Blind Adaptive Mask to Improve Intelligibility of Non-Stationary Noisy Speech

F. Farias, *Student Member, IEEE*, and R. Coelho, *Senior Member, IEEE*

***Abstract*—This letter proposes a novel blind acoustic mask (BAM) designed to adaptively detect noise components and preserve target speech segments in time domain. A robust standard deviation estimator is applied to the non-stationary noisy speech to identify noise masking elements. The main contribution of the proposed solution is the use of this noise statistics to derive an adaptive information to define and select samples with lower noise proportion. Thus, preserving speech intelligibility. Additionally, no information of the target speech and noise signals statistics is previously required to this non-ideal mask. The BAM and three competitive methods, Ideal Binary Mask (IBM), Target Binary Mask (TBM), and Non-stationary Noise Estimation for Speech Enhancement (NNESE), are evaluated considering speech signals corrupted by three non-stationary acoustic noises and six values of signal-to-noise ratio (SNR). Results demonstrate that the BAM technique achieves intelligibility gains comparable to ideal masks while maintaining good speech quality.**

***Index Terms*—Acoustic mask, adaptive methods, speech intelligibility, nonstationarity.**

I. Introduction

MOST everyday listening experiences are in the presence of acoustic noise such as car noise, people talking in the background, construction noise, rain and other natural phenomenons. These effects may add unwanted content to a target speech signal while diminish its intelligibility [1] and its quality [2]. Applications such as speaker recognition, speech to text and source localization exhibit lower accuracy when the signal is corrupted by additive noise. Thus, the mitigation of this acoustic interference in noisy speech is an important research topic. The solutions proposed in the literature are mainly twofold: speech enhancement methods to increase quality, and binary acoustic masks to improve intelligibility.

Speech enhancement schemes mitigate the masking interference to improve the noisy signal quality, usually estimating the noise statistics. These noise estimation approaches generally consider the frequency or the time domain. Methods such as the Spectral Subtraction (SS) [3] and the Optimally Modified Log-Spectral Amplitude (OMLSA) [4] use some transform to represent signals in the frequency domain and then estimate the noise components. Methods in the time domain usually estimate noise statistics with statistical estimators as the present in Non-stationary Noise Estimation for Speech Enhancement (NNESE) [5] or time-frequency decomposition, e.g., in the Empirical Mode Decomposition (EMD)-Based filtering with Hurst exponent (EMDH) [6], [7]. Although speech enhancement algorithms successfully improve speech quality, they are not designed to achieve intelligibility gain. This is particularly challenging in non-stationary environments. In some cases, the suppression of noisy components causes a distortion that hinders intelligibility [8].

Acoustic masks [9]–[11] are defined to emulate the capacity of the human auditory system to segregate a specific sound of interest even in the presence of many others. This is also known as *cocktail party* effect [12]. Consequently, improving the intelligibility of the target speech signal present in a variety of applications. Acoustic binary masks can be classified as ideal or blind. Ideal masks are constructed using information of the clean speech, as well as the noisy speech. The Ideal Binary Mask (IBM) [13] is built comparing the energy of the signal and of the noise in each Time-Frequency (T-F) region. The Target Binary Mask (TBM) [14] builds its mask comparing the energy of the clean signal with the Speech Shaped Noise (SSN). Blind masks are made based on an estimation of clean speech characteristics from the noisy signal [15], [16]. However, the accuracy of this estimation usually depends on extensive training using neural networks and large databases. Binary masked speech may present high objective quality scores, but the abrupt difference between the retained and discarded regions of a binary masked signal often cause musical noise, which can lead to quality loss [17].

This letter proposes a blind acoustic mask in the time domain to improve speech intelligibility. The main idea of this strategy is to estimate noise components and the proportion of the speech signal in each short-time frame. This information is used to delimit a set of samples proportional to the presence of the target signal in each frame, so if the frame is mostly comprised by the target signal, these samples take most of the frame. While the frame is processed to mitigate the effects of noise, the samples in the delimited set are preserved, thus maintaining speech intelligibility. Additionally, as a blind mask it avoids the usage of prior information from the clean speech and noise.

