PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

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

Simone GRAETZER[1]; Carl HOPKINS[1]

[1] Acoustics Research Unit, School of Architecture, University of Liverpool, UK

ABSTRACT

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) that are typically 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 low SNRs (from -29 to -5 dB). These comprise an Ideal Binary Mask (IBM) with a local criterion set to 0 dB and an Ideal Ratio Mask (IRM). The performance of Short-Time Objective Intelligibility (STOI), and a STOI variant (termed STOI+), 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 STOI+ is a suitable metric for speech mixed with WGN at low SNRs and processed by IBMs with $LC = 0$, even when the speech is high-pass filtered to flatten the spectral tilt.

Keywords: Speech, Intelligibility, Enhancement

1. 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 masks are estimated with access to the clean speech signal such that the user-defined rule specifies a Local Criterion (*LC*). For example, with $LC = 0$ dB, the gain is set to one when the difference between the clean speech and the noise 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. In the case of signals mixed with broadband noise at very low SNRs, the enhanced signal effectively becomes binary-gated noise, *i.e.*, noise on which 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 an Ideal Ratio Mask (IRM); this determines the mask value from the estimated ratio of the target and the mixture signal energy (1). Assessments of IBMs and IRMs in the literature tend to use SNRs between -5 and +10 dB, for which IRMs are shown to have potential advantages in terms of speech quality and intelligibility (e.g., see 2,3). In this paper, the aim is to assess the intelligibility benefits of both binary and ratio ideal masks for speech mixed with white Gaussian noise (WGN) for low SNRs, which range from -29 to -5 dB.

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 STOI (4) have been shown to perform relatively well in predicting intelligibility and, in particular, the effects on intelligibility of non-linear speech processing (e.g., for noise suppression, in which context 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 application of time-frequency varying gain functions. It is suited to mixtures of speech and noise that have been subjected to non-linear processing, where the

[1] Current email addresses: s.n.graetzer@salford.ac.uk, carl.hopkins@liverpool.ac.uk

noise or degradation is not additive. STOI tends to outperform traditional objective metrics for ideal TFS-processed speech at least for sentences for which intelligibility scores are ≥ 20% and mixture SNRs exceed -10 dB (5). Recent work by the authors (6) on the evaluation of STOI for speech mixed with noise at low SNRs after enhancement with IBMs indicates that an improved STOI-based metric (referred to in this paper as STOI+) would not use clipping; hence this STOI variant is also assessed.

The experiments reported in this paper compare the performance of the IBM with $LC = 0$ and the IRM for WGN and low SNRs, and compare the performance of STOI+ with STOI, non-intrusive STOI (NI-STOI) (7), the Normalised Covariance Metric (NCM) and the Coherence Speech Intelligibility Metric or *CSII* (8,9). This study aimed 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 and the proposed STOI variant, termed STOI+, in predicting the percentages of words correctly identified for speech processed by an IBM algorithm with $LC = 0$.

2. EXPERIMENTAL PROCEDURE

2.1 Objective metrics

In this paper, STOI was calculated using publicly available code (4). The STOI+ metric is the same as STOI but with the clipping function removed and where only finite intermediate d values are averaged to obtain the final d value. If clipping is not performed, the correlation coefficients are independent of any normalisation. NI-STOI is similar to STOI except that it estimates clean signal envelopes from the degraded speech envelopes. The true clean signal is only used to determine which frames include speech via a voice activity detector (*i.e.*, the voice activity detector is 'ideal'). 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' (7, p. 5086). The clipping process is not applied. The NCM is calculated according to Goldsworthy and Greenberg's method (10). CSII is computed using short-time segments with a window length of 30 ms and a 7.5 ms frame shift. Only the results for the mid-level region, $CSII_m$, are reported.

Separate logistic mappings are computed for talker gender, filter and masking algorithm conditions. Transformation via the rotationally symmetric logistic function, $f(x) = 100/(1+e^{ax+b})$, allows the calculation of parametric *figures of merit*: linear correlation coefficients (Pearson's ρ) and estimates of the prediction error based on these coefficients (σ_e; see 4).

