T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Speaker Separation in Multi-Channel Environment Using Deep Learning
著者(和文)	Liu Conggui, 井上 中順, 篠田浩一
Authors(English)	Conggui Liu, Nakamasa Inoue, Koichi Shinoda
出典(和文)	情報処理学会研究報告, vol. 115, no. 11, pp. 1-6
Citation(English)	Technical Reports of IPSJ SLP, vol. 115, no. 11, pp. 1-6
発行日 / Pub. date	2017, 2
権利情報 / Copyright	ここに掲載した著作物の利用に関する注意: 本著作物の著作権は（社）情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。 The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof.

Powered by T2R2 (Tokyo Institute Research Repository)

Speaker separation in multi-channel environment using deep learning

Conggui Liu[1,a)] Nakamasa Inoue[1] Koichi Shinoda [1]

Abstract: This paper addresses multi-channel speaker separation based on a deep delay-and-subtraction beamformer. Deep neural network(DNN) is first applied to estimate the delay time between speakers and microphones , and then speakers' speech is recovered from mixed signals by using a delay-and-subtraction algorithm. We evaluated our method by using simulated data made from WSJCAM0 database. The proposed method achieved high precision source localization, and about 62% relative improvement on word error rate (WER) over a delay-and-sum (DS) beamformer.

Keywords: speech separation, multiple speakers, multi-channel, deep neural network, beamforming

1. Introduction

Speaker separation aims to reconstruct individual speakers' speech from mixed speech in which more than one speaker is talking simultaneously. It can be used in many applications such as hearing aid and speech recognition in meetings. Speaker's speech can be differentiated by using linguistics and speakers' spatial information. Many methods have been proposed to address the problem of multi-talker speech separation with one channel input [1][2][3]. However, it is difficult to separate multiple speakers' speech from only one mixed signal, because this problem is ill-posed, which may have more than one solution. With microphones getting cheaper and cheaper, multi-channel speech separation methods [4][5][6] have become popular. They can utilize speakers' spatial information. In this paper, we focus on the speech separation with multi-channel inputs.

Many methods have been proposed for multi-channel speech separation using the statistical properties of signals and speakers' spatial clues in mixed speech, e.g., blind source separation techniques [7][8][9], beamforming techniques [12][13].

Blind speech separation aims to extract all sources from mixed signals. Its examples include independent component analysis-based (ICA) methods [7][8] and time-frequency masking-based methods [9]. In the ICA-based approach, a separation filter is estimated by using the ICA algorithm [10][11] and the high-order statistics of signals. Then, the filter is used to extract independent sources from mixed signals. The time-frequency masking approach applies a mask for each point in the time-frequency domain of mixed signals to select each speaker's speech signal.

However, beamforming techniques [12][13] can separate the source in the specific direction from mixed signals by using a spatial filter. In most of the beamforming approaches, a source localization algorithm is first applied to estimate the time difference of arrival (TDOA) of each speaker among microphones [14]. Then, the estimated TDOA and the statistical properties of signals can be used to estimate beamformer weights or a spatial filter. While this approach has shown good performance in terms of interference suppression, it has several weaknesses. For example, it is difficult to separate speakers' speech from signals when the number of active speakers changes over time. This problem can be solved by using all speakers' locations for each speaker's speech separation. Also, the TDOA of some speakers may be lost due to false detection generated from outliers. The false detections can be reduced by applying the deep neural network to the source localization, though it is not emphasized in the related literatures [15][16].

Deep neural network (DNN) has been successfully applied to localize only one speaker with multi-channel inputs [16][17][18], owing to its powerful generalization capabilities. A multi-layer perceptron neural network (MLP) is applied to classify the direction of arrival (DOA) in [16], while it fails when speaker' location is unknown. A complex-value beamforming weight vector is predicted by using the DNN in the frequency domain [17][18], but it is difficult to be applied in multiple speakers localization, when the size of the estimated weight vector is large. It has not widely been applied to localize multiple speakers, because this task becomes more difficult with the growth of the number of speakers.

In this paper, we propose a delay-and-subtraction beamformer by combining a delay-and-subtraction algorithm with DNN. The DNN is discriminatively trained to predict speakers' delay time between speakers and microphones. Then, the predicted delay time and mixed signals are used to recover all speakers' speech by using a delay-and-subtraction algorithm. Our approach can work even when the activation of speakers changes. We demonstrate its potential by evaluating it on simulated data sets made from WSJCAM0 database.