Several experiments are conducted to evaluate the proposed mask in terms of speech intelligibility and quality. The noisy scenario is composed by three background acoustic noises with six different SNR values. Three objective measures are adopted for intelligibility evaluation. The Short Time Objective Intelligibility (STOI) [18] is the state-of-the-art in intelligibility

Manuscript received May 21, 2021; accepted May 27, 2021. Date of publication June 3, 2021; date of current version June 21, 2021. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grant 308155/2019, in part by the Fundação de Amparo ã Pesquisa do Estado do Rio de Janeiro (FAPERJ) under Grant 203075/2016, and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under Grant Code 001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Z. Jane Wang. *(Corresponding author: R. Coelho.)*

The authors are with the Laboratory of Acoustic Signal Processing, Military Institute of Engineering, Rio de Janeiro, RJ 22290-270, Brazil (e-mail: felipe.farias@ime.eb.br; coelho@ime.eb.br).

Digital Object Identifier 10.1109/LSP.2021.3086405

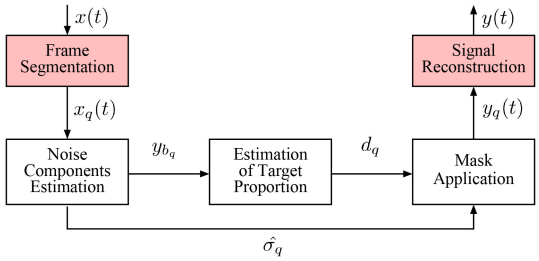

Fig. 1. Schematic of the proposed Blind Acoustic Mask.

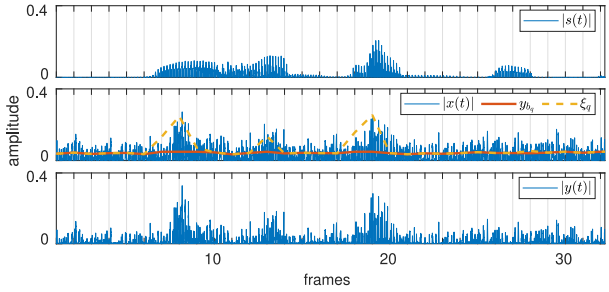

Fig. 2. Clean speech signal $|s(t)|$, corresponding speech signal corrupted with Babble noise at SNR $= -5$ dB $|x(t)|$, lower threshold y_{b_q}, and upper threshold ξ_q.

prediction. The Approximated Short-Time Speech Intelligibility Index (ASII) [19] and Extended Speech Intelligibility Index (ESII) [20] are designed to deal with non-stationary distortions. Objective quality evaluation is performed using the Perceptual Evaluation of Speech Quality (PESQ) [21] and the overall quality composite measure (OQCM) [22]. The Index of Non-Stationarity (INS) [23] is also selected to analyse the effect of the proposed and baseline methods.

II. BAM: Blind Acoustic Mask

The schematic of the proposed blind acoustic mask is depicted in Fig. 1. It consists of three main steps: First, the signal is separated into non-overlapping short-time frames with T samples each. In each frame, the noise standard deviation is detected using a robust estimator. In the second step, this information is used to derive the parameter d_q that refers to the proportion of speech that is present in the q-th frame. Then the adaptive mask is applied in each frame, according to the parameter d_q, and the processed frames are concatenated to reconstruct the processed signal. This mask employs the noise statistics estimation used in [5]. The aim is to improve intelligibility of speech corrupted by additive noise, while maintaining the quality gain obtained using a speech enhancement technique.

A. Step 1: Noise Components Estimation

This step begins with the segmentation of the signal in non-overlapping short-time frames. In each frame, the d-Dimensional Trimmed Estimator (DATE) [24] is used to detect the noise standard deviation. This estimator was first defined to work on signals mixed with Gaussian noise over the entire signal. However, as shown in [5] it also works with different noise distributions.

