Learning soft mask with DNN and DNN-SVM for multi-speaker DOA estimation using an acoustic vector sensor

Disong Wang[a], Yuexian Zou[a,*], Wenwu Wang[b]

[a] ADSPLAB, School of ECE, Peking University, Shenzhen, 518055, China

[b] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, United Kingdom

[*]Corresponding author. Email: zouyx@pkusz.edu.cn

Abstract

Using an acoustic vector sensor (AVS), an efficient method has been presented recently for direction-of-arrival (DOA) estimation of multiple speech sources via the clustering of the inter-sensor data ratio (AVS-ISDR). Through extensive experiments on simulated and recorded data, we observed that the performance of the AVS-DOA method is largely dependent on the reliable extraction of the target speech dominated time-frequency points (TD-TFPs) which, however, may be degraded with the increase in the level of additive noise and room reverberation in the background. In this paper, inspired by the great success of deep learning in speech recognition, we design two new soft mask learners, namely deep neural network (DNN) and DNN cascaded with a support vector machine (DNN-SVM), for multi-source DOA estimation, where a novel feature, namely, the tandem local spectrogram block (TLSB) is used as the input to the system. Using our proposed soft mask learners, the TD-TFPs can be accurately extracted under different noisy and reverberant conditions. Additionally, the generated soft masks can be used to calculate the weighted centers of the ISDR-clusters for better DOA estimation as compared with the original center used in our previously proposed

AVS-ISDR. Extensive experiments on simulated and recorded data have been presented to show the improved performance of our proposed methods over two baseline AVS-DOA methods in presence of noise and reverberation.

Keywords: DOA estimation; tandem local spectrogram block; soft mask; deep neural network; support vector machine.

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 wide-range applications such as video conferencing and intelligent robots for identifying speech source locations swiftly and accurately [1]. Among them, Acoustic Vector Sensor (AVS) is a promising candidate providing great convenience in configuration and portability [2]. Different from the conventional arrays with omnidirectional microphones, an AVS contains one pressure sensor and three orthogonal velocity sensors that are collocated at a point geometry in space, and has a smaller size but provides more directional information [3, 4]. Recently, several AVS based DOA estimation algorithms have been proposed [5-11], including those for the under-determined DOA estimation problem [7-11], where the number of sources is greater than the number of sensors. In these studies, a common assumption has been made that the target speech dominated TF points (TD-TFPs) can be extracted based on the sparseness of speech signals [12]. In [8, 9], the subspace characteristics of the local TF covariance matrix have been exploited to determine the TD-TFPs to estimate the DOAs. However, the ambient noise and reverberation may corrupt the signal subspace

[13], which leads to the performance degradation when using the method based on the selection of the TFPs with high Signal to Noise Ratio (HSNR).

In our previous work [7], the DOA estimation of multi-sources has been addressed by clustering the inter-sensor data ratios of single acoustic vector sensor (AVS-ISDR), where the Sinusoidal Tracks Extraction (SinTrE) method [12] is introduced to extract the reliable TD-TFPs by exploiting the harmonic structure of speech. Then the ISDRs that contain DOA cues are calculated at the extracted TD-TFPs and clustered by the Kernel Density Estimation (KDE) method [14]. As a result, the DOAs are estimated using the centers of the ISDR-clusters. The AVS-ISDR was shown to be effective in estimating the DOAs for up to seven speech sources under low noise and reverberation conditions. However, experimental results also show that the performance of the TD-TFPs extraction by SinTrE deteriorates as the level of noise and reverberation increases, resulting in performance degradation in the DOA estimation. Clearly, the reliable extraction of TD-TFPs is crucial for the AVS-ISDR method to obtain good DOA estimation performance under different noisy and reverberant conditions.

To obtain the reliable extraction of TD-TFPs, in this paper, we perform our study from the following aspects. First, we get some insights from the perceptual mechanism of the human auditory system that the target speech and interferers are separated in local TF regions [15]. Second, we evaluate the local spectrogram block (LSB) of the received signals for four channels of the AVS under different noise and reverberation levels. Experiments showed that the LSBs centered by the TD-TFPs are distinguishable from those centered by the interferers (noise or reverberation) dominated TFPs (ID-TFPs). Third, the LSBs of TD-TFPs and ID-TFPs can be considered as two different patterns, and hence can be learned in a supervised manner.

Based on the above findings, we firstly propose a novel tandem LSB (TLSB) feature, which is defined as the LSBs of the four channels of AVS in tandem that are centered by the same time-frequency point, as the input to the training system. Then, we design two different soft mask learners to extract TD-TFPs:

(1) Making use of the powerful learning ability of deep neural network (DNN) [16] with large scale training dataset, a DNN is trained by mapping the TLSB feature to the Idea Binary Mask (IBM) [17] for each TFP. Then in the testing phase, the received signals of the AVS can be transformed to TLSB features and then decoded by the well-trained DNN to generate the soft mask, which represents the probability of a TFP being considered as TD-TFP. By comparing the soft masks with a predefined threshold, the TD-TFPs can be accurately extracted.

(2) The last hidden layer representations (LHLR) of DNN are taken as the feature for training the linear support vector machine (SVM), which is motivated by the following reasons: 1) DNN can be viewed as a hierarchical feature detector, and each hidden layer of DNN is a different representation of the original feature, where the LHLRs with high dimension are more linearly separable and therefore useful for classification [18]; 2) SVM can tackle the high dimensional data classification problems [19], and is currently one of the best performers for a number of classification tasks in speech applications [18, 20-22]. In addition, the linear separability of LHLRs facilitates the performance of linear SVM with lower computational complexity as compared with kernel SVMs. Similarly, the soft masks can also be obtained via the decision function of SVM.

Following our previously proposed AVS-ISDR algorithm, the soft masks are also used to calculate the weighted centers of the ISDR-clusters, for further improving the DOA estimation accuracy.

The remainder of this paper is organized as follows. The formulation of the AVS-ISDR algorithm is illustrated in Section 2. In Section 3, we present our proposed soft mask learning algorithms for DOA estimation in details, and experiments and analysis are given in Section 4 before we conclude the paper.

2. Formulation of AVS-ISDR

2.1. Data model for AVS

Assume the acoustic signal is sampled by one single AVS in a noisy and reverberant environment. The signal observed by the AVS at the discrete time instance t can be modeled as

$$\mathbf{x}(t)=\sum_{i=1}^{I}\mathbf{h}_i(t)*s_i(t)+\mathbf{n}(t) \tag{1}$$

where $\mathbf{x}(t)=[x_u(t), x_v(t), x_w(t), x_o(t)]^T$ represents the received signal at three bidirectional sensors (*u*-, *v*-, *w*-sensors) and one omnidirectional sensor (*o*-sensor) respectively, the superscript T denotes the vector transpose. I is the number of speech sources, $s_i(t)$ is the ith source, $\mathbf{h}_i(t)=[h_{ui}(t), h_{vi}(t), h_{wi}(t), h_{oi}(t)]^T$ $(1\le i\le I)$ is the impulse response sample vector from the ith source to the corresponding sensor, $*$ denotes convolution and $\mathbf{n}(t)=[n_u(t), n_v(t), n_w(t), n_o(t)]^T$ is defined as the noise components. By taking the short-time Fourier transform (STFT), Eqn. (1) can be written as

$$\mathbf{X}(k,m)=\sum_{i=1}^{I}\mathbf{H}_i(k)S_i(k,m)+\mathbf{N}(k,m) \tag{2}$$

