MODULATION WIENER FILTER FOR IMPROVING SPEECH INTELLIGIBILITY

Chung-Chien Hsu, Kah-Meng Cheong, Jen-Tzung Chien, and Tai-Shih Chi

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30013, R.O.C.

ABSTRACT

This paper presents a single-channel high-dimensional Wiener filter in the spectro-temporal modulation domain. Unlike other conventional noise reduction techniques, the proposed algorithm not only reduces noise but also enhances the "textures" of the speech signal. A non-iterative decision-directed noise estimation method is adopted to estimate the modulation SNR for the modulation-domain Wiener filter. The efficacy of the proposed algorithm in enhancing speech intelligibility is assessed using the short-time objective intelligibility (STOI) measure. Statistical analysis results demonstrate that our proposed algorithm can improve STOI scores in speech-shape noise (SSN) and white noise conditions, but not in babble noise condition, while the conventional Wiener filter fails to improve STOI scores in all three noise conditions.

Index Terms— spectro-temporal modulation, Wiener filter, speech enhancement, speech intelligibility

1. INTRODUCTION

Speech is the most important biosignal for human communication. Nowadays, many speech-application related devices have been developed to facilitate our daily lives. The applications can be grouped into two categories: for machine to listen and for human to listen. One of the important applications for human listening is to help hearing-impaired (HI) patients. Noise reduction is a critical element in hearing-aid devices, which are supposedly developed to improve perceived speech quality and intelligibility for HI patients. However, subjective listening tests showed that conventional single-channel noise reduction algorithms do not improve speech intelligibility either of English [1, 2] or of Chinese and Japanese [3]. Therefore, developing a better single-channel noise reduction algorithm for hearing aids is still a challenge for researchers.

Speech contains rich information in both spectral and temporal domains. The fluctuations of speech signal across time and frequency axes are referred to as modulations. Psychoacoustic experiments show the slow temporal energy modulations ($\leq$ 16 Hz) of speech are highly related to speech intelligibility [4]. Indeed, the temporal modulations reflect changes of the vocal tract through time and encode lots of linguistic information. Considering temporal modulations has inspired many engineering approaches such as the bi-frequency (acoustic and modulation frequencies) representation for audio coding [5], the temporal-modulation incorporated front-end feature extraction for automatic speech recognition (ASR) [6], and the temporal-modulation domain estimator for speech enhancement [7].

In addition to psychoacoustic experiment results, neurophysiological evidences also suggest that neurons of the auditory cortex (A1) respond to both spectral and temporal modulations of the input sounds. A computational auditory model was proposed accordingly [8]. Later on, psychoacoustic experiments were also conducted to determine which spectro-temporal modulations are critical for speech comprehension [9]. Not surprisingly, the concept of using spectro-temporal modulation analysis has shown in many applications, such as speech intelligibility assessment [10] and robust feature extraction for ASR [11]. As for frequency modulation, a psychoacoustic study demonstrated that frequency modulations significantly enhance human speech recognition in noisy environment [12]. This study supports our idea that the frequency modulation energy associated with harmonics of speech can be used as a robust feature, especially effective against non-stationary noise, for voice activity detection (VAD) [13].

In our previous work, we proposed a spectro-temporal modulation subband Wiener filter for Fourier spectrograms and demonstrated its capability in improving speech quality [14]. However, for HI patients, enhancing speech intelligibility is just as important. Most of the conventional noise reduction methods tend to improve the signal-to-noise ratio (SNR) of the speech signal. It has been shown that SNR improvement is not totally correlated to intelligibility improvement [15]. Since spectro-temporal modulations have been shown critical for speech comprehension [12] and speech intelligibility can be measured by assessing spectro-temporal modulation contents [10], it is reasonable to assume spectro-temporal modulations are highly related to speech intelligibility. Therefore, we postulate that enhancing modulation SNR would improve speech intelligibility. In this paper, we propose a direct *priori* modulation SNR estimator for our previously developed modulation subband Wiener filter and show its capability of improving speech intelligibility.

The rest of the paper is organized as follows. Section

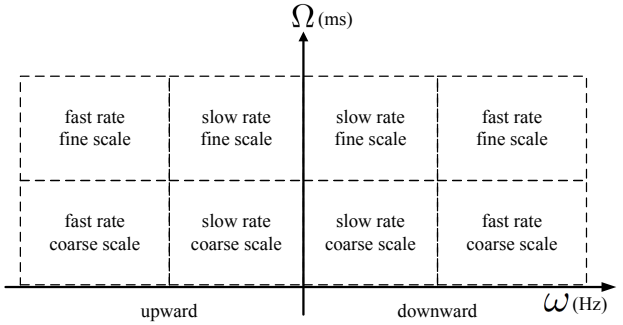

Fig. 1. ω-Ω (rate-scale) space.