First, the samples $x_q(t)$ from the q-th frame are sorted from lower to higher absolute values $\|Y_1\| \leq \|Y_2\| \leq \cdots \leq \|Y_T\|$, where $\|Y_t\| \in [0, 1], t = 1, \ldots, T$. Then a value t_{min} is computed according to [24], such that samples whose norms are lower than $\|Y_{t_{min}}\|$ are considered as only noise. Let b_q be the lowest value of t that is higher than t_{min} and obeys the relation $\|Y_{t-1}\| \leq \frac{c\sum_{i=1}^{T}\|Y_i\|}{T} \leq \|Y_{t+1}\|$. If such value of t does not exist, than $b_q = t_{min}$. The detection threshold c is defined in [24], and it is here computed as $c = 4.3542$ following the procedure in [5]. Then, the noise components standard deviation for that frame is estimated as

$$\hat{\sigma_q} = \frac{c\sum_{i=1}^{b_q}\|Y_i\|}{b_q}. \tag{1}$$

In this step it is also defined the value y_{b_q}, amplitude from the vector Y_i associated to the value b_q. Any sample with amplitude lower than y_{b_q} is considered as noise.

B. Step 2: Estimation of Target Proportion d_q

The removal of samples with amplitude values below y_{b_q} may yield an improvement in quality. However, some of those frames are mainly composed by the target signal. The modification of samples from those frames usually compromises intelligibility. Thus, a parameter d_q is defined to identify frames where the target signal is prevalent:

$$d_q = \frac{|\sigma_{q_{ny}} - \hat{\sigma_q}|}{|\sigma_{q_{ny}} + \hat{\sigma_q}|} \tag{2}$$

where $\hat{\sigma_q}$ is the estimated noise standard deviation of the frame q and $\sigma_{q_{ny}}$ is the noisy signal standard deviation.

C. Step 3: Mask Application

The proposed mask defines which samples are left unaltered in each frame. These samples correspond to a region where d_q is greater than the lower amplitude y_{b_q}. The lower bound of this region is defined by y_{b_q} and the upper bound is defined through an adaptive threshold ξ_q.

$$\xi_q = max(y_{b_q}, d_q). \tag{3}$$

Each sample of the q-th frame of the processed signal is then given by

$$|y_q(t)| = \begin{cases} |x_q(t)|, & \text{if } y_{b_q} < |x_q(t)| < \xi_q; \\ |x_q(t)| - \alpha\,\hat{\sigma_q}, & \text{if } |x_q(t)| \geq \xi_q; \\ \beta\,|x_q(t)|, & \text{otherwise.} \end{cases} \tag{4}$$

where α is the over-subtraction factor for the speech signal reconstruction and β is the flooring factor to avoid negative amplitude values. The sign information of the original sample $x_q(t)$ is applied to obtain the final sample $y_q(t)$. Finally, all frames are concatenated to form the processed signal $y(t)$.

The region in each frame is illustrated in Fig. 2, which depicts the clean signal $s(t)$, the noisy signal $x(t)$, the lower bound y_{b_q} and upper threshold ξ_q of the masked region for each frame q. Note that frames 18-21 are mainly composed by speech. Thus, the mask preserves most of the signal in these frames.

TABLE I
INTELLIGIBILITY MEASURES FOR UNP SPEECH SIGNALS

	STOI					
SNR (dB)	**-6**	**-5**	**-3**	**0**	**3**	**5**
Babble	0.355	0.356	0.388	0.447	0.499	0.529
Cafeteria	0.357	0.381	0.419	0.458	0.503	0.540
Factory	0.436	0.476	0.506	0.557	0.601	0.654
	ASII_{ST}					
SNR (dB)	**-6**	**-5**	**-3**	**0**	**3**	**5**
Babble	0.348	0.365	0.391	0.433	0.497	0.519
Cafeteria	0.364	0.377	0.410	0.446	0.504	0.523
Factory	0.398	0.415	0.446	0.501	0.555	0.586
	ESII					
SNR (dB)	**-6**	**-5**	**-3**	**0**	**3**	**5**
Babble	0.306	0.329	0.361	0.416	0.497	0.525
Cafeteria	0.327	0.344	0.386	0.432	0.505	0.528
Factory	0.371	0.394	0.433	0.503	0.570	0.608