where m is the time frame index and k is the frequency bin index, $S_i(k,m)$ is the STFT of $s_i(t)$. $\mathbf{X}(k,m)$, $\mathbf{H}_i(k)$ and $\mathbf{N}(k,m)$ are the 4-by-1 STFT coefficient vector of $\mathbf{x}$(t), $\mathbf{h}_i(t)$, and $\mathbf{n}(t)$ respectively, which are given by

$$\mathbf{X}(k,m)=[X_u(k,m),X_v(k,m),X_w(k,m),X_o(k,m)]^T \tag{3}$$

$$\mathbf{H}_i(k)=[H_{ui}(k),H_{vi}(k),H_{wi}(k),H_{oi}(k)]^T \tag{4}$$

$$\mathbf{N}(k,m)=[N_u(k,m),N_v(k,m),N_w(k,m),N_o(k,m)]^T \tag{5}$$

With the reverberation, $\mathbf{H}_i(k)$ ($1\le i\le I$) can be decomposed into [9]

$$\mathbf{H}_i(k)=\mathbf{H}_i^d(k)+\mathbf{H}_i^r(k) \tag{6}$$

where $\mathbf{H}_i^d$ (k) and $\mathbf{H}_i^r$ (k) are the direct-path component and reflection component respectively, which are denoted as

$$\mathbf{H}_i^d(k)=e^{-j\omega_k\tau_i}\mathbf{a}_i,\quad \mathbf{H}_i^r(k)=\sum\nolimits_q \alpha_i^q e^{-j\omega_k\tau_i^q}\mathbf{a}_i^q \tag{7}$$

where τ_i is the direct-path time delay, ω_k is the kth discrete angular frequency, and $\mathbf{a}_i$ is the manifold vector for speech source $s_i(t)$ with the elevation $\theta_i\in[0°, 180°]$ and azimuth $\varphi_i\in[0°, 360°)$, which has the form

$$\mathbf{a}_i=[u_i,v_i,w_i,1]^T \tag{8}$$

where u_i, v_i and w_i are given by

$$u_i=\sin\theta_i\cos\varphi_i, v_i=\sin\theta_i\sin\varphi_i, w_i=\cos\theta_i \tag{9}$$

$\mathbf{a}_i^q=[u_i^q, v_i^q, w_i^q, 1]^T$ is the manifold vector pointing towards the qth reflection component, τ_i^q and α_i^q are the time delay of the reflection and attenuation due to absorption at surfaces of the room. Therefore, the problem of DOAs estimation of multi-sources is converted into the estimation of $[u_i, v_i, w_i]$ ($1\leqslant i\leqslant I$).

2.2. Inter-sensor data ratio model

The inter-sensor data ratios (ISDR) of the AVS are defined as [7]

$$r_{fo}(k,m)=\frac{X_f(k,m)}{X_o(k,m)},(f=u,v,w) \tag{10}$$

where r_{uo}, r_{vo}, r_{wo} are the ISDRs between *u*- and *o*-sensor, *v*- and *o*-sensor, *w*- and *o*-sensor respectively. Based on the Eqn. (2)-(9), for the *f*-sensor (*f*=*u*, *v*, *w*), the ISDR can be represented as

$$r_{fo}(k,m)=\frac{\sum_{i=1}^{I}H_{fi}(k)S_i(k,m)+N_f(k,m)}{\sum_{i=1}^{I}H_{oi}(k)S_i(k,m)+N_o(k,m)},(f=u,v,w) \tag{11}$$

If the time-frequency point **X**(*k*,*m*) is a TD-TFP, which is assumed to be dominated by the *i*th sources and the direct-path component is significantly larger than the reflection and noise components, **X**(*k*,*m*) can be appproximated by

$$\mathbf{X}(k,m)\approx\mathbf{H}_i^d(k)S_i(k,m)+\mathbf{N}(k,m) \tag{12}$$

where $\mathbf{H}_i^d(k)$ is the direct-path component defined in (7), then ISDRs can be transformed into

$$r_{fo}(k,m)\approx\frac{e^{-j\omega_k\tau_i}f_iS_i(k,m)+N_f(k,m)}{e^{-j\omega_k\tau_i}S_i(k,m)+N_o(k,m)}=\frac{f_i+\varepsilon_{n_fs_i}(k,m)}{1+\varepsilon_{n_os_i}(k,m)}=f_i+e_{fo}(k,m),(f=u,v,w) \tag{13}$$

where $\varepsilon_{n_fs_i}(k,m)=e^{j\omega_k\tau_i}N_f(k,m)/S_i(k,m)$, $\varepsilon_{n_os_i}(k,m)=e^{j\omega_k\tau_i}N_o(k,m)/S_i(k,m)$, and $e_{fo}(k,m)$ is the residual error caused by ambiet noise, reverberation and model mismatch.

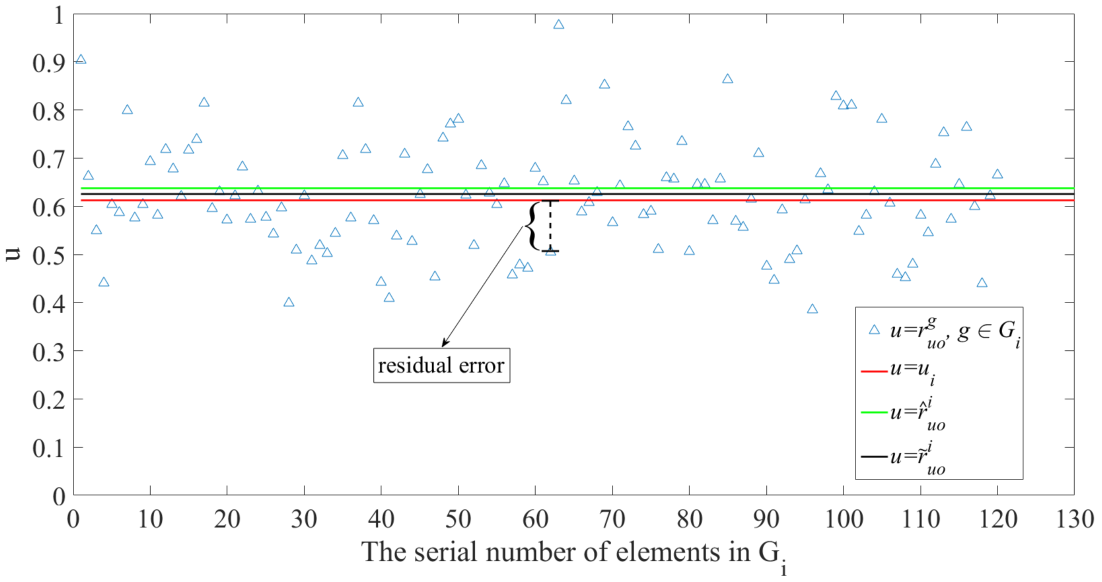


Fig. 1. Illustration of the ISDRs between the *u*-sensor and *o*-sensor, where '△' is the ratio r^g_{uo} in the *i*th ISDR-cluster, the red line is the true ratio u_i, the green line is the average of r^g_{uo} ($g \in G_i$), and the black line is the weighted average of r^g_{uo} ($g \in G_i$).

