Interspeech 2018
2-6 September 2018, Hyderabad

Joint Noise and Reverberation Adaptive Learning for Robust Speaker DOA Estimation with An Acoustic Vector Sensor

Disong Wang, Yuexian Zou

ADSPLAB/Intelligent Lab, School of ECE, Peking University, Shenzhen, 518055, China

1501213965@pku.edu.cn, zouyx@pkusz.edu.cn (Corresponding author)

Abstract

Deep neural network (DNN) based DOA estimation (DNN-DOAest) methods report superior performance but the degradation is observed under stronger additive noise and room reverberation conditions. Motivated by our previous work with an acoustic vector sensor (AVS) and the great success of DNN based speech denoising and dereverberation (DNN-SDD), a unified DNN framework for robust DOA estimation task is thoroughly investigated in this paper. First, a novel DOA cue termed as sub-band inter-sensor data ratio (Sb-ISDR) is proposed to efficiently represent DOA information for training a DNN-DOAest model. Second, a speech-aware DNN-SDD is presented, where coherence vectors denoting the probability of time-frequency points dominated by speech signals are used as additional input to facilitate the training to predict complex ideal ratio masks. Last, by stacking the DNN-DOAest on the DNN-SDD with a joint part, the unified network is jointly fine-tuned, which enables DNN-SDD to serve as a pre-processing front-end to adaptively generate 'clean' speech features that are easier to be correctly classified by the following DNN-DOAest for robust DOA estimation. Experimental results on simulated and recorded data confirm the effectiveness and superiority of our proposed methods under different noise and reverberations compared with baseline methods.

Index Terms: direction-of-arrival estimation, speech denoising and dereverberation, deep neural network, joint adaptive learning, acoustic vector sensor

1. Introduction

Direction of arrival (DOA) estimation of acoustic sources with a microphone array of small size has drawn much attention due to its low cost, compact physical size and possible wide-range applications such as service robots [1]. Among them, Acoustic Vector Sensor (AVS) is a promising candidate [2], since an AVS contains one pressure sensor and two or three orthogonal velocity sensors that are collocated at a point geometry in space, and has a smaller size but provides more directional information [3]. The AVS based DOA estimation algorithms could be traced back to the early 1990s, where two DOA estimators based on the intensity and velocity-covariance were firstly proposed [4]. Then many conventional DOA estimation methods had been applied to AVS, such as Multiple Signal Classification (MUSIC) [5], Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [6, 7], beamforming and Capon [3].

Though significant progress has been made, DOA estimation under noisy and reverberant environment is still a challenging task. To improve DOA performance, maximum steered response power (MSRP) [8] and maximum likelihood (ML) [9] methods were introduced, and the effects of noise and reverberation on these DOA estimators have been analyzed. Besides, under the assumption that the target speech dominated time-frequency points (TD-TFPs) can be extracted based on the sparseness of speech signal [10], DOA estimation could be improved when performed on TD-TFPs [10-12] or low-reverberant-single-source (LRSS) zones [13]. However, when the noise and reverberation become severe, less reliable TD-TFPs or LRSS zones can be determined, which causes the performance degradation in DOA estimation.

In recent years, due to the powerful learning ability of deep neural network (DNN) for speech techniques [14], in our previous work [15], a classification DNN based DOA estimation (DNN-DOAest) model was proposed, where the inter-sensor data ratios (ISDR) calculated on TD-TFPs are employed as effective DOA cues. Experiments show the significant improvement of DNN-DOAest approach compared with other non-learning based methods. Whereas, we also noted that the generalization capability of DNN-DOAest is limited since the DOA cues could be corrupted by unseen noise and reverberation, especially when noise and reverberation become severer. Motivated by the success of the DNN based speech denoising and dereverberation (DNN-SDD) serving as a front-end for other DNN based tasks, e.g., speech recognition [16-18] and voice activity detection (VAD) [19], in this work, a unified DNN framework is thoroughly investigated for robust speaker DOA estimation, which mainly contains the following contributions: First, inspired by the widely-used interaural level difference (ILD) which considers the sub-band energy for effectively inferring the DOA in binaural speech source localization [20], we propose to use sub-band ISDRs (Sb-ISDR), which are obtained by using ISDRs in each mel-scale sub-band, as effective DOA cues to build a DNN-DOAest model trained with large scale data synthesized under different noisy and reverberant conditions. Second, we present a speech-aware DNN-SDD employing the coherence vector [21] that denote the probability of time-frequency points (TFP) dominated by speech signals as additional input, which facilitates the training for prediction of complex ideal ratio masks (cIRM) [22]. Last, a unified network is developed by stacking the DNN-DOAest on top of the DNN-SDD with a joint part, and trained by a joint noise and reverberation adaptive learning strategy, which enables the DNN-SDD to adaptively generate denoised and anechoic speech features that are easier to be accurately judged by the DNN-DOAest for robust DOA estimation. Extensive experimental results under different noisy and reverberant conditions demonstrate the superiority of our proposed methods.

