SPECTRAL MASK ESTIMATION USING DEEP NEURAL NETWORKS FOR INTER-SENSOR DATA RATIO MODEL BASED ROBUST DOA ESTIMATION

W. Q. Zheng[1]*, Y. X. Zou*[1]**, C. Ritz*[2]

[1]ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen, 518055, China
[2]School of ECTE, University of Wollongong, Wollongong, 2522, Australia.
*Corresponding author: zouyx@pkusz.edu.cn

ABSTRACT

Accurate DOA estimation based on clustering the inter-sensor data ratios (ISDRs) of a single acoustic vector sensor (AVS), referred as AVS-ISDR, relies on reliable extraction of time-frequency points with high local signal-to-noise ratio (HLSNR-TFPs) and its performance degrades in noisy environments. This paper investigates deep neural networks (DNNs) trained with noisy-clean speech pairs under different SNR levels and noise types to improve the performance of AVS-ISDR in noise conditions. The DNNs is trained to learn characteristics reflecting the level of speech information at different TFPs, which helps to generate a reliable spectral mask for obtaining a noise-reduced spectral. Correspondingly, a robust DOA estimation algorithm named as AVS-DNN-ISDR has been developed. Experimental results verify the proposed DNN-based spectral mask improves the reliable HLSNR-TFPs extraction at different SNR levels. Results from simulations and real AVS recordings further validate AVS-DNN-ISDR achieving high DOA estimation accuracy even when the SNR is lower than 0dB.

Index Terms— Direction of arrival estimation, acoustic vector sensor, deep neural networks, spectral mask estimation, inter-sensor data ratios

1. INTRODUCTION

Direction of arrival (DOA) estimation of the spatial speech source is a key technique in many applications such as video conferencing and service robots for identifying the speech source localization swiftly and accurately [1]. The acoustic vector sensor (AVS), with special properties to provide more information than the commonly used scalar sensor arrays, has previously been successfully applied in source localization applications [2].

In previous work, Li *et al*. [3] introduced a high resolution DOA estimation method using an AVS array under a spatial sparsity representation (SSR) framework by making use of the relationship between the received data model of the AVS array and its subarray manifold. In [3], 8 AVS units with spacing of half of the source wavelength were used for data capture, which limits its application when the geometry size is the main concern. To reduce the size of AVS array, Zou *et al*. [4] developed the inter-sensor data ratios (ISDRs) for multisource DOA estimation with a single AVS. In [4], the Sinusoidal tracks extraction (SinTrE) method was applied to extract the time-frequency points with high local signal-to-noise ratio (HLSNR-TFPs) by exploring the harmonic structure of speech. Since ISDRs of an AVS are independent on the source frequencies, there is no need to consider the spatial aliasing problem [5], and it is easy to estimate the elevation and azimuth angles at the same time by kernel density estimation (KDE) on the ISDRs at the time-frequency points (TFPs) with HLSNR-TFPs. However, HLSNR-TFPs extraction is sensitive to noise, which corrupts the ISDRs and degrades the DOA estimation performance.

Deep neural networks (DNNs) acts a nonlinear mapping with strong learning capability [6]. Recent research [7] using a greedy layer-wise unsupervised learning procedure has successfully applied DNNs to automatic speech recognition [8], speech enhancement [9] and other related tasks [10], outperforming the state-of-art systems. A recent study [11] adopted DNNs to denoise acoustic features at each TFP for speech separation. This idea motivates us to learn the relationship between the energy of speech and noise from the noisy-clean data pairs to generate the spectral mask, helping to extract the HLSNR-TFPs more robustly and accurately.

Following the study by Zou *et al*. [4], this paper raises a novel method to help extract the HLSNR-TFPs using DNNs for the ISDRs model. Specifically, we train a DNN model with pairs of noisy and clean speech signals, and then the recorded speech signal is decoded by the trained DNN model to generate the spectral mask, which is applied to weight each TF point derived for all AVS channels. In this way, the high energies of speech are kept, while background noise and high frequency information are removed. Hence, reliable HLSNR-TFPs for ISDRs are extracted by SinTrE method to obtain a robust DOA estimation.

