Interspeech 2018
2-6 September 2018, Hyderabad

Using Shifted Real Spectrum Mask as Training Target for Supervised Speech Separation

Yun Liu, Hui Zhang, Xueliang Zhang*

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Inner Mongolia University, Hohhot, China

liuyun.nogizaka@qq.com, alzhu.san@163.com, cszxl@imu.edu.cn

Abstract

Deep learning-based speech separation has been widely studied in recent years. Most of these kind approaches focus on recovering the magnitude spectrum of the target speech, but ignore the phase estimation. Recently, a method called shifted real spectrum (SRS) is proposed. Unlike the short-time Fourier transform (STFT), the SRS contains only real components which encode the phase information. In this paper, we propose several SRS-based masks and use them as the training target of deep neural networks. Experimental results show that the proposed target outperforms the commonly used masks computed on STFT in general.

Index Terms: shifted real spectrum, deep neural networks, training targets, speech separation

1. Introduction

Many speech applications, such as robust automatic speech recognition (ASR) and voice communication, need to acquire speech signals for further processing, but the signal of interest may be corrupted by additive background noise sometimes. To fight against the noise, speech separation aims to extract the target speech signal from a noisy speech. However, in the real environment, the speech separation performance is far from satisfactory, especially in the case of non-stationary noise and monaural conditions. This study focuses on monaural speech separation and non-stationary noise.

It is challenging that monaural speech separation uses only a single microphone to capture speech signal, while there are also many valuable methods have been proposed. Speech enhancement approaches [1, 2], such as spectral subtraction [3], estimate clean speech from noisy speech by estimating the noise firstly. In order to estimate the noise, speech enhancement approaches typically assume that the noise is stationary, therefore these methods cannot deal with the non-stationary noise. Computational auditory scene analysis (CASA) [4] tries to simulate the processing of the human auditory system which can solve the speech separation problem easily. CASA uses the ideal binary mask (IBM) [5] as the basic computational target. IBM performs well in both of the stationary and non-stationary noise conditions. Taking the IBM as the computational target leads the speech separation to be formalized as a supervised learning problem which could be solved with deep learning algorithms. Recently, with the development of supervised methods, the speech separation systems have achieved considerable performance improvements [6].

A typical supervised speech separation system usually learns a mapping function from noisy features to a training target with a supervised model, such as a deep neural network (DNN). The input features have been well-studied [7]. Amplitude modulation spectrogram (AMS) [8], mel-frequency cepstral coefficient (MFCC) [9], gammatone frequency cepstral coefficient (GFCC) [10], perceptual linear prediction (PLP) [11], and relative spectral transforms PLP (RASTA-PLP) [12] are commonly used features, and a complementary feature set included AMS, MFCC, GFCC and RASTA-PLP has been recommended in [7] and then been applied in many studies. The training targets have also been well-studied [13]. There are mainly two groups of training targets: mapping-based targets and masking-based targets. Mapping-based targets are the spectral representations of clean speech, such as the short-time Fourier transform (STFT) magnitude spectrum. Masking-based targets describe the relationships between clean speech and background interference in the time-frequency (T-F) domain, such as the IBM, ideal ratio mask (IRM) [14], FFT-mask [13], target binary mask (TBM) [15].

Most of these training targets only focus on the magnitude spectrum and ignored phase spectrum, because the early studies suggested that the phase is unimportant [16, 17], while recent studies suggest that phase is also important for perceptual quality [18]. Ignoring the phase will lead to degradation in the speech separation performance. The phase-sensitive mask (PSM) [19] and the complex ideal ratio mask (cIRM) [20] take the phase into their consideration and show better separation performance than the training targets without phase information. The cIRM is a complex mask, whose elements are complex numbers. Note that PSM is the real part of the cIRM.