2. Data Model

In this paper, focusing on DOA estimation with small-sized microphone array, an AVS is used as acoustic transducer which contains one omnidirectional sensor (*o*-sensor) and two orthogonally oriented directional sensors (*u*-sensor and *v*-

10.21437/Interspeech.2018-1135

sensor respectively) [15]. Then the signal observed by the AVS at the discrete time instance t can be modeled as

$$\mathbf{x}(t)=\mathbf{h}^d(t)*s(t)+\mathbf{h}^r(t)*s(t)+\mathbf{n}(t) \quad (1)$$

where $\mathbf{x}(t)=[x_u(t), x_v(t), x_o(t)]^T$ represents the received signal at u-, v- and o-sensor respectively, the superscript T denotes the vector transpose, $s(t)$ is the speech source, $\mathbf{h}^d(t)$ and $\mathbf{h}^r(t)$ are 3-by-1 impulse responses of the direct sound and reflections respectively. $*$ denotes the convolution operation and $\mathbf{n}(t)$ is defined as 3-by-1 noise components. By taking the short-time Fourier transform (STFT), Eqn. (1) yields

$$\mathbf{X}(k,l)=\mathbf{H}^d(k)S(k,l)+\mathbf{H}^r(k)S(k,l)+\mathbf{N}(k,l) \quad (2)$$

where l ($1\leqslant l\leqslant L$) is the frame index and k ($1\leqslant k\leqslant K$) is the frequency bin index, $\mathbf{X}(k, l)=[X_u(k, l), X_v(k, l), X_o(k, l)]^T$, $\mathbf{H}^d(k)$, $\mathbf{H}^r(k)$ and $\mathbf{N}(k, l)$ are the 3-by-1 STFT coefficient vectors of $\mathbf{x}(t)$, $\mathbf{h}^d(t)$, $\mathbf{h}^r(t)$ and $\mathbf{n}(t)$ respectively, $S(k, l)$ is the STFT of $s(t)$. Specifically, $\mathbf{H}^d(k)$ and $\mathbf{H}^r(k)$ can be denoted as [13]

$$\mathbf{H}^d(k)=e^{-j\omega_k\tau}\mathbf{a}, \quad \mathbf{H}^r(k)=\sum\nolimits_q \alpha^q e^{-j\omega_k\tau^q}\mathbf{a}^q \quad (3)$$

where τ is the direct-path time delay, ω_k is the kth discrete angular frequency, and $\mathbf{a}=[u, v, 1]^T$ is the manifold vector for speech source $s(t)$ with the azimuth φ. For single AVS, $u=\cos\varphi$ and $v=\sin\varphi$. $\mathbf{a}^q=[u^q, v^q, 1]^T$ is the manifold vector pointing towards the qth reflection component, τ^q and α^q are the time delay of the reflection and attenuation due to absorption at surfaces of the room. It is obvious that the direct sound component contains the true DOA information that can be represented by the inter-sensor data ratio (ISDR) [10], which is defined at (k, l) as

$$\mathbf{p}(k,l)=\left(\Re\left(\frac{X_u(k,l)}{X_o(k,l)}\right),\Re\left(\frac{X_v(k,l)}{X_o(k,l)}\right)\right) \quad (4)$$