Relation to prior work: This work focuses on a robust single source DOA estimation based on the ISDR model of a single AVS, and first applies the DNNs based spectral mask to improve the DOA estimation performance. Compared to common microphone array based techniques for DOA estimation [2], the AVS has a smaller size and a spatial compact structure which makes it attractive for mobile speech applications [12, 13]. Following the properties of AVS and the time-frequency (TF) sparsity of the speech, Zou *et al*. [4] proposed a multisource DOA estimation algorithm using ISDRs of single AVS, which relies on the reliable HLSNR-TFP extraction and estimating the mean value of ISDRs using clustering method.

Recent studies show that an estimated mask value at each TFP independently has been used to suppress reverberation and noise for speech intelligibility [14, 15]. Wang successfully estimated the ideal binary mask (IBM) using DNNs for monaural speech separation [16, 17]. Li *et al*. proposed a robust spectral masking system based on DNNs trained on the same filter-bank features for improving the performance of acoustic modeling in noise-robust speech recognition [18].

The remaining of this paper is organized as follows. Sect.2 provides the DOA estimation with single AVS. We then describe our proposed DNNs based spectral mask estimation and DOA estimation system in Sect.3 and Sect.4, respectively. Experimental results are presented in Sect.5. We conclude our study at last.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

2. DOA ESTIMATION WITH SINGLE AVS

For completeness, we firstly give a brief illustration of DOA estimation with single AVS based on TF domain sparsity. Here we term this DOA algorithm as AVS-ISDR in short.

2.1. Data Model for AVS

Generally, an AVS composes of four sensors, where the omnidirectional sensor and three directional sensors are defined as the *o*-, *u*-, *v*- and *w*-sensor, respectively. As the four-sensors AVS, the manifold vector for the spatial speech *s*(*t*) with DOA of (θ_s,ϕ_s) can be denote as [4]

$$\boldsymbol{a}(\theta_s,\phi_s)=[u_s,v_s,w_s,1]^T,\boldsymbol{a}\in R^{4\times 1} \quad (1)$$

where $\theta_s\in[0,180°]$, $\phi_s\in[0,360°)$ are the elevation and azimuth angle respectively, and $[.]^T$ denotes the matrix transposition. Elements u_s, v_s and w_s are named the *x*-, *y*-, *z*-axis direction cosines respectively, given by

$$u_s=\sin\theta_s\cos\phi_s, v_s=\sin\theta_s\sin\phi_s, w_s=\cos\theta_s \quad (2)$$

The data captured by the AVS at time *t* is expressed as [4]

$$\boldsymbol{x}(t)=\boldsymbol{a}(\theta_s,\phi_s)s(t)+\boldsymbol{n}(t) \quad (3)$$

where $\boldsymbol{x}(t)=[x_u(t), x_v(t), x_w(t), x_o(t)]^T$ denotes the output of the AVS, consists of the *u*-, *v*-, *w*- and *o*-sensor, respectively. $\boldsymbol{n}(t)=[n_u(t), n_v(t), n_w(t), n_o(t)]^T$ are the additive zero-mean Gaussian noise at the *u*-, *v*-, *w*- and *o*-sensor respectively, which are assumed uncorrelated to the speech source and uncorrelated to each other.

2.2. Inter-Sensor Data Ratio Model

It is commonly accepted that speech signals have sparsity in TF domain [5], which indicates that only one speech source with highest energy dominates and the contributions from other sources can be negligible at a specific TF point (τ, ω). Following this assumption, taking the short-time Fourier transform (STFT) of (3) in a compact form gives

$$\boldsymbol{X}(\tau,\omega)=\boldsymbol{a}(\theta_s,\phi_s)S(\tau,\omega)+\boldsymbol{N}(\tau,\omega) \quad (4)$$

where $\boldsymbol{X}(\tau,\omega)=[X_u(\tau,\omega), X_v(\tau,\omega), X_w(\tau,\omega), X_o(\tau,\omega)]^T$. $S(\tau,\omega)$ is STFT of $s(t)$ and $\boldsymbol{N}(\tau,\omega)=[N_u(\tau,\omega), N_v(\tau,\omega), N_w(\tau,\omega), N_o(\tau,\omega)]^T$ is the STFT of $\boldsymbol{n}(t)$.

