IMPROVING THE PERCEPTUAL QUALITY OF IDEAL BINARY MASKED SPEECH

Leo Lightburn[1], Enzo De Sena[2], Alastair Moore[1], Patrick A. Naylor[1], Mike Brookes[1]

[1]Dept. of Electrical and Electronic Engineering, Imperial College London, SW7 2BX, UK
[2]Institute of Sound Recording, University of Surrey, GU2 7XH, UK

ABSTRACT

It is known that applying a time-frequency binary mask to very noisy speech can improve its intelligibility but results in poor perceptual quality. In this paper we propose a new approach to applying a binary mask that combines the intelligibility gains of conventional binary masking with the perceptual quality gains of a classical speech enhancer. The binary mask is not applied directly as a time-frequency gain as in most previous studies. Instead, the mask is used to supply prior information to a classical speech enhancer about the probability of speech presence in different time-frequency regions. Using an oracle ideal binary mask, we show that the proposed method results in a higher predicted quality than other methods of applying a binary mask whilst preserving the improvements in predicted intelligibility.

Index Terms— Binary mask, speech quality, speech intelligibility, speech enhancement, speech presence probability

1. INTRODUCTION

At Signal-to-Noise Ratios (SNRs) below about $0\,\mathrm{dB}$ both the intelligibility and the perceived quality of noisy speech degrade sharply. Classical enhancement algorithms [1, 2, 3] process noisy speech by transforming it into the Time-Frequency (TF) domain and multiplying it by a TF gain before transforming back into the time domain for resynthesis. The TF gain is typically chosen to minimize the expected squared error in the Spectral Amplitudes (SAs) or Log Spectral Amplitudes (LSAs) for an assumed statistical model. Unfortunately, while these algorithms are able to give substantial improvements in both SNR and perceptual quality, they are normally unable to improve the intelligibility of the speech [4, 5]. More recently, algorithms that estimate the TF gain (also called the "ratio mask") using Deep Neural Networks (DNNs) [6, 7] or Long Short-Term Memory (LSTM) recurrent neural networks [8, 9] have been successful in improving both quality and intelligibility. As would be expected, the performance of these algorithms degrades somewhat when the nature of the interfering noise differs from that used in training the neural network [6].

Numerous studies have shown that the intelligibility of noisy speech can be improved by applying a binary-valued TF gain (or "binary mask") [10, 11, 12]. The two mask values are most commonly 1 and 0 although sometimes a non-zero gain, such as 0.1, is used instead of 0. These studies have led to the development of enhancement algorithms that determine a binary mask by classifying each TF cell as "speech-dominated" or "noise-dominated" using features extracted from the noisy speech [13, 14, 15]. Although these algorithms are able to improve intelligibility, they introduce distortion artefacts into the speech that make it unpleasant to listen to. The most widely used target for training the classifier is the so-called Ideal Binary Mask (IBM) which is obtained by thresholding the local SNR in each TF cell of the noisy speech used for training. A model that is able to predict the intelligibility of the binary-masked speech as a function of the SNR of the noisy speech and the threshold (denoted the Local Criterion or LC) used to define the IBM was presented in [16]. The authors proposed that the intelligibility gain of binary-masked speech arises from the introduction of spectrotemporal noise modulation that matches the TF energy distribution of the target speech. This insight led them to propose the Target Binary Mask (TBM) which is calculated directly from the clean speech and is independent of the noise in each TF cell. Tests have shown that the TBM gives similar intelligibility improvements to the IBM but, since it is not dependent on the noise, it has been suggested that it might be a better training target [15]. Related masks are the speaker-independent Universal Target Binary Mask (UTBM) [13] and the STOI-optimal Binary Mask (SOBM) [17] which explicitly optimizes the Short-Time Objective Intelligibility Measure (STOI) intelligibility metric [18].

In order to improve the quality of binary masked speech, a number of studies have experimented with modifying the binary mask before applying it to the noisy speech. In [19] the authors evaluated a number of mask modifications including adding dither to the mask and the application of temporal smoothing to the cepstrum of the mask as suggested in [20]. They concluded that the best results were obtained by applying the mask in the conventional way using gains of 1 and 0.1 for the two mask values. In [21], the estimated mask and also its complement were used to obtain intermediate estimates of the speech and noise. These estimates were then combined to derive a continuous-valued TF gain function which was applied to the original noisy speech. A final processing stage then imposed temporal continuity on the sequence of TF spectral magnitudes. The authors found that this processing was

able to improve the quality of the enhanced speech while preserving its intelligibility.