where $\Re(\cdot)$ denotes the real part of a complex number. In our previous work [15], ISDRs calculated on the target speech dominated time-frequency points (TD-TFPs) have been proven to be the effective DOA cues for building a classification DNN-DOAest model. However, when noise and reverberation become severe, less reliable TD-TFPs can be determined and ISDRs will be corrupted [15], leading to the performance degradation of DNN-DOAest. Therefore, more reliable DOA cues should be investigated. Besides, adding a speech denoising and dereverberation module, e.g., DNN-SDD, to enhance received signals is a straightforward idea to alleviate the above issues, but this module inevitably introduces distortions or mismatches which may deteriorate the performance of DNN-DOAest. An optional strategy is the joint training of the DNN-DOAest and DNN-SDD for reducing distortions to improve the DOA estimation precision, and this strategy is called joint noise and reverberation adaptive learning in our study.

3. Proposed Method

The proposed DOA estimation flowchart is shown in Figure 1, where the training stage can be divided into three steps. First, sub-band ISDR (Sb-ISDR) features of noisy and reverberant AVS signals are employed as DOA cues to train a classification DNN-DOAest model. Second, a speech-aware DNN-SDD is trained with log-power spectral (LPS) and coherence vectors as input and cIRMs as output. Last, a unified network is developed via joint adaptive learning of DNN-SDD and DNN-DOAest. In the DOA estimation stage, after the feature extraction of the testing AVS signals, with the voting strategy, the DOA corresponding to the maximum frequency of occurrence in

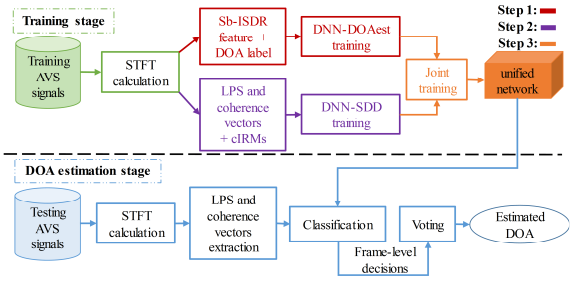

Figure 1: *Our proposed DOA estimation flowchart*

frame-level decisions made by the unified network is judged as the final estimated DOA. The details of DNN-DOAest, DNN-SDD and the unified network are elaborated below.

3.1. Classification DNN-DOAest

The DNN-DOAest is a classification DNN where the output denotes the posterior probabilities of 360 DOA angles (0°:1°:359°). Inspired by the widely-used ILD features that consider sub-band energy to infer the DOA in binaural speech source localization [20], we propose a Sb-ISDR formulated as follows

$$\mathbf{g}(f,l)=\left(\Re\left(\frac{\sum_{k\in\Omega_f}C_f(k)X_u(k,l)}{\sum_{k\in\Omega_f}C_f(k)X_o(k,l)}\right),\Re\left(\frac{\sum_{k\in\Omega_f}C_f(k)X_v(k,l)}{\sum_{k\in\Omega_f}C_f(k)X_o(k,l)}\right)\right) \quad (5)$$

where C_f is the transfer function of the f-th (f=1, 2, …, F) mel-scaled rectangle filter, and Ω_f is the support of C_f. Compared with ISDRs calculated on TD-TFPs, Sb-ISDRs omit TD-TFPs extraction which is unreliable under severe noise and reverberations, and the frequency-based decomposition allows the sub-band analysis by exploring local areas that are prone to contain more DOA information, which enables Sb-ISDRs to be reliable DOA cues. To incorporate more useful cues, the energy thresholding based voice activity detection (VAD) [23] is used to select the frame containing speech and DOA information. Then Sb-ISDRs of the speech-frames are normalized, e.g., for lth frame, $\mathbf{G}(l)=\{\mathbf{g}(f,l)/||\mathbf{g}(f,l)||, f=1,2,\ldots,F\}$, and utilized as the input of DNN-DOAest, where $||\cdot||$ denotes the norm of a vector. Then, the classification DNN-DOAest is trained by minimizing the cross-entropy criterion E_{ce}.