2.3. ISDRs clustering based DOA estimation

From Eqn. (13), the ISDRs $r_{uo}(k,m)$, $r_{vo}(k,m)$ and $r_{wo}(k,m)$ can be viewed as random variables in TF domains with the mean of u_i, v_i, and w_i respectively [7]. It is noted that the residual error $e_{fo}(k,m)$ is small for the TD-TFP, while large for the ID-TFP. To accurately estimate $[u_i, v_i, w_i]$ ($1 \leqslant i \leqslant I$), it is crucial to extract reliable TD-TFPs for the calculation of ISDRs. Specifically, assuming there are J TD-TFPs that are associated with I sources, then the ISDRs $\{[r^j_{uo}, r^j_{vo}, r^j_{wo}]_{1 \le j \le J}\}$ can be obtained and clustered into I classes where each represents one source. To illustrate this, we take the ISDRs between the *u*-sensor and *o*-sensor as an example, and plot the ratios r^g_{uo} ($g \in G_i$) in Fig 1, where G_i is the index set of the elements in the *i*th ISDR-cluster. As shown in Fig 1, the ratios

r_{uo}^{g} fluctuate up and down around the true ratio u_i (red line), thus it is a good choice to select the average (green line) of r_{uo}^{g} to approximate u_i for DOA estimation. Based on Eqn. (9), the centers of each ISDR-cluster can be calculated by taking the average of the points within the cluster and used for DOA estimation as follows:

$$\hat{r}_{fo}^{i} = \frac{\sum_{g \in G_i} r_{fo}^{g}}{|G_i|}, (i = 1, 2, ..., I, f = u, v, w) \tag{14}$$

$$\hat{\theta}_i = \cos^{-1} \hat{r}_{wo}^{i}, \quad \hat{\varphi}_i = \tan^{-1}(\hat{r}_{vo}^{i} / \hat{r}_{uo}^{i}), \quad (i = 1, 2, ..., I) \tag{15}$$

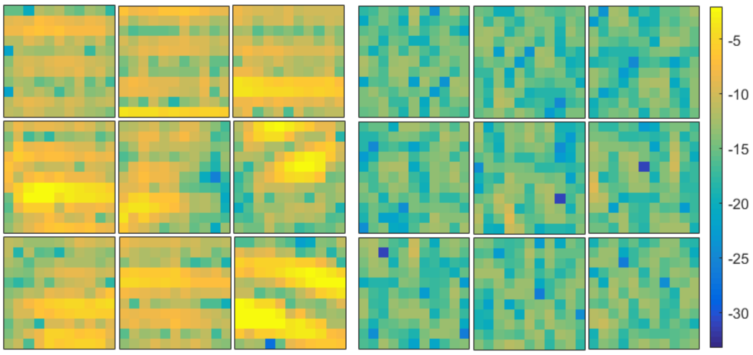
where $\{\hat{r}_{uo}^{i}, \hat{r}_{vo}^{i}, \hat{r}_{wo}^{i}\}$ is the center of the *i*th ISDR-cluster, and | • | denotes the number of elements in the set. $\hat{\theta}_i$ and $\hat{\varphi}_i$ are the estimated elevation and azimuth for the *i*th source.

As we can see from (13) and Fig 1, the biases in DOA estimation by AVS-ISDR mainly come from the residual errors $\{e_{uo}(k,m), e_{vo}(k,m), e_{wo}(k,m)\}$, since the large residual errors increase the estimation errors of the centers of the clusters for DOA estimation. In an effort to overcome this problem, two strategies have been exploited:

1) The TD-TFPs with low residual errors (in terms of a pre-defined threshold) are identified and extracted.
2) The weighted centers of ISDR-clusters are used to replace the original centers (14) by assigning the ISDRs having large residual errors with small weights, and the ISDRs having small residual errors with large weights.

3. Our proposed DOA estimation methods

In this section, the proposed novel TLSB features, which show different patterns for TD-TFPs and ID-TFPs, are firstly presented. Then, we present the details of soft mask learning by DNN and DNN-SVM in a supervised manner to extract reliable TD-TFPs. Finally, our proposed robust DOA estimation methods, by using the weighted centers of the ISDR-clusters (WISDR), termed in short as AVS-WISDR-DNN and AVS-WISDR-DNN-SVM, are introduced.



(a) (b)

Fig. 2. LSBs of 11×11 size that are randomly selected from the spectrogram of the received signal at the *o*-sensor with SNR level being 5dB and reverberation time being 350ms: (a) Nine local spectrogram blocks of TD-TFPs (TD-TFP-LSBs); (b) Nine local spectrogram blocks of ID-TFPs (ID-TFP-LSBs)

3.1. Extraction of the tandem local spectrogram block

According to above discussions, here we use the log-power STFT, $Y_f(k,m)=10\log_{10}(\|X_f(k,m)\|)$ ($f=u,v,w,o$), where $\|\bullet\|$ denotes the Euclidean norm. Then the shape of LSB centered by the TFP (k, m) of the f-sensor is defined as

$$\mathrm{LSB}_{(k,m)}^{f}=Y_f(k\pm b,m\pm c),(b=1,2,...,B,c=1,2,...,C) \tag{16}$$

where B and C are the row and column offset respectively, which are found empirically in our experiments.

To give some insights, an example is given here to show the patterns of LSBs, where the SNR level of Gaussian noise is set at 5dB and reverberation time at 350ms, the room size is 6m×6m×4m, the AVS is located at [3m, 3m, 1.3m], and two speech sources are placed 1.7m away from the sensor with DOA at (60°, -45°) and (80°, 120°) respectively. Then the spectrogram is obtained by taking the log-power STFT on the received signal of the AVS. The offsets *B* and *C* are all set to be 5 (the size of LSB is 11×11). Taking the LSBs of the *o*-sensor as an example, TD-TFP-LSBs and ID-TFP-LSBs are shown in Fig. 2 (a) and (b) respectively. From Fig. 2 (a) and (b), we can observe the following properties: 1) most TFPs in TD-TFP-LSBs have relatively high energy; 2) those TFPs in TD-TFP-LSBs with high energy constitute parallel “stripes”; 3) TD-TFP-LSBs contain more TD-TFPs. It is noted that similar patterns can be observed at other sensors (*u*-, *v*-, *w*-sensor) and in other noisy and reverberant enviroments.

Above observations motivate us to use the LSB as a cue to estimate the TF mask. Based on the structure of AVS, we propose to make use of the LSBs from all the 4 channels of the AVS, as illustrated in Fig. 3, where LSBs centered by the same TFP are vectorized and cascaded to form a 484 (4×11×11)-dimension vector termed as tandem LSB (TLSB).

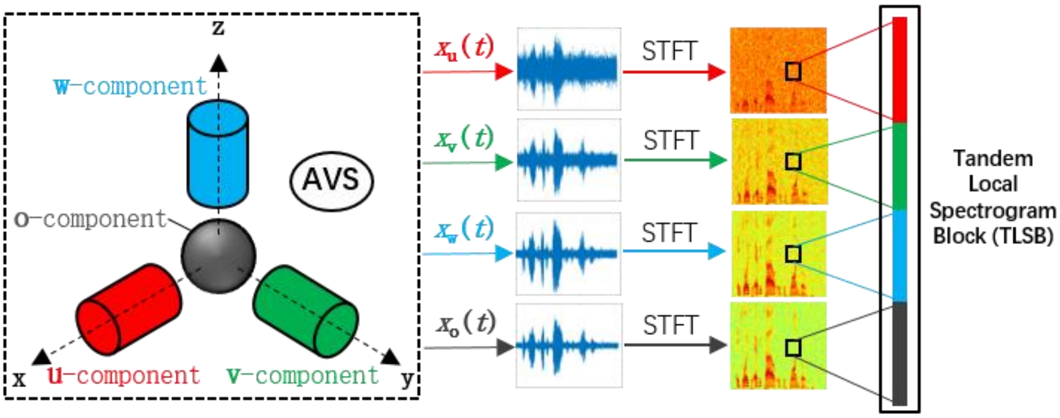


Fig. 3. Tandem local spectrogram block extraction

3.2. Design of the soft mask learner

A TD-TFP means the signal-to-noise ratio (SNR) of the TFP is larger than a local SNR where idea binary mask (IBM) has been suggested as a criterion as follows [17]

$$\mathrm{IBM}(k,m)=\begin{cases}1, & \mathit{if}\ \mathrm{SNR}(k,m)>10^{\eta}\\ 0, & \mathit{otherwise}\end{cases} \tag{17}$$

where η is a constant that is set to be 0.5 in this paper. Clearly, the IBM is 1 for TD-TFP and 0 for ID-TFP. It is noted that the IBM can only be used to determine whether the TFP is TD-TFP or not. To obtain a center that is closer to the true center as shown in Fig 1, the soft mask can be utilized. The soft mask, denoting the probability of a TFP being TD-TFP, can be used to determine the TD-TFPs and used as the weights to calculate the weighted centers of ISDR-clusters for better DOA estimation. Therefore, two soft mask learners have been proposed in the following subsections.