The rest of this paper is organized as follows. Related studies are reviewed in Section 2. Section 3 describes the details of

[1] Tokyo Institute of Technology, Tokyo,Japan
[a)] conggui@ks.cs.titech.ac.jp

© 2013 Information Processing Society of Japan

the proposed method. Section 4 shows and discusses experiment results. Section 5 concludes this paper.

2. Related studies

Speaker separation can be done by applying beamforming techniques to estimate a spatial filter. There are a lot of beamforming approaches [19][20][21][22]. A delay-and-sum beamformer [19] is the simplest one, in which mixed signals from all channels are realigned and then summed to generate separated speech. This beamformer is data-independent because only speakers' location information (e.g., TDOA) is applied to estimate the spatial filter. On the contrary, data-dependent beamformers attempt to use the speakers' location information and the statistical properties of the signals. In [20], a minimum mean square error (MMSE) beamformer is applied to estimate the spatial filter by minimizing the expectation of the squared error of signals. It has no explicit constraints about the source direction which may be useful for speech separation. A constraint on the source direction is applied in a minimum-variance distortionless response (MVDR) beamformer in [21]. Both the MMSE beamformer and the MVDR beamformer use the statistical properties of source signals. However, the source signals are unavailable in real environments. A minimum power distortionless response (MPDR) beamformer [22] is proposed to estimate the spatial filter by using minimum mean square output as its optimum criteria. It directly uses the statistical properties of the mixed speech for the beamformer optimization.

There are several difficulties in applying the above discussed techniques. For example, the number of active speakers may change over time in real environments. To handle real cases, a signal-to-interference ratio (SIR) beamformer [23] is proposed. In this method, speech activities are detected and then clustered for all speakers. Then, clustered speech is used to estimate the spatial filter by maximizing the signal-to-interference ratio. However, separated signals may have different distortions corresponding to different frequencies since there is no constraint on sources' directions. This may deteriorate the performance of speech recognition. We propose a delay-and-subtraction algorithm in which each speaker's speech can be recovered by canceling the speech from the other directions.

For multi-source localization, there are mainly two types of methods for estimating the TDOA of each source among microphones. One is clustering-based approach. In this approach, points of mixed signals in the time-frequency domain are first clustered for each source by using current TDOA estimation. Then, the clustered points are used to re-estimate the TDOA of sources. This approach is very sensitive to the initial TDOA. The other approach is based on the angular spectrum [24][25], in which peaks are corresponding to the TDOA of sources. In [24], a method with generalized cross-correlation with phase transform (GCC-PHAT) is proposed to generate the angular spectrum by using the covariance matrix of the mixed speech. In order to reduce false detection of TDOA, a non-linear function is used to emphasize the angular spectrum in the GCC-PHAT (GCC-NON) in [25].

In the above localization techniques, the TDOA of some speakers may be lost due to false detection generated from outliers. The false detections can be reduced by applying the deep neural network to the source localization, though it is not emphasized in the related literatures [16] [17]. DNN is also explored to localize multiple speakers. For example, the DNN is applied to classify multi-speaker locations by using the eigenvectors of the complex correlation matrix of signals [26]. However, the number of the DNN output nodes and training data patterns is large. In addition, there is a large gap between the performance of the data with known speakers' locations and the data with unknown speakers' locations. On the contrary, we apply DNN to localize multiple speakers by predicting the delay time between speakers and microphones.

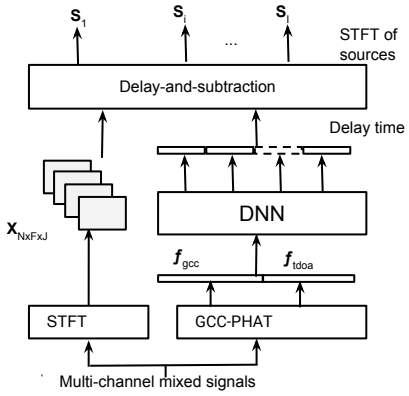

Fig. 1 The overview of proposed speech separation system.

3. Deep delay-and-subtraction beamformer

In this section, a DNN-based speaker separation method is discussed. The overview of the whole system is shown in Figure 1. The DNN is first applied to estimate the delay time between speakers and microphones. The estimated delay time and multi-channel inputs are then used to recover speakers' speech by applying a delay-and-subtraction algorithm. There are four parts described in this section : 1) mixing model, 2) input of deep neural network, 3) DNN structure, and 4) delay-and-subtraction (DSB) algorithm.

