DECEMBER 13 2022

Comparison of ideal mask-based speech enhancement algorithms for speech mixed with white noise at low mixture signal-to-noise ratios

Simone Graetzer ; Carl Hopkins

J Acoust Soc Am 152, 3458–3470 (2022)
https://doi.org/10.1121/10.0016494

JASA ARTICLE

Comparison of ideal mask-based speech enhancement algorithms for speech mixed with white noise at low mixture 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 literature shows that the intelligibility of noisy speech can be improved by applying an ideal binary or soft gain mask in the time-frequency domain for signal-to-noise ratios (SNRs) between –10 and +10 dB. In this study, two mask-based algorithms are compared when applied to speech mixed with white Gaussian noise (WGN) at lower SNRs, that is, SNRs from −29 to –5 dB. These comprise an Ideal Binary Mask (IBM) with a Local Criterion (LC) set to 0 dB and an Ideal Ratio Mask (IRM). The performance of three intrusive Short-Time Objective Intelligibility (STOI) variants—STOI, STOI+, and Extended Short-Time Objective Intelligibility (ESTOI)—is compared with that of other monaural intelligibility metrics that can be used before and after mask-based processing. The results show that IRMs can be used to obtain near maximal speech intelligibility (>90% for sentence material) even at very low mixture SNRs, while IBMs with $LC = 0$ provide limited intelligibility gains for SNR < -14 dB. It is also shown that, unlike STOI, STOI+ and ESTOI are suitable metrics for speech mixed with WGN at low SNRs and processed by IBMs with $LC = 0$ even when speech is high-pass filtered to flatten the spectral tilt before masking.
 https://doi.org/10.1121/10.0016494
(Received 16 June 2022; revised 11 November 2022; accepted 21 November 2022; published online 13 December 2022)
[Editor: John H. L. Hansen] Pages: 3458–3470

I. INTRODUCTION

Degraded speech signals, such as those transmitted along a noisy channel, can be enhanced by means of time-frequency segregation (TFS). In this approach, signals can be decomposed into time-frequency (T–F) units and, on the basis of a local estimate of the signal-to-noise ratio (SNR), a user-defined rule sets the gain of each unit to one or zero to form an Ideal Binary Mask (IBM). Ideal binary masks are estimated with access to the clean speech signal where the user specifies a Local Criterion (LC). For example, with $LC = 0$ dB, the gain is set to one when the (local) SNR is at least 0 dB. The degraded signal is enhanced by multiplying it by the IBM, such that where the interference dominates the target speech, the signal energy is discarded. This is a traditional definition of a binary mask matrix where $LC = 0$ is implied. An IBM with $LC = 0$ has been considered theoretically optimal[1] and is used in this paper. In the case of signals mixed with broadband noise at very low SNRs, the enhanced signal effectively becomes binary-gated noise, i.e., noise onto which crude approximations of speech temporal envelopes have been imposed. An unwanted feature of IBMs can be tone-like artefacts in the enhanced signal; this distortion is often referred to as musical noise. An alternative mask-based speech enhancement algorithm that minimises or avoids these artefacts is the Ideal Ratio Mask (IRM), in which mask gain values are estimates of the target and mixture signal energy ratios.[2] When IRMs are applied to very noisy speech, the modulated noise is essentially noise-vocoded speech, where there is a continuous range of modulation gains (e.g., Shannon *et al.*,[3] Souza and Rosen[4]). The ideal mask approach might be considered similar to the noise vocoder approach when the global SNR of the mixed signal is not so low that the signal is essentially pure noise, but there is an important difference. When noise is vocoded, even when the target speech envelope is reconstructed perfectly from the noise signal, the target temporal fine structure will be entirely absent, and this is likely to affect the quality of the enhanced speech, if not the intelligibility (see, e.g., Kates and Arehart[5,6]).

Assessments of IBMs and IRMs in the literature tend to use SNRs between –10 and +10 dB, for which IRMs have been shown to have potential advantages relative to IBMs in terms of speech quality and intelligibility.[7,8] In this paper, the aim is to assess the intelligibility benefits of both ideal binary and ideal ratio masks for speech mixed with white Gaussian noise (WGN) at low SNRs that range from –29 to –5 dB. The benefits can be estimated as a difference from the baseline—where the speech has not been enhanced—for SNRs from –26 to –5 dB. A baseline intelligibility score of 0% at $SNR = -29$ dB is extrapolated from the value of 0% at $SNR = -26$ dB because the value at –26 dB has already fallen to 0%. The baseline conditions and results are

[a]Present address: Acoustics Research Centre, School of Science, Engineering and Environment, University of Salford, Salford, M5 4NT, UK.
[b]Author to whom correspondence should be addressed: carl.hopkins@liverpool.ac.uk

https://doi.org/10.1121/10.0016494

described in a previous paper that is hereafter described as the "baseline."[9]

Listening tests provide information about the improvement in intelligibility that can be achieved by masking algorithms. However, it is not always feasible to run these tests. Correlation-based "objective" or instrumental metrics such as Short-Time Objective Intelligibility (STOI)[10] have been shown to perform relatively well in predicting relative intelligibility and, in particular, the effects on intelligibility of non-linear speech processing, where the noise or degradation is not additive (e.g., in the context of noise suppression, where a measure must be able to assess intelligibility accurately when the enhanced signal contains no target speech fine structure). STOI was introduced for intelligibility prediction before and after the application of T-F varying gain functions, or TFS. STOI tends to outperform traditional objective metrics for ideal TFS-processed speech at least for sentences for which intelligibility scores are $\geq 20\%$ and mixture SNRs that exceed –10 dB.[11]

STOI was defined by Taal *et al.*[10] 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 T–F units. Recent work by the current authors on the evaluation of STOI for speech mixed with noise at low SNRs, both before and after enhancement with IBMs, indicates that an improved STOI-based metric—referred to in this paper as STOI+—for SNRs equal to or lower than 0 dB would not use clipping.[9,12]

In several other investigations or extensions of STOI, the clipping procedure has been removed.[13–15] Jensen and Taal[16] developed Extended Short-Time Objective Intelligibility (ESTOI) to improve STOI performance for highly fluctuating or modulated noise sources (also indicating that the metric is suitable for non-modulated noise signals), which does not incorporate the clipping procedure. Steinmetzger and colleagues[17] evaluated whether STOI could predict the masker periodicity benefit and commented that STOI's underprediction of the benefit appears to originate from the clipping procedure. Specifically, the authors argued that the clipping algorithm "discards a substantial portion of the subtle envelope differences between the aperiodic and periodic maskers. Consequently, some of the acoustic properties of the masker are not represented in STOI." (p. 2568). The only published comparison of results with and without clipping that is known to the authors is in Graetzer and Hopkins.[9]

As noted in the baseline study,[9] stationary noise (e.g., WGN) can be used for masking in speech security applications where the focus is on the risk of only a few words being intelligible; this tends to occur when SNR < -10 dB. The experiments reported in this paper build on the baseline study[9] by comparing the performance of the IBM with $LC = 0$ and the IRM for WGN and low SNRs relative to the baseline, and compare the performance of STOI with other intrusive forms of STOI—STOI+ and ESTOI—and with Non-Intrusive Short-Time Objective Intelligibility (NI-STOI),[18] the Normalised Covariance Metric (NCM)[19] and the Coherence Speech Intelligibility Metric (CSII),[20,21] and the Normalised Sub-band Envelope Correlation metric (NSEC).[22] The aims of the study were to (a) identify differences in the percentages of words correctly identified in speech degraded by additive WGN and enhanced by two masking algorithms, IBM with $LC = 0$ and IRM, and (b) to evaluate the performance of STOI, STOI+, and ESTOI relative to a non-intrusive form of STOI, NI-STOI, and some well known measures, in predicting the percentages of words correctly identified for speech processed by an IBM algorithm with $LC = 0$.

In Sec. II, the experimental procedures will be outlined, including a brief discussion of how STOI+ differs from conventional STOI. In Sec. III, the effects of SNR, high-pass filter (HPF), and talker gender on intelligibility scores, and the performance of metrics in estimating those scores, will be reported. It will be shown that both STOI+ and ESTOI perform better than STOI for high-pass (HP) filtered speech at the low SNRs considered. In Sec. IV, the reasons for the variable performance of the HPF, and the relative performance of STOI, STOI+, ESTOI, and the other metrics are discussed.