To compute the mean metric bias, b, the measured scores, y, were subtracted from the corresponding predicted scores, x, as follows, where N is the number of measured scores: $b = 1/N \cdot \sum(x - y)$. The interquartile range of the bias indicates the reliability of the predictions, with smaller ranges indicating higher reliability.

2.2 Talkers and speech material

For the speech material, the 720 IEEE sentences (11) were produced with a normal vocal effort by twelve talkers (six male, six female) in an anechoic chamber. The talkers were native speakers of British English and their speech was close to Received Pronunciation. The speech signals were high-pass filtered to remove energy below 60 Hz using a Finite Impulse Response (FIR) filter with a Kaiser window, and low-pass filtered to attenuate energy above 9 kHz (predominantly electrical background noise). These recordings are freely available in the ARU speech corpus (12).

2.3 Subjects and method

The experiment received prior approval from the University of Liverpool ethics committee. Twenty-four untrained listeners between the ages of 18 and 45 were recruited as subjects (12 male,12 female). All listeners used English as a first language. Before the experiment, each subject underwent an audiometric screening test according to ISO 8253-1 (13) to determine their thresholds of hearing between 125 Hz and 8 kHz. Only listeners with a hearing loss less than 20 dB hearing level (HL) took part in the experiment.

Each listener heard sentences produced by four of the twelve talkers (two male, two female), who were selected pseudo-randomly. All listeners were assigned nine SNRs ϵ [-29 -26 -23 -20 -17 -14 -11 -8 -5] dB, two filter conditions (with a high-pass filter, HPF, and without, non-HPF; defined in Section 2.4) and one masking algorithm. Each SNR and filter combination was used for one word list for each of the four talkers. Each listener therefore participated in a total of 72 listening conditions (4 talkers

x 9 SNRs x 2 filter conditions x 1 masking algorithm). The total testing time per listener was approximately 2.5 hours, which was divided into two sessions. The subject was asked to type the words they identified into a text box in a MATLAB custom graphical user interface (GUI) within a time limit of 15 s per sentence. The subject was asked not to guess any words and to check their spelling before submitting words. Incorrect spelling was identified and assessed after the experiment using the following rules: (a) allow misspellings using "a" instead of "e," (b) ignore punctuation such as apostrophes, (c) allow homonyms, and (d) allow either American English or British English spelling.

2.4 Signal processing

The clean speech signals are termed 'non-HPF'. Duplicates of these signals were high-pass filtered by means of a linear-phase equiripple type FIR filter, which was used to flatten the spectral tilt by increasing the magnitude of the speech signal as the normalised frequency approached a value of one and decreasing it as the normalised frequency approached zero. This 'pre-emphasis' filter condition is termed 'HPF'. Long-term average spectra per filter condition and talker gender are shown in Figure 1. The active speech levels of all signals were equalised using the procedure in ITU-T P.56 (14). A pseudo-randomly selected segment of WGN was added to the speech signal to obtain the required SNRs based on the active speech level (14). This additive noise was gated on and off 1 s before and after the signal respectively.

