SUPERVISED SPEECH ENHANCEMENT WITH REAL SPECTRUM APPROXIMATION

Yun Liu[1], *Hui Zhang*[*1], *Xueliang Zhang*[1] *and Linju Yang*[2]

[1]Department of Computer Science, Inner Mongolia University, Hohhot, China
[2]Beijing Unisound Information Technology Co. Ltd., China

liuyun.nogizaka@qq.com, alzhu.san@163.com, cszxl@imu.edu.cn, ylj0203@qq.com

ABSTRACT

Speech enhancement aims to separate a target speech from background noise. Recently, speech enhancement has been formulated as a supervised learning problem, in which a learning machine is trained to estimate the target spectrum denoted as mapping-based method or a time-frequency mask denoted as masking-based method. Signal approximation methods indirectly estimate the target spectrum via the mask estimation, which combines the advantages of both mapping based and masking based methods. Moreover, conventional methods usually ignore the phase which is also important to the speech quality. To consider the phase, the complex number spectrum needs to be modeled. However, modeling may be difficult. In this work, a pure real number spectrum is used as an alternative representation of the complex number spectrum, and a signal approximation method is used for speech enhancement. Experimental results show that the proposed method outperforms other commonly used methods.

Index Terms— Speech Enhancement, Mask Estimation, Real Spectrum, Signal Approximation

1. INTRODUCTION

Speech enhancement involves recovering a target signal from nonspeech background noise [1]. It has many applications, such as robust automatic speech recognition (ASR), mobile speech communication and hearing aids. Speech enhancement problem could be addressed by many traditional methods, such as spectral subtraction [2], Wiener filtering [3], statistical model based methods [4] and nonnegative matrix factorization [5]. After introducing the time-frequency (T-F) masking, speech enhancement problem is treated as a supervised learning problem. However, to estimate the ideal binary mask (IBM), speech enhancement becomes a binary classification problem. The IBM classifies T-F units into speech-dominant and noise-dominant units. The estimated IBM can improve the speech intelligibility [6]. With the rapid development of deep learning, many learning machines have been employed to learn mappings from noisy acoustic features to different T-F masks or the target speech itself [7].

There are mainly two groups of training targets: masking-based targets and mapping-based targets. Masking-based targets describe the time-frequency relationships between the target speech and background noise, and they include IBM, target binary mask (TBM) [8], ideal ratio mask (IRM) [9], spectral magnitude mask (SMM) [10], complex ideal ratio mask (cIRM) [11], and so on. Mapping-based targets are the spectral representations of the target speech, which include short-time Fourier transform (STFT) magnitude spectrum, mel spectrum, and so on. Signal approximation (SA), which is a combination method, estimates a mask but optimizes a mapping-based loss function. For example, in the magnitude spectrum approximation (MSA) [12], a learning machine is used to predict the IRM, and then the estimated IRM is used to reconstruct the estimated spectrum. The learning machine is not only used to estimate the IRM, but is also trained to minimize the mean-square error (MSE) between the estimated and target magnitude spectrum. The signal approximation method can combine the advantages of both the masking-based and mapping-based approaches.

In earlier studies, it was found that phase is unimportant to the speech quality [13, 14], and partly because of this, most previous speech enhancement methods only enhance the magnitude and ignore the phase. However, reconstructing the estimated speech requires combining the estimated magnitude and the phase from the noisy speech. A recent study [15] reveals that better phase estimation improves speech perceptual quality. This study has led some researchers to develop phase enhancement algorithms for speech enhancement [11, 16–19]. In [19], the phase-sensitive spectrum approximation (PSA) method is proposed. It evaluates both the magnitude and phase error in a complex number spectrum with a phase-sensitive loss function. The cIRM defines a complex number mask on the complex number spectrum, and it can simultaneously enhance magnitude and phase. By taking phase estimation into consideration, these methods result in better performance than the methods that ignore phase.

Both the PSA and cIRM works on the complex number

This research was supported in part by the China national nature science foundation (No. 61876214, No. 61866030).

