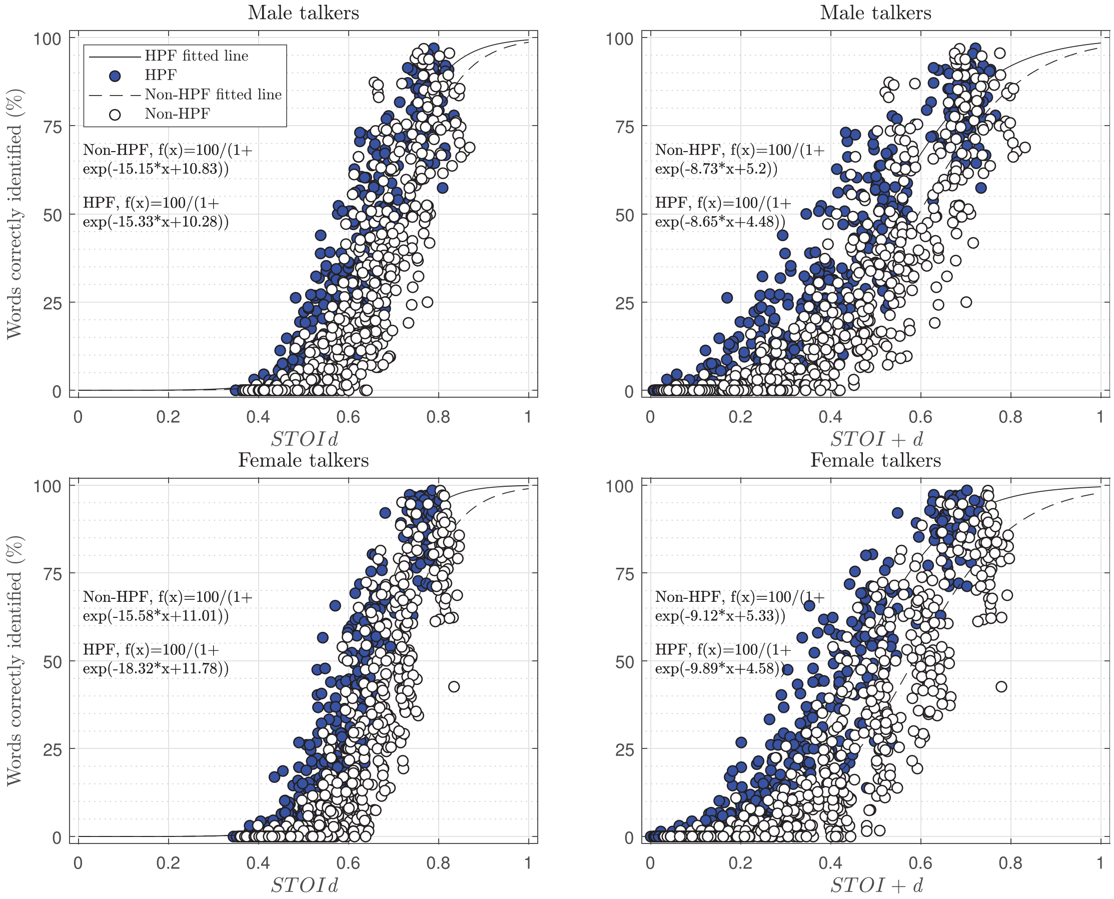

FIG. 5. (Color online) Scatterplots of STOI and STOI+ by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function per talker gender (males, upper; females, lower). Markers represent average metric values over the ten sentences within each IEEE word list.

In case the inclusion of large numbers of intelligibility scores at or close to zero affected relative metric performance, this comparison of metrics was repeated using only SNRs from −17 to 0 dB (these values being identical to the logistic model SNR values). Relative performance was nearly identical with the exception that STI performance tended to improve slightly in the non-HPF filter condition. However, it was still amongst the worst performers. NSEC, NCM, and STOI+ were associated with the lowest prediction error across both analyses.

Prediction bias and reliability (as described in Sec. II D) is shown for each metric across talkers and SNRs in Fig. 9. For these experimental conditions, bias is typically positive, with the exception of STI, in which case the interquartile range spans zero. In the non-HPF condition, NSEC and especially STI are shown to be relatively unreliable for prediction purposes, as indicated by the large interquartile ranges, while for both male and female talkers, STOI+ is associated with the lowest median and mean bias, although STOI, ESTOI, NCM, NSEC, and $\mathrm{CSII}_{\mathrm{High}}$ are also associated with relatively low median bias. In the HPF condition, NSEC and $\mathrm{CSII}_{\mathrm{High}}$ are associated with the lowest median and mean bias, $\mathrm{CSII}_{\mathrm{Mid}}$ and STI with the highest mean bias, and $\mathrm{CSII}_{\mathrm{Mid}}$ with the highest median bias. ESTOI bias is also relatively high. STI is least reliable for prediction (i.e., it has the largest interquartile range), and NCM is most reliable. Overall, regarding bias and reliability, performance tends to be poorest for STI, NSEC, and $\mathrm{CSII}_{\mathrm{Mid}}$ in the non-HPF condition and STI and $\mathrm{CSII}_{\mathrm{Mid}}$ in the HPF condition.