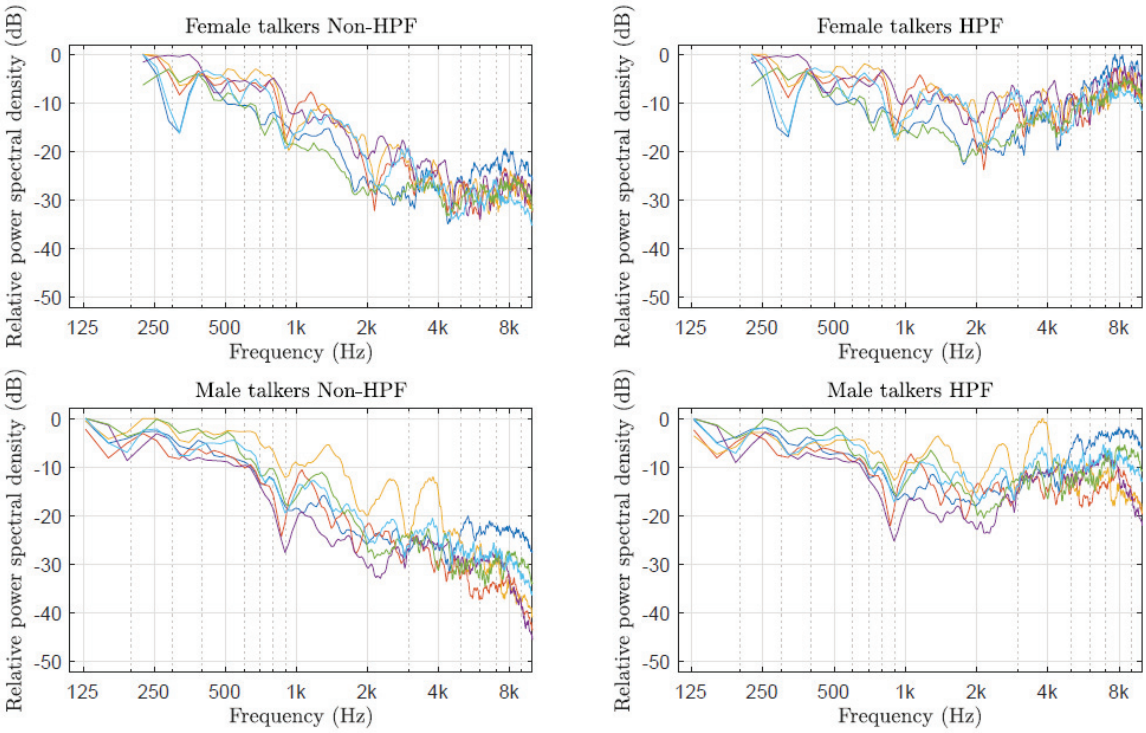

Figure 1 – Long-term average spectra for male and female talkers in the HPF and non-HPF conditions.

As has been mentioned, in this experiment, two speech enhancement algorithms are evaluated: IBM with $LC = 0$ and IRM. Other masking algorithms were also evaluated in the larger experiment but are not presented here due to lack of space. Algorithms were based on publicly available code from Wang (15). The frequency range for both types of mask is 80 Hz to 12 kHz, where the lower limit 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 is half the sampling rate (after downsampling to 24 kHz).

The process of speech enhancement via 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 and M, the masker. This function involves assessing whether, in each channel and frame, the difference between the energy in the target cochleagram and in the scaled masker cochleagram is greater than $LC = 0$ dB.

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

To obtain the signals, a fourth-order gammatone filterbank was used with 128 channels, a gammatone filter length of 128 ms, and a frequency range of 80 Hz to 12 kHz with frequencies equally spaced on the ERB rate scale, with middle-ear loudness-based gain normalisation. The output of the gammatone filterbank was used to generate a cochleagram with a window length of 20 ms with a 10 ms frame shift. The IRM is obtained on the basis of the gain function shown in Eq. 2, where the 'tuning' or compression constant, $\beta = 0.5$.

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

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 Figure 2.