In this study, we proposed a new training target with phase information. It is mainly inspired by the shifted real spectrum (SRS) [21] which is a spectral representation method in the real number field instead of STFT in the complex number field. On the basis of SRS, we proposed SRS-masks. Because SRS is in the real number field, all of the elements of SRS-mask are real numbers. Following the definition of cIRM and IRM, we define two versions of the SRS-mask. One is cIRM-like SRS-mask, called $cIRM_{srs}$. It contains the same information as the cIRM. Therefore, as the cIRM, $cIRM_{srs}$ is an optimal mask, we can perfectly reconstruct the speech signal from the $cIRM_{srs}$ and the noisy speech. Another version is IRM-like SRS-mask, called IRM_{srs}. As IRM, IRM_{srs} assumes the noise and speech are independent, so that the IRM_{srs} varies from 0 to 1, which makes it easy to model. Experimental results show that using the proposed SRS-mask can achieve better performance than the IRM which lacks of phase information, and achieve comparable performance compared to the cIRM and PSM which contain phase information. Further analysis indicates the IRM-like IRM_{srs} is a good tradeoff between the accuracy and modeling difficulty.

This research was supported in part by the China national nature science foundation (No. 61365006).

10.21437/Interspeech.2018-1650

2. Time-frequency Representations and Masks

Given a time-domain signal x, we can decompose it into time-frequency domain via discrete time Fourier transform (DTFT). The result is a complex spectrum X. For a complex number, it has a real part and an imaginary part. We denote the real part of X as X_R which contains real part of all elements. Similarly, we denote the imaginary part of X as X_I. Thus, $X = X_R + j \cdot X_I$, where $j = \sqrt{-1}$, is the imaginary unit. The signal x can be represented as (X_R, X_I) without loss.

The real part X_R stands for a composition of a series of cosine basis functions, which is an even function. Similarly, X_I stands for a composition of a series of sine basis functions, which is an odd function. This means every time-domain x can be represented in terms of its even and odd part:

$$x = x_{even} + x_{odd} \tag{1}$$

where $x_{even} = IDTFT(X_R)$ and $x_{odd} = IDTFT(j \cdot X_I)$, $IDTFT$ is the inverse DTFT.

For special signal x, X_R or X_I can be ignored when it is a special value or can be retrieved from the left one. For example, if x is an even function, we have $x_{even} = x = IDTFT(X_R)$, and $x_{odd} = IDTFT(j \cdot X_I) = 0$. If x is an odd function, we have $x_{even} = IDTFT(X_R) = 0$, and $x_{odd} = x = IDTFT(j \cdot X_I)$.

In SRS, signal x is padded with zeros to make $x(t) = 0$, when $t \leq 0$, where t is the time index. As shown in Fig. 1, this type of signal x can be decomposed into x_{even} and x_{odd}, where the x_{even} and x_{odd} have the following relationship:

$$x_{even}(t) = x_{odd}(t) = \frac{1}{2}x(t) \qquad \text{if} \quad t > 0 \tag{2}$$

$$x_{even}(t) = -\, x_{odd}(t) \qquad \text{if} \quad t \leq 0 \tag{3}$$

Because:

$$x(t) = x_{even}(t) + x_{odd}(t) = 0 \qquad \text{if} \quad t \leq 0 \tag{4}$$

$$x(t) = x_{even}(t) + x_{odd}(t) \qquad \text{if} \quad t > 0 \tag{5}$$

$$x_{even}(t) = x_{even}(-t) \quad \text{and} \quad x_{odd}(t) = -x_{odd}(-t) \tag{6}$$

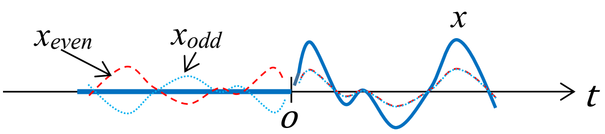


Figure 1: *Zero-padded signal and its even and odd parts.*

Therefore, with proper zero-padding, the signal x can be represented as X_R or X_I without loss. The original signal x can be recovered as $x = 2 \cdot IDTFT(X_R) = 2 \cdot IDTFT(j \cdot X_I)$. The zero-padding should make sure that $x(t) = 0$ when $t \leq 0$. It means we need to pad zeros $m + 1$ times at least, where m is the signal length. To make the padded signal length even, we pad zeros $m + 2$ times as shown in Fig. 2. In summary, we get the time-frequency representation as follows: firstly, the signal is separated into windowed frames. Secondly, we apply DTFT to these windowed frames. Thirdly, we take the real part (or imaginary part) as the representation of signal. We discard the left steps in the original paper [21].