3.2. Speech-aware DNN-SDD

The DNN-SDD is a regression model that maps noisy and reverberant speech features, e.g., log-power spectral (LPS), to their clean versions [24] or ideal/binary ratio mask (IRM/IBM) [16]. In our study, the DNN-SDD is designed to simultaneously enhance signals of multi-channel of the AVS. Therefore, the LPS of u-, v-, o-sensor are cascaded to be the input of DNN-SDD. Besides, the coherence matrix is employed as additional feature. The coherence matrix **CM** has the same dimension with each spectrogram derived from signals of each channel, and the element $\mathbf{CM}(k, l)$ is calculated via the coherence test [21]

$$\mathbf{CM}(k,l)=\frac{1}{6}\sum_{a\neq b}\frac{\left|\mathbf{T}_{(a,b)}(k,l)\right|^2}{\mathbf{T}_{(a,a)}(k,l)\mathbf{T}_{(b,b)}(k,l)} \quad (6)$$

where $\mathbf{T}_{(a,b)}(k, l)$ is (a, b) element of $\mathbf{T}(k, l)$, which is the covariance matrix calculated across time bins as follow

$$\mathbf{T}(k,l)=\frac{1}{2Z+1}\sum_{z=-Z}^{Z}\mathbf{X}(k,l-z)\mathbf{X}^H(k,l-z) \quad (7)$$

where the time shift Z is set to 2 in this paper. The coherence value $\mathbf{CM}(k, l)$ ranges from 0 to 1 and can be regarded as the

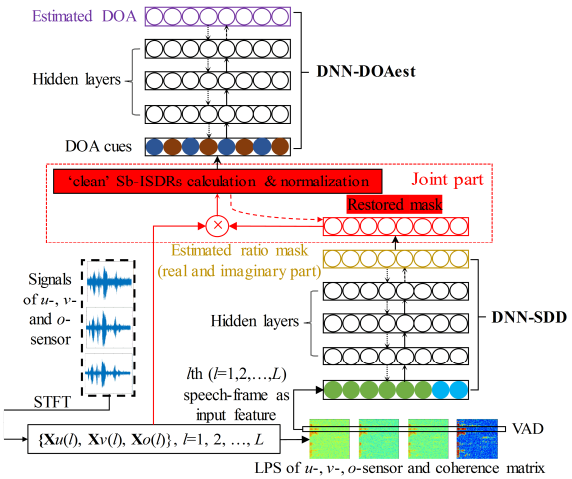

Figure 2: *Unified network trained with joint noise and reverberation adaptive learning*

probability of TFP (k, l) dominated by direct sound [21], which makes **CM** similar to IRM and provide complementary information. Therefore, the coherence vector at each frame can be combined with LPS features as the input vector for training, which is termed as speech-aware DNN-SDD training. Besides, different from the adoption of multiple adjacent frames as the input of DNN-SDD [22], no frame expansion is used in our experiments to reduce the complexity of DNN-SDD.

From Eqn. (5), it is noted that the calculation of Sb-ISDRs requires real and imaginary components of STFT of AVS signals, thus the cIRM [22] is employed and defined as the complex ratio between the STFT of direct and mixing signals

$$\mathbf{M}(k,l)=\frac{\mathbf{D}(k,l)}{\mathbf{X}(k,l)} \tag{8}$$

where $\mathbf{D}(k,l)=\mathbf{H}^d(k)S(k,l)$ is the direct sound in (2). Following [22], the real part $\mathbf{M}_r(k,l)$ and imaginary part $\mathbf{M}_i(k,l)$ of $\mathbf{M}(k,l)$ are normalized to be in range (-1,1) through the hyperbolic tangent operation before training and used as the output of the DNN-SDD. Then, the mean-square error E_{mse} is employed as the cost function for training.