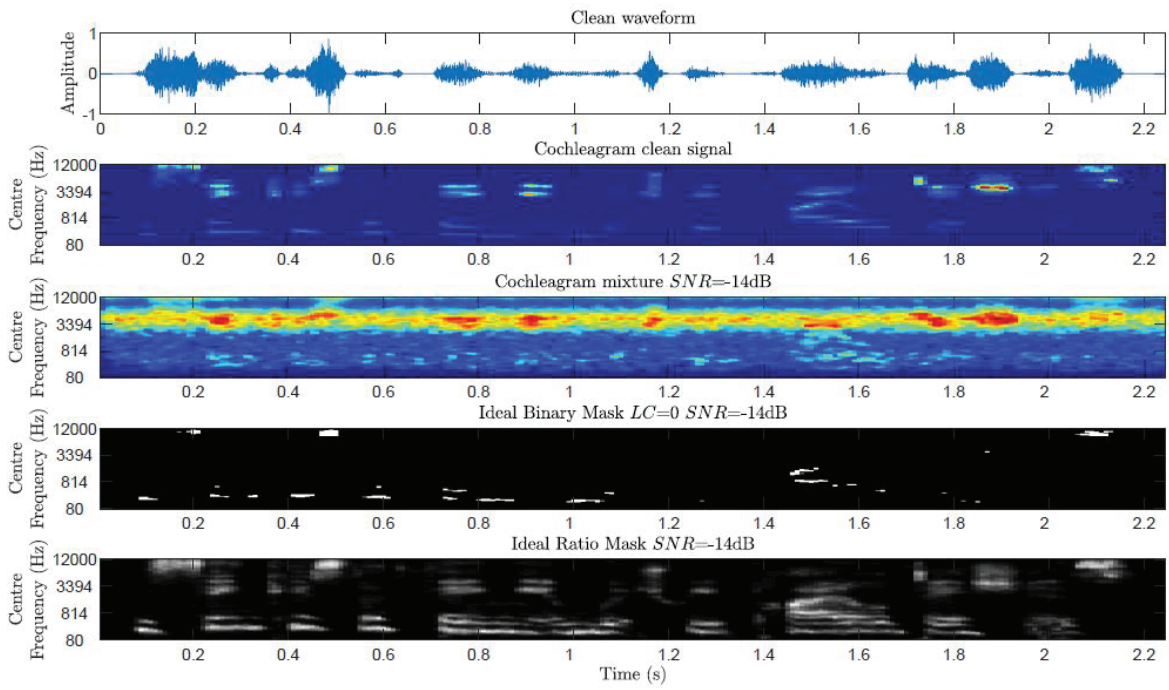

Figure 2 – Example Ideal Binary and Ratio Masks with mixture *SNR* = -14 dB.

2.5 Signal presentation

The experiment took place in a sound-attenuated booth in which the background noise level at the listener position was 22 dB L_{Aeq}. The stimuli were presented diotically to the subjects using a playback system that consisted of Beyer Dynamic DT770 Pro headphones (electronically limited to 95 dB output) and a PC located outside the sound-attenuated booth. The PC ran a MATLAB GUI. The audio output of the playback system (headphones and PC) was calibrated using a B&K type 4100 head-and-torso simulator (HATS) (Type 4189 microphones in each ear canal). The background noise measured at the entrance to the ear canal with the HATS wearing headphones connected to the PC was also 22 dB L_{Aeq}.

The sampling rate of the signals presented was 24 kHz with 16-bit resolution. The presentation level was set by the subjects at the beginning of the experiment to 70 or 75 dB L_{Aeq}. When setting the presentation level, listeners heard sentences processed by their assigned masking algorithm with a mixture SNR of -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 assigned to the listener.

3. RESULTS

3.1 Intelligibility scores

Figure 3 shows the intelligibility scores per masking algorithm expressed as words identified correctly per word list by each listener per talker, SNR and filter condition. 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 measured as 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) was <5%. For IBM with LC = 0, 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. The results show that 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. Unlike the IBM with LC = 0, the performance of the IRM did not depend on SNR.

For IBM with LC = 0, a mixed-effects logistic regression model was fit to the intelligibility scores expressed as successes and failures out of trials and with listener, talker and word list as random effects and SNR and filter condition as fixed effects. The non-HPF condition was the filter reference condition. The model was chosen based on nested model comparisons using likelihood ratio tests with a Chi-squared test statistic. The results are shown in Table 1. There was an effect of SNR such that for every 1 dB increase in SNR, there was a 0.29 increase in the log odds (O = 1.33) of identifying a word correctly. There was a (main) effect of filter on intelligibility scores such that the HPF was beneficial to intelligibility when the SNR was 0 dB; there was an estimated 0.43 increase in the log odds of identifying a word correctly (O = 1.54). However, as the focus in this paper is on $SNR \leq$ -5 dB, it is the interaction of filter and SNR that is of interest: at the highest SNRs the HPF tended to improve intelligibility, while it was detrimental at lower SNRs. The intelligibility benefits of the HPF filter condition were evident at SNRs between -11 and -5 dB for male talkers and -8 and -5 dB for female talkers.