TABLE II
NORMALIZED MEAN PROCESSING TIME

NNESE	IBM	TBM	BAM
0.5	9.4	9.2	1.0

III. EXPERIMENTS AND DISCUSSION

The proposed technique is evaluated in terms of intelligibility and quality considering several noisy conditions. It is compared to baseline speech enhancement NNESE [5], and acoustic masks TBM [14] and IBM [12]. The speech enhancement is considered the baseline for quality gain, as the IBM for intelligibility improvement. For the objective evaluation, a subset of 20 utterances from the TIMIT database [25] is randomly selected to compose each scenario, leading to 120 tests per method. Each segment is sampled at 16 kHz and has average time duration of 3 seconds. The Babble and Factory noises are selected from the RSG-10 [26] database while the Cafeteria noise was collected from DEMAND [27]. Speech signals are corrupted considering six SNRs: -6 dB, -5 dB, -3 dB, 0 dB, 3 dB and 5 dB.[1]

NNESE and the proposed mask are applied on a 32 ms frame-by-frame basis considering the parameters α and β set to $\alpha = 0.35$ and $\beta = 0.65$. IBM and TBM separate signals in 20 ms Time-Frequency regions, with 10 ms overlapping. Frequency separation is performed through a 64-channel gammatone filterbank with center frequencies ranging from 50 to 8000 Hz according to the Equivalent Rectangular Bandwidth [28]. The IBM Relative Criterion is set to -5 dB, as recommended in [29]. The TBM is set to detect 99% of the speech energy in each frame.

The intelligibility scores of the unprocessed signals are presented in Table I. The SNR values were chosen such as the STOI score of the UNP signals vary between 0.45 and 0.75, which are considered the threshold of poor and good intelligibility, respectively [30].

Table II indicates the computational complexity, here represented by the processing time required for each method evaluated for 512 samples per frame. These values are obtained with an Intel(R) Core(TM) i5-8400 CPU with six threads and 8 GB RAM and are normalized by the execution time of the proposed BAM. It can be noted that the proposed mask needs no prior information of the corrupting noise and requires about only 10% of the computational time of binary masks.

[1]Some example noisy and processed files are available at http://lasp.ime.eb.br/index.php?vPage=downloads.

TABLE III
AVERAGE $INS_{\max}$ OF UNP AND MASKED SIGNALS

noise	UNP	IBM	TBM	BAM
Babble	106	2160	2065	142
Cafeteria	79	2398	2275	120
Factory	48	1933	1837	90
Overall	78	2164	2059	117

A. Index of Non-Stationarity

The Index of Non-Stationarity (INS) [23] is here adopted to objectively evaluate the non-stationarity of noisy speech signals. This measure is obtained comparing the target signal with a set of stationary references called *surrogates* at different time scales T_h/T, where T_h is a short-time analysis window and T is the total duration of the signal. A threshold γ is defined for each window length T_h, considering a confidence degree of 95%. Therefore, the signal is considered non-stationary whenever INS $> \gamma$.

Fig. 3 depicts spectrograms and INS values for a speech signal, the corresponding signal with Factory noise and SNR of 3 dB, the noisy signal processed by the IBM, TBM, and the proposed BAM. Note that the noise modifies the temporal and spectral characteristics of the speech signal, reducing its non-stationary behavior. In this example, the maximum INS ($INS_{\max}$) changes from 320 in speech signal to 55 in noisy speech. The proposed BAM recovers some of the spectral and temporal characteristics of the signal. This can be seen near 0.5 s, where the separation between formants is lost due to noise and retrieved by the BAM. Additionally, the proposed mask yields an $INS_{\max}$ value of 0.95, which means that it restores some of the non-stationary behavior of the clean signal. Unlike the proposed BAM, the IBM and TBM masks lead to $INS_{\max}$ greater than 5000 and 1200, respectively.

Table III presents the average $INS_{\max}$ results computed from the speech signals corrupted by the three acoustic noises. The average $INS_{\max}$ obtained from the clean speech signals is 570. Again, the overall $INS_{\max}$ obatined with BAM is closer to that obtained by clean speech. This indicates that the proposed BAM was able to restore some of the speech signal characteristics. Similar conclusions cannot be drawn from the other masks, that lead to $INS_{\max}$ values exceptionally different from the original signal. These unusual values are explained by the binary effect particularly for zeroing masking condition.

B. Objective Intelligibility Evaluation

Three objective measures are adopted to evaluate intelligibility gain of the proposed and competing masks. STOI [18] was developed to predict the intelligibility of noisy speech processed by T-F weighting masks, such as speech enhancement methods. ASII_{ST} [19] and ESII [20] are based on the classic Speech Intelligibility Index (SII) [31], designed to deal with the non-stationarity of speech and its distortions. Both measures are based on the weighted SNR of the signal. The three measures vary between 0 and 1, in which 1 represents a fully intelligible sentence. The STOI objective measure is normalized by the intelligibility achieved for the target signal corrupted by SSN noise at 10 dB, considered here as a good intelligibility reference.