3.2.1. Soft mask learning by DNN

With the TLSB as input, we propose to employ the DNN to learn the soft TF mask for each TFP, which involves the training phase and test phase.

In the training phase, we create a training dataset of TLSBs that are extracted from the spectrograms of an AVS in different noisy and reverberant environments (details are given in Section 4), and the IBM of each TFP is used as the ground truth. With the training dataset $\{(\mathrm{TLSB}_d, l_d), d=1, 2, ..., D\}$, where D is the number of TLSB samples and l_d is the label (IBM) corresponding to the dth TLSB, the DNN is firstly pre-trained via a deep generative model of TLSBs by a stack of multiple restricted Boltzmann machines (RBMs) in an unsupervised fashion by using the contrastive divergence (CD) algorithm [23]. Then following the learning rate annealing and early stopping strategies used in the BP process [16], the DNN is fine-tuned using a stochastic gradient descent (SGD) algorithm by maximizing the cross-entropy between the true IBM and the predicted probability.

In the test phase, with the test TLSB at (k, m), the trained DNN is used to generate the soft mask (i.e. a posterior probability, which is the output of DNN) for the TFP as

$$p = P(\text{IBM=1}|\text{TLSB}(k,m)) \tag{18}$$

Then any TFP with the soft mask larger than a predefined value (set to be 0.9 empirically) is taken as a TD-TFP, which is used for DOA estimation.

3.2.2. Soft mask learning by DNN-SVM

With the well-trained DNN, in a generative manner, the last hidden layer representations (LHLR) of DNN can be obtained by using the TLSB as the input

$$\text{LHLR}_d = \Gamma(\text{TLSB}_d), (d = 1, 2, ..., D) \tag{19}$$

where $\Gamma(\cdot)$ is the mapping from the input to the last hidden layer of DNN. As discussed above, LHLRs have the linear separability in favour of the linear SVM. Thus, the new training dataset {(LHLR$_d$, l_d), d=1, 2, ..., D} can be obtained by Eqn. (19) and used for training a linear SVM, which has the following decision function [24]

$$L(\boldsymbol{z}) = \sum_{i=1}^{N_s} \omega_i \boldsymbol{\xi}_i^T \boldsymbol{z} + \omega_0 \tag{20}$$

where $\boldsymbol{z}$ is the test LHLR, $\boldsymbol{\xi}_i$ is the ith support vector associated with the weight ω_i, N_s is the total number of support vectors, and ω_0 is the bias term. It is noted that, when the decision function $L(\boldsymbol{z})$ is positive, the TFP corresponding to the test LHLR is judged to be a TD-TFP. Intuitively, when $L(\boldsymbol{z})$ has a larger positive value, the TFP is determined as a TD-TFP with a higher confidence, and vice versa. Therefore, similar to the relevance vector machine (RVM) [25] that has the identical function of SVM but provides probabilistic classification, the soft mask based on SVM can be defined by wrapping Eqn. (20) in a sigmoid squashing function

$$p = \frac{1}{1+e^{-L(\boldsymbol{z})}} \tag{21}$$

Then any TFP with the soft mask larger than 0.5 ($L(\boldsymbol{z})$ is positive) is taken as a TD-TFP.

3.3. DOA estimation via weighted ISDR centers

Following the ISDR model presented in [7], we propose a weighted ISDR (WISDR) model for DOA estimation. Specifically, take the J TD-TFPs determined by DNN as an example, assume the corresponding soft masks are $\{p_1, p_2, \ldots, p_J\}$. Then the ISDRs $\{[r_{uo}^j, r_{vo}^j, r_{wo}^j]_{1\le j\le J}\}$ can be calculated by Eqn. (10) and clustered into I classes by using the kernel density estimation (KDE) as used in [7]. The soft mask represents the probability of the TFP being considered as a TD-TFP, and as a result, it becomes useful for estimating the centers of the clusters. As shown in Fig 1, the center $\hat{r}_{uo}^i$ of the ratio r_{uo}^g ($g \in G_i$) in the ith ISDR-cluster is severely impacted by the r_{uo}^g with high residual errors. By assigning each r_{uo}^g with the corresponding soft mask as the weight, the weighted center $\tilde{r}_{uo}^i$ of r_{uo}^g ($g \in G_i$) is able to approximate the true ratio u_i more closely as compared with the center $\hat{r}_{uo}^i$. Thus, different from Eqn. (14), we take the weighted average as the center of the ith cluster as follows

$$\tilde{r}_{fo}^i = \frac{\sum_{g\in G_i} p_g r_{fo}^g}{\sum_g p_g}, (i = 1, 2, \ldots, I,\ f = u, v, w) \tag{22}$$

Similiar to (15), by replacing the original centers (14) with the weighted centers (22), the DOA can be estimated by

$$\tilde{\theta}_i = \cos^{-1} \tilde{r}_{wo}^i, \quad \tilde{\varphi}_i = \tan^{-1}(\tilde{r}_{vo}^i / \tilde{r}_{uo}^i), \quad (i = 1, 2, \ldots, I) \tag{23}$$

To distinguish from the baseline AVS-ISDR algorithm, we term the proposed algorithms in short as AVS-WISDR-DNN and AVS-WISDR-DNN-SVM respectively，which are summarized in Tables 1 and 2.

Table 1. Summary of our proposed AVS-WISDR-DNN algorithm

Algorithm 1. AVS-WISDR-DNN

```
1. DNN training:
    1) Construct the training dataset {(TLSB$_d$, $l_d$), $d$=1, 2, ..., $D$} by extracting the TLSB feature and corresponding IBM from the spectrograms of an AVS in different noisy and reverberant environments;
    2) Train the DNN with the TLSBs and corresponing IBMs as training pairs;
    3) Save the DNN model.
2. DOA estimation stage:
    1) Transform the received signal of the AVS to the spectrograms;
    2) Extract the TLSBs from the spectrograms as the input of DNN;
    3) Compute the soft masks by using TLSBs as the input of DNN (18), and determine the TD-TFPs with values of the soft masks larger than a predefined threshold (e.g., 0.9);
    4) Compute and cluster the ISDRs (10) of TD-TFPs into $I$ classes by KDE;
    5) Perform DOA estimation (23) on the weighted centers of ISDR-clusters (22).
```