II. EXPERIMENTAL PROCEDURES

The experiment received prior approval from the University of Liverpool Committee on Research Ethics.

A. Speech signal processing

1. Speech recordings

Twelve talkers (six male, six female) between 21 and 47 years of age were recorded in an anechoic chamber using a 0.5 in. Brüel & Kjær (B&K) type 4190 microphone (Brüel & Kjær, Nærum, Denmark) at 1 m on axis, a B&K type 2669 preamplifier and a LANXI 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).

The speech recordings are identical to those described in the baseline study.[9] Talkers produced the Institute of Electrical and Electronics Engineers (IEEE) sentences,[23] 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" in order 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.[24]

2. Signal processing

The signal processing prior to the creation of T–F masks was identical to that described in the baseline paper.[9]

As in the baseline paper, the spectral tilt was flattened using a HPF. This filter was designed to give amplitudes of zero and one at normalised frequencies between zero and one (Nyquist), with an approximately linear relationship between amplitude and normalised frequency. The spectral tilt is removed because in speech security, there is usually a need to assess worst-case scenarios, and 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[25]).

In the present study, two speech enhancement algorithms were evaluated: IBM with $LC = 0$ and IRM. Algorithms were based on publicly available code.[26] The frequency range for both mask types was from 80 Hz to 12 kHz. The lower limit of this range was decided on the basis of where the background noise below 100 Hz approaches the long-term average male speech spectra and the upper limit equals half the sampling rate (after downsampling to $f_s = 24$ kHz).

The process of speech enhancement by means of IBM algorithms involves multiplying signals by binary gain values based on the local SNR in each T–F unit. The gain function is given in Eq. (1), where T denotes the target speech and M, the masker in dB, t denotes time, and f, frequency. This function involves assessing whether, in each channel and frame, the difference between the energy in the target cochleagram and the scaled masker cochleagram is greater than LC in dB,

$$IBM(t,f) = \begin{cases} 1 & \text{if} \quad T(t,f) - M(t,f) > LC \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

To obtain the signals, a fourth-order gammatone filterbank was designed with 128 channels, a gammatone filter length of 2048 samples (85 ms for a 24 kHz sampling rate), and a frequency range of 80 Hz to 12 kHz with frequencies equally spaced on the ERB rate scale,[27] with middle-ear loudness-based gain normalisation. The output of the gammatone filterbank was used to generate a cochleagram that had a window length of 320 sample points (approximately 13.3 ms for a 24 kHz sampling rate) with a frame shift equal to half the window size.

The IRM is obtained on the basis of the gain function shown in Eq. (2),

$$IRM(t,f) = \left(\frac{|T(t,f)|^2}{|T(t,f)|^2 + |M(t,f)|^2} \right)^{\beta}. \tag{2}$$

The "tuning" or compression constant is set to $\beta = 0.5$. Wang and colleagues[8] "experimented with different β values and found $\beta = 0.5$ to be the best choice" (p. 1851). A preliminary investigation by the present authors was consistent with this finding. Wang *et al.*[8] pointed out that when $\beta = 0.5$, Eq. (2) becomes similar to the square-root Wiener filter. Note that in a comparison of different algorithms for noise suppression, a Wiener filter was found to give a similar performance to envelope restoration (i.e., matching the envelope of the noisy speech to the clean speech) where the latter provided the highest predicted intelligibility.[28]

Within each channel and frame, a raised cosine window is multiplied by the output of the gain function plus weights to create weighted mask values. The filterbank output is multiplied by the mask values to obtain the synthesised signal. An example speech signal with associated cochleagrams and both IBM and IRM masks is shown in Fig. 1. This illustrates the hard gating used by the IBM compared to the softer decisions made by the IRM, which lets through signal energy over a wider range of frequencies.

B. Listening tests

Twenty-four listeners (12 male and 12 female) took part in the experiment. These listeners had not been exposed previously to the speech material. They were pseudo-randomly selected from a larger set of 48 untrained listeners aged between 18 and 45 years (full dataset, $\mu = 30.0$ years, $\sigma = 8.0$ years; male subset, $\mu = 25.7$ years, $\sigma = 6.4$ years; female subset, $\mu = 31.8$ years, $\sigma = 7.6$ years). (The other listeners were assigned to other mask types, which are not discussed here.) Six males and six females were assigned pseudo-randomly to each mask type. All listeners used British English as a first language and had a good (self-reported) spelling ability. Their hearing thresholds were tested using an Otovation Amplitude T3 audiometer (AMPT3-0111-042) (Otovation LLC, King of Prussia, PA) with supra-aural Telephonics audiometric headphones (TDH-39P 296D000-1) (Telephonics, Farmingdale, NY) and did not exceed 20 dB hearing level (HL) between 125 Hz and 8 kHz.[29] The listening test procedure, conditions, and equipment were identical to those used in the baseline study.[9]

The listening experiment was conducted in a sound-attenuated booth with diotic presentation using Beyer Dynamic DT770 Pro circumaural headphones (Beyer Dynamic GmbH & Co. KG, Heilbronn, Germany). The audio output of the system was calibrated using the Head-and-Torso Simulator (HATS) with Brüel & Kjær (Nærum, Denmark) type 4189 microphones in each ear canal. The background noise at the entrance to the ear canal during testing was estimated to be 22 dB L_{Aeq} by using the Brüel & Kjær type 4100 HATS (Brüel & Kjær, Nærum, Denmark) wearing the headphones when they were connected to the PC that was running the experiment from a MATLAB graphical user interface.

The presentation level was chosen by the subjects at the beginning of the experiment to be 70 or 75 dB L_{Aeq}. When setting the presentation level, listeners heard sentences processed by their assigned masking algorithm with a mixture $SNR = -5$ dB. In the familiarisation stage, listeners heard one clean sentence and four enhanced speech sentences at SNRs equal to –5, –8, –11, or –14 dB. These sentences were selected pseudo-randomly from the talkers that were assigned to the listener.

Listeners were asked to identify as many words as possible in each sentence and were able to correct their spelling. They had ≈15 s after the sentence played to enter the words

https://doi.org/10.1121/10.0016494

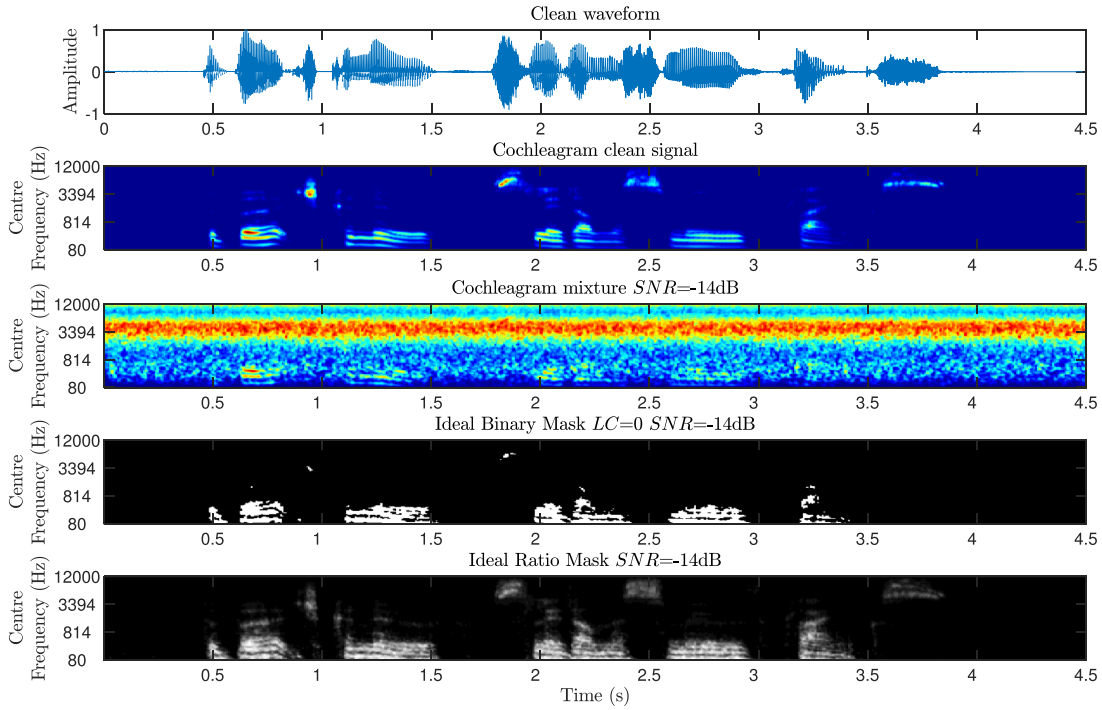

FIG. 1. (Color online) Example of speech with mixture $SNR = -14$ dB when processed with an IBM and an IRM.

they had heard into the text box and make any corrections. They were allowed to pause the test at any time and were offered breaks of up to five minutes every ≈30 min.