STFT spectrum. Training a learning machine for complex numbers is difficult; therefore, PSA works on the real part of the spectrum, while cIRM works on both the real and imaginary parts. In addition, a method for reconstructing the complete target speech from solely real spectrum has been proposed, and it is called shifted real spectrum [20]. Speech enhancement based on the shifted real spectrum can be addressed with a pure real spectrum [21]. In this study, we aim to estimate a real spectrum based mask with a signal approximation method called real spectrum approximation (RSA) which can directly estimate the complete speech spectrum in the real number domain.

2. BACKGROUND

2.1. Time-frequency Decomposition

For a time-domain signal $x[n]$, we can transform it into the time-frequency domain by discrete time Fourier transform (DTFT), which is the STFT spectrum. At a given time t and frequency f, the spectrum $X_{t,f}$ can be divided into two parts: the magnitude and the phase in the polar coordinates or the real and imaginary parts in the Cartesian coordinates, as shown in Eq. (1).

$$X_{t,f} = |X_{t,f}|e^{i\theta_{X_{t,f}}} = \mathrm{Re}(X_{t,f}) + \mathrm{Im}(X_{t,f}) \quad (1)$$

where $|X_{t,f}|$ is the magnitude, $\theta_{X_{t,f}}$ is the phase, $\mathrm{Re}(X_{t,f})$ is the real part and $\mathrm{Im}(X_{t,f})$ is the imaginary part of the complex number spectrum $X_{t,f}$.

The representations using polar and Cartesian coordinates can be interconnected:

$$|X_{t,f}| = \sqrt{\mathrm{Re}(X_{t,f})^2 + \mathrm{Im}(X_{t,f})^2} \quad (2)$$

$$\theta_{X_{t,f}} = \arctan\frac{\mathrm{Im}(X_{t,f})}{\mathrm{Re}(X_{t,f})} \quad (3)$$

$$\mathrm{Re}(X_{t,f}) = |X_{t,f}|\cos(\theta_{X_{t,f}}) \quad (4)$$

$$\mathrm{Im}(X_{t,f}) = |X_{t,f}|\sin(\theta_{X_{t,f}}) \quad (5)$$

From Eq. (4) and (5), we can conclude that $\mathrm{Re}(X_{t,f})$ is even because cosine function is an even function and $\mathrm{Im}(X_{t,f})$ is odd because sine is an odd function. Therefore, the real part of the spectrum, $\mathrm{Re}(X_{t,f})$, is also the even part, and the imaginary part of the spectrum, $\mathrm{Im}(X_{t,f})$, is also the odd part. This property is used to build a pure real spectrum. We drop t, f in the rest of this paper.

2.2. Time-frequency Mask

Assuming noisy speech y is a sum of the speech s and the noise n, then $y = s + n$. Taking them into the T-F domain and considering their spectrums $Y = S + N$.

The speech spectrum S can be obtained from Y by an ideal mask M:

$$S = M \cdot Y \qquad \text{where} \quad M = \frac{S}{S+N} \quad (6)$$

Taking different approximation of M, we obtain a series of T-F masks:

$$IRM = \sqrt{\frac{|S|^2}{|S|^2 + |N|^2}} \quad (7)$$

$$SMM = \frac{|S|}{|Y|} \quad (8)$$

The IRM and SMM only apply to the magnitude spectrum.

$$PSM = \frac{|S|}{|Y|}\cos(\theta^S - \theta^Y) \quad (9)$$

The PSM only applies to the real part of the spectrum.

$$cIRM = \frac{S}{Y} \quad (10)$$

The cIRM is the M itself, and it is an optimal mask which can recover the speech perfectly.

2.3. Signal Approximation Methods

Masking-based methods train a learning machine to estimate these T-F masks, and their loss function is commonly the MSE of the estimated and target masks. However, mapping-based methods train a learning machine to estimate the speech spectrum directly, and their loss function is commonly the MSE of the estimated and target spectrum. It also can be the MSE of the estimated and the target magnitude spectrum, if the phase is ignored.

The signal approximation methods combine the masking-based and the mapping-based methods. Here a learning machine is trained to estimate these T-F masks. But their loss function is the MSE of the estimated and target spectrum. In the MSA, the loss function is the MSE of the estimated and target magnitude spectrum: $E_{MSA} = |\hat{M} \cdot |Y| - |S||^2$, where $\hat{M}$ is the estimated mask. In the PSA, the loss function is the MSE of the real part of the estimated and target spectrum: $E_{PSA} = |\hat{M} \cdot |Y| - |S|cos(\theta)|^2$.