Table 1 – Mixed-effects logistic regression model for IBM LC = 0.

	Estimate	SE	z	p
(Intercept)	1.87	0.2	9.48	< 0.0001
SNR	0.29	0	61.1	< 0.0001
Filter HPF	0.43	0.08	5.66	< 0.0001
Filter HPF:SNR	0.05	0.01	6.54	< 0.0001

3.2 Metric evaluation

Ideally, any metric should make full use of its defined range, *i.e.*, [0,1]. $CSII_m$ values were found to extend down to zero for unintelligible speech signals, whereas STOI+ d and NCM did not extend down to zero for non-HPF speech but did for HPF speech, and STOI did not extend below 0.2 for either filter condition (see Figure 4). NI-STOI is unsuitable for prediction, as it has a range of only 0.13 with a minimum of 0.78 for non-HPF speech and only assigns values > 0.75 for speech with intelligibility scores above zero.

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 2. STOI+, but not STOI, performs as well as other metrics (ρ confidence intervals overlap). STOI grossly overestimates the intelligibility of HPF speech mixed with WGN at very low SNRs and processed using an IBM with LC = 0 (Figure 4). The overestimation is likely to be due to STOI correlating the clean speech signal with itself when mixture SNRs are low, the noisy speech is highly degraded, and the processed signal is sparse, which generates spurious intermediate d values equal to one.

STOI is positively biased and unreliable in the HPF condition. STOI+, NCM, and NI-STOI have a relatively small bias and high reliability in both filter conditions, with a median close to zero and

relatively small interquartile ranges. $CSII_m$ is slightly more positively biased than STOI+, NCM, and NI-STOI but is relatively reliable.

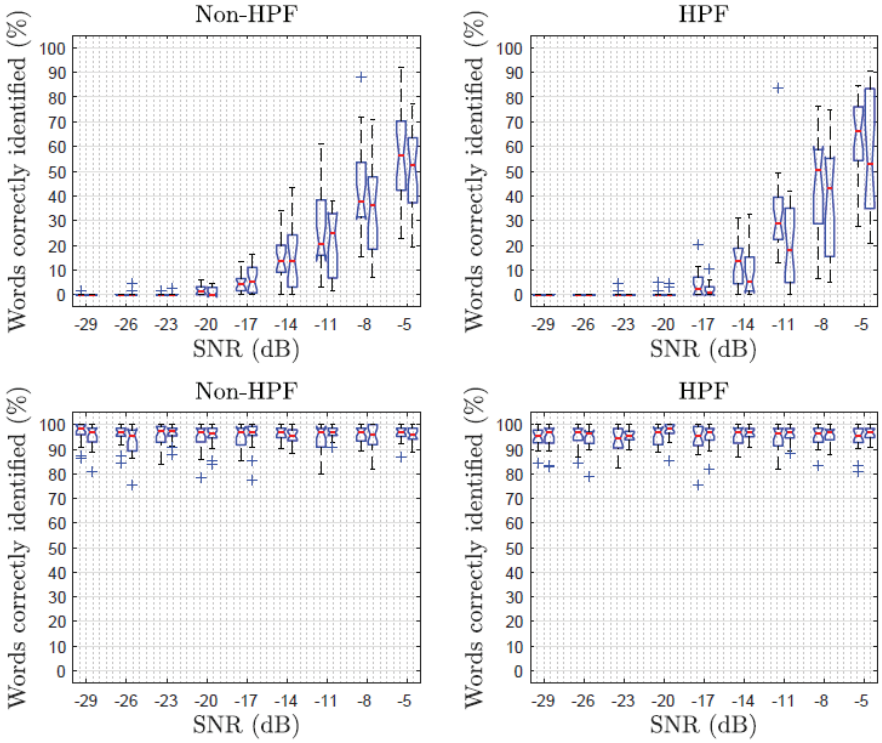

Figure 3 – Boxplots of speech intelligibility scores for IBM with $LC = 0$ (upper plots) and IRM (lower plots). At each SNR, male talkers are shown to the left, and female talkers to the right.