3.3. Joint adaptive learning for the unified network

By stacking the DNN-DOAest on top of DNN-SDD with a joint part, a unified network is developed as illustrated in Figure 2. Specifically, using the speech-frame features as input, the real and imaginary parts of cIRM for u-, v-, o-sensor are estimated by the DNN-SDD. At the joint part shown in red in Figure 2, the inverse normalization [22] is applied to obtain restored masks $\hat{\mathbf{M}}_r(k,l)=[\hat{M}_{ur}(k,l),\hat{M}_{vr}(k,l),\hat{M}_{or}(k,l)]^T$ and $\hat{\mathbf{M}}_i(k,l)=[\hat{M}_{ui}(k,l),\hat{M}_{vi}(k,l),\hat{M}_{oi}(k,l)]^T$, which are used to estimate the direct sound

$$\hat{\mathbf{D}}(k,l)=\left(\hat{\mathbf{M}}_r(k,l)+j\hat{\mathbf{M}}_i(k,l)\right)\mathbf{X}(k,l) \tag{9}$$

Then the 'clean' Sb-ISDRs are calculated as

$$\hat{\mathbf{g}}(f,l)=\left(\Re\left(\frac{\sum_{k\in\Omega_f}C_f(k)\hat{D}_u(k,l)}{\sum_{k\in\Omega_f}C_f(k)\hat{D}_o(k,l)}\right),\Re\left(\frac{\sum_{k\in\Omega_f}C_f(k)\hat{D}_v(k,l)}{\sum_{k\in\Omega_f}C_f(k)\hat{D}_o(k,l)}\right)\right) \tag{10}$$

From Figure 2, it is clear to see that the normalized $\hat{\mathbf{G}}(l)=\{\hat{\mathbf{g}}(f,l)/||\hat{\mathbf{g}}(f,l)||, f=1, 2, \ldots, F\}$ are used as the input of the DNN-DOAest. It is noted that the Sb-ISDRs calculation (10) can be realized in the real domain, e.g., the first element of $\hat{\mathbf{g}}(f,l)$ can be rewritten as

$$\frac{U_r(k,l)O_r(k,l)-U_i(k,l)O_i(k,l)}{O_r^2(k,l)+O_i^2(k,l)} \tag{11}$$

where $U_r(k,l)$ and $U_i(k,l)$ are the real and imaginary parts of $\sum_{k\in\Omega_f}C_f(k)\hat{D}_u(k,l)$

$$U_r(k,l)=\sum_{k\in\Omega_f}C_f(k)\left(\hat{M}_{ur}(k,l)X_{ur}(k,l)-\hat{M}_{ui}(k,l)X_{ui}(k,l)\right) \tag{12}$$

$$U_i(k,l)=\sum_{k\in\Omega_f}C_f(k)\left(\hat{M}_{ur}(k,l)X_{ui}(k,l)+\hat{M}_{ui}(k,l)X_{ur}(k,l)\right) \tag{13}$$

where $X_{ur}(k,l)$ and $X_{ui}(k,l)$ are the real and imaginary parts of $X_u(k,l)$ respectively. $O_r(k,l)$ and $O_i(k,l)$ are the real and imaginary parts of $\sum_{k\in\Omega_f}C_f(k)\hat{D}_o(k,l)$, and can be derived in a similar manner. Therefore, all weights of the unified network can be updated in the real domain. In our experiments, a combination of the MSE for DNN-SDD and the cross-entropy for DNN-DOAest is used as the objective function for joint adaptive learning of the unified network

$$E_{unified}=\lambda E_{mse}+(1-\lambda)E_{ce} \tag{14}$$

where λ is a constant and set to 0.5 empirically.