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. Homophones and some alternative spellings were allowed and the words "a" and "the" were removed from the analysis. See Graetzer and Hopkins[9] for details.

C. Implementation of metrics

In this paper, STOI, STOI+, ESTOI, NCM, CSII, NSEC, and NI-STOI are evaluated for two mask-based speech intelligibility enhancement methods.

1. STOI and STOI+

STOI[10] is based on the correlation between the envelopes of a clean, x, and a degraded or processed speech signal, y, decomposed into regions that are approximately 384 ms (30 samples) in length (where $fs = 10$ kHz). The procedure used was identical to that described in Taal *et al.*[10] The output, d, is typically positive, limited to values ≤ 1, and ideally has a monotonic relationship with speech intelligibility scores. Previous studies (such as those by the authors[9] and Tang *et al.*[30]) suggest that STOI d rarely falls below 0.3, even for signals associated with 0% intelligibility scores, and where the Speech Intelligibility Index (SII),[31] NCM, and CSII are zero. In testing, it was found that STOI and STOI+ varied at the second decimal place with pseudo-randomly selected segments of additive WGN. Hence, correlations were computed based on objective indices that are averaged over sentences within word lists. For consistency, for all metrics considered, scores are averaged over sentences within word lists.

As the relationship between STOI-based measures and intelligibility scores is monotonic, STOI-based values were converted to mapped values, i.e., predicted intelligibility scores, *via* a transfer function, specifically, a logistic function, to linearise the relationship between STOI-based measures and intelligibility scores and therefore report linear correlation coefficients (ρ values) and determine the distribution of prediction errors. The logistic function was used to map the variable d (representing STOI or STOI+ values) with the free parameters, a (slope) and b (centre), as shown in Eq. (3). Free parameters a and b were fitted to the data using a nonlinear least squares procedure and with starting values derived from Taal *et al.*[10] As in the case of Taal *et al.* the (a,b) parameter values are derived from a single data set; there are no separate training, validation, and test datasets,

$$f(d) = \frac{100}{1 + \exp(ad + b)}. \tag{3}$$

2. ESTOI

ESTOI was calculated using the procedure described in the baseline paper.[9] Jensen and Taal[16] proposed ESTOI as a measure to improve on STOI in the case of highly modulated noise sources. 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[16] was used in this

study. 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 was mapped using Eq. (3).

3. NCM

NCM[19] was calculated using publicly available code.[32] 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 Speech Transmission Index (STI)[33] method to derive a final, band-weighted, value of between 0 and 1. The procedure was identical to that described in the baseline paper.[9] Logistic mapping was performed after Taal *et al.*[10] using Eq. (3).

4. CSII

CSII was originally developed by Kates and Arehart[21] for predicting the speech intelligibility of peak- or centre-clipping distortions, such as those associated with hearing aids. 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.[34] The $\mathrm{CSII_{High}}$ index is associated with segments at or above the overall rms level of the signal; the $\mathrm{CSII_{Mid}}$ index, at or up to 10 dB below the same level, and the $\mathrm{CSII_{Low}}$ index, 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 $\mathrm{CSII_{Mid}}$ can be combined linearly and transformed with a logistic function to derive a fourth measure, termed *I3*. In this paper, as in the baseline paper,[9] the short-time CSII implementation[32] was used in which CSII is averaged over short-time segments of 30 ms in length with a 25% window skip rate. In this paper, the critical band weighting function of NCM and CSII was set to ANSI S3.5[31] weighting for "Short Passages." Preliminary testing indicated that $\mathrm{CSII_{Low}}$ performs poorly and $\mathrm{CSII_{I3}}$ performs no better than $\mathrm{CSII_{Mid}}$ under the present experimental conditions, and so are 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-based metrics [Eq. (3)], the second function provided by Taal *et al.*[35] as shown in Eq. (4) and a linear fit.

$$f(x) = \frac{100}{1 + (ax + b)^c}. \tag{4}$$

The prediction error indicated that Eq. (3) tended to perform as well as or better than these alternatives; hence, the logistic function was also fit to the CSII metrics.

5. NSEC

Boldt and Ellis[22] developed NSEC based on the correlation of the envelopes of the original speech, x, and the degraded speech, y, after T–F decomposition, frequency equalisation, amplitude compression, and DC removal. In this implementation, the energy envelopes are derived *via* a 16 channel Gammatone filterbank with center frequencies from 80 to 8000 Hz, equally spaced on the ERB scale, and with a window length of 0.08 s with a 50% overlap. For more information, see Boldt and Ellis.[22] The original mapping function proposed by the authors is given in Eq. (5). However, Taal *et al.*[36] obtained better performance with Eq. (4). Hence, the latter function is applied to NSEC scores in this paper, as in the baseline paper,[9]

$$f(x) = \frac{1}{1 + e^{(b-x)/a}}. \tag{5}$$

6. NI-STOI

NI-STOI[18] is a non-intrusive variant of the STOI algorithm in which clean speech envelopes are estimated from the degraded/processed signal envelopes. In the method applied here, the true clean speech signal is used only to determine which frames contain speech *via* a voice activity detector. A faint noise signal is added to the degraded signal to allow NI-STOI to predict the intelligibility of signals "where aggressive speech processing renders the presented signal almost inaudible" (Andersen *et al.*,[18] p. 5086). Further details are available in Andersen *et al.*[18]

D. Evaluation procedures

The experiment conducted evaluates the effectiveness of objective measures for the prediction of ideal TFS-processed speech intelligibility with reference to the clean signal by comparing predicted and measured intelligibility scores. These measures are compared on the basis of summary statistics such as minimum and maximum values, correlation coefficients, estimates of the prediction error, and estimates of metric bias and reliability. The distribution of metric values relative to intelligibility scores is 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). As in the baseline paper,[9] the prediction error was calculated as $\sigma_e = \sigma_d\sqrt{1-\rho^2}$ where σ_d is the standard deviation of the percentage words correct. Figures of merit ρ and σ_e are applied to the mapped objective scores, while τ is rank based and therefore independent of the mapping.

https://doi.org/10.1121/10.0016494

To compute metric bias, b, the measured scores, v, were subtracted from the corresponding predicted scores, w. Similarly, the mean bias, $\bar{b}$, was calculated according to Eq. (6), where N is in this case the number of measured scores,

$$\bar{b} = \frac{1}{N} \cdot \sum (w - v). \tag{6}$$

Predicted scores are mapped metric values, 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 indicate the reliability of the predictions, with smaller boxes or shorter whiskers indicating higher reliability. The position of the box plus whiskers, and especially the horizontal line marking the median, 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 v.3.5.1[37] with percentages of words identified correctly (measured intelligibility 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 model was chosen based on nested model comparisons using likelihood ratio tests with a Chi-squared test statistic.

The resulting logistic regression model can be described as shown in Eq. (7) where 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 $SNR = -5$ dB, 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 zero, only SNR levels equal to or greater than –17 dB were included in the model. 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 –5 dB is used only in the logistic regression model,

$$logit(p) = \beta_0 + \beta_1 SNR + \beta_2 Filter + \beta_3 Gender + \beta_4 SNR \cdot Filter + e. \tag{7}$$

III. RESULTS

A. Word recognition scores

Figure 2 shows the intelligibility scores for each masking algorithm, IBM with $LC = 0$ and IRM, expressed as words identified correctly per word list by each listener per talker, SNR, and filter condition. Noisy speech medians[9] for male and female talkers are also shown in the figure to indicate the effect of the IBM or IRM.