Table 2. Summary of our proposed AVS-WISDR-DNN-SVM algorithm

```
Algorithm 2. AVS-WISDR-DNN-SVM
1. DNN-SVM training:
    1) Construct the training dataset {(TLSB$_d$, $l_d$), $d$=1, 2, ..., $D$} by extracting the TLSB feature and corresponding IBM from the spectrograms of an AVS in different noisy and reverberant environments;
    2) Train the DNN with the TLSBs and corresponing IBMs as training pairs;
    3) Extract the LHLRs with the well-trained DNN as shown in (19);
    4) Train the SVM with the LHLRs and corresponing IBMs as training pairs
    5) Save the DNN and SVM models.
2. DOA estimation stage:
    1) Transform the received signal of the AVS to the spectrograms;
    2) Extract the TLSBs from the spectrograms as the input of DNN;
    3) Extract the LHLRs with the well-trained DNN as shown in (19)
    4) Compute the soft masks by using LHLRs as the input of SVM (20) (21), and determine the TD-TFPs with values of soft masks larger than a predefined threshold (e.g., 0.5);
    5) Compute and cluster the ISDRs (10) of TD-TFPs into $I$ classes by KDE;
    6) Perform DOA estimation (23) on the weighted centers of ISDR-clusters (22).
```

Table 3. Configurations used for TLSB generation

Speech	50 randomly selected sentences from TIMIT [23]
DOA (°)	θ and φ randomly sampled from 0~180 and 0~360
SNR (dB)	-5 to 20 with 5 step
T_{60} (s)	0.15 to 0.75 with 0.1 step
Room size (m)	4×5×3 (small), 8×10×3.5 (medium), 15×18×4 (large)
Position of AVS	in the center of the room with the height to be 1.5m
Distance (m)	near (1) and far (3, 6, 9 for small, medium, large)

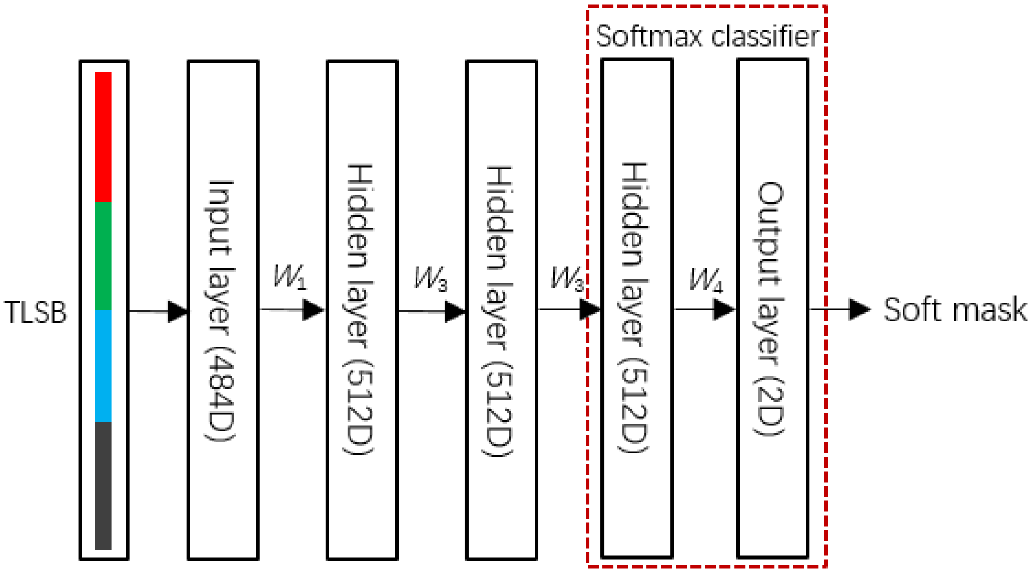


Fig. 4. The architecture of DNN used in our work

4. Experiments and analysis

4.1. Experimental settings

To create the dataset for training the DNN, the received signal **x**(*t*) of the AVS is generated according to Eqn. (1) where the room impulse responses **h**(*t*) are simulated following the image method proposed in [26], and **n**(*t*) is of Gaussian distribution. To obtain TLSBs in a variety of conditions, we simulate **x**(*t*) with different DOAs, room size, source to AVS distances, noise and reverberation levels, where the detailed configuration is summarized in Table 3. In each configuration, the elevation and azimuth are randomly sampled from [0°, 180°] and [0°, 360°) respectively. We simulate 3 types of room size: small (4m×5m×3m), medium (8m×10m×3.5m), and large (15m×18m×4m). In each room, the AVS is all placed in the center with the height of 1.5m. 50 sentences randomly selected from the TIMIT corpus [27] are used as the original speech sources, and each sentence is repeatedly used for different simulation configurations. The signals are sampled at 8kHz. The Hamming window of 256 samples is used to compute the spectrograms $Y_f(k,m)$ ($f=u,v,w,o$), with a 50% overlap between the neighbouring windows. To create a proper dataset, for the spectrograms obtained in each configuration, we extract TLSBs which can be divided into 3 parts:

1) TLSBs of TD-TFPs are all extracted and preserved, and the label (IBM) is set to be 1.
2) TLSBs of those ID-TFPs that lie in the LSBs of TD-TFPs are extracted and preserved, and the label is set to be 0.
3) By dividing the spectrogram into LSBs of size 11×11 without overlap across time frames and frequency bins, TLSBs of ID-TFPs are extracted and preserved, and the label is set to be 0.

Totally 7 million training samples are obtained, where 5 million training samples $\{(\mathrm{TLSB}_d, l_d), d=1, 2, ..., 5\times10^6\}$ are randomly selected to train the DNN, as we find the DNN has better performance with a large dataset and the performance is almost saturated with 5 million training samples. It is noted that the training dataset is generated under one-source condition, since the TLSBs under multi-source conditions have similar patterns.

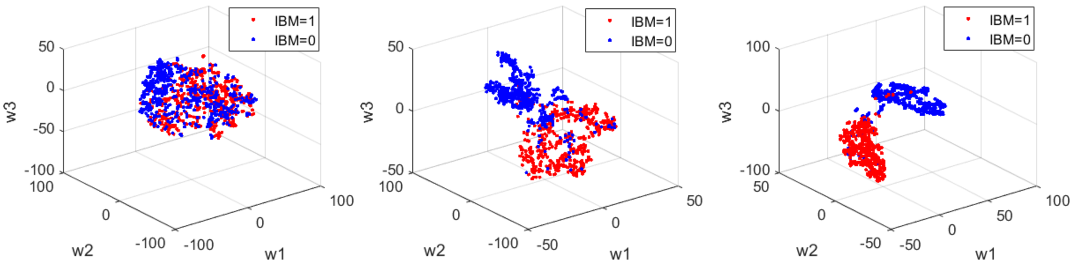
As for DNN, the architecture we adopted is demonstrated in Fig. 4, where the DNN contains one input layer (484-dimension, the block shape is the same as that in Section 4.1), three hidden layers with 512 units per layer and one output layer (2-dimension), and the last two layers constitute a softmax classifier. It is noted that the number of hidden layers of the DNN is determined with the cross-validation experiments by setting it as 2, 3, 4 and 5, where the DNN with 3 hidden layers gives the best performance in terms of the cross-validation classification accuracy. As a result, we choose the DNN with 3 hidden layers in our experiments. When the DNN is well-trained with the created dataset {(TLSB$_d$, l_d), d=1, 2, ..., D}, the corresponding LHLR dataset {(LHLR$_d$, l_d), d=1, 2, ..., D} can be obtained. As for the linear SVM, we randomly select 10^4 LHLR samples from the LHLR dataset, and use the default settings in the LIBSVM [28] package to train a linear SVM. In the test phase, the unused utterances selected from the TIMIT database are used as speech sources, the room size and the location of AVS are set to be 6m×6m×4m and [3m, 3m, 1.3m], and distances between the AVS and sources are all set to be 1.7m. The AVS-ISDR method [7] and the method by *Wu et al.* [9] (here termed as AVS-LRSS) are taken as baselines, where the settings of AVS-LRSS are the same as [9]. The root mean squared error (RMSE) is used as the performance metric

$$\mathrm{RMSE} = 0.5\sqrt{\frac{\sum_{l=1}^{L}\sum_{i=1}^{I}\left(\left(\theta_{il}-\theta_i\right)^2+\left(\varphi_{il}-\varphi_i\right)^2\right)}{I\times L}} \tag{24}$$

where L is the total number of trials, θ_{il} and φ_{il} are the estimation of θ_i and φ_i in the lth trial respectively.