Then the ISDRs of the AVS in the TF domain can be defined as [4]

$$I_{jo}(\tau,\omega)=X_j(\tau,\omega)/X_o(\tau,\omega), j=u,v,w \quad (5)$$

where $I_{uo}(\tau,\omega)$, $I_{vo}(\tau,\omega)$, $I_{wo}(\tau,\omega)$ are the ISDRs between *u*- and *o*-sensor, *v*- and *o*-sensor, *w*- and *o*-sensor, respectively.

With the HLSNR-TFP (τ,ω) effectively extracted by the SinTrE method, (5) can be reformulated as

$$I_{jo}(\tau,\omega)=b_j+\varepsilon_{jo}(\tau,\omega), j=u,v,w \quad (6)$$

where b_j is specified as u_s, v_s and w_s for *j*= *u*, *v*, and *w*, respectively. $\varepsilon_{uo}(\tau,\omega)$, $\varepsilon_{vo}(\tau,\omega)$ and $\varepsilon_{wo}(\tau,\omega)$ can be viewed as the residual error caused by additive noise, room reverberation and model mismatch. A detailed inference is proven in [4].

2.3. ISDRs Clustering Based DOA Estimation

From the description above, it is clear that the ISDRs $I_{uo}(\tau,\omega)$, $I_{vo}(\tau,\omega)$, $I_{wo}(\tau,\omega)$ can be viewed as random variables in TF domain with mean of u_s, v_s and w_s, respectively [4]. Specifically, the DOA estimation task is to estimate the cluster center at (u_s, v_s, w_s) by clustering the ISDRs corresponding to all HLSNR-TFPs. This can be achieved using kernel density estimation (KDE) [19, 20] applied to the ISDRs to give the clustering result ($\hat{u}_s$, $\hat{v}_s$, $\hat{w}_s$) associated with the extracted HLSNR-TFPs. Finally, the DOA information can be estimated by

$$\hat{\theta}_s=\cos^{-1}\hat{w}_s, \qquad \hat{\phi}_s=\tan^{-1}(\hat{v}_s/\hat{u}_s) \quad (7)$$

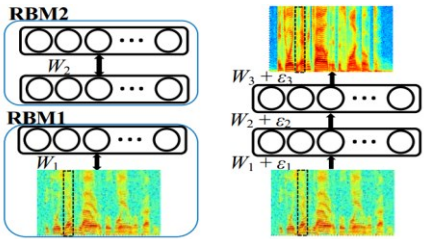

Fig. 1. Left: the RBM pre-training. Right: fine-tuning stage [9].

3. PROPOSED DEEP NEURAL NETWORKS BASED SPECTRAL MASK ESTIMATION

As mentioned above for the AVS-ISDR algorithm, HLSNR-TFPs extraction is key to DOA estimation performance. To some extent, the HLSNR-TFPs extraction is sensitive to the strong noise and reverberation. Hence, we propose spectral mask estimation using DNNs to help identify reliable HLSNR-TFPs.

3.1. DNNs Training

The DNNs training schedule includes an unsupervised pre-training phase and a supervised fine-tuning stage [21] with a collection of pairs of noisy and clean speech represented by the log-power spectra features [9]. We first pre-train a DNNs using a deep generative model of noisy log-spectra by a stack of multiple restricted Boltzmann machines (RBMs) in an unsupervised fashion [6]. The left part of Fig.1 describes the RBM pre-training from noisy data [9]. The first one denotes a Gaussian-Bernoulli RBM with one visible layer of linear variables, connected to a hidden layer. Then a pile of Bernoulli-Bernoulli RBMs is stacked behind the Gaussian-Bernoulli RBM. Afterwards, they can be trained layer-by-layer in an unsupervised greedy fashion [7]. During that, the contrastive divergence (CD) algorithm is used to update the parameters of each RBM [6].