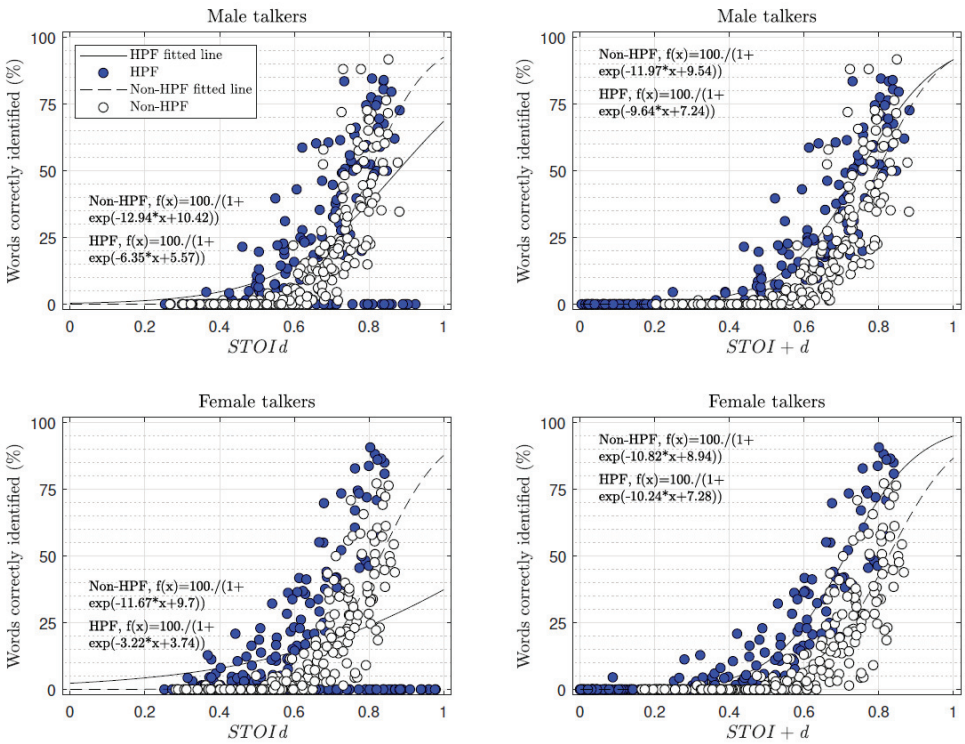

Figure 4 – Scatterplots of STOI and STOI+ by intelligibility scores with fitted lines deriving from the rotationally symmetric logistic function.

Table 2 – Metric performance for IBM $LC = 0$.

		Non-HPF			HPF		
Talkers	Metric	ρ	τ	σ_e	ρ	τ	σ_e
Male	STOI	0.85	0.75	11.98	0.60	0.46	20.01
	STOI+	0.86	0.76	11.73	0.93	0.75	9.36
	NI-STOI	0.89	0.75	10.66	0.92	0.74	9.85
	NCM	0.90	0.76	10.26	0.92	0.75	9.64
	$CSII_m$	0.90	0.77	10.27	0.92	0.76	9.72
Female	STOI	0.87	0.72	10.41	0.36	0.24	21.83
	STOI+	0.86	0.72	10.44	0.95	0.75	7.35
	NI-STOI	0.90	0.74	9.11	0.97	0.78	5.98
	NCM	0.91	0.73	8.73	0.96	0.77	6.52
	$CSII_m$	0.89	0.71	9.58	0.90	0.74	10.05

4. 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. The results presented here are consistent with the claim that soft masks and continuous gain functions result in speech with higher intelligibility – and higher quality – than binary masks (2).