2 gives a review of our spectro-temporal analysis and synthesis framework and demonstrates modulation contents of the speech signals. Then, a modulation Wiener filter is proposed by incorporating a direct modulation SNR estimation method. The performance on improving speech intelligibility is demonstrated in section 4. We end in section 5 with conclusion and future work.

2. MULTIRESOLUTION SPECTRO-TEMPORAL ANALYSIS AND SYNTHESIS FRAMEWORK

2.1. Spectrotemporal analysis and synthesis

Since speech can be assumed quasi-stationary, it is analysed using short-term Fourier transform (STFT).

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)e^{-j2\pi kl/N} \quad (1)$$

where $w(n)$ is an analysis window function and k refers to the frequency index.

For any input magnitude spectrogram $\mid X(n,k) \mid$, the 4-dimensional multi-resolution representation can be obtained as follows:

$$C_{\pm}(n,k,\omega,\Omega) = \mid X(n,k) \mid *_{nk} STIR_{\pm}(n,k;\omega,\Omega) \quad (2)$$

where $STIR_{\pm}(n,k;\omega,\Omega)$ is the spectro-temporal impulse response of the 2D modulation filter tuned to ω and Ω; $*_{nk}$ denotes two-dimensional convolution along the time and frequency axes. The $rate$ parameter ω (in Hz, as frequency) reflects how fast the local envelope of the magnitude spectrogram varies along the time axis. The $scale$ parameter Ω (in ms, as quefrency) reflects how broad the local envelope of the magnitude spectrogram distributed along the frequency axis. They are defined as the Fourier domains of the time and the frequency dimensions, respectively.

To reduce the computational costs of the 2D convolution, eq. (2) is reformulated as:

$$C_{\pm}(n,k,\omega,\Omega) = \mathcal{F}_{2D}^{-1}\{\mathcal{F}_{2D}\{\mid X(n,k) \mid\} \cdot STMF_{\pm}(\omega,\Omega)\} \quad (3)$$

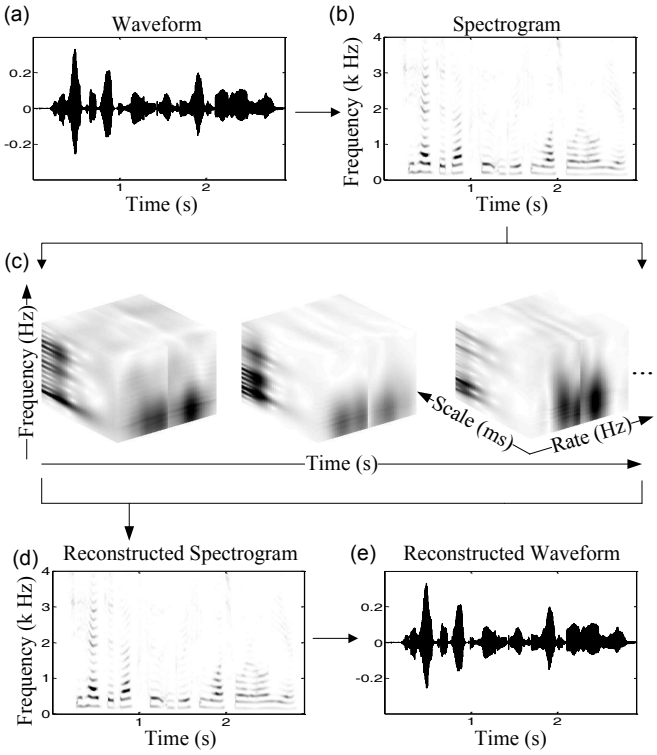

Fig. 2. Spectro-temporal analysis and synthesis process. (a) Time waveform; (b) its spectrogram; (c) the 4-D (scale-rate-frequency-time) output; (d) reconstructed spectrogram; (e) reconstructed waveform.

where $\mathcal{F}_{2D}$ and $\mathcal{F}_{2D}^{-1}$ denote the 2-D Fourier transform and the inverse 2-D Fourier transform; and $STMF_{\pm}(\omega,\Omega)$ denotes the spectro-temporal frequency responses of the 2D modulation filters. Detailed in designing these modulation filters can be accessed in [14]. In addition, the sign $(\pm)$ represents the sweeping direction of the filters (positive rate refers to the downward direction and negative rate refers to the upward direction). As shown in Fig. 1, the frequency response of a complex downward/upward filter is located in the first/second quadrant of the ω-Ω space.