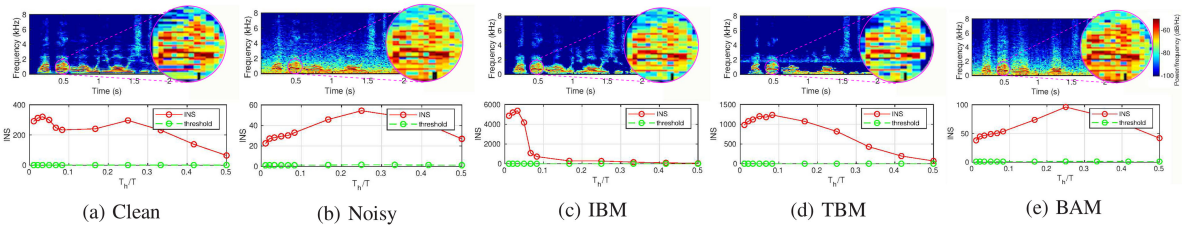

Fig. 3. Spectrograms and INS of (a) clean speech, (b) unprocessed speech corrupted with Factory noise at SNR = 3 dB, and noisy speech processed with (c) IBM, (d) TBM, and (e) the proposed BAM.

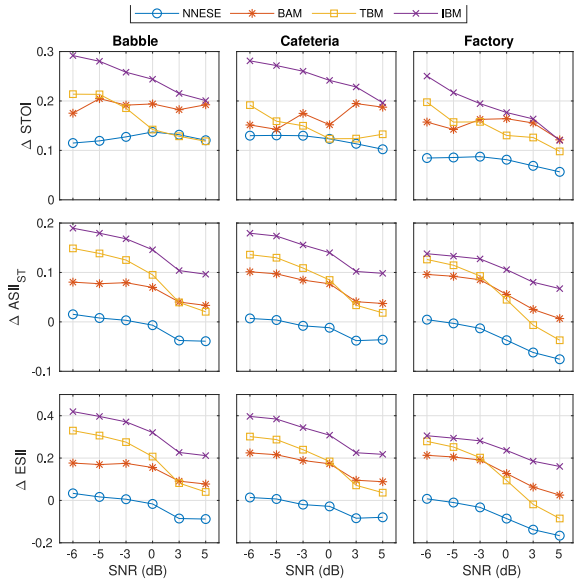

Fig. 4. Average improvement in intelligibility for noisy speech with Babble, Cafeteria and Factory.

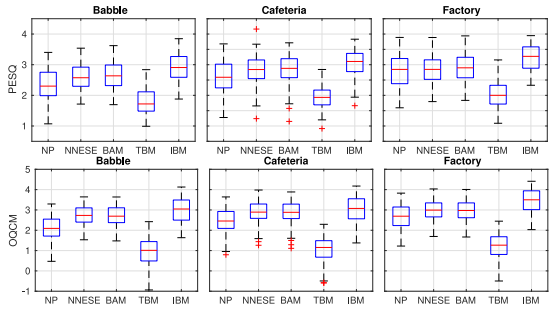

Fig. 5. Box-plots of PESQ and OQCM scores for noisy speech with Babble, Cafeteria and Factory.

The average improvement in terms of STOI is presented in Fig. 4. It can be noted that the IBM presents the highest ΔSTOI results: 0.25 for Babble noise, followed by 0.19 for the proposed BAM and 0.17 for the proposed TBM. ΔSTOI results for the Cafeteria noise are similar to those obtained for Babble. For the Factory noise, however, the intelligibility improvement is lower: 0.19 for IBM, 0.15 for BAM, and 0.14 for TBM. This is due to the higher STOI scores of speech signals corrupted by the Factory noise (refer to Table I), which makes the intelligibility improvement to be more challenging for this specific noise source. Additionally, the proposed BAM outperforms the NNESE for most of the noise conditions.