As the SNR decreases from −17 to −26 dB, the differences between the metrics in prediction bias increase: prediction bias is particularly large for $\mathrm{CSII}_{\mathrm{Mid}}$ and STI, which overpredict intelligibility. In the case of STOI, there is less reliability at $SNR < -17$ dB than for other metrics.

IV. DISCUSSION

A. Effect of SNR and high-pass filtering of speech on intelligibility scores

The results confirm that the intelligibility of noisy speech decreases as a sigmoidal function of mixture SNR. The maximum score is 98% with or without HPF, and at $SNR = 0$ dB, scores exceeded 80%. In the context of speech security, the acceptable percentage of words that are correctly identified tends to be between 0 and 20%. In this work, the median intelligibility scores achieve or exceed 20% at $SNR = -8$ dB, which confirms the need to extend the evaluation of metrics to SNRs below −10 dB.

It was noted that even at SNRs of −26 and −23 dB, words were identified in the non-HPF condition: 1.6%, or one word. In a security context, these low percentages require consideration. These words occurred near the

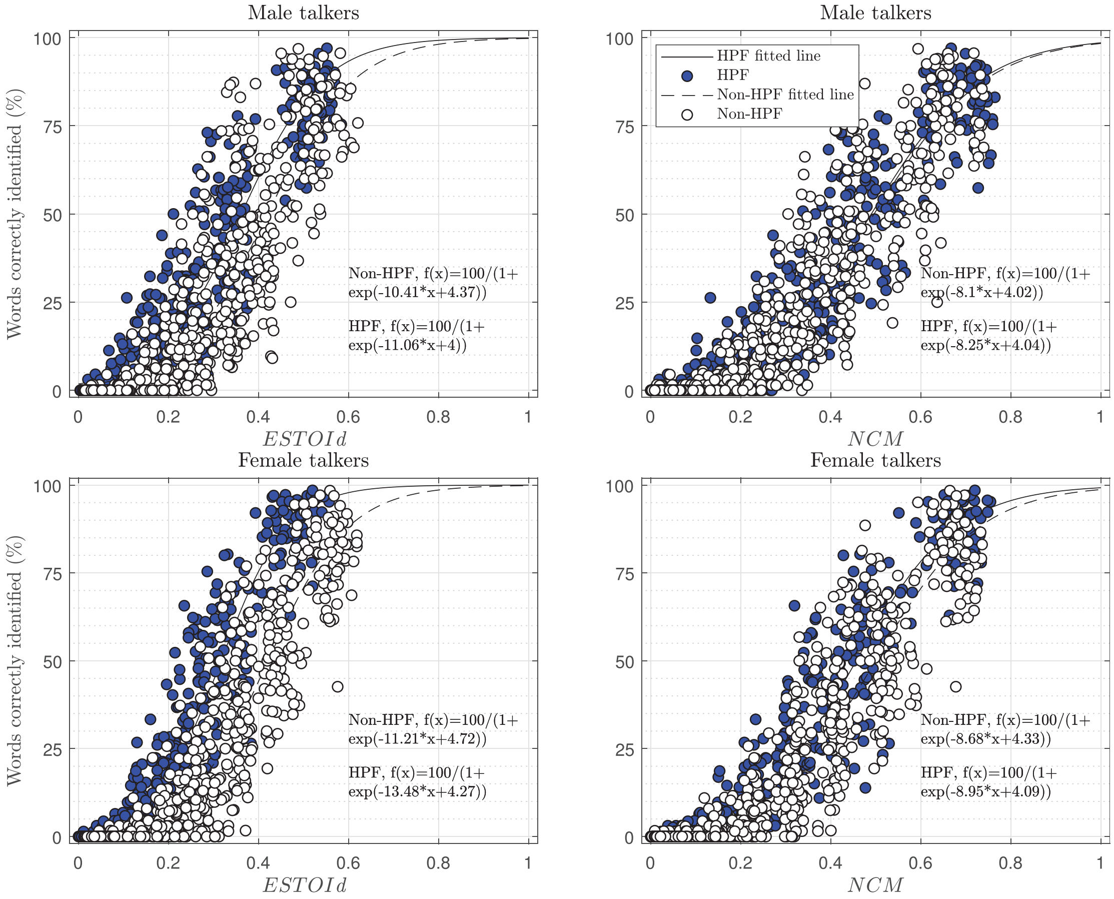

FIG. 6. (Color online) Scatterplots of ESTOI and NCM by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function per talker gender. Markers represent average metric values over the ten sentences within each IEEE word list.

beginning of the sentence within a noun phrase in subject position in the relevant sentences and are monosyllabic, so they take prominence/stress in British English, which is cued by loudness and length. These factors, local SNR and duration, are likely to have allowed the listeners to obtain "glimpses" of these words in the presence of the competing white Gaussian noise.