4. Experiments and analysis

To generate the training data, simulations are conducted in an 8m×6m×3m room with an AVS located at [4m, 3m, 1.5m]. Similar to [13], room impulse responses are generated using the image method [25] by varying the reverberation time (T_{60}) at 0.3s, 0.6s and 0.9s. White Gaussian noise with different signal-to-noise ratios (SNR) varied from 0dB to 20dB with 5dB interval are added to each of three channels. The source is placed 1m away from the AVS by changing the DOA from 0° to 359° with 1° interval. For each angle, the simulation is repeated 3 times. At each time, 1 sentence sampled at 16kHz randomly selected from 4620 training sentences of the TIMIT dataset [26] is used as the source signal. Besides, the STFTs are realized by 512-point fast Fourier transform (FFT) and using hamming window of 512 samples, with a 50% overlap between neighboring windows. Therefore, the dimension of the input and output of DNN-SDD is 1028 (257×4) and 1542 (257×6) respectively. Following [24], the DNN-SDD has three hidden layers with 2048 units per layer and sigmoid activation function is used. For the calculation of Sb-ISDRs, the number of mel-scaled rectangle filters is set to 40. Thus, the dimension of input and output for DNN-DOAest is 80 (40×2) and 360 respectively. Additionally, the DNN-DOAest has three hidden layers with 720 units per layer and tanh activation function is used. The RMSProp optimization method [27] is applied with 15 epochs for the training of DNN-DOAest, DNN-SDD and the unified network with 5e-4, 5e-4 and 1e-6 as the initial learning rate halved every 5 epochs, respectively, which are all performed using the TensorFlow [28].

The testing is conducted in both simulation environment and real scenario. For simulation environment, the room is 6m ×7m×4m with the AVS located at [3m, 3.5m, 1.5m], and 100 testing sentences are randomly selected from the TIMIT dataset. At each simulation, we randomly choose 1 testing sentence as the test source which is 2m away from the AVS. For real scenario, experiments are conducted by using the AVS data capturing system developed by ADSPLAB [10] in an 8.5m× 3m×5m room with uncontrolled reverberation and background noise such as air conditioning and computer servers, and the distance between the speaker and the AVS is 1m. The AVS-ISDR method [10] that implements DOA estimation on ISDRs of TD-TFPs, and the AVS-LRSS method [13] that realizes DOA estimation on low-reverberant-single-source (LRSS)

Table 1: *MAE/RMSE (°) of different methods under different SNR conditions with T_{60} fixed at 0.5s*

Methods	SNR					
	-5dB	0dB	5dB	10dB	15dB	20dB
AVS-ISDR	17.20/22.76	16.94/22.06	16.86/21.63	15.79/19.01	15.36/18.99	14.34/18.58
AVS-LRSS	16.46/21.70	12.53/16.03	10.12/14.96	9.52/11.87	8.93/11.69	8.48/10.78
AVS-DNN	15.71/20.62	10.96/13.86	9.13/11.37	8.78/10.98	8.50/10.79	8.44/10.64
AVS-UN	13.68/19.23	9.31/11.47	8.33/10.25	7.65/9.68	6.88/8.57	6.91/8.58
AVS-UN-SA	**12.63/18.91**	**9.28/11.32**	**7.61/9.28**	**7.40/9.15**	**6.85/8.47**	**6.43/7.91**

Table 2: *MAE/RMSE (°) of different methods under different T_{60} conditions with SNR fixed at 5dB*

Methods	T_{60}					
	0.2s	0.4s	0.6s	0.8s	1.0s	1.2s
AVS-ISDR	4.96/5.75	10.56/12.90	20.63/26.85	28.41/37.63	33.97/45.63	39.54/53.59
AVS-LRSS	**1.82/2.37**	8.98/10.83	14.08/26.18	22.27/38.31	26.39/46.58	34.76/56.51
AVS-DNN	6.48/8.06	8.10/10.27	9.89/12.88	11.53/14.70	12.65/16.52	13.73/19.07
AVS-UN	5.55/7.16	7.69/9.58	8.68/11.08	10.33/12.97	11.78/15.41	12.78/17.71
AVS-UN-SA	5.09/6.73	**7.05/8.74**	**8.41/10.31**	**9.97/12.35**	**11.49/14.86**	**12.60/17.48**

Table 3: *MAE/RMSE (°) of different methods in real scenario*

Methods	MAE	RMSE
AVS-ISDR	9.35	11.55
AVS-LRSS	9.96	11.75
AVS-DNN	8.30	12.32
AVS-UN	6.88	10.28
AVS-UN-SA	**5.68**	**9.05**

zones are taken as baseline methods, where the settings are the same as those in [10, 13]. Besides, three proposed systems are implemented for performance comparison. The first is the DNN-DOAest trained without the DNN-SDD module (denoted as AVS-DNN). To show the effectiveness of the coherence vector for predicting cIRMs, another two systems are the unified network with speech-aware (denoted as AVS-UN-SA) and without speech-aware (denoted as AVS-UN) respectively. All methods are evaluated by the mean absolute error (MAE) and root mean square error (RMSE).