Scores for noisy speech were close to 0% below –11 dB, with a speech reception threshold (associated with intelligibility scores of 50%) close to –5 dB. The size of the intelligibility gains, in terms of the difference between noisy and enhanced speech medians, tended to be largest for IBM $LC = 0$ at SNRs between –11 and –8 dB. Below $SNR = -11$ dB, mask density (defined as the number of ones in the mask as a proportion of the total number of values) was $< 5\%$. For the IBM with $LC = 0$ condition, there was a linear relationship between mask density and intelligibility scores, such that a decrease in density was associated with a decrease in intelligibility. For the IRM, the size of the gains tended to be constant with SNRs between –29 and –11 dB and then gradually decreased as the SNR increased from –11 to –5 dB. The IRM gave consistently high numbers of words correctly identified compared to the IBM. For example, performance at $SNR \leq -20$ dB improved from 0% to 5% with IBM $LC = 0$ to close to 100% correct when the IRM was applied. As is evident in Fig. 2, unlike the IBM with $LC = 0$, the performance of the IRM did not depend on SNR. As the percentages of words correctly identified for the IRM are close to 100%, in this paper, analysis of the metrics is confined to the IBM mask type only.

For the IBM with $LC = 0$, a logistic regression model was fit to the intelligibility scores. This was expressed as successes and failures out of trials with SNR, filter condition, talker gender, and the interaction of SNR and filter condition as fixed effects. The reference conditions were –5 dB SNR, the non-HPF condition, and male talker gender. The results are shown in Table I. There was an effect of SNR such that when moving from –5 to –8 dB SNR, there was a 0.62 decrease in the log odds ($O = 0.54$) of identifying a word correctly, and the decrease in the log odds became larger as the SNR decreased, as would be expected. There was a main effect of filter on intelligibility scores such that the HPF was beneficial to intelligibility when the SNR was –5 dB; there was an estimated 0.27 increase in the log odds of identifying a word correctly ($O = 1.31$). However, the benefit of the HPF tended to be more apparent for male talkers than female talkers. There was a main effect of talker gender such that, at the reference level of $SNR = -5$ dB, there was a slight decrease in the odds of identifying a word correctly when the talker was female ($O = 0.73$). However at –14 and –17 dB SNR, there was no observable difference between talker genders.

The HPF tended to improve intelligibility at SNRs between –11 and –5 dB for male talkers and –8 and –5 dB for female talkers. However, the use of the HPF was detrimental at lower SNRs (–14 and –17 dB SNR). 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 –5 dB,

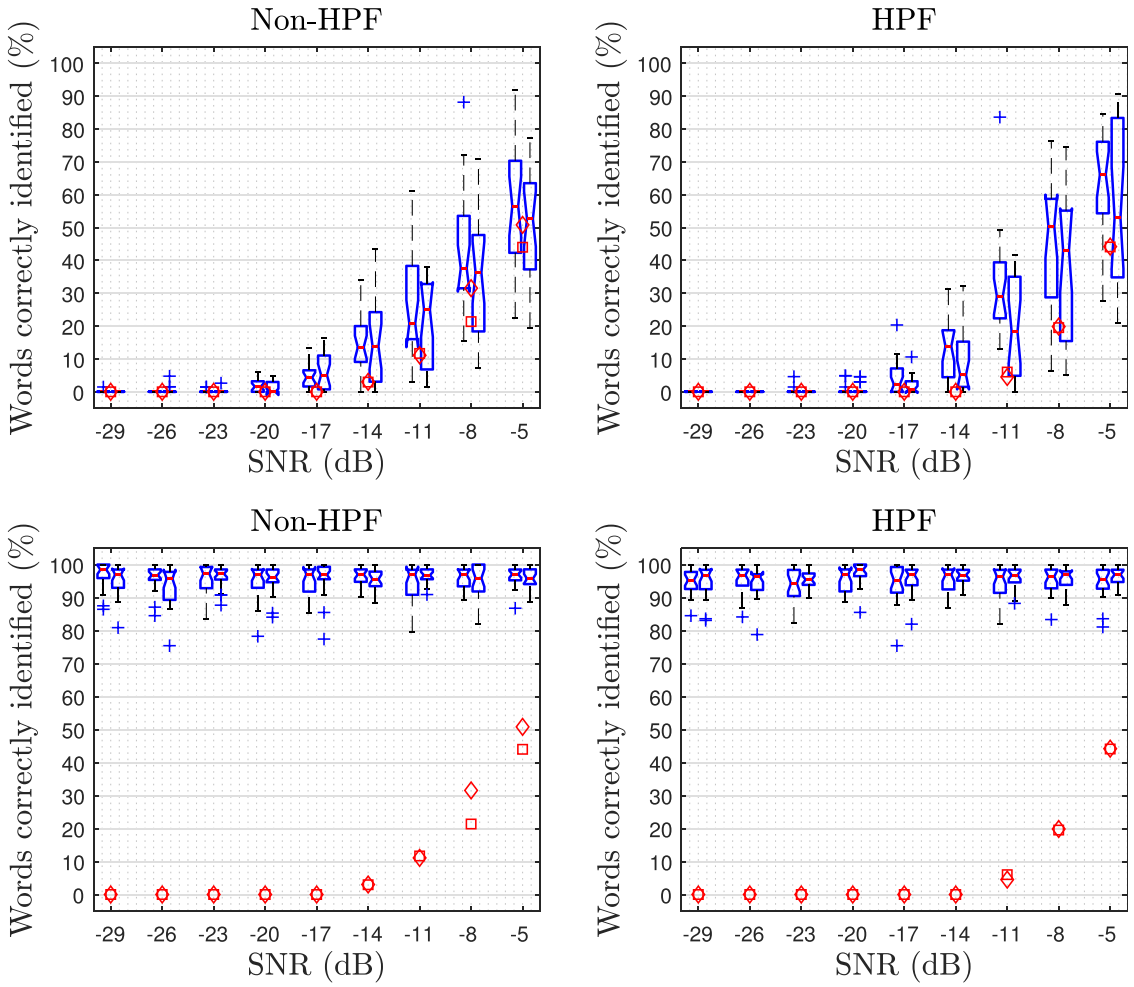

FIG. 2. (Color online) Boxplots of speech intelligibility scores for IBM with $LC = 0$ (upper plots) and IRM (lower plots). Corresponding noisy speech medians are shown as red squares for male talkers and red diamonds for female talkers (including the value of 0% words correctly identified at $SNR = -29$ dB, which is extrapolated from the value of 0% at $SNR = -26$ dB using shape-preserving piecewise cubic interpolation). At each SNR, the boxes that correspond to male and female talkers are shown on the left, and right, respectively.

the log odds of identifying a word correctly were lower in the HPF condition than in the non-HPF condition, with the log odds decreasing as the SNR is lowered.

The result for $SNR = -5$ dB has already been reported. At $SNR = -8$ and -11 dB, there is no difference between filter conditions ($p = 1$). At $SNR = -14$ dB, the log odds decreased by 0.35 ($SE = 0.08$, $z = -4.57$, $p < 0.001$). At $SNR = -17$ dB, the log odds decreased by 0.64 ($SE = 0.13$, $z = -4.97$, $p < 0.0001$). In sum, the HPF does not improve the intelligibility of speech mixed with WGN and processed by an IBM with $LC = 0$ at $-17 \leq SNR \leq -8$ dB. The approximate R^2 derived from the full model deviance and the null model deviance is 0.62.

TABLE I. Logistic regression model output for WGN mixed with speech at SNRs between –17 and –5 dB inclusive and processed by an IBM with $LC = 0$. The interaction of SNR and Filter is discussed in Sec. III A.

	Estimate	*Odds*	*SE*	*z*	*p*
(Intercept)	0.31	1.36	0.04	8.12	<0.0001
–17 dB SNR	–3.00	0.05	0.08	–35.27	<0.0001
–14 dB SNR	–1.94	0.14	0.06	–31.34	<0.0001
–11 dB SNR	–1.32	0.27	0.05	–24.17	<0.0001
–8 dB SNR	–0.62	0.54	0.05	–12.22	<0.0001
Filter HPF	0.27	1.31	0.05	5.29	<0.0001
Talker female	–0.31	0.73	0.03	–11.22	<0.0001

B. Objective intelligibility metric results

Table II provides a summary of the metric statistics for the IBM mask type only. All metrics except STOI extend

TABLE II. Summary of metric statistics for the non-HPF and HPF conditions for the IBM mask type only.