3. SINGAL APPROXIMATION ON REAL SPECTRUM

3.1. Real spectrum

Any time-domain signal $x[n]$ can be represented with its even and odd parts, which is similar to a function:

$$x[n] = x[n]_{even} + x[n]_{odd} \quad (11)$$

If $x[n]$ is real and causal, which means the signal $x[n] = 0$ for all $n < 0$.

$$x[n] = x[n]_{even} + x[n]_{odd} \qquad \text{if} \quad n \geq 0 \quad (12)$$

$$0 = x[n]_{even} + x[n]_{odd} \qquad \text{if} \quad n < 0 \quad (13)$$

From Eq. (13), we can get:

$$x[n]_{even} = -x[n]_{odd} \quad \text{if} \quad n < 0 \tag{14}$$

and the properties of the even and odd functions:

$$x_{even}[-n] = x_{even}[n] \tag{15}$$

$$x_{odd}[-n] = -x_{odd}[n] \tag{16}$$

Thus:

$$x[n]_{even} = x[n]_{odd} \quad \text{if} \quad n \geq 0 \tag{17}$$

Substitute Eq. (17) into Eq. (12), we can get

$$x[n] = \ 2x[n]_{even} \ \text{ if } \ n \geq 0 \tag{18}$$

Therefore, we can reconstruct the real and causal signal $x[n]$ only by its even part. In this work, we use the even part; however, the odd part can be used in a similar manner. By taking the DTFT of $x[n]$, we can get:

$$X = 2X_{even} = 2\mathrm{Re}(X) = 2|X|\cos(\theta_X) \tag{19}$$

In practice, any real framed signal can be converted to a real causal signal by padding equal samples of zero to make the negative part become 0. In this work, we pad $m+2$ zeros to the framed speech to make the padded frame have even length, where m denotes the frame length.

We can obtain the real spectrum X^{R} of a time-domain signal $x[n]$, by framing, padding, and applying the Fourier transform and taking its real part. The real spectrum X^{R} is an equivalent representation of the original signal $x[n]$ because $x[n]$ can be recovered from X^{R} without loss.

3.2. Real Spectrum Approximation

Based on the real spectrum, we define the real spectrum mask (RSM):

$$RSM = \frac{S^{\mathrm{R}}}{Y^{\mathrm{R}}} \tag{20}$$

Like the cIRM, RSM is an optimal mask, which can recover the speech perfectly. We apply the signal approximation concept to the real spectrum, and call this method RSA. We train a learning machine to estimate the RSM and use the MSE of the estimated and target real spectrum as a loss function: $E_{RSA} = |\hat{M} \cdot Y^{\mathrm{R}} - S^{\mathrm{R}}|^2$. The loss function is the exact MSE between the estimated and target spectrum, and it does not omit the phase as MSA does or the imaginary part as PSA does.

4. EXPERIMENTS

4.1. Model Setup

We use a bidirectional long short-term memory (BLSTM) recurrent neural network as the model to estimate the mask. The BLSTM contains 2 layers. There are 384 cells in each layer. A fully connected layer is used as the output layer. The Adam optimizer is used and the initial learning rate is 0.001. MSE is used as the loss function.

The input features is a complementary set that includes four features are extracted from a 64-channel Gammatone filterbank: amplitude modulation spectrogram (AMS), mel-frequency cepstral coefficient (MFCC), relative spectral transforms perceptual linear prediction (RASTA-PLP) and cochleagram response, as well as their deltas [11]. Mean and variance normalization is applied to these features before feeding them into the BLSTM.

The used real spectrum is generated by sampling signals into 16 kHz, and then divided into frames using a 20 ms Hamming window with 10 ms overlap. We use a 320-point FFT analysis, thus the STFT magnitude spectrum in each frame is a 161-D vector. Due to the padding of zeros, the real spectrum in each frame is a 322-D vector.