In this paper we propose an alternative approach to applying a binary mask that preserves the intelligibility gains of conventional binary masking whilst addressing the issue of poor speech quality. We are motivated by the model from [16], described above, that the intelligibility gains of binary masked speech arise because the mask identifies the TF cells containing significant speech energy. Accordingly, in our proposed approach we do not use the mask directly as a TF gain but instead use it to supply prior information about the probability of speech presence to a classical speech enhancer [3] that minimizes the expected squared error in the LSAs. To evaluate this approach we have used an oracle IBM as the binary mask since this is commonly used in other studies.

2. PROPOSED ENHANCEMENT SCHEME

2.1. Optimally-modified log-spectral amplitude estimator

Here we present a brief overview of the Optimally Modified Log-Spectral Amplitude Estimator (OM-LSA) algorithm from [3]. The noisy speech is first converted into the Short Time Fourier Transform (STFT)-domain using overlapping Hamming analysis windows. Let $X(k,m)$, $N(k,m)$ and $Y(k,m)$ denote the complex STFT coefficients of the clean speech, the noise and noisy speech respectively in frequency bin k of frame m. Speech is absent in this bin under the hypothesis $H_0(k,m)$ and present under the hypothesis $H_1(k,m)$. The STFT coefficients of the speech and noise are modelled as statistically independent Gaussian random variables. We want the gain function $G(k,m)$ in frequency bin k of frame m which satisfies

$$G(k,m)\,|Y(k,m)| = \exp\left\{E\left[\log|X(k,m)| \mid Y(k,m)\right]\right\}$$

where $E[\cdot]$ is the expectation operator. Under the constraint that $G(k,m)$ is larger than a threshold G_{min} when speech is absent, it is shown in [3] that

$$G(k,m) = \left\{G_{H_1}(k,m)\right\}^{p(k,m)} G_{min}^{1-p(k,m)}$$

where

$$G_{H_1}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)} \exp\left(\frac{1}{2}\int_{v(k,m)}^{\infty} \frac{e^{-t}}{t} dt\right)$$

is the gain under hypothesis $H_1(k,m)$. The conditional speech presence probability $p(k,m) \triangleq P\left(H_1(k,m) \mid Y(k,m)\right)$ is computed as

$$p(k,m) = \left\{1 + \frac{q(k,m)}{1-q(k,m)}\left(1+\xi(k,m)\right)\exp\left(-v(k,l)\right)\right\}$$

where $q(k,m) \triangleq P\left(H_0(k,m)\right)$ is the *a priori* probability of speech absence. In the expression for $p(k,m)$,

$$\xi(k,m) \triangleq E\left[|X(k,m)|^2 \mid H_1(k,m)\right] / E\left[|N(k,m)|^2\right]$$

is the *a priori* SNR, and

$$v(k,m) \triangleq \gamma(k,m)\xi(k,m) / \left(1+\xi(k,m)\right),$$

where

$$\gamma(k,m) \triangleq |Y(k,m)|^2 / E\left[|N(k,m)|^2\right]$$

is the *a posteriori* SNR. An estimate $\hat{\xi}(k,m)$ of $\xi(k,m)$ is obtained using a modified version of the decision-directed approach from [1],

$$\hat{\xi}(k,m) = \alpha G_{H_1}^2(k,m-1)\gamma(k,m-1) + (1-\alpha)\max\left\{\gamma(\mathrm{k,m})-1, 0\right\}$$

where α is a smoothing parameter.

2.2. Speech presence probability prior

In [3] an estimator $\hat{q}(k,m)$ was used to obtain the probability of speech absence, $q(k,m)$, from $\hat{\xi}(k,m)$. We propose to instead obtain $q(k,m)$ from a binary mask, $d(k,m)$. The *a priori* probability parameter $q(k,m)$ from [3] is set to