The DNNs is initialized after pre-training, and then taking the noisy speech represented by the log-power spectra features as the input layer of DNNs and its corresponding clean features as the output layer [9]. The procedure of fine-tuning is described in the right part of Fig.1. The back-propagation algorithm with maximizing the across-entropy learning criterion between the target and the predicted output is used for fine-tuning in supervised manner [21].

3.2. DNNs Based Spectral Mask Estimation

From the output of DNNs, it is interestingly noted that the energy of speech and noise can be easily differentiated when the signal frequency is less than 1500Hz. This phenomenon can be associated with the fact that the majority of energy of speech is distributed in the fundamental frequency and a series of harmonic frequencies [22], which mainly range to no more than 1500Hz. Motivated by this, we propose a binary energy spectral mask to weight each TFP representation based on the output of DNNs, which can be defined as

$$m(\tau,\omega)=\begin{cases}1 & if\, P(\tau,\omega)>\eta\\ 0 & otherwise\end{cases} \quad (8)$$

where $P(\tau, \omega)$ denotes the output of DNNs. In this study, the threshold η is set to 0.5 by empirical results. With this spectral mask, we have kept the high energy of speech and reduced the high frequency information and background noise using:

$$\hat{\boldsymbol{X}}(\tau,\omega)=m(\tau,\omega)\boldsymbol{X}(\tau,\omega)^T \quad (9)$$

where $\hat{\boldsymbol{X}}(\tau,\omega)$ is the masked spectrum, with which the reliable HLSNR-TFPs can be extracted by the SinTrE method.

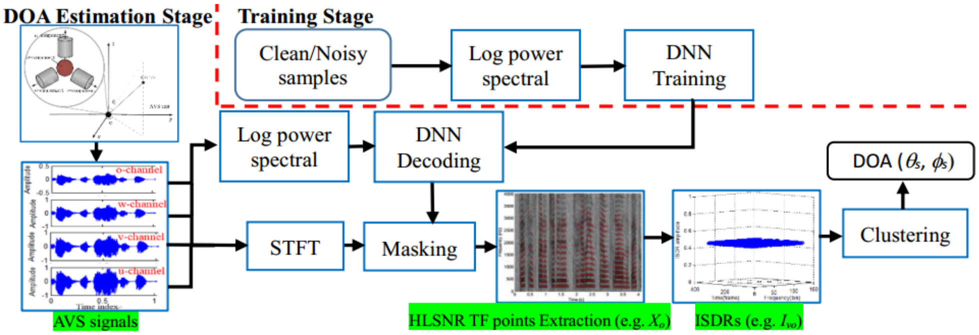

Fig. 2. Block diagram of the proposed AVS-DNN-ISDR system. Above the red dotted line is the stage for training the DNN model.

4. THE PROPOSED DOA ESTIMATION SYSTEM

To illustrate the task of DOA estimation effectively, we describe the proposed DOA estimation system shown in Fig.2. The proposed algorithm is termed as AVS-DNN-ISDR in short, which is developed under the cluster of ISDR data using a single AVS. The system consists of two stages.

DNN training stage: 1) Extract the log-power spectra features of clean and noisy speech set with the overlap Hamming window; 2) Train the DNNs with the noisy log-power spectra features as the input layer and its corresponding clean features as the output layer; 3) Save the DNNs model with the parameters of each layer.

DOA estimation stage: 1) Extract the log-power spectra features of the *o*-sensor data and calculate the STFT of the AVS output data; 2) Get the binary energy mask through the DNN decoding; 3) Add the mask on the STFT of all four sensors in each TFP; 4) Extract HLSNR-TFPs by the SinTrE method; 5) Compute the ISDRs by (5) on the HLSNR-TFPs; 6) Estimate the DOA via (7) by the clustering result derived using KDE [19, 20].