Metric	Minimum	Median	Mean	Maximum	Interquartile range	Range
STOI	0.25	0.60	0.60	0.98	0.27	0.73
STOI+	0	0.52	0.48	0.88	0.39	0.88
ESTOI	–0.09	0.21	0.25	0.73	0.44	0.81
NCM	0	0.45	0.46	0.90	0.35	0.90
$CSII_{High}$	0	0.47	0.44	0.85	0.42	0.85
$CSII_{Mid}$	0	0.16	0.20	0.56	0.33	0.56
NSEC	0.09	0.71	0.66	0.92	0.26	0.83
NI-STOI	0.07	0.85	0.82	0.92	0.07	0.85

https://doi.org/10.1121/10.0016494

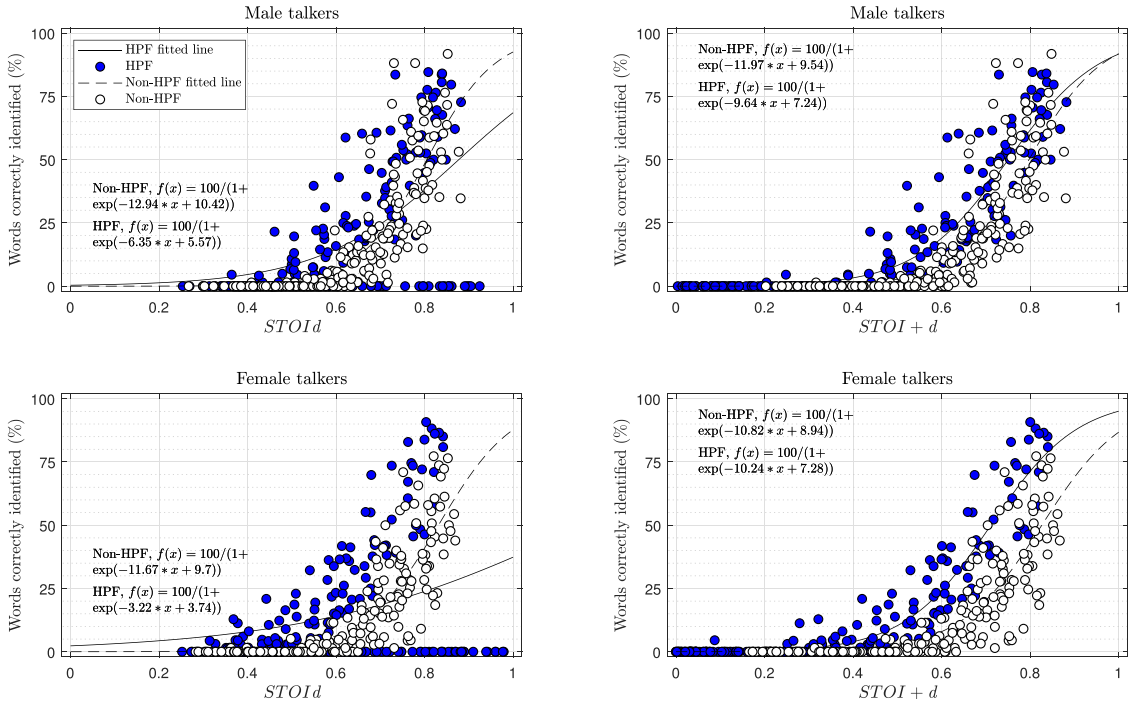

FIG. 3. (Color online) IBM with $LC = 0$: scatterplots of STOI and STOI+ by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function.

down to approximately zero, while STOI extends down to 0.25 for these data. $CSII_{Mid}$ and STOI have the smallest ranges at 0.56 and 0.73, while STOI+, ESTOI, NCM, $CSII_{High}$, NSEC, and NI-STOI have a range of 0.8–0.9. NI-STOI has the smallest interquartile range of 0.07.

For the IBM mask type, the distribution of metric values according to intelligibility scores is shown in Fig. 3 for STOI and STOI+, Fig. 4 for ESTOI and NCM, Fig. 5 for $CSII_{High}$ and $CSII_{Mid}$, and Fig. 6 for NSEC and NI-STOI. The logistic function parameter values are reported within

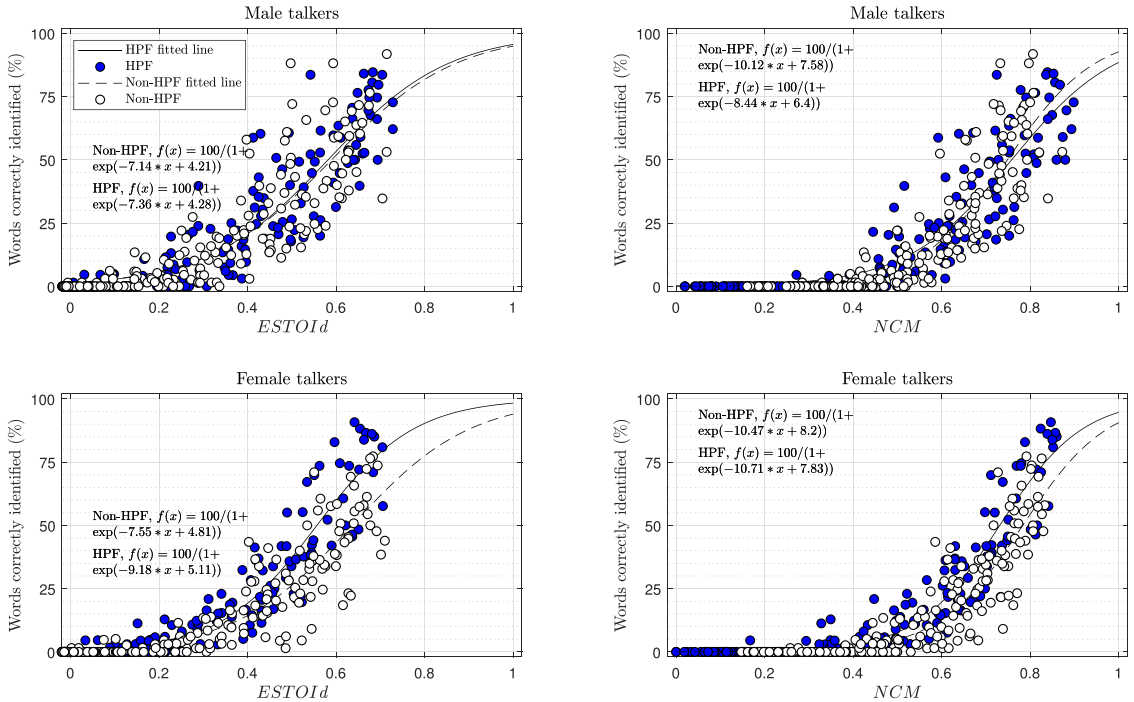

FIG. 4. (Color online) IBM with $LC = 0$: scatterplots of ESTOI and NCM by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function.