One aim of the study was to determine whether the HPF improves the intelligibility of speech for $SNR < -10$ dB. Recall that the HPF flattens the speech spectrum but does not strongly attenuate low frequencies ($f < 300$ Hz), unlike the traditional high-pass Butterworth filter method (e.g., Skowronski and Harris[30]) In this study, a logistic regression model and associated *post hoc* tests indicate that when $SNR = 0$ dB, there is no reliable effect of the HPF on speech intelligibility. Likewise, median intelligibility scores close to zero for $SNR < -17$ dB indicate that the HPF has no effect at these SNRs. However, the HPF is detrimental to speech intelligibility for $-17 < SNR < -5$ dB. These results suggest that, when speech is mixed with WGN at these global SNRs, the local SNR is not sufficiently improved by the HPF at higher speech frequencies, i.e., within the range of the second and third formants, to increase intelligibility for the average listener.

As suggested in Sec. I, the HPF increases the energy in the mid- to high-frequency range (1–4 kHz) relative to the low frequency range (less than 1 kHz). An increase in the proportion of speech energy in the mid- to high-frequency range relative to the low frequency range is known to increase intelligibility in noise. However, WGN masks the mid- and high-frequency components of speech, and the ear integrates more noise energy per auditory band at higher frequencies than at lower frequencies for this noise type. Hence, at relatively low SNRs ($SNR < 0$ dB), the HPF does not provide an intelligibility benefit.

Skowronski and Harris[30] found that their high-pass filter improved speech intelligibility at $SNR = -10$ dB for 6 of their 16 speakers. However, they used speech materials that consisted of closed sets of two, four, or ten confusable items rather than open sets, as in the current study. Hence, an SNR of -10 dB in their study is not equivalent to the same SNR in the current study.

B. Evaluation of intelligibility metrics

For the purposes of speech security, the fitted curve for a metric should ideally have a slowly rising exponential curve from the point at which the intelligibility score is zero, leading to a shallow slope for the linear region where there are intermediate intelligibility scores. In addition, narrower prediction bounds are preferred. These requirements are satisfied by STOI+, NCM, NSEC, and $\mathrm{CSII}_{\mathrm{High}}$, of which NSEC has the lowest upper prediction bound