$$q(k,m) = \begin{cases} Q^1 & d(k,m) = 1 \\ Q^0 & d(k,m) = 0 \end{cases}$$

where Q^1 and Q^0 are free parameters. Similarly, the value of G_{min} is set to

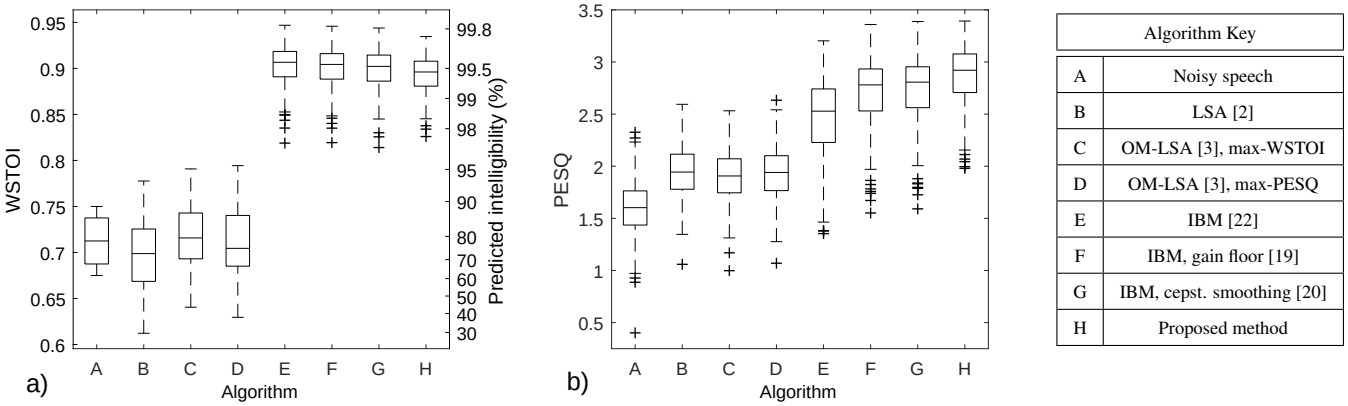
$$G_{min} = \begin{cases} G^1 & d(k,m) = 1 \\ G^0 & d(k,m) = 0 \end{cases}$$

where G^1 and G^0 are free parameters.

By using the value of the binary mask to control the probability of speech absence in this way, the algorithm softly imposes on the enhanced speech the spectrotemporal modulations that are encapsulated in the mask and that are important for speech intelligibility [16, 18]. At the same time, the algorithm improves the SNR and the perceived quality of the speech by applying an SNR-dependent time-frequency gain, $G(k,m)$.

3. EXPERIMENTAL PROCEDURES

The enhancement scheme outlined in the previous section was tested on 80 TIMIT [23] utterances mixed with extracts of babble and speech shaped (SS) noise from the RSG.10 [24] database. All signals were resampled to 10 kHz. The noisy utterance SNRs were chosen so that the Weighted-STOI (WSTOI) [25] objective intelligibility metric gave scores of $\{0.675,\ 0.700,\ 0.725,\ 0.750\}$ which correspond to average SNRs of $\{-3.6,\ -1.5,\ -1.3,\ 3.8\}$ dB for babble noise and $\{-4.4,\ -3.0,\ -1.5,\ -0.3\}$ dB for SS noise.



Algorithm Key	
A	Noisy speech
B	LSA [2]
C	OM-LSA [3], max-WSTOI
D	OM-LSA [3], max-PESQ
E	IBM [22]
F	IBM, gain floor [19]
G	IBM, cepst. smoothing [20]
H	Proposed method

Fig. 1: Boxplots of a) WSTOI and b) PESQ, for noisy speech utterances after processing with different enhancement algorithms. Methods E-H use an IBM which is computed using oracle knowledge of the clean speech.

The STFT used to compute the IBM used 50% overlapping Hanning analysis windows of length 25.6 ms.

A total of eight enhancement methods were evaluated (labelled A through H). The proposed method (H) was compared with the LSA estimator (B) [2], the OM-LSA estimator (C, D) [3], the IBM (E) [22] and two methods of improving the intelligibility of IBM-masked speech: cepstral smoothing of the IBM (G) [20], and replacing mask zeros with a minimum gain, $\mathcal{M}_{\min}$, (F) [19]. The IBMs used in methods E-H were computed using oracle knowledge of the clean speech. The algorithm parameters listed in Table 1 were optimised on a separate training set of 80 noisy speech utterances. The parameters of methods F and G were chosen to maximise the Perceptual Evaluation of Speech Quality (PESQ) objective quality metric. The parameters of method H were chosen to maximise the sum of a normalised predicted intelligibility score and a normalised predicted Mean Opinion Score (MOS), where each PESQ score was mapped to a predicted MOS using the mapping from [26] and each WSTOI score was mapped to a predicted intelligibility using the mapping from [25]. The OM-LSA algorithm parameter was chosen to optimize either WSTOI (method C) or PESQ (method D). For all algorithm parameters other than those listed in Table 1, the default values from [27, 20, 3] were used. The l_{env}, l_{low} and l_{high} cepstral smoothing parameters in [20] were adjusted to account for the 10 kHz sample rate. The LSA (B), the OM-LSA (C, D) and the proposed method (H) used the noise estimator from [28, 27].