4.1. DOA estimation under simulation environment

Table 1 gives the DOA estimation results under different noisy conditions by changing the SNR from -5dB to 20dB with 5dB interval and fixing the T_{60} at 0.5s. Under each condition, the DOA is varied from 0° to 359° with 1° interval. As the SNR increases, the MAE and RMSE of all methods decrease as expected, and both lower MAE and RMSE can be obtained by AVS-DNN, AVS-UN and AVS-UN-SA, which shows the robustness of our proposed systems to noise. It is noted that AVS-UN-SA and AVS-UN outperform the AVS-DNN, which demonstrates the importance of DNN-SDD module for denoising and dereverberation that facilitate the following DOA classification by DNN-DOAest module. Besides, the AVS-UN-SA achieves better performance than the AVS-UN, showing the effectiveness of the coherence vector for predicting cIRMs.

Table 2 presents the accuracy of DOA estimation under various reverberant conditions by varying the T_{60} from 0.2s to 1.2s with 0.2s interval and fixing the SNR at 5dB. Similar results can be observed except for that AVS-ISDR and AVS-LRSS methods achieve lower error when the T_{60} is 0.2s. Since under the moderate condition, the extraction of TD-TFPs and LRSS zones is reliable and accurate, which enables AVS-ISDR and AVS-LRSS methods to obtain lower DOA estimation error. However, as the T_{60} increases, our proposed systems outperform the AVS-ISDR and AVS-LRSS methods, which shows the robustness of our proposed systems to reverberation. Compared with the AVS-UN, it is interesting to see that, under strong reverberations (e.g., $T_{60} \geqslant 0.6$s), less improvements can be obtained by the AVS-UN-SA, as the coherence vector is affected by strong reverberations and provides less reliable complementary information to predict accurate cIRMs.

4.2. DOA estimation in a real scenario

Table 3 illustrates the DOA estimation results in a real scenario, where the DOA of the speaker varies from 0° to 315° with 45° interval and 5 trials are repeated at each angle. Compared with baseline methods, it is observed that the lower MAE is achieved by the AVS-DNN, but the RMSE is higher, which means the AVS-DNN is not stable enough in the real and complex environment. However, it is encouraging to see that the AVS-UN-SA still offers the best performance with the lowest MAE and RMSE followed by AVS-UN, which further validates the effectiveness of the DNN-SDD module for enhancing signals and the robustness of our proposed DOA estimation systems in the real scenario.

5. Conclusions

To improve the DOA estimation performance in challenging environments with small-sized microphone array, we propose a unified network for robust DOA estimation by using an AVS. A novel DOA cue of AVS (Sb-ISDR) and coherence vector are investigated and employed to train a classification DNN-DOAest and a speech-aware DNN-SDD separately. Then a unified network is developed by stacking the DNN-DOAest on the DNN-SDD with a joint component. Via the joint noise and reverberation adaptive learning, the DNN-DOAest is able to obtain robust DOA estimation under different simulated and real noisy and reverberant conditions by using 'clean' speech features, which are adaptively produced by the DNN-SDD. Besides, our proposed systems have the potential to perform robust DOA estimation under various noise conditions, including directional interferences and non-directional noise, which will be studied in our future work.

6. Acknowledgements

This work was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ 20170306165153653 & JCYJ20170817160058246).

7. References

[1] F. Ribeiro, C. Zhang, D. A. Florêncio *et al.*, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, no. 7, pp. 1781-1792, 2010.

[2] 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, no. 1, pp. 608-619, 2006.

[3] M. Hawkes, and A. Nehorai, "Acoustic vector-sensor beamforming and Capon direction estimation," *IEEE Transactions on Signal Processing,* vol. 46, no. 9, pp. 2291-2304, 1998.