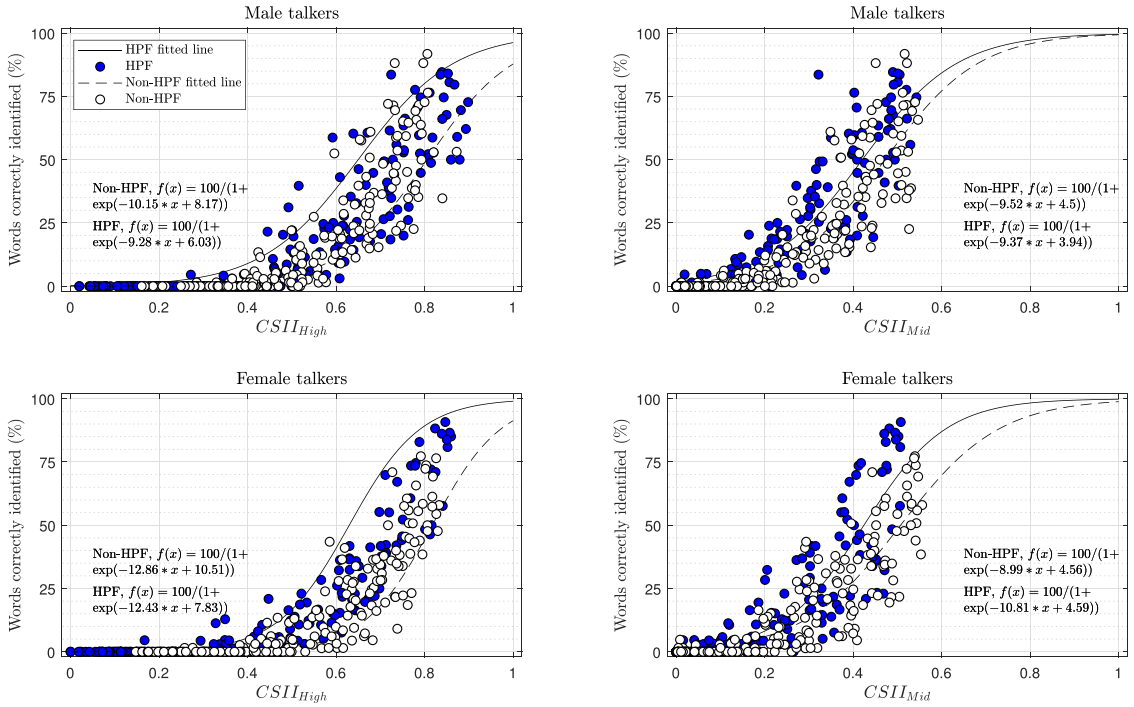

FIG. 5. (Color online) IBM with $LC = 0$: scatterplots of $CSII_{High}$ and $CSII_{Mid}$ by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function.

the figures; these values and their confidence intervals are provided in the Appendix (Tables IV and V).

In previous work by the authors on noisy speech,[9] there were found to be discontinuities in the relationship between the values of two metrics, NCM and $CSII_{mid}$, and intelligibility scores; however, these are not evident for the IBM with $LC = 0$. There are no obvious discontinuities in the relationship between metric values and intelligibility scores for any of the metrics shown in Figs. 3–6. There is evidence of a problem with the STOI metric in Fig. 3 where there are intelligibility scores of 0% associated with STOI d values of close to 1 in the HPF condition. These data points increase

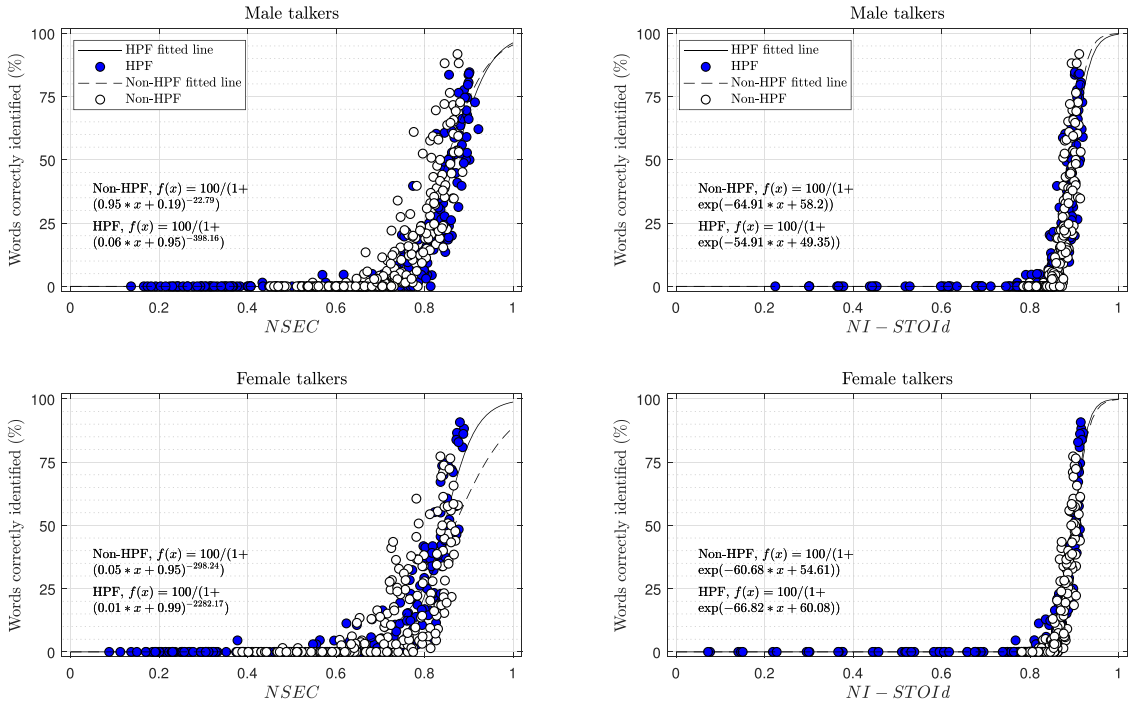

FIG. 6. (Color online) IBM with $LC = 0$: scatterplots of NSEC and NI-STOI by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function.

https://doi.org/10.1121/10.0016494

the shallowness of the slope of the fitted line relative to the non-HPF condition. This pattern is not observed for STOI+. The relationship between metric value and percentage of words correctly identified is similar for STOI+ (Fig. 3) and ESTOI and NCM (Fig. 4). Aside from STOI, those metrics that have a higher centre value in the logistic function (i.e., $b > 0.6$), tend not to extend down to zero in the non-HPF condition. As shown in Fig. 6, NI-STOI has the steepest slope and the highest centre values, with a long left tail of zero values in the HPF condition. This pattern may have negative implications, at least when the range of metric values is large, as in the case of the HPF condition. This is, first, because there may be an expectation that the whole range from 0 to 1 will be used, and that values along that scale will respond to a simple intelligibility rating, e.g., "bad," "fair," or "good," as in the case of STI. However, mapping potentially avoids the need for a metric to span from zero to one. Second, as the slope of the fitted line is extremely steep (as is reflected in the very small interquartile range), prediction of a single intelligibility score given a single metric value may be difficult.

Conventional figures of merit—correlation coefficients (ρ, τ) and the standard deviation of the prediction error (σ_e)—are reported for IBM $LC = 0$ in Table III. For ρ, 95% confidence intervals are reported to allow the identification of significant differences between metrics.

In the non-HPF condition, for male speakers, $CSII_{High}$ does not perform as well as NCM or $CSII_{Mid}$, while the other metrics perform similarly. For female speakers, NSEC does not perform as well as STOI, ESTOI, NCM, $CSII_{High}$, and $CSII_{Mid}$ or NI-STOI, while STOI and STOI+ do not perform as well as NCM (although there is only a difference of 0.02 in Kendall's τ); NCM and ESTOI perform similarly.

In the HPF condition, for male speakers, STOI+, ESTOI and NSEC perform similarly well and greatly outperform STOI. For female speakers, STOI+, ESTOI, NCM and NI-STOI perform similarly well, and greatly outperform STOI. In this condition, STOI performs less well than the other metrics both for male and female speakers due to an overestimation of the intelligibility of HP-filtered speech mixed with WGN at very low SNRs (Fig. 3). However, STOI+, which excludes the normalisation and clipping procedures, performs as well as other metrics, i.e., ρ confidence intervals overlap. Likewise, ESTOI performs well.

With regard to metric bias and reliability, in the HPF condition, STOI is positively biased and highly unreliable (Fig. 7). The STOI+ median bias is closer to 0% than the median biases of STOI, ESTOI, NCM, and $CSII_{Mid}$. The performance is similar for STOI+, NSEC, and NI-STOI. Across both filter conditions, STOI+, NCM, and NI-STOI have a relatively small bias and high reliability, with a median close to zero and relatively small interquartile ranges. Of the three intrusive STOI metrics, STOI+ has the smallest bias.

IV. DISCUSSION

While the use of the IRM algorithm resulted in scores close to 100% for all SNRs between –29 and –5 dB, IBM $LC = 0$ resulted in limited intelligibility gains for SNRs at and below –17 dB. In fact, median intelligibility scores were 0% below –17 dB SNR. The results presented here support the claim[7] that soft masks and continuous gain functions result in speech with higher intelligibility than binary masks. They are also consistent with the finding by Kjems *et al.*[38] that when $LC = 0$, the intelligibility of binary-masked speech will decrease as the SNR decreases.