The improvement obtained for the ASII $_{ST}$ measure is also shown in Fig. 4. Note that the proposed mask improves intelligibility in almost every condition, but this effect is more accentuated at low SNR values. The proposed BAM leads to average Δ ASII$_{ST}$ of 0.08 for SNR = -3 dB, and 0.02 for SNR = 3 dB. It can also be observed that the proposed mask outperforms the TBM for SNR $>$ 0 dB.

Similarly to ΔASII $_{ST}$ results, the proposed BAM shows interesting ESII improvement for SNR values lower than 3 dB. The maximum Δ ESII is 0.08 for the Babble noise at SNR = -6 dB, 0.10 for Cafeteria noise at SNR = -6 dB, and 0.09 in Factory noise at the same SNR. Furthermore, the average improvement is comparable to that achieved by the TBM.

C. Objective Quality Evaluation

Speech quality is evaluated using the objective PESQ [21] and OQCM [22] measures. PESQ was developed to assess quality of narrow-banded speech and handset telephony. A signal is considered of fair quality when its PESQ score is above 2, given this objective metric aims to predict Mean Opinion Scores (MOS) [32]. The OQCM combines the PESQ, the weighted spectral slope (WSS), and the log-likelihood ratio (LLR): $\text{OQCM} = 1.594 + 0.805\,\text{PESQ} - 0.512\,\text{LLR} - 0.007\,\text{WSS}$. The idea is to maximize the correlation with subjective scores in terms of overall speech quality.

The PESQ and OQCM scores are presented in Fig. 5. The best PESQ results are obtained using the IBM for all noise sources, while the proposed BAM presents the highest scores among the remaining techniques. In terms of OQCM, the BAM and NNESE methods achieve quite similar values: 2.7 for Babble, 2.9 for Cafeteria, and 3.0 for Factory. In average, these values are 0.5 greater than those achieved with the noisy signals. These results reinforce that the proposed BAM substantially improves speech intelligibility, while its quality gain is similar to the NNESE speech enhancement solution.

IV. Conclusion

This letter introduced a time-domain blind acoustic mask to improve intelligibility of speech signals corrupted by non-stationary noises. The proposed BAM needs no prior information of the corrupting noise and requires about only 10% of the computational time of binary masks. The proposed mask achieved an average gain of 0.17 in terms of the STOI measure. Additionally, it outperformed the blind TBM mask, especially for the highest SNR values. Finally, BAM obtained interesting speech quality scores while achieved significant speech intelligibility improvement when compared to the NNESE.

REFERENCES

[1] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2336–2347, 2009.

[2] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[4] 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.

[5] R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 6–10, Jan. 2016.

[6] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with emd and hurst-based mode selection," in *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 899–911, May 2014.

[7] R. Coelho and L. Zão, "Empirical mode decomposition theory applied to speech enhancement," *Proc. Signals Images: Adv. Results Speech, Estimation, Compress., Recognit., Filter. Process.*, 2015.

[8] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[9] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, no. 3, pp. 230–239, 2009.

[10] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1010–1013, Dec. 2010.

[11] L. Zão, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 620–624, 2014.

[12] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Boston, MA: Springer, 2005, pp. 181–197.

[13] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[14] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear Hear.*, vol. 27, no. 5, pp. 480–492, 2006.

[15] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[16] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 133, no. 3, pp. 1607–1614, 2013.

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

[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, Sep. 2011.

[19] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[20] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[21] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. P.862 Int. Telecommu. Uni.*, 2001.

[22] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[23] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3459–3470, Jul. 2010.

[24] D. Pastor and F.-X. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1545–1555, Apr. 2012.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Tech. Rep. n, vol. 93, 1993.

[26] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 Noise Data-base," *Report IZF 1988-3, TNO Institute for Perception*, Soesterberg, The Netherlands, 1988.

[27] J. Thiemann, N. Ito, and E. Vincent, "Demand: A collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust.*, 2013, pp. 1–6.

[28] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Proc. Meeting IOC Speech Group Audit. Modelling RSRE*, vol. 2, no. 7, 1987.

[29] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[30] B. Sauert and P. Vary, "Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1. 2006, pp. I-I.

[31] *American National Standards Institute*, American national standard: Methods for calculation of the speech intelligibility index, New York, NY, USA, 1997.

[32] E. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 225–246, Jun. 1969.