4.2. Experimental results

4.2.1. Visualization of hidden layer representations of DNN



(a) (b) (c)

Fig. 5. 3-dimensional projection (w1, w2, w3) of hidden layer representations (HLR): (a) First HLRs; (b) Second HLRs; (c) Third HLRs (LHLRs)

To illustrate the distribution of learned hidden layer representations (HLR) via DNN, Fig. 5 shows the 3-dimensional projection of representations of 3 hidden layers of DNN. The projection is achieved by the t-SNE algorithm [29] and 10^3 TLSB samples are randomly selected from $\{(\mathrm{TLSB}_d, l_d), d=1, 2, ..., D\}$. In a generative manner, the first, second and third HLRs can be obtained via DNN with the TLSB samples as input. From Fig. 5, it can be observed that the HLRs become more separable as the depth of hidden layers increases, and the third HLRs, namely LHLRs, provide the best capability to discriminate the most TD-TFPs (IBM=1) and ID-TFPs (IBM=0). These results demonstrate that DNN is able to extract the LHLR features from the raw TLSB features which help to distinguish whether the TFP is a TD-TFP or ID-TFP.

4.2.2. Performance comparison for TD-TFPs extraction

Table 4. Average F1 scores versus reverberation time T_{60}, with SNR = 5dB

method	Reverberation time T_{60}				
	0.15s	0.25s	0.35s	0.45s	0.55s
SinTrE	0.242	0.082	0.038	0.024	0.021
Coherence test	0.478	0.108	0.056	0.054	0.036
DNN	0.669	0.357	0.280	0.230	0.192
DNN-SVM	**0.701**	**0.382**	**0.313**	**0.258**	**0.225**

To verify the effectiveness of TLSB based DNN and DNN-SVM for extracting TD-TFPs, as compared with the existing SinTrE [12] and coherence test [9] method, we generate the test TLSB dataset that is synthesized under different reverberation levels with the SNR fixed at 5dB, where the F1 score is used

$$\mathrm{F1} = \frac{2P_r R_e}{P_r + R_e} \tag{25}$$

where P_r is the precision, which is the number of correctly predicted positive (IBM=1) results divided by the number of all predicted positive results, and R_e is the recall, which is the number of correctly predicted positive results divided by the number of all true positive results. Under each reverberant condition, 100 trials have been conducted and the average F1 score is used as the evaluation metric, and the results are shown in Table 4. From Table 4, we can see that, as expected, the average F1 scores of all methods decrease when the reverberation time T_{60} is increased, and our proposed methods have significant improvements over the SinTrE and the coherence test methods, where the DNN-SVM gives the best performance with the highest average F1 scores, since the SVM gives better classification performance than the softmax of DNN [30].

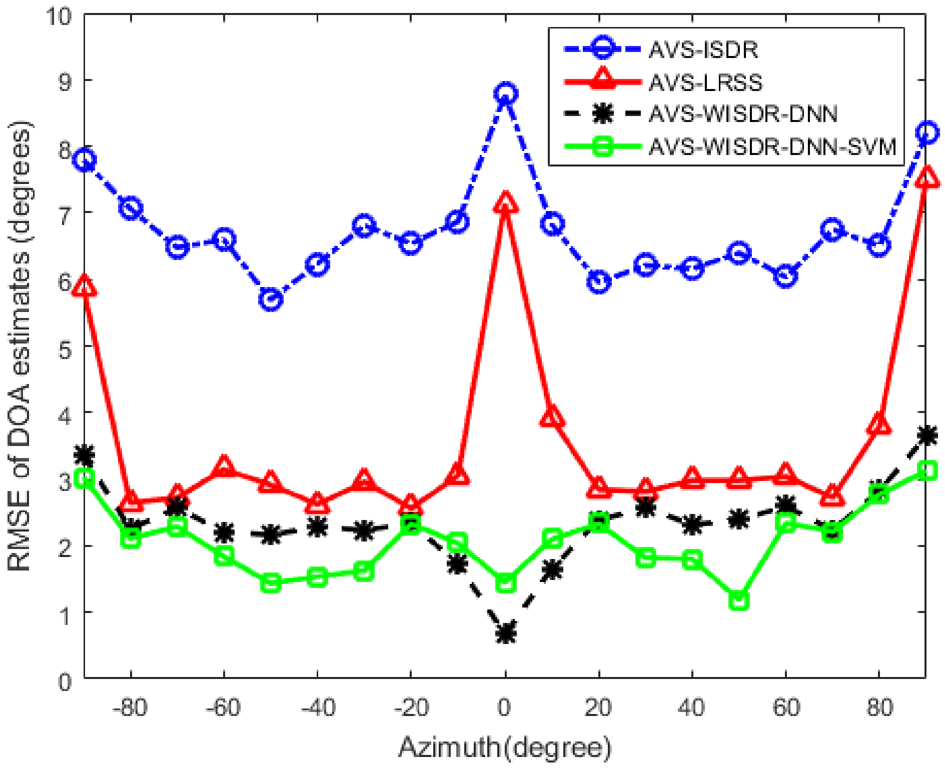


Fig. 6. RMSE versus different azimuth

4.2.3. DOA estimation accuracy versus azimuth

This experiment aims to evaluate the performance of DOA estimation versus different azimuth, where the elevation is fixed at 60°, the azimuth is varied from -90° to 90° with 10° step, and the SNR and T_{60} are fixed at 5dB and 0.35s, respectively. 100 trials have been repeated for each azimuth, and the results are shown in Fig. 6. It can be clearly seen that AVS-LRSS outperforms the AVS-ISDR for all azimuths, and both have the degraded performance when the azimuth is -90° , 0° and 90° . However, it is promising to see that AVS-WISDR-DNN and AVS-WISDR-DNN-SVM achieve better performance for all azimuths, which confirms the effectiveness of the TLSBs used for soft mask estimation based on DNN and DNN-SVM.