In addition to evaluating the methods using binary masks with the full STFT frequency resolution, binary masks at four reduced resolutions were also evaluated. The reduced resolution masks were computed from the SNR in bands formed by merging non-overlapping ranges of contiguous STFT bins. Within the enhancement algorithm, a single mask value was used for all of the STFT bins within the range that was used to compute it. The R reduced resolution bands had centre frequencies uniformly spaced on the Equivalent Rectangular Bandwidth (ERB) scale [29] between 100 Hz and 5 kHz. The values $R = \{20,\ 40,\ 60,\ 80\}$ were included in the tests. The algorithm parameters listed in Table 1 were optimized on the training set separately for each mask resolution.

Algorithm	Parameter	Optimal value
C: OM-LSA max-WSTOI	$G_{\min}$	-10 dB
D: OM-LSA max-PESQ	$G_{\min}$	-16 dB
F: IBM, gain floor	$\mathcal{M}_{\min}$	-34 dB
G: IBM, cepst. smoothing	$\mathcal{M}_{\min}$	-34 dB
	β_{pitch}	0
	β_{peak}	0.14
H: Proposed method	G^1	-3 dB
	G^0	-43 dB
	Q^1	0.7
	Q^0	1

Table 1: Summary of trained parameters and their optimal values. For the IBM-based methods, the optimal values are displayed for the case where the IBM has the full STFT frequency resolution.

4. RESULTS

Figures 1a and 1b show the WSTOI and PESQ scores, respectively, for the noisy speech utterances after processing with different enhancement methods. The LSA (B) and OM-LSA (C, D) algorithms resulted in an improvement in PESQ of about 0.2 compared with the unprocessed noisy speech. However, the effect of the algorithms on WSTOI varied significantly between utterances and both of the algorithms severely damaged the WSTOI scores of some utterances as can be seen from the long box plot whiskers in Fig. 1a. The standard IBM (E) resulted in a median PESQ score of 2.5 and almost full in-

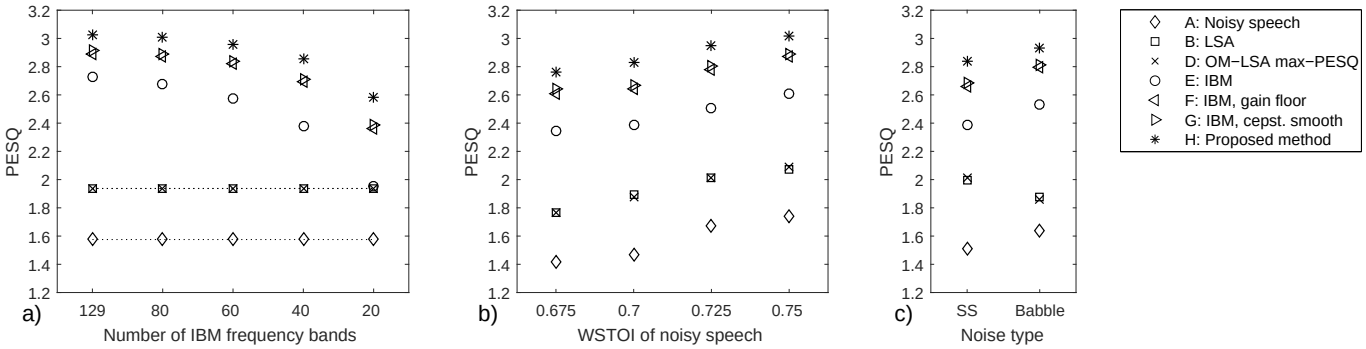


Fig. 2: PESQ scores of the noisy speech and the noisy speech after processing with different methods. PESQ against a) the number of frequency bands in the IBM, b) the WSTOI score of the unprocessed noisy speech, and c) the noise type.

telligibility as indicated by the scale on the rightmost edge of Fig. 1a. The three other methods (F, G, H) that are based on the IBM resulted in very similar improvements in predicted intelligibility to the IBM, but gave higher PESQ scores. The proposed method (H) resulted in the largest improvement in PESQ. The next best improvement came from the IBM with cepstral smoothing, followed closely by the IBM with a gain floor. We emphasize that methods E, F, G and H all make use of an IBM computed using oracle knowledge of the SNR in each time-frequency bin.