After obtaining the time-frequency representation, we can build up a time-frequency mask. In the STFT representation,

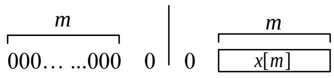


Figure 2: *Zero-padding.*

we consider Y, N and S as the complex spectrum of the time-domain noisy speech y, noisy n and speech s signals. Commonly used masks included IRM, cIRM and PSM are listed in Tab. 1. Where $|x|$ is the modulus of x. In PSM, $\theta = \theta^S - \theta^Y$, where θ^S and θ^Y are the phase angle of speech S and noisy speech Y. Take $S = |S| \cdot e^{j \cdot \theta^S}$ and $Y = |Y| \cdot e^{j \cdot \theta^Y}$, then

$$cIRM = \frac{S}{Y} = \frac{|S| \cdot e^{j \cdot \theta^S}}{|Y| \cdot e^{j \cdot \theta^Y}} = \frac{|S|}{|Y|} e^{j \cdot (\theta^S - \theta^Y)} \tag{7}$$

$$= \frac{|S|}{|Y|} \left(\cos(\theta^S - \theta^Y) + j \cdot \sin(\theta^S - \theta^Y) \right) \tag{8}$$

Thus PSM is the real part of the cIRM.

Table 1: *Mask and its Formula.*

Mask	**Formula**
IRM	$\sqrt{\frac{\lvert S\rvert^2}{\lvert S\rvert^2 + \lvert N\rvert^2}}$
$cIRM$	$\frac{S}{Y}$
PSM	$\frac{\lvert S\rvert}{\lvert Y\rvert} \cos \theta$
IRM_{srs}	$\sqrt{\frac{\lvert S_{srs}\rvert^2}{\lvert S_{srs}\rvert^2 + \lvert N_{srs}\rvert^2}}$
$cIRM_{srs}$	$\frac{S_{srs}}{Y_{srs}}$

Given the estimated time-frequency mask $\hat{M}$, we can recover the estimated speech $\hat{S}$ by multiplying the mask with the frequency-domain representation of the noisy speech. In the STFT domain, $\hat{S} = \hat{M} \otimes Y$, where $\otimes$ denotes the element-wise multiplication.

Following the definition of cIRM and IRM, in SRS representations, we define two versions of SRS-mask: cIRM-like $cIRM_{srs}$, and IRM-like IRM_{srs} as in Tab. 1, where Y_{srs}, N_{srs} and S_{srs} are the SRS representations of the time-domain noisy speech y, noise n and speech s signals. Given the estimated time-frequency SRS-mask $\hat{M}_{srs}$, we can obtain the estimated speech $\hat{S}$ by multiplying the mask with the SRS representation of the noisy speech: $\hat{S} = \hat{M}_{srs} \otimes Y_{srs}$.

Spectrogram plots of masks, clean speech and noisy speech are given in Fig. 4. As shown in the figures, we find the proposed masks have a clear structure which indicates they can be learned easily.

3. Experiments and Results

3.1. Dataset and System setup

We select 600 utterances as training utterances from the IEEE database randomly, which consists of 720 spoken utterances by a single male speaker. The rest of 120 utterances are selected as test utterances. We use five types of noise as our training and test noise, including speech-shaped noise (SSN), babble noise (babble), factory noise (factory), destroyerengine noise (engine) and destroyerops noise (operating). SSN is a stationary noise, while the other noises are non-stationary and each signal is around 4 minutes long. The first 2 minutes of each noise are random cut and mixed with 600 selected training utterances at SNRs of -3, 0, 3, and 6 dB, resulting in 12000 (600 signals $\times$ 4 SNRs $\times$ 5 noises) mixtures as training set. The last 2 minutes of each noise are random cut and mixed with 120 test utterances at the same 4 SNRs, resulting in 2400 (120 signals $\times$ 4 SNRs $\times$ 5 noises) mixtures as test set. The noise is divided into two sections to ensure that the test noise is not repeated in the training set. All utterances are downsampled to 16 kHz. We have divided the speech signal into frames using 20 ms hamming window with 10 ms overlap.