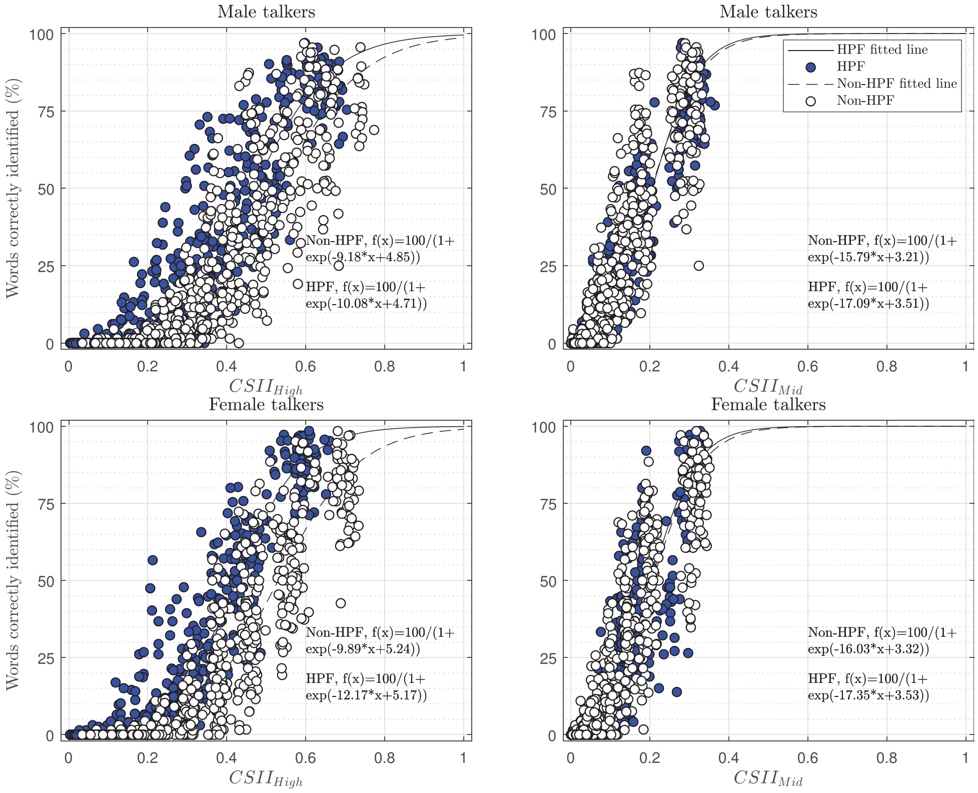

FIG. 7. (Color online) Scatterplots of $\mathrm{CSII}_{\mathrm{High}}$ and $\mathrm{CSII}_{\mathrm{Med}}$ by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function per talker gender. Markers represent average metric values over the ten sentences within each IEEE word list.

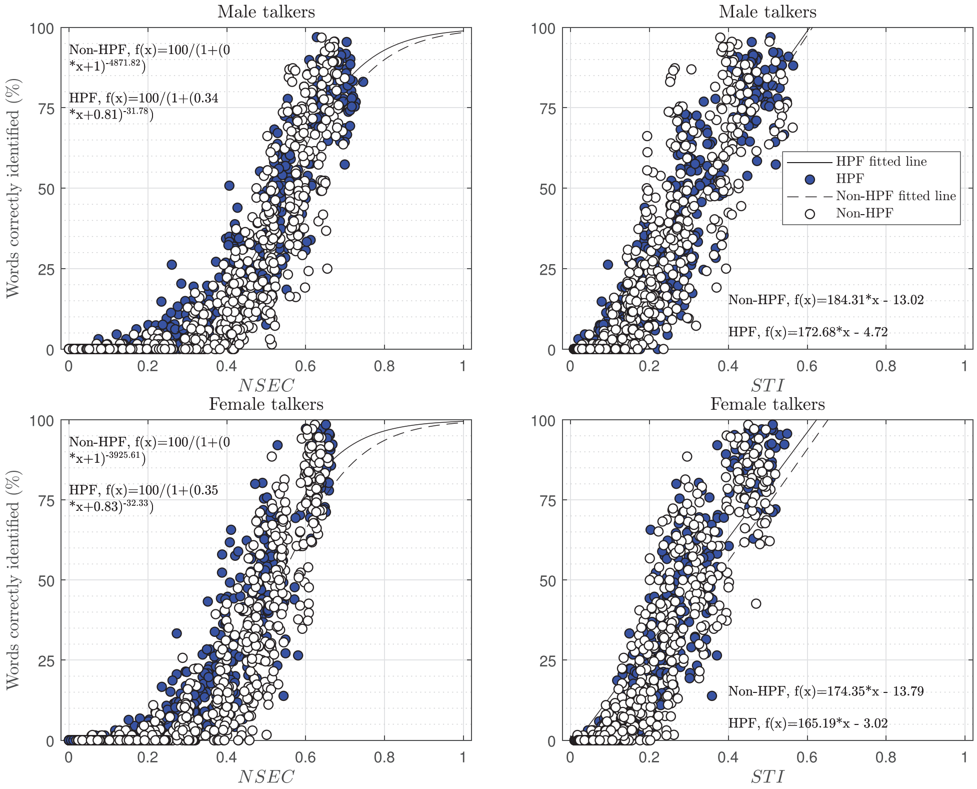

FIG. 8. (Color online) Scatterplots of NSEC and speech-based STI by intelligibility scores with fitted lines deriving from a rotationally symmetric logistic function and a simple linear function, respectively, per talker gender. Markers represent average metric values over the ten sentences within each IEEE word list.

when the metric is zero (see Fig. 4). The prediction bounds for these metrics tend to be narrowest for the linear region and widest where intelligibility scores are below 20%; in contrast, STI (with a non-sigmoidal fit) has relatively uniform prediction bounds across the range of metric values.

When comparing metrics on the basis of summary statistics and the distribution of metric values relative to intelligibility scores, STOI has one of the smallest ranges and the highest minimum value (0.34). ESTOI, $\mathrm{CSII}_{\mathrm{Mid}}$, and STI also have relatively small ranges (see Table II). In contrast, STOI+ varies from 0 to 0.83. Of course, higher STOI and STOI+ values would be expected to occur when $SNR > 0\,\mathrm{dB}$.

Under some experimental conditions, Payton and Shrestha[32] found that their STI ranged from zero to one. However, in this study, STI did not exceed 0.56. This discrepancy may be due to the fact that they evaluated their method only at $SNR = 0\,\mathrm{dB}$, whereas the current study uses $SNR \leq 0\,\mathrm{dB}$.