Fig. 2a shows the PESQ scores of the different methods plotted against the frequency resolution of the IBM. The results of method C were omitted for clarity as they were very similar to those of method D. As the resolution decreases, the PESQ score of all the IBM-based methods decreases. The PESQ decrease becomes sharper when fewer than 60 bins are used but the relative improvement of the proposed method is preserved or increased at low resolutions. The corresponding plot for WSTOI has been omitted because the WSTOI scores are almost independent of the frequency resolution. Fig. 2b shows the PESQ scores of the different methods against the WSTOI scores of the unprocessed noisy speech. This shows that the improvement in PESQ of the proposed method was largely independent of the predicted intelligibility of the unprocessed noisy speech. Fig. 2c shows the PESQ scores of the different methods for the two noise types separately. With both noise types, the proposed method resulted in the highest predicted quality.

Fig. 3 shows a histogram of the differences in PESQ resulting from processing the noisy speech utterances with the proposed method (H) and the second best performing method, IBM with cepstral smoothing. In 375 out of 400 sample pairs (93.75%) the proposed method resulted in a higher predicted quality than the IBM with cepstral smoothing, meaning the improvement was statistically significant with $p \ll 10^{-3}$ using a 1-sided sign test.

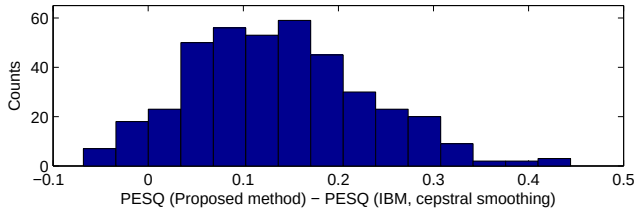


Fig. 3: Distribution of the difference in PESQ resulting from processing the noisy speech utterances with the proposed method and the IBM with cepstral smoothing.

5. CONCLUSION

We have presented a new approach to applying a binary mask that preserves the intelligibility gains given by conventional binary masking but also incorporates a speech enhancer's ability to improve perceptual quality. The binary mask is not applied directly as a TF gain but is instead used to supply prior information to a classical speech enhancer about the probability of speech presence in different TF regions. The proposed method resulted in a statistically significant improvement in PESQ compared with other methods of applying an oracle binary mask whilst preserving the improvements in predicted intelligibility.

6. ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M026698/1] and the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)", funded by the Seventh Framework Programme of the European Commission under Grant Agreement no. 316969.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[3] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[4] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, pp. 1777–1786, 2007.

[5] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: dependency on signal-to-noise ratios," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 531–539, 2012.

[6] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," vol. 23, no. 12, pp. 1–12, Dec. 2015.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 708–712.

[9] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," *Proc. IEEE Global Conf. Signal and Information Processing (GlobalSIP)*, pp. 577–581, Dec. 2014.

[10] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear & Hearing*, vol. 27, pp. 480–492, 2006.

[11] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[12] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[13] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, May 2014.

[14] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Causal binary mask estimation for speech enhancement using sparsity constraints," in *Proc. Intl. Congress on Acoustics*, Montreal, June 2013.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[16] 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.*, vol. 126, no. 3, pp. 1415–1426, Sept. 2009.

[17] L. Lightburn and M. Brookes, "SOBM - a binary mask for noisy speech that optimises an objective intelligibility metric," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5078–5082.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[19] T. Stokes, C. Hummersone, and T. Brookes, "Reducing binary masking artifacts in blind audio source separation," in *Audio Engineering Society Convention 134*, May 2013, number 8853.

[20] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 45–48.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *J. Acoust. Soc. Am.*, vol. 136, pp. 892–902, 2014.

[22] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197. Kluwer Academic, 2005.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Corpus LDC93S1, Linguistic Data Consortium, Philadelphia, 1993.

[24] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise data-base," Tech. Rep. IZF 1988–3, TNO Institute for perception, 1988.

[25] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[26] ITU-T, "Mapping function for transforming P.862 raw result scores to MOS-LQO," Recommendation P.862.1, Nov. 2003.

[27] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997–2015.

[28] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[29] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.