Regarding the use of a high-pass filter, at all SNRs except –5 dB, the filter applied in the present study provided no significant benefit to intelligibility. At $SNR = -14$ and

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

		Males			Females		
		ρ (CIl–CIu)	τ	σ_e	ρ (CIl–CIu)	τ	σ_e
Non-HPF	*STOI*	0.85(0.81–0.88)	0.75	11.83	0.86(0.83–0.89)	0.73	10.1
	STOI+	0.86(0.82–0.89)	0.75	11.61	0.86(0.82–0.89)	0.73	10.13
	ESTOI	0.88(0.85–0.91)	0.75	10.64	**0.91(0.88–0.93)**	**0.76**	**8.49**
	NCM	**0.89(0.86–0.92)**	**0.76**	**10.26**	**0.91(0.89–0.93)**	**0.75**	**8.27**
	$CSII_{High}$	0.82(0.77–0.86)	0.73	12.92	0.88(0.85–0.91)	0.73	9.41
	$CSII_{Mid}$	**0.89(0.86–0.92)**	**0.77**	**10.19**	0.89(0.86–0.91)	0.72	9.18
	NSEC	0.88(0.85–0.91)	0.74	10.58	0.79(0.73–0.83)	0.65	12.33
	NI-STOI	0.88(0.85–0.91)	0.75	10.57	**0.90(0.87–0.92)**	**0.75**	**8.79**
HPF	*STOI*	0.61(0.52–0.69)	0.50	19.56	0.37(0.25–0.48)	0.26	21.45
	STOI+	**0.93(0.91–0.94)**	**0.76**	**9.24**	**0.95(0.93–0.96)**	**0.76**	**7.35**
	ESTOI	**0.93(0.91–0.94)**	**0.75**	**9.18**	**0.95(0.93–0.96)**	**0.77**	**7.26**
	NCM	**0.92(0.90–0.94)**	**0.75**	**9.73**	**0.96(0.95–0.97)**	**0.76**	**6.53**
	$CSII_{High}$	0.90(0.87–0.92)	0.74	10.89	0.93(0.91–0.95)	0.75	8.39
	$CSII_{Mid}$	**0.92(0.89–0.94)**	**0.75**	**9.89**	0.90(0.87–0.92)	0.74	10.21
	NSEC	**0.93(0.90–0.94)**	**0.75**	**9.28**	0.93(0.91–0.94)	0.75	8.58
	NI-STOI	**0.92(0.90–0.94)**	**0.75**	**9.73**	**0.96(0.95–0.97)**	**0.77**	**6.11**

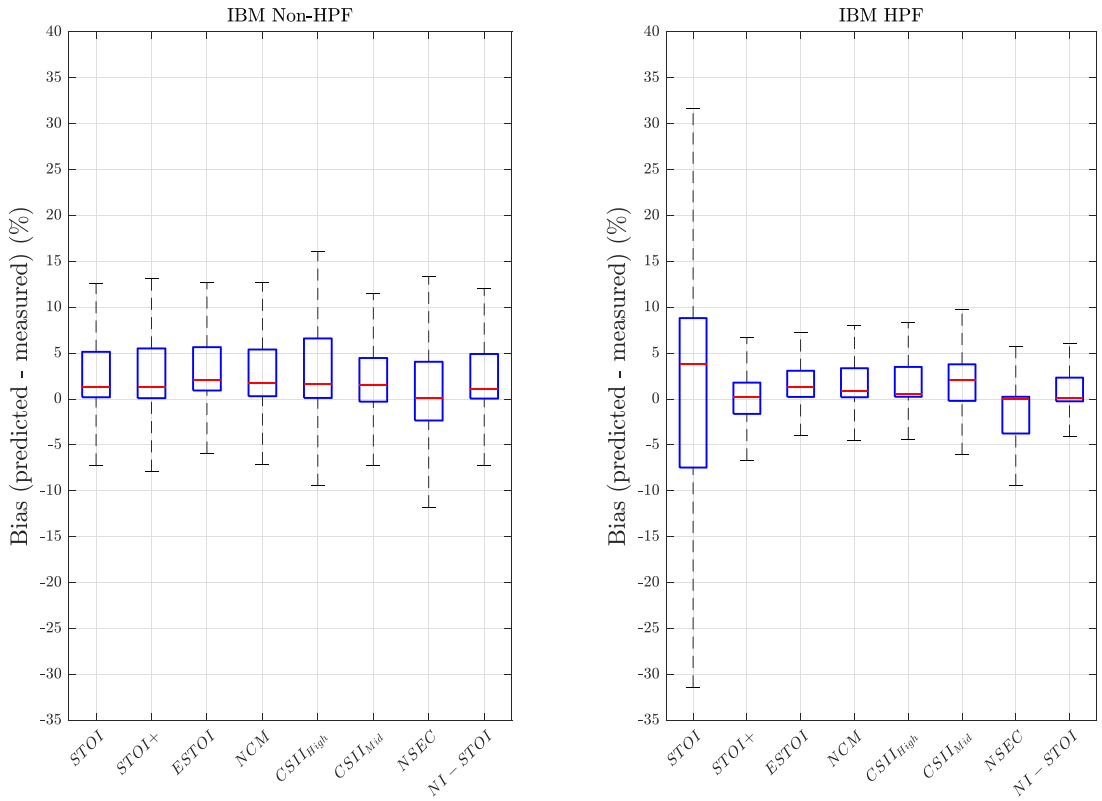

FIG. 7. (Color online) Prediction bias and reliability for the eight different metrics across talkers and SNRs for non-HPF (left) and HPF (right) speech enhanced by an IBM mask with $LC = 0$. The bias is typically positive, except for NSEC.

–17 dB, the lower intelligibility scores for the HPF condition compared to the non-HPF condition reflect the fact that at these low SNRs, when processed by IBMs with $LC = 0$, HPF signals consist of isolated short bursts of high frequency energy, musical noise, and less energy than non-HPF signals. This is because mask density is low. At $SNR = -5$ dB, any intelligibility gains in the HPF relative to the non-HPF condition are likely to be due to the preservation of high, as well as low to mid, frequency energy in the enhanced signal.

In the present study, logistic regression modelling indicated that speech intelligibility for signals processed by IBM $LC = 0$ was lower for female than male talkers at the reference SNR of –5 dB, but this difference was not apparent at $SNR = -17$ dB. The reason for this is not known but in the baseline study, when no speech enhancement method was applied, speech intelligibility was slightly higher for female than male talkers at the lower SNR of –17 dB.

The performance of STOI in predicting speech intelligibility at low mixture SNRs was evaluated and compared with that of other STOI-based intrusive metrics, STOI+ and ESTOI, in addition to NCM, $\text{CSII}_{\text{High}}$, CSII_{Mid}, NSEC, and NI-STOI. STOI+ outperformed STOI for the HPF condition using IBM with $LC = 0$ and performed similarly to more complex metrics, NCM and CSII_{Mid}. STOI+ performed particularly well relative to other metrics when the speech was HP-filtered, both in terms of conventional figures of merit and in terms of prediction bias and reliability. However, in the non-HPF condition, STOI and STOI+ performed similarly. NI-STOI also performed well on conventional figures of merit across filter conditions, despite the fact that it has no access to the clean reference signal except in voice activity detection. The reason for the poor performance of STOI in the HPF condition is evident in Fig. 3: STOI overestimates the intelligibility of some signals associated with 0% words correctly identified. This occurs when the application of the binary mask results in zero energy within a large number of frequency bands and regions because, due to the normalisation and clipping procedure, in these bands and regions, x is correlated with itself. The present and baseline studies indicate that when predicting speech intelligibility for normal-hearing listeners both before and after mask-based speech enhancement, if mixture SNRs are relatively low, STOI+, ESTOI, and NCM would be a suitable choice of objective metric. An advantage of ESTOI is that it can be applied to speech mixed with both modulated and unmodulated noise sources.[16]

Taal *et al.*[10] assessed STOI with IBMs using only 15 listeners and speech material from a single female Danish talker. Previous studies typically used one to three, and at most five, SNRs per noise type. In contrast, the present study used recordings of 12 British English talkers with nine SNRs ranging from –29 to –5 dB, with 24 listeners. However, as only WGN is considered, caution should be taken in extending the findings to fluctuating and/or narrow-band noise sources.

https://doi.org/10.1121/10.0016494

V. CONCLUSIONS

Two ideal masking algorithms—IBM $LC = 0$ and IRM—have been evaluated using listening tests with normal-hearing listeners. The commonly used IBM with $LC = 0$ performs poorly relative to the IRM for WGN and SNRs at and below –17 dB. The results for the IRM demonstrated large improvements in intelligibility over IBM with $LC = 0$ even at very low SNRs, i.e., –26 and –29 dB.