NCM and $\mathrm{CSII}_{\mathrm{Mid}}$ have clear discontinuities in the distribution when plotted against measured intelligibility scores (Figs. 6 and 7). Discontinuities are potentially problematic for prediction; strict monotonicity is preferable, such that inverse mapping from metric values to intelligibility scores can be performed. However, these discontinuities occur where intelligibility scores are >20%; hence, for speech security, they are less problematic.

STOI+, NCM, and NSEC tend to perform better on the chosen figures of merit than $\mathrm{CSII}_{\mathrm{High}}$, $\mathrm{CSII}_{\mathrm{Mid}}$, and STI (Table III). Regarding prediction bias and reliability, while all metrics tended to have a positive bias, the bias tends to be largest for $\mathrm{CSII}_{\mathrm{Mid}}$ and STI and lowest for STOI+ in the non-HPF condition and largest for $\mathrm{CSII}_{\mathrm{Mid}}$ and lowest for NSEC in the HPF condition (Fig. 9). In general, STOI+, ESTOI, NCM, NSEC, and $\mathrm{CSII}_{\mathrm{High}}$ perform well in terms of median bias. However, NSEC and STI are shown to be least reliable for prediction purposes.

Overall, the proposed method, STOI+, performs at least as well as the other metrics considered here and, under some conditions, better than STOI, ESTOI, STI, NSEC, $\mathrm{CSII}_{\mathrm{Mid}}$, and $\mathrm{CSII}_{\mathrm{High}}$. STOI+ and NCM are shown to be associated with the lowest prediction error and bias and the greatest reliability for intelligibility prediction for WGN maskers at SNRs from −26 to 0 dB. Both of these metrics use a wide range of values between zero and one and are robust to high-pass filtering. The speech-based STI method used in this paper appears to be less suitable for SNRs below 0 dB.

V. CONCLUSIONS

An assessment is made of two short-time methods to evaluate the intelligibility of speech mixed with white Gaussian noise over a wide range of SNRs from −26 to 0 dB. These are STOI and a variant, STOI+, which are compared with ESTOI, NCM, NSEC, $\mathrm{CSII}_{\mathrm{High}}$, $\mathrm{CSII}_{\mathrm{Mid}}$, and speech-based STI. This study extends previous comparisons of STOI and STOI-based metrics with other invasive intelligibility metrics by using speech from 12 talkers, 6 male and 6 female, rather than the typical 1–3, and 9 SNRs, rather than the typical 3–5.

While the normalisation and clipping procedures have been discarded in several published studies, no comparison of