A complementary set of four features is provided as the input to the DNN which contains AMS, MFCC, RASTA-PLP and cochleagram response. These features are extracted from a 64-channel gammatone filterbank, and their deltas are used. After extracting features from the noisy speech, we apply mean and variance normalization to these features. At last, auto-regressive moving average (ARMA) [22] filters are applied on these features. ARMA filter can smooth each feature dimension to reduce background noise interference. ARMR has already been used in speech separation and achieved good results [23]. We splice the ARMA-filtered features with a five-frames context windows (two previous and two following frames) into an input feature vector.

DNN is used to estimate the mask. The DNN has three hidden layers and each hidden layer has 1024 units. Hidden units use the rectified linear unit (ReLU) as activation function. Dropout regularization is used for network training to prevent overfitting and dropout rate is 0.2. Adam is used for optimization and the mean-square error (MSE) loss function is used in the backpropagation algorithm to update the DNN weights.

3.2. Comparison Methods

We compare the proposed training targets IRM_{srs} and $cIRM_{srs}$ with IRM, cIRM and PSM. We evaluated the estimated speech from every mask with two objective metrics: short-time objective intelligibility (STOI) score [24] and the perceptual evaluation of speech quality (PESQ) score [25]. STOI shows an objective intelligibility between clean and separated speech by calculating the correlation of short-time temporal envelopes. The range of STOI is $[0, 1]$, while the higher score means better performance. PESQ is calculated by comparing the separated speech with the corresponding clean speech. The range of PESQ is $[-0.5, 4.5]$, higher score also means better performance.

The value range of $cIRM_{srs}$, cIRM and PSM is $(-\infty, \infty)$, which is not suitable for estimating, so we compress them with the following hyperbolic tangent function.

$$M_x = K\frac{1 - e^{-C\cdot m_x}}{1 + e^{-C\cdot m_x}} \tag{9}$$

where m_x is the unprocessed mask. This compression restricted mask value to $[-K, K]$ and C controls its steepness. Experiment from [20] has proved that $K = 10$ and $C = 0.1$ are used to train the DNN. We use the following inverse function to recover the uncompressed mask. Where O_x is the output from the DNN.

$$\hat{M}_x = -\frac{1}{C}\log\left(\frac{K - O_x}{K + O_x}\right) \tag{10}$$

3.3. Experiment Results

Firstly, we show that we can retrieve the original signal from the SRS representation without loss. Fig. 3 shows a signal, the reconstruction of it via STFT and SRS and their reconstruction errors. It can be observed that reconstruction error using both approaches is very less (less than 15-16 orders of magnitude). Hence, the SRS representation gives reconstruction without loss.

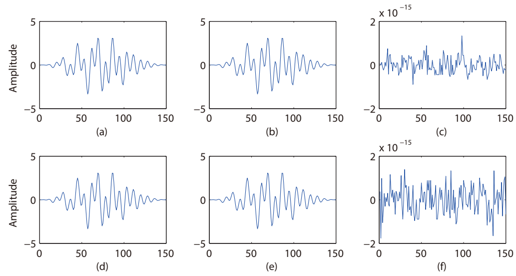


Figure 3: *(a) original signal, (b) signal reconstructed using STFT, and (c) reconstruction error. Similarly, (d) same signal, (e) signal reconstructed using SRS, and (f) reconstruction error.*

Secondly, we show the proposed ideal SRS-mask, $cIRM_{srs}$ and IRM_{srs}, can reconstruct the clean speech properly. Tab. 2 lists the separation performances of different ideal mask. It can be seen that both ideal IRM_{srs} and $cIRM_{srs}$ can achieve reasonable performance of clean speech reconstruction in STOI and PESQ. The proposed SRS-mask can improve the speech quality with the phase information.