4.2. Dataset

The dataset evaluated in our experiment is derived from the TIMIT dataset [22]. The dataset generation method is the same as used in [10]. In our experiment, 2000 utterances from TIMIT training set are randomly chosen as the target speech for training. All 192 utterances from the TIMIT core test set are selected for test. Five types of noises are used for training: a speech-shaped noise (SSN) and four other types of noises chosen from NOISEX database [23], which include babble noise, factory noise, destroy engine noise, and destroyer operations room noise. In addition to these five types of noises used for noise-matched condition test, other four types of new noises from the CHiME-4 dataset are also used for noise-unmatched condition test, which consists of bus noise, cafe noise, street noise and pedestrian noise. These noises include highly non-stationary talking and music noises, which make the speech enhancement become a challenging task. The training set is built by mixing all the target speech and the noises at -5 and 0 dB signal-to-noise ratio (SNR). The test set is built by mixing all the target speech and the noises at -5, 0, and 5 dB SNR, where 5 dB SNR is the SNR-unmatched condition. For the noise-matched condition test data, the first half of each noise is used in the training set, and the second half is used in the test set, to ensure that test noises are different from training.

4.3. Comparison of Methods and Evaluation Metrics

To evaluate performance, we compare the proposed RSA method with the mapping-based, masking-based method, and other signal approximation methods. The mapping-based method estimates the STFT magnitude spectrum which is called MAP. The masking-based comparison methods estimate the IRM, SMM, and cIRM. The signal approximation methods include the MSA and the PSA, as described in Section 2.3.

Table 1. Speech enhancement performance of different methods in noise-matched conditions at -5,0,5 dB.

Target	SNR(dB)											
	-5				0				5			
	STOI(%)	PESQ	SNR	SDR	STOI(%)	PESQ	SNR	SDR	STOI(%)	PESQ	SNR	SDR
mix	56.1	1.36	-5.00	-4.78	67.5	1.68	0.00	0.11	78.0	2.04	5.00	5.08
MAP	74.6	1.97	3.58	2.18	82.4	2.28	5.01	4.81	86.9	2.52	6.04	6.76
IRM	74.0	2.04	4.63	4.34	82.7	2.40	7.87	8.10	88.3	2.74	11.38	11.92
SMM	**75.6**	2.02	4.12	4.09	**84.3**	2.39	7.16	7.75	**89.4**	2.71	10.30	11.37
cIRM	72.7	2.07	5.85	5.96	82.4	2.49	8.44	9.38	88.4	2.83	10.95	12.62
MSA	75.1	**2.13**	5.20	5.13	83.7	2.48	8.40	8.84	89.1	2.79	11.59	12.45
PSA	72.6	2.07	5.65	5.95	82.5	2.48	8.78	9.58	88.5	2.81	11.78	12.73
RSA	73.0	2.09	**5.93**	**6.14**	82.8	**2.51**	**8.97**	**9.68**	88.7	**2.85**	**12.01**	**13.00**

Table 2. Speech enhancement performance of different methods in noise-unmatched conditions at -5,0,5 dB.

Target	SNR(dB)											
	-5				0				5			
	STOI(%)	PESQ	SNR	SDR	STOI(%)	PESQ	SNR	SDR	STOI(%)	PESQ	SNR	SDR
mix	64.1	1.58	-5.00	-4.78	73.4	1.94	0.00	0.11	81.8	2.29	5.00	5.08
MAP	76.8	2.03	3.95	2.88	83.2	2.31	5.22	5.19	87.0	2.51	6.07	6.78
IRM	77.2	2.17	4.72	4.52	84.3	2.55	8.06	8.30	89.1	2.90	11.69	12.21
SMM	**78.7**	2.17	4.20	4.68	**85.6**	2.54	7.19	8.20	**90.2**	2.88	10.32	11.80
cIRM	76.1	2.20	6.35	6.80	83.9	2.59	9.13	10.14	88.9	2.92	11.58	13.59
MSA	78.1	**2.26**	6.01	6.21	85.4	2.61	9.22	9.86	90.1	2.93	12.34	13.33
PSA	76.2	2.22	6.47	6.96	84.3	2.61	9.73	10.68	89.5	2.93	12.64	14.01
RSA	76.8	2.24	**6.67**	**7.08**	84.6	**2.63**	**9.89**	**10.75**	89.6	**2.95**	**12.89**	**14.11**

All the compared methods use the same BLSTM structure as the model. The activation function of output layer is selected to fit the requirement of the estimation target. Real spectrum values are clipped to $[-1, 1]$ when training, and a $\tanh$ activation function is used in the mask output layer. We compress the cIRM with a hyperbolic tangent function as in [11], while MAP follows the practice of [10].