Eq. (3) shows that the spectro-temporal analysis is a pure linear operation such that the reconstructed magnitude spectrum $\mid X'(n,k) \mid$ can be obtained from the four-dimensional representation $C_{\pm}(n,k,\omega,\Omega)$ by

$$\mid X'(n,k) \mid = \mathcal{R}\{\mathcal{F}_{2D}^{-1}\{\frac{\sum_{\omega,\Omega}\mathcal{F}_{2D}\{C_{\pm}(n,k;\omega,\Omega)\}}{\sum_{\omega,\Omega} STMF_{\pm}(\omega,\Omega)}\}\} \quad (4)$$

where $\mathcal{R}\{\cdot\}$ denotes the real-part operator. Finally, the speech sound is synthesized using the overlap-and-add (OLA) method [16]. Note, the original modulation phase and acoustic phase are applied in inverse operations. Fig. 2 shows the spectro-temporal analysis and synthesis process for a sample utterance and results at different stages.

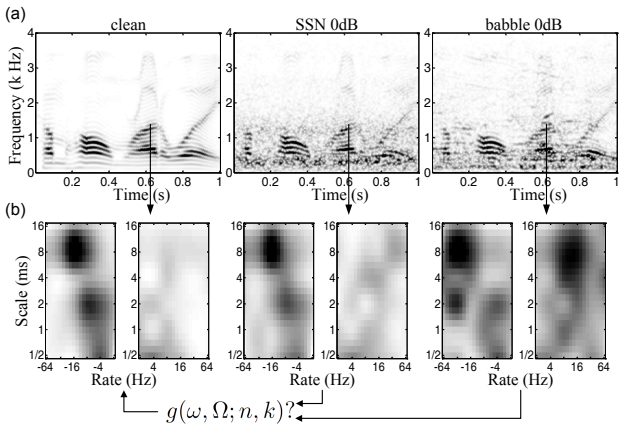

Fig. 3. (a) Clean and noisy spectrograms; (b) spectro-temporal modulation contents of the same T-F unit.

2.2. Rate-Scale representation

The multi-resolution analysis process can capture the prominent spectro-temporal "texture" of speech, such as pitch, harmonicity, formant, amplitude modulation (AM), frequency modulation (FM) and onset/offset. Fig. 3(a) shows the Fourier spectrograms of a clean and two noisy utterances, which are corrupted by speech-shape noise (SSN) and babble noise at 0 dB SNR, respectively. Fig. 3(b) demonstrates the corresponding modulation contents $\mid C(\omega, \Omega; n, k) \mid$ of a particular time-frequency (T-F) unit. The peak in the R-S plot indicates that the harmonics is moving upward with 8$\sim$16 Hz temporal modulation and 8$\sim$10 ms spectral modulation (100$\sim$125 Hz harmonic spacing) around the particular T-F unit. In the noisy conditions, the R-S plots are damaged, especially in the babble noise. The textures of speech (AM, FM, etc.) are destroyed such that it becomes less intelligible. Therefore, we want to find a Wiener-gain function which can suppress noise and enhance the underlying textures of speech.

3. PROPOSED SPEECH ENHANCEMENT ALGORITHM

For common speech enhancement algorithm, the observed noisy speech signal $y(n)$ is formulated as:

$$y(n) = x(n) + n(n) \tag{5}$$

where $x(n)$ is the clean speech signal and $n(n)$ is the noise. It is easy to derive the Wiener filter in the frequency domain as:

$$H(k) = \frac{\xi(k)}{\xi(k) + 1} \tag{6}$$

where $\xi(k)$ denotes the *a priori* SNR at frequency k. The filter preserves the spectrum at high SNR ($\xi(k) \rightarrow \infty, H(k) \rightarrow 1$) and attenuates the spectrum at low SNR ($\xi(k) \rightarrow 0, H(k) \rightarrow 0$). A simple and efficient approach was proposed to directly estimate the *a priori* SNR $\xi(k)$ using a recursive update rule to combine the past and present estimates [17].