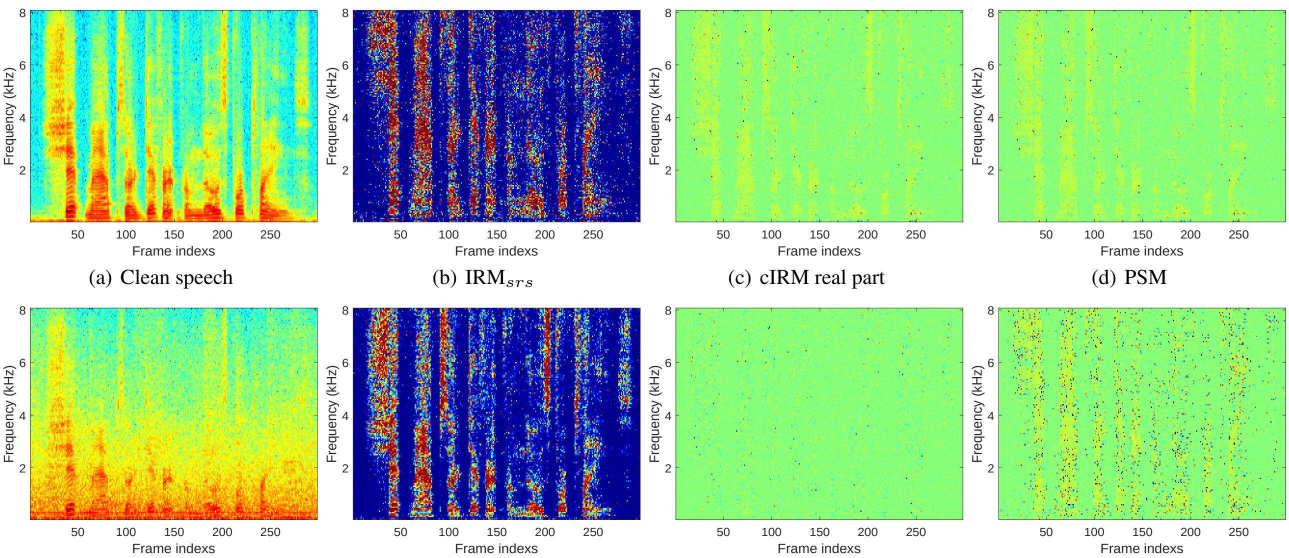
Table 2: *A example of ideal mask performances.*

Mask	**STOI**	**PESQ**
IRM	0.95	3.42
PSM	0.95	3.62
$cIRM$	1.00	4.50
IRM_{srs}	0.95	3.53
$cIRM_{srs}$	1.00	4.50

Lastly, we show the difficulty in estimating these masks. The separating results under different SNRs and noise types are given in Tab. 3. Bold means the best results on one kind of noise. Theoretically, cIRM, PSM, IRM_{srs} and $cIRM_{srs}$ take phase information into account, which may lead to a better performance than IRM. As shown in Tab. 3, the masks with phase information perform better than the IRM in general. Compared the cIRM, $cIRM_{srs}$ performs worse, although both of them are optimal mask. Because the big mask value in $cIRM_{srs}$ is much more than cIRM, which is not easy to be estimated. Compared the cIRM, PSM is sub-optimal but performs better. It indicates PSM is easier to estimate. Finally, the proposed IRM_{srs} obtains comparable performance as PSM, and achieves the best scores in the terms of speech intelligibility. It indicates that the IRM_{srs} can be learned easily. IRM_{srs} is a good tradeoff between the accuracy and modeling difficulty.

Table 3: *Average performance scores between different masks on -3,0,3,6 dB noisy speech.*