Regarding the use of a high-pass filter, at lower SNRs, lower intelligibility scores for the HPF condition than the non-HPF condition reflect the fact that at these 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. At higher SNRs, 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.

The performance of STOI in predicting speech intelligibility at low mixture SNRs was evaluated and compared with that of proposed variant STOI+, NI-STOI, NCM and $CSII_m$. STOI+ outperformed STOI for the HPF condition using IBM with $LC = 0$ and performed similarly to more complex metrics, NCM and $CSII_m$. NI-STOI also performed well on conventional figures of merit but was found to be unsuitable for use in intelligibility prediction. The results extend the assessment of STOI to British English as Taal *et al*. (4) assessed STOI with IBMs using only 15 listeners and speech material from a single female Danish talker and previous studies typically consider one to three, and at most five, SNRs per noise type. In contrast, this 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 narrowband noise sources.

5. CONCLUSIONS

Two masking algorithms have been evaluated using listening tests with normal-hearing listeners. It was shown that the commonly used IBM with $LC = 0$ dB performs relatively poorly for WGN and SNRs at and below -17 dB. The results for the IRM demonstrated significant improvements in intelligibility over IBM with $LC = 0$ even at extremely low SNR levels, *i.e.*, -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 $SNR \geq -8$ 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 grossly overestimated intelligibility while STOI+ performed relatively well under all conditions. 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.

REFERENCES

1. Srinivasan S, Roman N, Wang DL. Binary and ratio time-frequency masks for robust speech recognition. Speech Communication 2006;48:1486–1501.
2. Hummersone C, Stokes T, Brookes, T. On the ideal ratio mask as the goal of computational auditory scene analysis. In Naik GR, Wang W, editors. Blind source separation. Springer, Berlin, Heidelberg; 2014. p. 349-368.
3. Wang Y, Narayanan A, and Wang DL. On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 2014;22:1849–1858.
4. Taal CH, Hendriks RC, Heusdens R, Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio Speech and Language Processing 2011;19(7):2125-2136.
5. Taal CH, Hendriks RC, Heusdens R, Jensen, J. On predicting the difference in intelligibility before and after single-channel noise reduction. In Proc. Int. Workshop Acoust. Echo Noise Control 2010.
6. Graetzer S, Hopkins C. Evaluation of STOI for speech at low signal-to-noise ratios after enhancement with ideal binary masks. Proc 25th International Congress on Sound and Vibration (ICSV), Hiroshima, 2018.
7. Andersen AH, de Haan JM, Tan ZH, Jensen, J. A non-intrusive short-time objective intelligibility measure. Proc Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference 2017. p. 5085-5089.
8. Holube I, Kollmeier B. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J Acoust Soc Am 1996;100(3):1703-1716.
9. Kates JM, Arehart KH. Coherence and the speech intelligibility index. J Acoust Soc Am 2005;117(4): 2224-2237.
10. Goldsworthy RL, Greenberg JE. Analysis of speech-based speech transmission index methods with implications for non-linear operations, J Acoust Soc Am 2004;116:3679–3689.
11. IEEE. Recommended practice for speech quality measurements, IEEE Transactions on Audio and Electroacoustics 1969;17(3):227-246.
12. Hopkins C, Graetzer S, Seiffert, G. 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, 2019. Available from: http://dx.doi.org/10.17638/datacat.liverpool.ac.uk/681
13. BS EN ISO 8253-1. Acoustics. Audiometric test methods Part 1: Pure-tone air and bone conduction audiometry, 2010.
14. ITU-T P.56. Objective measurement of active speech level. Recommendation P.56. International Telecommunication Union, 2011.
15. Wang, DL. MATLAB code, 1995-2014. Available from: http://web.cse.ohio-state.edu/pnl/software.html