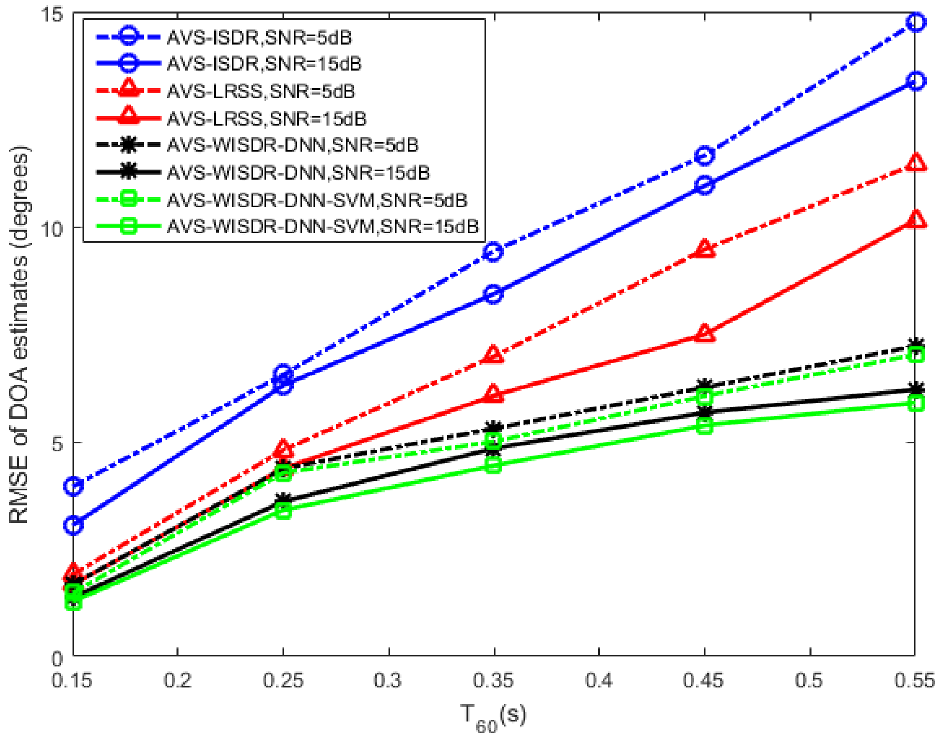
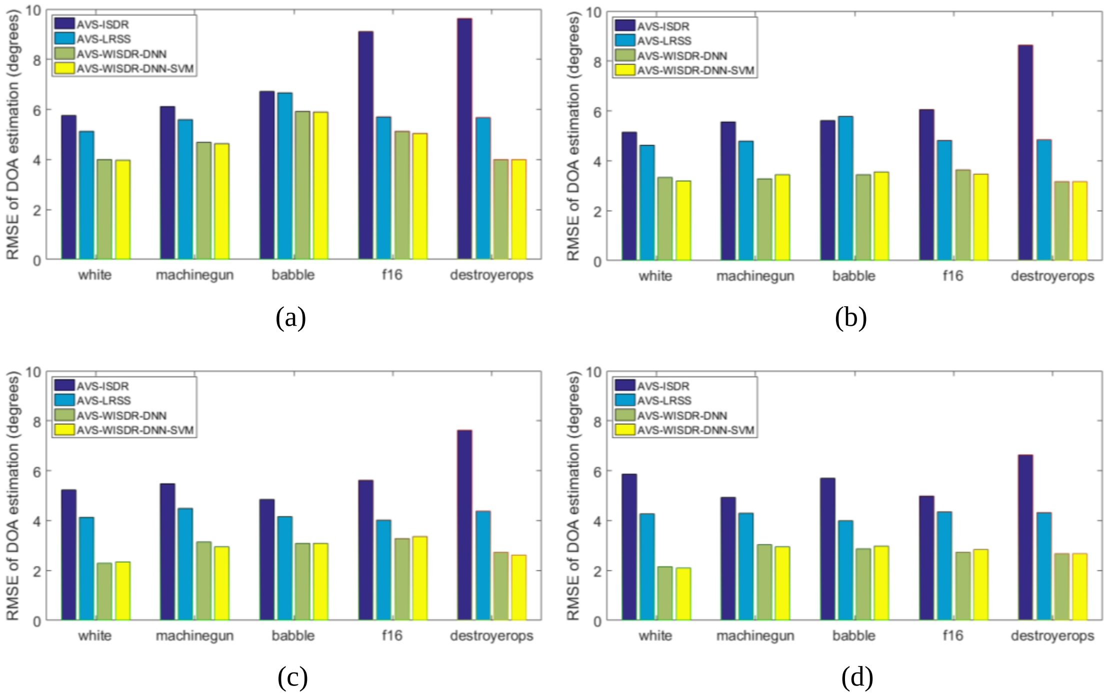


Fig. 7. RMSE versus different noise and reverberation levels with 2 sources located at (60°, -45°) and (80°, 120°)

4.2.4. DOA estimation of multi-sources

Fig. 7 shows the performance of DOA estimation of two sources located at (60°, -45°) and (80°, 120°) in different noisy and reverberant environments and L=100. It can be seen that the performance of all methods degrades with increasing levels of noise and reverberation, however our proposed methods still achieve better performance under all conditions, followed by AVS-LRSS and AVS-ISDR, which demonstrates the advantage of the proposed method in noisy and reverberant environments. In addition, since the DNN and DNN-SVM are trained by the dataset generated under different noisy and reverberant conditions, our methods are less sensitive and more robust to noise and reverberation.



(a)

(b)

(c)

(d)

Fig. 8. Performance evaluation under different noise conditions with different SNR levels: (a) 0dB; (b) 5dB; (c) 10dB; and (d) 15dB

4.2.5. Performance evaluation under different noise conditions

The DNN and DNN-SVM used in our work are aimed for predicting the type of the time-frequency points (TD-TFPs or ID-TFPs), which shows good performance under the white noise condition. To analyze the performance of our proposed algorithms under different noise conditions, we conducted experiments under 5 types of noise: white, machniegun, babble, f16 and destroyerops noise, which are seclected from the NOISEX-92 corpus [31]. We used one source, varied the SNR from 0dB to 15dB with 5dB interval and fixed T_{60} at 0.35s. Then, 100 trials have been repeated for each SNR level, and the DOA is randomly generated for each trial. The experimental results are shown in Fig. 8. From the results shown in Fig. 8, we have the following observations. 1) With the increase in SNR, our proposed methods give lower DOA RMSE results for each noise-type. 2) For a certain type of noise (f16 as an example), our proposed methods (green and yellow color bars) outperform the AVS-ISDR and AVS-LRSS algorithms. 3) For a certain SNR (0dB as an example), our proposed algorithms give the lowest DOA RMSE results for white noise while they give highest DOA RMSE results for babble noise. Such performance degradation is expected since the training data of DNN for our algorithms is only constructed by mixing the clean speech with white noise. 4) The mismatch of the noise condition between the test data and the training data leads to the performance degradation of our proposed algorithms. These observations also suggest that a large scale training dataset that encompasses many possible the combinations of speech and noise conditions, are helpful for enhancing the generality of our proposed DNN-based DOA estimation methods.

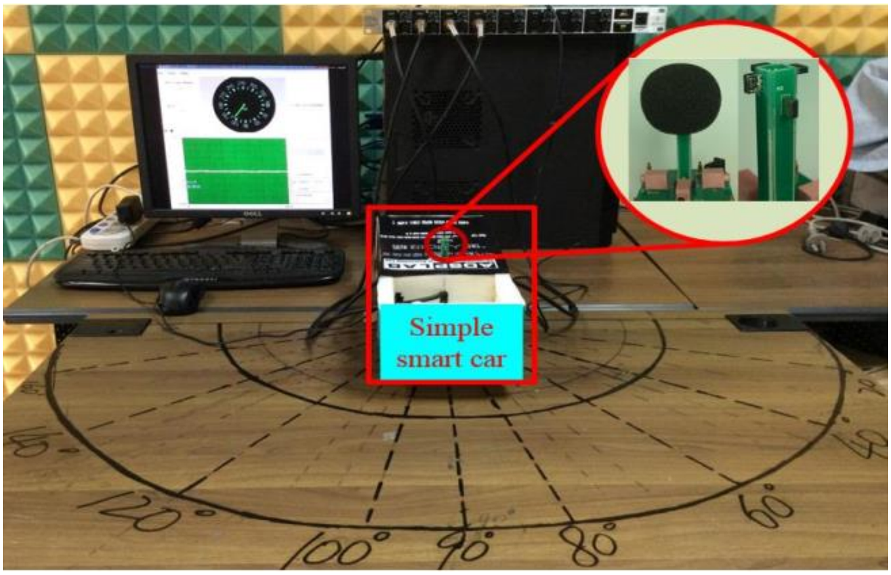


Fig. 9. Experimental device of DOA estimation system based on single AVS

Table 5. DOA estimation in a real scenario

DOA of speaker 1		[90°, 0°]				ART(s)
DOA of speaker 2		[90°, 45°]	[90°, 90°]	[90°, 135°]	[90°, 180°]	
RMSE (°)	AVS-ISDR	8.29	6.64	5.93	5.90	**0.486**
	AVS-LRSS	5.44	5.64	5.53	5.47	9.206
	AVS-WISDR-DNN	5.37	4.76	5.00	4.77	1.481
	AVS-WISDR-DNN-SVM	**5.12**	**4.48**	**4.73**	**4.41**	3.039

4.2.6. DOA estimation in a real scenario

Finally, we conduct an experiment in a real scenario using the AVS data capturing system developed by ADSPLAB as shown in Fig. 9, where a single AVS is placed on top of the smart car to capture the signals, and the room has a size of about 8.5m×3m×5m with uncontrolled reverberation and background noise from air conditioner and computer servers. Specifically, the experimental settings for the data recording are as follows: two speakers are used as the sources, the DOA of one speaker is fixed at [90^o, 0^o], and the elevation of the other speaker is fixed at 90^o, while the azimuth varies from 45^o to 180^o with a 45^o interval, which, therefore, results in 4 types of combinations. Besides, the distance between the speakers and the AVS is all set as 1m, and 10 trials have been conducted for each combination.