5. EXPERIMENTS AND RESULT ANALYSIS

In the DNN training stage, 2000 utterances from the training set of TIMIT database [23] are taken randomly. Additive Gaussian white noise (AWGN) and Babble noise with SNR levels varying from -5 to 30 dB increased by 5dB are added to the utterances to build a training set. This generated training set, which is a collection of large noisy training data (including one case of clean training data), is used to train the DNN models. The speech is down-sampled to 8kHz with the frame length of 256 samples (32 msec) and a frame shift of 128 samples. A short-time Fourier analysis is used to compute the DFT of each overlapping Hamming windowed frame. The number of epoch for each layer of RBM pre-training is 20. Learning rate of pre-training is set to 0.001. We use the cross entropy for the loss function of fine-tuning. The mini-batch size is set to 256. Input features of DNNs are the 129 dimensions log-power spectra features [9], which are normalized to zero mean and unit variance. The DNNs consists of 3 hidden layers with 512 units in each layer. In each DOA estimation trial, the utterance is selected from the TIMIT test set with combination of noise types and SNR levels. Due to the page limitation, only the results under white noise at different SNR levels are given, but the similar results have been observed under Babble noise. The performance of the proposed AVS-DNN-ISDR algorithm, GMDA-Laplace algorithm [5] and AVS-ISDR algorithm [4] are evaluated. The root mean squared error (RMSE) metric is used as performance measure for DOA estimation accuracy:

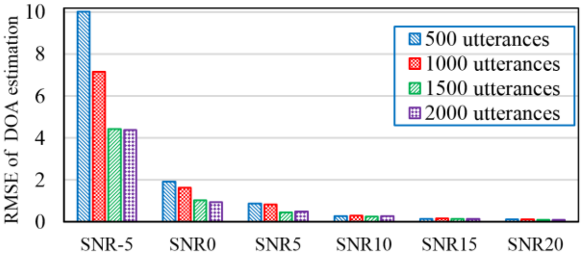

Fig. 3. RMSE of DOA estimation on the test set at different SNRs using different training set size of DNNs.

$$RMSE = 0.5\sqrt{\sum_{i=1}^{N_T}\left((\hat{\theta}_i-\theta)^2+(\hat{\phi}_i-\phi)^2\right)/N_T} \qquad (10)$$

where N_T is the number of independent trials. θ and ϕ are the target angle. $\hat{\theta}_i$ and $\hat{\phi}_i$ are the estimated angle on the i^{th} trial.

1) Effect of the training set size

Fig.3 presents the RMSE of DOA estimation on the test set at different SNRs using different training set size of DNNs. Poor results are obtained under strong noise if the training data size is only 500 utterances, which indicates that sufficient training samples are very important to obtain a more generalized model. When the training data size becomes larger, the DNN based spectral mask provides more reliable HLSNR-TFP extraction and greatly improves the performance of DOA estimation in a low SNR environment. Even up to 2000 utterances across two noise types (nearly 27 hours), the performance is not saturated.

2) Evaluation of HLSNR-TFP extraction

This section evaluates the proposed method of combining the DNNs based spectral mask and SinTrE method to achieve reliable HLSNR-TFP extraction. This experiment is conducted based on the previous experiment setting. We compare two methods for HLSNR-TFP extraction. One is the SinTrE method, the other is our proposed method (DNN-SinTrE in short). Fig.4 shows the HLSNR-TFP extraction plotted on the spectrograms of the *o*-sensor received data example of AVS by SinTrE and DNN-SinTrE methods on different SNR levels. Benefit from the learning spectral by DNNs, our proposed method can greatly reduce the pseudo HLSNR-TFPs in low SNR conditions, while the SinTrE method extracts more and more invalid data with the SNR falls. The results further verify the effectiveness and superiority of the HLSNR-TFP extraction by the DNN-SinTrE method, which indicates the robust DOA estimation in low SNR conditions.

3) DOA estimation accuracy

This simulation aims to evaluate the DOA estimation accuracy of our proposed AVS-DNN-ISDR algorithm at different angles. Specifically, the ϕ_s of spatial speech source changes from 0° to 180°, increased by 10°, under θ_s=60° and SNR=0dB without

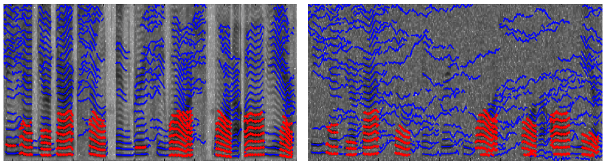

Fig. 4. HLSNR-TFP extraction: by SinTrE (blue lines) and DNN-SinTrE (red lines). Case 1(Left): clean speech; Case 2 (Right): AGWN noise at SNR=10dB.