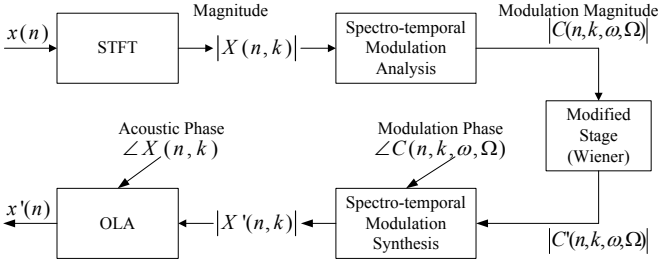

Fig. 4. Block diagram of the proposed modulation Wiener filter to enhance speech intelligibility.

Now, we extend the method into the modulation domain to estimate the *a priori* modulation SNR as follows:

$$\hat{\xi}(n,k,\omega,\Omega) = a \cdot \frac{\mid \hat{C}_X(n-1,k,\omega,\Omega) \mid^2}{\mid C_N(n-1,k,\omega,\Omega) \mid^2} + (1-a) \cdot \max\left(\frac{\mid C_Y(n-1,k,\omega,\Omega) \mid^2}{\mid C_N(n-1,k,\omega,\Omega) \mid^2} - 1, 0\right) \tag{7}$$

where a is a smoothing constant; $\hat{C}_X(n-1,k,\omega,\Omega)$ denotes the enhanced modulation spectrum at time frame $n-1$ and frequency bin k; $C_Y(n-1,k,\omega,\Omega)$ and $C_N(n-1,k,\omega,\Omega)$ denote the noisy and noise modulation spectra, respectively. Then the general form of the parametric Wiener gain in the modulation domain is defined as :

$$g(n,k,\omega,\Omega) = \left(\frac{\hat{\xi}(n,k,\omega,\Omega)}{\hat{\xi}(n,k,\omega,\Omega) + \alpha}\right)^{\beta} \tag{8}$$

where α and β are attenuation parameters. By varying the parameters, we can obtain different Wiener filters with different attenuation gains, which control the trade-off between speech distortion and noise reduction of the Wiener filters. Basically, the conventional Wiener gain $g(n,k)$ in the frequency domain only modifies a particular T-F unit, while the gain $g(\omega, \Omega; n, k)$ in the modulation domain modifies "local" modulations of that T-F unit to enhance underlying "textures" around that unit. Note that the degrees of the "local" are characterized by different widths of the impulse responses of different 2D modulation filters parameterized by (ω, Ω). Our proposed modulation Wiener filter to enhance speech intelligibility is summarized in Fig. 4.

4. EVALUATION AND RESULTS

For evaluations, we used the wideband clean samples from NOISEUS corpus [15], which contains thirty phonetically-balanced sentences spoken by three male and three female

Table 1. Mean and standard deviation of STOI scores for each enhancement method and noise type in the 0 dB SNR condition.

	SSN	babble	white
noisy	0.70 (0.05)	0.64 (0.05)	0.74 (0.05)
Wiener	0.72 (0.05)	0.63 (0.06)	0.76 (0.05)
Proposed	0.77 (0.04)	0.66 (0.05)	0.80 (0.05)
IdBM	0.86 (0.02)	0.85 (0.04)	0.86 (0.03)

speakers (five sentences per speaker). The clean speech signals were first downsampled to 16 kHz sampling frequency and three types of noise (speech-shaped noise (SSN), cafeteria babble, and white noise) were added to corrupt the clean signals with 0 dB SNR. The SSN and cafeteria babble noise were extracted from the Noise Recordings [15] and white noise was extracted from NOISEX-92 [18]. To determine the α and β parameters of the proposed modulation Wiener filter, we conducted pilot experiments under the SSN condition to calculate the average STOI score and found $\alpha = 8$ and $\beta = 0.5$ gives the best score. The STOI calculates the short-time temporal envelope correlation between clean and degraded speech signals and has been shown *highly* correlated with subjective intelligibility scores [19]. The STOI ranges from 0 to 1, and a higher value means more intelligible. For a fair comparison with the conventional Wiener filter, its parameters were also optimally selected as $\alpha = 1$ and $\beta = 0.5$, which give the highest STOI score in each noise condition. The noise estimation for the conventional Wiener filter was done using the non-iterative decision-directed noise estimation method as well [17].

The enhanced speech signal using ideal binary mask (IdBM) was also generated as an upper bound on the STOI measure. The IdBM is derived with *a priori* knowledge of energies of the target and interference sounds. Specifically, the mask retains/removes the T-F unit when its SNR is greater/smaller than a predefined local criterion (LC). The IdBM has been shown carrying critical speech intelligibility information not only for normal-hearing (NH) listeners but also for HI patients [20, 21]. In this work, the LC was set to -5 dB. The mean and standard deviation of the STOI scores are shown in Table 1 for each enhancement method and each noise type. The standard deviations are listed in parentheses.