The RMSE results of DOA estimation are shown in Table 5. It can be seen that the proposed AVS-WISDR-DNN-SVM offers the best performance with the lowest RMSE for each source combination, followed by AVS-WISDR-DNN, AVS-LRSS and AVS-ISDR, which further demonstrates the effectiveness and superiority of our proposed methods. It is noted that the DNN and DNN-SVM are trained without performing any matching from the training dataset to the real test environment. Our proposed methods offer better performance due to the generalization ability of DNN and DNN-SVM to other unseen conditions. We will study the possibility of matching a training dataset to the given test environment for better DOA estimation in our future work.

Through quantitative analysis, by limiting the recorded data to be 3s for each trial, we also record the average running time (ART) of each algorithm in Table 5, where the AVS-ISDR has the smallest ART and AVS-LRSS has the largest ART. In essence, the DOA estimation of AVS-LRSS is based on the multiple signal classification (MUSIC) algorithm, which involves the MUSIC spectrum search to determine the elevation and azimuth simultaneously, and thus has a higher computational load. In contrast, the AVS-ISDR performs DOA estimation on the TD-TFPs with ISDRs that can be simply calculated with much lower complexity, which therefore has lower computational loads. Finally, our proposed methods provide a tradeoff between the DOA estimation accuracy and speed (running time), where the computational costs for TD-TFPs extraction by DNN and DNN-SVM are higher than those for the SinTre used in AVS-ISDR and the coherence test used in AVS-LRSS, however their TD-TFPs extraction accuracy is much higher, as shown in Table 4. In addition, due to the use of a number of support vectors, the computational cost of SVM tends to be higher than that of the softmax of DNN, as a result, the DNN-SVM is slower than DNN. Similiar to AVS-ISDR, our proposed methods are much faster than AVS-LRSS for DOA estimation.

5. Conclusion

In this paper, we have presented two soft mask learning methods for DOA estimation of multi-sources using DNN and DNN-SVM. The methods are based on the analysis of a previous method, i.e. AVS-ISDR algorithm, which we proposed earlier. The performance of this previous method largely depends on the reliable extraction of TD-TFPs that could be affected significantly by the increasing levels of noise and reverberation. A novel TLSB feature, that is shown to be different for TD-TFPs and ID-TFPs has been presented. By training a DNN with a large scale dataset that is composed by TLSB and corresponding IBM under various noisy and reverberant conditions, the soft masks can be generated via DNN to determine reliable TD-TFPs and used to calculate the weighted centers of ISDR-clusters for better DOA estimation. Due to the scalability and flexibility of DNN, the LHLR features learned from TLSBs are shown to be more linearly separable and thus used to train a linear SVM with a lower computational complexity. We note that the DNN-SVM can also be used to generate the soft masks by mapping the outputs of SVM to posterior probability for DOA estimation. The proposed AVS-WISDR-DNN and AVS-WISDR-DNN-SVM methods have shown significant improvements over AVS-ISDR and AVS-LRSS methods, where AVS-WISDR-DNN-SVM offers the best performance among these compared methods.

Our future work aims to exploit the influence of the size and shape of local spectrogram blocks on soft masking and design other DNN architecture to further improve the estimation performance of the soft masks. Besides, the selection of LHLR samples to further improve the training of a linear SVM is also worth studying.

6. Acknowledgement

This work was partially supported by NSFC and SSRP of China (grant numbers: 61271309 and CXZZ20140509093608290) and the EPSRC of the UK (grant numbers: EP/L027119/2 and EP/K014307/2).

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] J. Cao, J. Liu, J. Wang *et al.*, “Acoustic vector sensor: reviews and future perspectives,” *IET Signal Processing*, 2016.

[5] 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.

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

[7] 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 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4011-4015, 2013.

[8] S. Zhao, T. Saluev, and D. L. Jones, “Underdetermined direction of arrival estimation using acoustic vector sensor,” *Signal Processing,* vol. 100, pp. 160-168, 2014.

[9] K. Wu, V. Reju, and A. W. Khong, “Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 444-448, 2015.

[10] W. Zheng, Y. Zou, and C. Ritz, “Spectral mask estimation using deep neural networks for inter-sensor data ratio model based robust DOA estimation,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 325-329, 2015.

[11] Y. H. Jin, and Y. Zou, “Robust speaker DOA estimation with single AVS in bispectrum domain,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3196-3200, 2016.

[12] W. Zhang, and B. D. Rao, “A two microphone-based approach for source localization of multiple speech sources,” *IEEE transactions on audio, speech, and language processing,* vol. 18, no. 8, pp. 1913-1928, 2010.

[13] D. Levin, E. A. Habets, and S. Gannot, “On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields,” *The Journal of the Acoustical Society of America,* vol. 128, no. 4, pp. 1800-1811, 2010.

[14] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, “Kernel density estimation via diffusion,” *The Annals of Statistics,* vol. 38, no. 5, pp. 2916-2957, 2010.

[15] J. B. Allen, “How do humans process and recognize speech?,” *IEEE Transactions on Speech and Audio Processing,* vol. 2, no. 4, pp. 567-577, 1994.

[16] 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.

[17] N. Roman, and J. Woodruff, “Intelligibility of reverberant noisy speech with ideal binary masking,” *The Journal of the Acoustical Society of America,* vol. 130, no. 4, pp. 2153-2161, 2011.

[18] Y. Wang, and D. Wang, “Towards scaling up classification-based speech separation,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 7, pp. 1381-1390, 2013.

[19] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery,* vol. 2, no. 2, pp. 121-167, 1998.

[20] N. Yang, R. Muraleedharan, J. Kohl *et al.*, "Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion," in Spoken Language Technology Workshop (SLT), 2012 IEEE, pp. 455-460, 2012.

[21] 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.

[22] C. J. Taylor, "2012 Benjamin Franklin Medal in Computer and Cognitive Science presented to Vladimir Vapnik," *Journal of the Franklin Institute,* vol. 352, no. 7, pp. 2579-2584, 2015.

[23] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation,* vol. 14, no. 8, pp. 1771-1800, 2002.

[24] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152, 1992.

[25] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research,* vol. 1, no. Jun, pp. 211-244, 2001.

[26] 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.

[27] 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, 1988.

[28] C.-C. Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, no. 3, pp. 27, 2011.

[29] L. v. d. Maaten, and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research,* vol. 9, no. Nov, pp. 2579-2605, 2008.

[30] Y. Tang, "Deep learning using support vector machines," *CoRR, abs/1306.0239,* vol. 2, 2013.

[31] A. Varga, and H. J. 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.