The speech enhancement performance is evaluated in terms of short-term object intelligibility (STOI) [24], perceptual evaluation of the speech quality (PESQ) [25], SNR, and the source-to-distortion ratio (SDR). For all metrics, a higher score means better performance. These evaluation metrics are commonly used to evaluate the speech intelligibility and quality. In [26], it is indicated that the SDR has an obvious correlation with the word error rate of speech recognition system.

4.4. Results

The experimental results are listed in Table 1 and Table 2, where Table 1 shows the noise-matched test condition, and Table 2 shows the noise-unmatched test condition. From Table 1, we observe that the RSA gives the highest SNR and SDR improvements at all SNR conditions. In terms of STOI, SMM performs best. The methods that ignore phase (MAP, IRM, SMM, MSA) are not worse than those using phase (cIRM, PSA, RSA), which indicates that intelligibility is not phase-sensitive. In terms of PESQ, the methods that ignore phase are worse than those using phase in general, which indicates that phase is related to the speech quality. In terms of PESQ and STOI, RSA and PSA perform better at 0 and 5 dB SNR conditions, while MSA performs best at the -5 dB SNR condition because phase estimation at lower SNR is difficult. In general, SA methods obtain higher SDR improvement, which corresponds with results of [19].

The results for the noise-unmatched test condition (Tabel 2) are similar to that for the noise-matched condition. This shows that the proposed RSA method has good generalization ability to new noise and SNR conditions. Compared with cIRM, RSA obtains a better performance. This indicates that in supervised learning with neural networks, it is easier to model the real spectrum representation than the STFT complex number spectrum representation.

5. CONCLUSION

In this study, we proposed a real spectrum approximation method that can estimate the optimal mask in real number domain and has an overall outstanding performance in general. The high SDR of this method prompts us to apply it to a speech recognition system in future studies. In practice, an appropriate speech enhancement training target is chosen by balancing the estimation difficulty and reconstruction error. Considering this , the real spectrum and real spectrum mask may be a better choice than the STFT and STFT-based masks, and the proposed RSA method is a good approach to estimate the optimal mask.

6. REFERENCES

[1] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. abs/1708.07524, 2017.

[2] Steven 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.

[3] Pascal Scalart et al., "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 629–632.

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

[5] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[6] Eric W Healy, Sarah E Yoho, Yuxuan Wang, and DeLiang Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[7] Yuxuan Wang and DeLiang Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[8] Ulrik Kjems, Jesper B Boldt, Michael S Pedersen, Thomas Lunner, and DeLiang 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.

[9] Soundararajan Srinivasan, Nicoleta Roman, and De Liang Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

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

[11] Donald S. Williamson, Yuxuan Wang, and De Liang 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.

[12] Felix Weninger, John R Hershey, Jonathan Le Roux, and Björn Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[13] 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.

[14] Yariv Ephraim and David 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.

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

[16] David Gunawan and Deep Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, 2010.

[17] Pejman Mowlaee, Rahim Saeidi, and Rainer Martin, "Phase estimation for signal reconstruction in single-channel source separation," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[18] Martin Krawczyk and Timo Gerkmann, "Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1931–1940, 2014.

[19] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan 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] Meet H Soni, Rishabh Tak, and Hemant A Patil, "Novel shifted real spectrum for exact signal reconstruction," *Proc. Interspeech 2017*, pp. 3112–3116, 2017.

[21] Yun Liu, Hui Zhang, and Xueliang Zhang, "Using shifted real spectrum mask as training target for supervised speech separation," in *Proc. Interspeech 2018*, 2018, pp. 1151–1155.

[22] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*, vol. 93, 1993.

[23] Andrew Varga and Herman 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, no. 3, pp. 247–251, 1993.

[24] Cees Taal, Richard Hendriks, Richard Heusdens, and Jesper 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.

[26] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.