One-way analysis of variance (ANOVA) tests were carried out to compare STOI scores of our proposed algorithm and of the conventional Wiener filter with STOI scores of original noisy signals in each noisy condition. Test statistics are shown in Table 2. From these results, one can observe our proposed algorithm demonstrates significant effects in improving STOI scores in SSN [$F(1, 58) = 19.236, p < 0.001$] and white [$F(1, 58) = 17.435, p < 0.001$] noisy conditions, but not in the babble noise condition [$F(1, 58) = 0.345$, $p = 0.559$]. The reason is that the babble noise consists of similar spectro-temporal modulations as speech such that our proposed method has intrinsic difficulty in reducing the babble noise in the modulation domain. In contrast, the conventional Wiener filter fails to improve STOI scores in all noisy conditions. It is consistent to the conclusion from the subjective test results reported in [1, 2, 3] that conventional Wiener filter does not improve speech intelligibility. Note, many supervised learning algorithms were proposed recently for enhancing speech intelligibility by demonstrating their capability in producing higher STOI scores [22]. Our proposed method is an unsupervised algorithm, which might be more attainable for small hearing assistive devices.

Table 2. Results of one-way ANOVA between enhanced speech (from proposed modulation Wiener filter and conventional Wiener filter) and noisy speech.

	Wiener	Proposed
SSN	$F(1, 58) = 2.290$, $p = 0.136$	$F(1, 58) = 19.236$, $p < 0.001$
babble	$F(1, 58) = 1.031$, $p = 0.314$	$F(1, 58) = 0.345$, $p = 0.559$
white	$F(1, 58) = 1.974$, $p = 0.165$	$F(1, 58) = 17.435$, $p < 0.001$

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a single-channel speech enhancement algorithm, which suppresses noise and enhances the textures of the speech in the modulation domain. The objective STOI scores demonstrate that our proposed algorithm can improve speech intelligibility under SSN and white noise conditions.

In this work, we only use a direct rule to estimate *a priori* modulation SNR. More generalized or iterative noise estimation mechanisms [15] could be tested in the future. In addition, different time and frequency sensitivities of different modulation filters should be considered in the noise estimation module. For instance, the weighted combination rule of the past and present estimates of $\xi(n, k, \omega, \Omega)$ could be dropped in low rate filters but set highly sensitive in high rate filters. In addition to refining the noise estimation module, we will conduct psychoacoustic experiments for HI patients in the near future to evaluate the efficacy of the proposed algorithm in improving Mandarin speech intelligibility for patients.

6. ACKNOWLEDGEMENTS

This research is supported by the Ministry of Science and Technology, Taiwan under Grant MOST 103-2220-E-009-003 and the Biomedical Electronics Translational Research Center, National Chiao Tung University.

7. REFERENCES

[1] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America*, vol. 122, pp. 1777–1786, 2007.

[2] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[3] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi, and P. C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," *Journal of the Acoustical Society of America*, vol. 129, pp. 3291–3301, 2011.

[4] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.

[5] J. K. Thompson and L. E. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," in *Proc. of IEEE ICASSP*, 2003, pp. V–397–400.

[6] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.

[7] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, pp. 282–305, 2012.

[8] T. Chi, P. Ru, and S. A. Shamma, "Multi-resolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[9] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Computational Biology*, vol. 3, pp. e1000302, 2009.

[10] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[11] R. M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 34–43, 2012.

[12] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. of the National Academy of Sciences of the United States of America*, vol. 102, pp. 2293–2298, 2005.

[13] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Voice activity detection based on frequency modulation of harmonics," in *Proc. of IEEE ICASSP*, 2013, pp. 6679–6683.

[14] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Spectro-temporal subband wiener filter for speech enhancement," in *Proc. of IEEE ICASSP*, 2012, pp. 4001–4004.

[15] P. C. Loizou, *Speech Enhancement : Theory and Practice, Second Edition*, CRC Press, 2013.

[16] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 236–243, 1984.

[17] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. of IEEE ICASSP*, 1996, pp. 629–632.

[18] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 19, pp. 2125–2136, 2011.

[20] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *Journal of the Acoustical Society of America*, vol. 123, pp. 1673–1682, 2008.

[21] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 125, pp. 2336–2347, 2009.

[22] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 22, pp. 1849–1858, 2014.