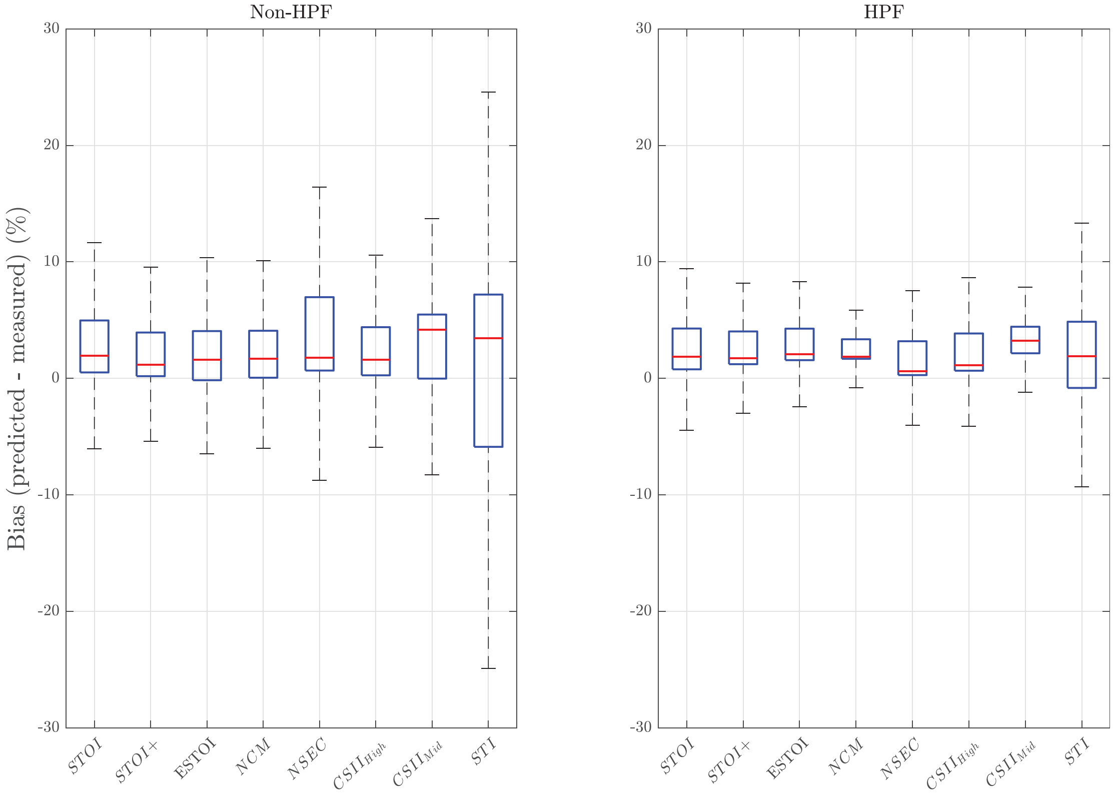

FIG. 9. (Color online) Prediction bias and reliability for the eight different metrics across talkers and SNRs for non-HPF (left) and HPF (right) conditions. The bias is typically positive, except for STI, which is also the least reliable for prediction.

TABLE III. Figures of merit for objective metrics for male and female talkers. For ρ, CI_l indicates the lower bound of the 95% confidence interval, while CI_u indicates the upper bound of the same. Boldface is used to indicate the better performing metric(s) within a given condition.

		Males			Females		
		ρ (CI_l-CI_u)	τ	σ_e	ρ (CI_l-CI_u)	τ	σ_e
Non-HPF	STOI	0.91 (0.90–0.92)	0.74	10.97	0.92 (0.91–0.93)	0.75	10.87
	STOI+	**0.92 (0.91–0.93)**	**0.76**	**10.54**	0.93 (0.92–0.94)	0.77	10.16
	ESTOI	0.90 (0.88–0.91)	0.74	11.67	0.93 (0.92–0.94)	0.76	10.15
	NCM	**0.92 (0.91–0.93)**	**0.77**	**10.46**	**0.94 (0.93–0.94)**	**0.77**	**9.75**
	NSEC	**0.93 (0.92–0.94)**	**0.77**	**9.96**	0.93 (0.92–0.94)	0.76	9.94
	CSII$_{\text{High}}$	0.90 (0.88–0.91)	0.75	11.74	0.93 (0.92–0.94)	0.77	10.45
	CSII$_{\text{Mid}}$	0.90 (0.89–0.91)	0.75	11.39	0.92 (0.91–0.93)	0.76	10.81
	STI	0.89 (0.87–0.90)	0.75	12.09	0.91 (0.90–0.92)	0.76	11.46
HPF	STOI	0.95 (0.94–0.95)	0.73	8.50	**0.96 (0.95–0.96)**	**0.70**	**7.73**
	STOI+	**0.95 (0.95–0.96)**	**0.75**	**8.02**	**0.96 (0.96–0.97)**	**0.75**	**7.30**
	ESTOI	0.94 (0.94–0.95)	0.74	8.64	**0.96 (0.95–0.96)**	**0.75**	**7.76**
	NCM	**0.96 (0.95–0.96)**	**0.76**	**7.81**	**0.96 (0.96–0.97)**	**0.76**	**7.43**
	NSEC	**0.96 (0.96–0.97)**	**0.76**	**7.12**	**0.96 (0.95–0.96)**	**0.75**	**7.91**
	CSII$_{\text{High}}$	0.92 (0.91–0.93)	0.72	10.16	**0.96 (0.95–0.96)**	**0.74**	**7.92**
	CSII$_{\text{Mid}}$	**0.95 (0.95–0.96)**	**0.76**	**7.87**	0.92 (0.91–0.93)	0.74	10.27
	STI	0.94 (0.93–0.94)	0.75	9.35	0.93 (0.93–0.94)	0.75	9.58

results with and without these procedures has been made previously. In this paper, it has been shown that normalisation and clipping increase STOI prediction error and reduce metric reliability when speech is mixed with white Gaussian noise at low global SNRs. When compared with STOI, ESTOI, CSII$_{\text{High}}$, CSII$_{\text{Mid}}$, NSEC, and speech-based STI, both NCM and STOI+ perform well for speech mixed with white Gaussian noise at SNRs from −26 to 0 dB—with or without high-pass filtering of the speech signal—in terms of prediction error, prediction bias, and reliability. In this study, logistic regression modeling demonstrated that high-pass filtering, which increases the proportion of high to low frequency energy, was detrimental to intelligibility for SNRs between −5 and −17 dB (inclusive). Whilst the results for NCM and STOI+ indicate their suitability for prediction, the upper bound for a 95% level of confidence is ≈20% when these metrics are in the range 0–0.2; hence, future work could investigate potential approaches to reduce this uncertainty for the purpose of speech security. Future work could also consider the efficacy of the metrics evaluated in this paper for speech that is mixed with additive noise and enhanced by means of mask-based algorithms.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Gary Seiffert of the Acoustics Research Unit, University of Liverpool, for his assistance with the speech recordings and with arranging the experimental setup for the listening tests.