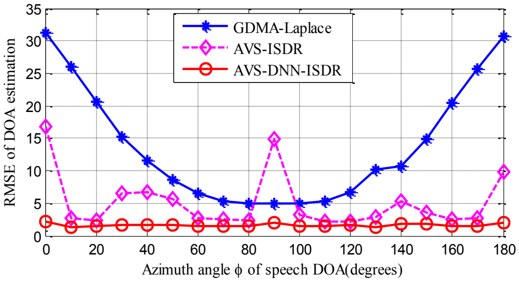

Fig. 5. RMSE versus different source DOA.

reverberation. The RMSE is obtained by 100 independent trials on each changed ϕ_s plotted in Fig.5. It is clear that the DOA estimation accuracy of our proposed AVS-DNN-ISDR is superior to that of other algorithms for all angles. Interestingly, our proposed algorithm can keep the RMSE less than 2° for all angles under the condition of 0dB, while the AVS-ISDR algorithm has a great error at some special angles (e.g. 0°, 90° and 180°).

4) Robustness of the DOA estimation

The spatial speech source is located at the DOA of (60°, 45°), and the SNR varies from -5dB to 30dB increased by 5dB without reverberation. The results are shown in Fig.6 by 100 independent trials for each SNR level. It is noted that the performance of our AVS-DNN-ISDR algorithm and the AVS-ISDR algorithm are of high accuracy (RMSE close to 0°) when SNR is larger than 10dB, outperforming that of GMDA-Laplace. In addition, it is encouraging to see that the AVS-DNN-ISDR still obtains higher DOA estimation accuracy when the SNR is smaller than 10dB, and the RMSE is much smaller than that of other two algorithms when the SNR is 0, which further verifies our proposed algorithm is more effective and robust under strong noise.

5) DOA estimation with different reverberation levels

In this experiment, the behavior of the AVS-DNN-ISDR under different reverberation levels is evaluated. The room impulse response is simulated by the image method [24] with the virtual room size of 10×5×4 m^3. Five different reverberation time (RT_{60}) conditions are considered. The speech source is set at the DOA of (60°, 45°) when the SNR is 0dB. The RMSE of DOA estimation shown in Fig.7 is conducted by 100 trials for each RT_{60}. It is clear that the curve of the AVS-DNN-ISDR method is approximately constant and keep a lower RMSE than that of AVS-ISDR for all RT_{60} conditions. This indicates that our proposed method is not sensitive to the room reverberation, which is a very favorable property since the performance of many exiting DOA estimation algorithms degrades when heavy room reverberation exists.

6) DOA estimation in a real scenario

We also conduct the DOA estimation by a simple smart car with an AVS data capturing system on it, developed by ADSPLAB (refer to Fig.8) [4]. In the case of given θ_s=90°, the car will move towards the speaker whose azimuth angle is estimated. Uncontrolled reverberation is present in the room of about 8.5×3×5 m^3 and background noise includes the noise from air conditioning and computer servers. The distance between the speaker and the AVS is 1m. Due to paper limitation, 5 different azimuth angles are estimated in Table I. It is clear that the proposed AVS-DNN-ISDR provides high DOA estimation accuracy in a real scenario, which further validate the assumptions and derivations of our proposed method.

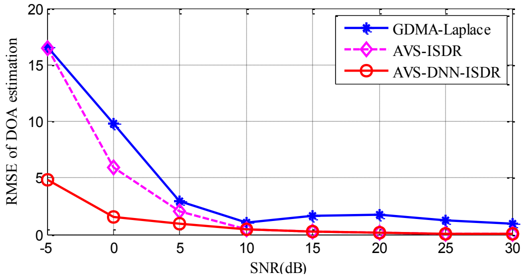

Fig. 6. RMSE versus different SNR levels.

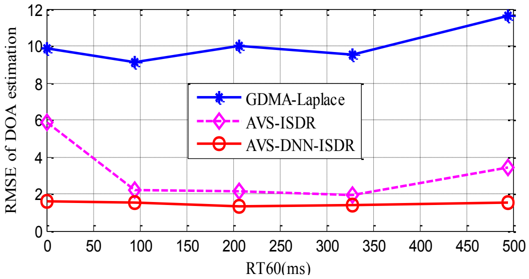

Fig. 7. RMSE versus different RT_{60}.