It was demonstrated that emphasising the higher frequencies of speech by means of a high-pass filter prior to mixing with WGN can make the speech more audible, hence intelligible, at these frequencies when *SNR* is relatively favourable, at $SNR = -5$ dB. However, when the mask density is very low, the IBM-processed signal is sparse, and intelligibility scores are low regardless of whether the filter has been applied.

When signals were high-pass filtered before mixing with WGN at low mixture SNRs and processed with IBM $LC = 0$, STOI overestimated intelligibility while STOI+ and ESTOI performed relatively well, as did NCM. However, caution should be taken when choosing intelligibility metrics to ensure that the metric has been validated for a specific, or at least similar, condition.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Gary Seiffert, particularly for his assistance with the speech recordings, and the other members of the Acoustics Research Unit. We also wish to thank the editors and reviewers of the current and baseline paper for their valuable contributions.

APPENDIX

The logistic function parameter values are reported within the figures; these values and their confidence intervals are provided in Tables IV and V.

TABLE IV. Free parameters for the logistic mapping of STOI, STOI+, ESTOI, NCM, $CSII_{High}$, $CSII_{Mid}$, and NI-STOI with 95% confidence intervals.

		Males		Females	
		a	*b*	*a*	*b*
Non-HPF	*STOI*	–12.94 (–14.76 to –11.12)	10.42 (9.02–11.81)	–11.67 (–13.27 to –10.06)	9.70 (8.45–10.95)
	STOI+	–11.97 (–13.62 to –10.32)	9.54 (8.30–10.78)	–10.82 (–12.33 to –9.31)	8.94 (7.78–10.10)
	ESTOI	–7.14 (–8.01 to –6.27)	4.21 (3.75–4.68)	–7.55 (–8.40 to –6.70)	4.81 (4.33–5.29)
	NCM	–10.12 (–11.32 to –8.92)	7.58 (6.73–8.42)	–10.47 (–11.62 to –9.31)	8.20 (7.35–9.05)
	$CSII_{High}$	–10.15 (–11.80 to –8.49)	8.17 (6.93–9.41)	–12.86 (–14.52 to –11.21)	10.51 (9.24–11.79)
	$CSII_{Mid}$	–9.52 (–10.65 to –8.40)	4.50 (4.01–4.98)	–8.99 (–10.04 to –7.94)	4.56 (4.09–5.03)
	NI-STOI	–64.91 (–72.50 to –57.31)	58.20 (51.45–64.95)	–60.68 (–67.31 to –54.05)	54.61 (48.71–60.51)
HPF	*STOI*	–6.35 (–7.89 to –4.80)	5.57 (4.41–6.73)	–3.22 (–4.59 to –1.85)	3.74 (2.74–4.74)
	STOI+	–9.64 (–10.67 to –8.62)	7.24 (6.49–7.98)	–10.24 (–11.16 to –9.33)	7.28 (6.66–7.91)
	ESTOI	–7.36 (–8.10 to –6.62)	4.28 (3.87–4.70)	–9.18 (–10.00 to –8.35)	5.11 (4.67–5.54)
	NCM	–8.44 (–9.34 to –7.53)	6.40 (5.74–7.05)	–10.71 (–11.57 to –9.85)	7.83 (7.23–8.44)
	$CSII_{High}$	–9.28 (–10.45 to –8.11)	6.03 (5.31–6.76)	–12.43 (–13.79 to –11.07)	7.83 (7.00–8.66)
	$CSII_{Mid}$	–9.37 (–10.36 to –8.37)	3.94 (3.55–4.34)	–10.81 (–12.16 to –9.46)	4.59 (4.05–5.13)
	NI-STOI	–54.91 (–61.66 to –48.16)	49.35 (43.31–55.38)	–66.82 (–72.06 to –61.58)	60.08 (55.39–64.77)

TABLE V. Free parameters for NSEC logistic mapping with 95% confidence intervals. Confidence intervals are not provided for *c* as these are unrealistic (one exception is males with non-HPF).

	Males			Females		
	a	*b*	*c*	*a*	*b*	*c*
Non-HPF	0.95 (–3.49–5.40)	0.19 (–3.59–3.97)	–22.79 (–133.21–87.64)	0.05(–5.37–5.48)	0.95 (–3.75–5.66)	–298.24
HPF	0.06 (–4.83–4.96)	0.95 (–3.31–5.20)	–398.16	0.01(–5.02–5.04)	0.99 (–3.26–5.24)	–2282.17

[1]Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," J. Speech Commun. **51**(3), 230–239 (2009).
[2]S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," J. Speech Commun. **48**(11), 1486–1501 (2006).
[3]R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304 (1995).
[4]P. Souza and S. Rosen, "Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech," J. Acoust. Soc. Am. **126**(2), 792–805 (2009).
[5]J. M. Kates and K. H. Arehart, "The hearing aid speech quality index (HASQI)," J. Audio Eng. Soc. **58**, 363–381 (2010).
[6]J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," J. Speech Commun. **65**, 75–93 (2014).
[7]C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation* (Springer, Berlin, 2014), pp. 349–368.
[8]Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1849–1858 (2014).
[9]S. Graetzer and C. Hopkins, "Intelligibility prediction for speech mixed with white Gaussian noise at low signal-to-noise ratios," J. Acoust. Soc. Am. **149**(2), 1346–1362 (2021).
[10]C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio. Speech Lang. Process. **19**(7), 2125–2136 (2011).
[11]R. Hendriks, C. Taal, R. Heusdens, and J. Jensen, "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Proceedings of the 2010 International Workshop on Acoustic Echo Noise Control*, Tel Aviv, Israel (August 30–September 2, 2010).
[12]S. Graetzer and C. Hopkins, "Evaluation of STOI for speech at low signal-to-noise ratios after enhancement with ideal binary masks," in *Proceedings of the 25th International Congress on Sound and Vibration (ICSV)*, Hiroshima, Japan (July 8–12, 2018).
[13]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)*, Bucharest, Romania (August 27–31, 2012), pp. 504–508.
[14]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)*, South Brisbane, Australia (April 19–24, 2015), pp. 5078–5082.
[15]A. H. Andersen, J. M. De Haan, Z. H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," IEEE/ACM Trans. Audio. Speech Lang. Process. **24**(11), 1908–1920 (2016).
[16]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. **24**(11), 2009–2022 (2016).
[17]K. Steinmetzger, J. Zaar, H. Relano-Iborra, S. Rosen, and T. Dau, "Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations," J. Acoust. Soc. Am. **146**(4), 2562–2576 (2019).
[18]A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA (March 5–9, 2017), pp. 5085–5089.
[19]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).
[20]I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**(3), 1703–1716 (1996).
[21]J. Kates and K. Arehart, "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**(4), 2224–2237 (2005).
[22]J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland (August 24–28, 2009), pp. 1849–1853.
[23]IEEE, "Recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 227–246 (1969).
[24]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, Liverpool, UK, http://dx.doi.org/10.17638/datacat.liverpool.ac.uk/681 (Last viewed April 22, 2022).
[25]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).
[26]D. Wang, "MATLAB toolbox for cochleagram analysis and synthesis," https://web.cse.ohio-state.edu/~wang.77/pnl/shareware/cochleagram/ (Last viewed April 22, 2022).
[27]B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753 (1983).
[28]J. M. Kates, "Modeling the effects of single-microphone noise-suppression," Speech Commun. **90**, 15–25 (2017).
[29]ISO 8253-1:2010: "Acoustics, audiometric test methods Part 1: Pure-tone air and bone conduction audoimetry" (International Organization for Standardization, Geneva, Switzerland, 2010).
[30]Y. Tang, R. J. Hughes, B. M. 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).
[31]ANSI S3.5 (R2007): *Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York, 1997).
[32]P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL, 2013).
[33]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*, Beijing, China (September 14–17, 1992), pp. H2–H6.
[34]K. H. Arehart, J. M. Kates, M. C. Anderson, and L. O. Harvey, "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).
[35]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. Am. **130**(5), 3013–3027 (2011).
[36]C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX (March 14–19, 2010), pp. 4214–4217.
[37]R Core Team, "R: A language and environment for statistical computing," https://www.R-project.org/ (Last viewed December 3, 2022).
[38]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. Am. **126**(3), 1415–1426 (2009).