SNR	**MASK**	**PESQ**					**STOI**				
		babble	engine	operating	factory	SSN	babble	engine	operating	factory	SSN
-3dB	Mixture	1.53	1.57	1.47	1.41	1.38	0.60	0.64	0.63	0.61	0.61
	IRM	1.82	2.27	2.23	2.01	2.11	0.70	0.84	0.80	0.76	**0.78**
	IRM-srs	1.90	**2.30**	**2.29**	2.10	**2.19**	**0.72**	**0.85**	**0.81**	**0.77**	**0.78**
	cIRM	1.75	2.20	2.25	2.04	2.12	0.69	0.82	0.78	0.75	0.75
	cIRM-srs	1.71	2.19	2.21	2.02	2.07	0.69	0.81	0.78	0.74	0.75
	PSM	**1.92**	2.25	2.28	**2.12**	2.17	**0.72**	0.84	0.80	0.76	**0.78**
0dB	Mixture	1.72	1.68	1.68	1.57	1.54	0.67	0.71	0.69	0.67	0.67
	IRM	2.15	2.52	2.50	2.31	2.40	**0.80**	0.88	0.85	0.83	0.82
	IRM-srs	2.17	**2.53**	2.52	**2.37**	**2.43**	**0.80**	**0.89**	**0.86**	**0.84**	**0.84**
	cIRM	2.13	2.47	2.49	2.35	2.42	0.78	0.87	0.85	0.82	0.83
	cIRM-srs	2.06	2.45	2.48	2.32	2.38	0.78	0.87	0.84	0.82	0.82
	PSM	**2.18**	**2.53**	**2.53**	**2.37**	2.40	**0.80**	0.88	0.85	0.83	**0.84**
3dB	Mixture	1.91	1.81	1.91	1.75	1.72	0.74	0.77	0.76	0.74	0.74
	IRM	2.43	2.75	2.73	2.56	2.65	0.85	0.91	**0.89**	0.87	0.88
	IRM-srs	2.46	**2.78**	**2.76**	2.61	2.68	**0.86**	**0.92**	**0.89**	**0.88**	**0.89**
	cIRM	2.45	2.71	2.76	2.62	**2.70**	0.86	0.90	0.88	0.87	0.88
	cIRM-srs	2.41	2.68	2.73	2.60	2.64	0.85	0.91	0.88	0.87	0.87
	PSM	**2.48**	**2.78**	**2.76**	**2.63**	2.65	0.86	0.91	**0.89**	**0.88**	**0.89**
6dB	Mixture	2.11	1.95	2.14	1.95	1.92	0.81	0.83	0.82	0.80	0.81
	IRM	2.70	2.95	2.94	2.81	2.88	0.90	**0.94**	0.91	0.91	0.91
	IRM-srs	2.73	**2.96**	2.96	**2.84**	2.91	**0.91**	**0.94**	**0.92**	**0.92**	**0.92**
	cIRM	2.72	2.90	2.95	2.83	**2.93**	0.90	0.93	**0.92**	0.91	0.91
	cIRM-srs	2.66	2.87	2.95	2.83	2.90	0.90	0.93	0.91	0.91	0.91
	PSM	**2.74**	2.90	**2.98**	2.83	2.91	0.91	0.93	**0.92**	0.91	0.91



(a) Clean speech (b) IRM$_{srs}$ (c) cIRM real part (d) PSM

(e) Noisy speech (f) IRM (g) cIRM imaginary part (h) cIRM$_{srs}$

Figure 4: *Spectrogram plots of clean speech,noisy speech and masks.*

4. Conclusions

For supervised speech separation systems, the computational targets have an important influence on its performance. In this study, we proposed a new training target which based on the SRS time-frequency representation. The proposed SRS-mask takes the phase into consideration, and improves the speech intelligibility and quality in the speech separation. The IRM-like IRM_{srs} gives a good tradeoff between the accuracy and modeling difficulty and may become an alternative choice of the commonly used IRM.

5. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[5] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *arXiv preprint arXiv:1708.07524*, 2017. [Online]. Available: https://arxiv.org/pdf/1708.07524.pdf

[7] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[8] G. Kim and P. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2080–2090, 2010.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[10] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2008 IEEE International Conference on*. IEEE, 2008, pp. 1589–1592.

[11] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[12] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[13] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *Audio Speech & Language Processing IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1849–1858, 2014.

[14] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[15] 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," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[16] D. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *Acoustics Speech Signal Processing IEEE Transactions on*, vol. 30, no. 4, pp. 679–681, 1982.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics Speech Signal Processing IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 2003.

[18] K. Paliwal, K. Wjcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

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

[20] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[21] M. H. Soni, R. Tak, and H. A. Patil, "Novel shifted real spectrum for exact signal reconstruction," *Proc. Interspeech 2017*, pp. 3112–3116, 2017.

[22] C. P. Chen and J. A. Bilmes, "Mva processing of speech features," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, 2006.

[23] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[24] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[25] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on*, 2001, pp. 749–752.