APPENDIX

See Table IV for logistic free parameter values for all metrics except NSEC and STI. See Table V for logistic free parameter values for NSEC.

TABLE IV. Free parameters for the logistic mapping of STOI, STOI+, ESTOI, NCM, CSII$_{\text{High}}$, and CSII$_{\text{Mid}}$ with 95% confidence intervals.

		Males		Females	
		a	b	a	b
Non-HPF	STOI	−15.15 (–15.99 to −14.32)	10.83 (10.25–11.41)	−15.58 (–16.42 to −14.74)	11.01 (10.44–11.59)
	STOI+	−8.73 (–9.21 to −8.26)	5.20 (4.93–5.47)	−9.12 (–9.59 to −8.65)	5.33 (5.06–5.60)
	ESTOI	−10.41 (–11.01 to −9.80)	4.37 (4.13–4.61)	−11.21 (–11.78 to −10.65)	4.72 (4.50–4.95)
	NCM	−8.10 (–8.51 to −7.69)	4.02 (3.83–4.21)	−8.68 (–9.09 to −8.26)	4.33 (4.13–4.52)
	CSII$_{\text{High}}$	−9.18 (–9.71 to −8.65)	4.85 (4.58–5.11)	−9.89 (–10.40 to −9.38)	5.24 (4.98–5.50)
	CSII$_{\text{Mid}}$	−15.79 (–16.61 to −14.98)	3.21 (3.06–3.36)	−16.03 (–16.83 to −15.23)	3.32 (3.17–3.47)
HPF	STOI	−15.33 (–16.00 to −14.66)	10.28 (9.84–10.71)	−18.32 (–19.11 to −17.53)	11.78 (11.29–12.27)
	STOI+	−8.65 (–9.01 to −8.29)	4.48 (4.30–4.66)	−9.89 (–10.29 to −9.50)	4.58 (4.40–4.75)
	ESTOI	−11.06 (–11.55 to −10.56)	4.00 (3.84–4.16)	−13.48 (–14.05 to −12.91)	4.27 (4.10–4.44)
	NCM	−8.25 (–8.59 to −7.92)	4.04 (3.89–4.20)	−8.95 (–9.31 to −8.59)	4.09 (3.93–4.24)
	CSII$_{\text{High}}$	−10.08 (–10.62 to −9.54)	4.71 (4.48–4.95)	−12.17 (–12.71 to −11.62)	5.17 (4.95–5.39)
	CSII$_{\text{Mid}}$	−17.09 (–17.75 to −16.43)	3.51 (3.39–3.64)	−17.35 (–18.23 to −16.47)	3.53 (3.37–3.69)

TABLE V. Free parameters for NSEC logistic mapping with 95% confidence intervals. Confidence intervals are not provided for *c* in the non-HPF condition as these are unrealistic.

	Males			Females		
	a	*b*	*c*	*a*	*b*	*c*
Non-HPF	0.00 (–0.76–0.77)	1.00 (0.56–1.44)	−4871.82	0.00 (–0.77–0.78)	1.00 (0.58–1.41)	−3925.61
HPF	0.34 (–0.22–0.90)	0.81 (0.50–1.12)	−31.78 (−85.27–21.72)	0.35 (–0.27–0.97)	0.83 (0.52–1.13)	−32.33 (−89.80–25.14)

[1] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**(4), 1725–1736 (1990).

[2] M. Robinson, C. Hopkins, K. Worrall, and T. Jackson, "Thresholds of information leakage for speech security outside meeting rooms," J. Acoust. Soc. Am. **136**(3), 1149–1159 (2014).

[3] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**(2), 1220–1233 (2008).

[4] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**(1), 90–119 (1947).

[5] ANSI S3.5 (R2007): *Methods for the* Calculation *of the* Speech Intelligibility Index (Acoustical Society of America, New York, 1997).

[6] T. Houtgast, H. J. M. Steeneken, and A. W. Bronkhorst, "Speech communication in noise with strong variations in the spectral or the temporal domain," in *Proceedings of the 14th International Congress on Acoustics* (1992), Vol. 3, pp. H2–6.

[7] B. N. Gover and J. S. Bradley, "Measures for assessing architectural speech security (privacy) of closed offices and meeting rooms," J. Acoust. Soc. Am. **116**(6), 3480–3490 (2004).