[4] A. Nehorai, and E. Paldi, "Performance analysis of two direction estimation algorithms using an acoustic vector sensor," in *Proc. ICASSP*, 1993, pp. 360-363.

[5] K. T. Wong, and M. D. Zoltowski, "Self-initiating MUSIC-based direction finding in underwater acoustic particle velocity-field beamspace," *IEEE Journal of Oceanic Engineering,* vol. 25, no. 2, pp. 262-273, 2000.

[6] M. Hawkes, and A. Nehorai, "Wideband source localization using a distributed acoustic vector-sensor array," *IEEE transactions on signal processing,* vol. 51, no. 6, pp. 1479-1491, 2003.

[7] P. Tichavsky, K. T. Wong, and M. D. Zoltowski, "Near-field/far-field azimuth and elevation angle estimation using a single vector hydrophone," *IEEE Transactions on Signal Processing,* vol. 49, no. 11, pp. 2498-2510, 2001.

[8] D. Levin, S. Gannot, and E. A. Habets, "Direction-of-arrival estimation using acoustic vector sensors in the presence of noise," in *Proc. ICASSP*, 2011, pp. 105-108.

[9] D. Levin, E. A. Habets, and S. Gannot, "Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor," *The Journal of the Acoustical Society of America,* vol. 131, no. 2, pp. 1240-1248, 2012.

[10] Y. X. Zou, W. Shi, B. Li *et al.*, "Multisource DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor," in *Proc. ICASSP*, 2013, pp. 4011-4015.

[11] Y. Jin, Y. Zou, and C. H. Ritz, "Robust speaker DOA estimation based on the inter-sensor data ratio model and binary mask estimation in the bispectrum domain," in *Proc. ICASSP*, 2017, pp. 3266-3270.

[12] D. Wang, Y. Zou, and W. Wang, "Learning soft mask with DNN and DNN-SVM for multi-speaker DOA estimation using an acoustic vector sensor," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1692-1709, 2017.

[13] K. Wu, V. Reju, and A. W. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," in *Proc. ICASSP*, 2015, pp. 444-448.

[14] G. Hinton, L. Deng, D. Yu *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.

[15] Y. Zou, R. Gu, D. Wang *et al.*, "Learning a robust DOA estimation model with acoustic vector sensor cues," in *Proc. APSIPA-ASC*, 2017, pp. 1688-1691.

[16] A. Narayanan, and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. ICASSP*, 2014, pp. 2504-2508.

[17] M. Mimura, S. Sakai, and T. Kawahara, "Joint Optimization of Denoising Autoencoder and DNN Acoustic Model Based on Multi-Target Learning for Noisy Speech Recognition," in *Proc. Interspeech*, 2016, pp. 3803-3807.

[18] B. Wu, K. Li, F. Ge *et al.*, "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 8, pp. 1289-1300, 2017.

[19] Q. Wang, J. Du, X. Bao *et al.*, "A universal VAD based on jointly trained deep neural networks," in *Proc. Interspeech*, 2015.

[20] G. R. Karthik, and P. K. Ghosh, "Subband selection for binaural speech source localization," in *Proc. Interspeech*, 2017, pp. 1929-1933.

[21] S. Mohan, M. E. Lockwood, M. L. Kramer *et al.*, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America,* vol. 123, no. 4, pp. 2136-2147, 2008.

[22] D. S. Williamson, and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 7, pp. 1492-1501, 2017.

[23] P. Renevey, and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Eurospeech*, 2001.

[24] Y. Xu, J. Du, L.-R. Dai *et al.*, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 23, no. 1, pp. 7-19, 2015.

[25] 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, no. 4, pp. 943-950, 1979.

[26] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD,* vol. 107, pp. 16, 1988.

[27] T. Tieleman, and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning,* vol. 4, no. 2, pp. 26-31, 2012.

[28] M. Abadi, P. Barham, J. Chen *et al.*, "TensorFlow: A System for Large-Scale Machine Learning," in *OSDI*, 2016, pp. 265-283.