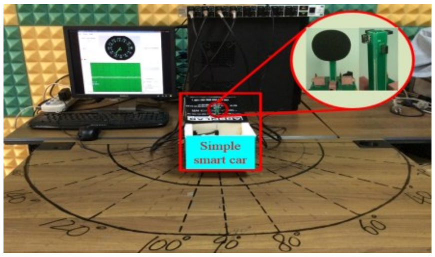

Fig. 8. A simple smart car with an AVS data capturing system

Table I. DOA estimation results in a real scenario

True DOA(°)	0	45	90	135	180
AVS-DNN-ISDR	0.78	45.18	90.35	137.10	180.26

6. CONCLUSION

In this paper, a novel deep neural networks based approach has been developed to generate an effective binary energy spectral mask, which greatly helps to obtain the enhanced speech spectrogram and results in a reliable HLSNR-TFP extraction. Accordingly, a robust DOA estimation algorithm termed as AVS-DNN-ISDR is proposed. Extensive experiments have been conducted to evaluate the performance of AVS-DNN-ISDR under different SNR levels and noise conditions. Results show that our proposed AVS-DNN-ISDR is able to obtain high DOA estimation accuracy under strong room reverberation and low SNR noisy conditions. Our future work will focus on the multisource DOA estimation by the learning DNNs in non-stationary noise situations.

7. ACKNOWLEDGEMENTS

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and Shenzhen Science & Technology Research Program (No: JC201105170727A).

12. REFERENCES

[1] F. Ribeiro, C. Zhang, D. A. Florêncio, and D. E. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 1781-1792, 2010.

[2] M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and Capon direction estimation," *Signal Processing, IEEE Transactions on,* vol. 46, pp. 2291-2304, 1998.

[3] B. Li and Y. X. Zou, "Improved DOA estimation with acoustic vector sensor arrays using spatial sparsity and subarray manifold," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 2557-2560, 2012.

[4] Y. X. Zou, W. Shi, B. Li, C. H. Ritz, M. Shujau, and J. Xi, "Multisource DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 4011-4015, 2013.

[5] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 1913-1928, 2010.

[6] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation,* vol. 18, pp. 1527-1554, 2006.

[7] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, pp. 504-507, 2006.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE,* vol. 29, pp. 82-97, 2012.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters,* vol. 21, pp. 65-68, 2014.

[10] X.-L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 853-857, 2013.

[11] Y. Wang and D. Wang, "Feature denoising for speech separation in unknown noisy environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7472-7476, 2013.

[12] M. E. Lockwood and D. L. Jones, "Beamformer performance with acoustic vector sensors in air," *The Journal of the Acoustical Society of America,* vol. 119, pp. 608-619, 2006.

[13] M. Shujau, C. Ritz, and I. Burnett, "Designing Acoustic Vector Sensors for localization of sound sources in air," in *Signal Processing Conference (EUSIPCO), 2009 Proceedings of the 17th European*, pp. 849-853, 2012.

[14] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," *IEEE Signal Processing Letters,* vol. 17, pp. 1010-1013, 2010.

[15] N. Roman and J. Woodruff, "Ideal binary masking in reverberation," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pp. 629-633, 2012.

[16] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 21, pp. 1381-1390, 2013.

[17] Y. Wang and D. Wang, "A structure-preserving training target for supervised speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6107-6111, 2014.

[18] B. Li and K. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* vol. 22, pp. 1296-1305, 2014.

[19] Z. Botev, J. Grotowski, and D. Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics,* vol. 38, pp. 2916-2957, 2010.

[20] http://www.mathworks.cn/matlabcentral/fileexchange/14034-kernel-density-estimator/content//kde.m

[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7092-7096, 2013.

[22] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 34, pp. 744-754, 1986.

[23] J. S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database NIST Tech Report, 1988.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America,* vol. 65, pp. 943-950, 1979.