[8] Institute of Electrical and Electronics Engineers, "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 227–246 (1969).

[9] ISO 60268-16:2011, *Sound System Equipment Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index* (International Organization for Standardization, Geneva, 2011).

[10] C. Ludvigsen, C. Elberling, G. Keidser, and T. Poulsen, "Prediction of intelligibility of nonlinearly processed speech," Acta Otolaryngol. Suppl. **109**, 190–195 (1990).

[11] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**(6), 3679–3689 (2004).

[12] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**(4), 2224–2237 (2005).

[13] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**(5), 3387–3405 (2009).

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

[15] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 5059–5063.

[16] C. C. Hsu, K. M. Cheong, J. T. Chien, and T. S. Chi, "Modulation Wiener filter for improving speech intelligibility," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 370–374.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in Proceedings of the *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2010), pp. 4214–4217.

[18] Y. Tang, R. J. Hughes, B. Fazenda, and T. J. Cox, "Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms," J. Speech Commun. **82**, 26–37 (2016).

[19] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in cochlear implants based on an intelligibility metric," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, August 27–31, 2012, Bucharest, Romania, pp. 504–508.

[20] L. Lightburn and M. Brookes, "SOBM—A binary mask for noisy speech that optimises an objective intelligibility metric," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5078–5082.

[21] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," IEEE Trans. Audio Speech Lang. Process. **24**(11), 1908–1920 (2016).

[22] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE Trans. Audio Speech Lang. Process. **24**(11), 2009–2022 (2016).

[23] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," IEEE Trans. Audio Speech Lang. Process. **22**(2), 430–440 (2014).

[24] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," IEEE SP Mag. **32**, 114–124 (2015).

[25] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in Mandarin," in *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4465–4469.

[26] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," IEEE-ACM Trans. Audio Speech Lang. Process. **26**(11), 2153–2166 (2018).

[27] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," Comput. Speech Lang. **35**, 73–92 (2016).

[28] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," Speech Commun. **51**(12), 1253–1262 (2009).

[29] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," IEEE Trans. Acoust. Speech Signal Process. **24**(4), 277–282 (1976).

[30] M. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," J. Speech Commun. **48**(5), 549–558 (2006).

[31] J. B. Boldt and D. P. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proceedings of the 2009 17th European Signal Processing Conference*, August 24–28, 2009, Glasgow, UK, pp. 1849–1853.

[32] K. Payton and M. Shrestha, "Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data," J. Acoust. Soc. Am. **134**(5), 3818–3827 (2013).

[33] C. Hopkins, S. Graetzer, and G. Seiffert, "ARU adult British English speaker corpus of IEEE sentences (ARU speech corpus) version 1.0 [data collection]," Acoustics Research Unit, School of Architecture, University of Liverpool, UK, http://dx.doi.org/10.17638/datacat.liverpool.ac.uk/681 (Last viewed 17 February 2021).

[34] Institute of Sound Recording (IoSR) Surrey matlab toolbox, available from https://github.com/IoSR-Surrey/MatlabToolbox/ (Last viewed 25 August 2020).

[35] ITU-T P.56:2011, "Objective measurement of active speech level," ITU-T Recommendation P.56 (2011).

[36] ISO 8253–1:2010, "Acoustics: Audiometric Test Methods Part 1: Basic Pure Tone Air and Bone Conduction Threshold Audiometry" (International Organization for Standardization, Geneva, 2010).

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

[38] K. H. Arehart, J. M. Kates, M. C. Anderson, and L. O. Harvey, Jr., "Effects of noise and distortion on speech quality judgments in normal-

hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **122**(2), 1150–1164 (2007).

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction," in *Proceedings of the Tenth Annual Conference of the International Speech Communication Association*, September 6–10, 2009, Brighton, UK.

[40] D. Cabrera, D. Jimenez, and W. L. Martens, "Audio and Acoustical Response Analysis Environment (AARAE): A tool to support education and research in acoustics," in *Proceedings of Internoise*, November 16–19, 2014, Melbourne, Australia.

[41] IEC 60268-16, *Sound System Equipment, Part 16: Objective Rating* of *Speech Intelligibility* by *Speech Transmission Index* (European Committee for Standardization, Brussels, 2011).

[42] A. Opsata, D. Cabrera, and, and M. Yadav, "Influence of time-varying talker directivity on the calculation of speech transmission index from speech in a room acoustical context," in *Proceedings of Internoise 2014*, November 16–19, 2014, Melbourne, Australia.

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

[44] R Core Team, "R: a language and environment for statistical computing," available from http://www.R-project.org (Last viewed 10 May 2020).

[45] ISO 9921:2003E, "Ergonomics: assessment of speech communication" (International Organization for Standardization, Geneva, 2003).