3.1 Mixing model

In a simple way, the mixed speech from each microphone can be expressed as the sum of delayed speakers' speech. Let I source signals be recorded by J microphones, here $J \geq I$. Then the mixed signal $\mathbf{x}_j$ in the j-th channel can be written as follows:

$$\mathbf{x}_j = \sum_{i \in I} \mathbf{s}_i(t - t_{ij}), \tag{1}$$

where $\mathbf{s}_i$ is the ith source, t is the time index, and t_{ij} is the delay time between the i-th source and the j-th microphone. Then, its short-time Fourier transform (STFT) can be described as:

$$\begin{aligned} X_j(\tau, f) &= \sum_{i \in I} S_i(\tau, f) H_{ij}(f) \\ &= \sum_{i \in I} S_i(\tau, f) e^{-j2\pi f t_{ij}}, \end{aligned} \tag{2}$$

where $H_{ij}(f)$ denotes a frequency response from the i-th source

© 2013 Information Processing Society of Japan

to the j-th microphone, and $S_i(\tau, f)$ is STFT of the i-th source $\mathbf{s}_i$ in a time-frequency point (τ, f). Since the transform function is only dependent of the frequency bin f and the delay time between microphones and speakers, speaker separation problem can be solved by estimating the delay time.

3.2 DNN input features

The DNN-based source localization aims to predict multiple speakers' positions. In order to achieve this, input features of the DNN should contain speakers' locations information and be easily used by the DNN.

The speakers' locations can be presented in several ways, such as the generalized cross-correlation (GCC) vectors [16][17] and the eigenvectors of the complex correlation matrix of mixed signals [18]. Here, we use GCC vectors as the input features of the DNN. Sources are assumed to be sparse in the time-frequency bins so that their locations can be estimated by using the spatial phase difference between two channels in the frequency domain. For random two channels, the covariance matrix of mixed speech $R(\tau, f) = X_1(\tau, f)X_2(\tau, f)^*$ is computed by using the STFT of mixed speech signals, $X_1(\tau, f)$ and $X_2(\tau, f)$. Then, the angular spectrum can be expressed as a function of the covariance matrix $R(\tau, f)$ as follows:

$$F_{\mathrm{gcc}}(\tau, f) = \frac{R(\tau, f)}{|R(\tau, f)|} \tag{3}$$

To obtain sufficient spatial features, the angular spectrum is transformed into GCC feature in the time domain. The GCC feature can be expressed as follows:

$$f_{\mathrm{gcc}}(\tau) = \Psi^{-1}(F_{\mathrm{gcc}}(\tau, f)), \tag{4}$$

where a function Ψ^{-1} denotes inverse short time Fourier transform. Since the silent parts of mixed speech signals including no speakers' location information are not helpful for the DNN training, the GCC feature is only extracted from the speech-active parts which are detected by using a voice activity detection algorithm.

After computation of the GCC features, it is worth describing how to calculate the dimension of DNN input features. In order to easily express this process, we take a circle array composed of 8 microphones with 20cm diameter, as an example. The sampling rate of speech is 16kHz. The GCC features are computed for each microphone pair. Since the maximum of sample delay is 10, only 21 samples of the GCC feature are selected. There are 28 possible pairs for 8 microphones. Then, the dimension of the selected GCC feature vector is summed to 588. In addition, the estimated TDOA by applying the GCC-PHAT method is concatenated to selected GCC features because the estimated TDOA also contains speakers' locations information.

3.3 DNN structure

The deep neural network aims to learn the capability of speaker localization. The DNN-based multi-speaker localization can be applied to predict speakers' location labels or a complex-value spatial weights. However, the number of the DNN output nodes increase quickly as the number of speakers grows. The delay time between speakers and microphones is a form of speakers' locations. Therefore, we apply the DNN to predict the delay time between speakers and microphones. Given a (GCC) feature vector $\mathbf{f}_{\mathrm{gcc}}(\tau)$ and a estimated TDOA vector $\mathbf{f}_{\mathrm{tdoa}}(\tau)$ for frame τ, the output vector $\mathbf{o}_l$ of the lth hidden layer ($1 \le l \le L$), where constant L denotes the number of hidden layers, can be expressed as a function of a input vector $\mathbf{o}_{l-1}$ as follows:

$$\mathbf{h}_l = \mathbf{W}_l\mathbf{o}_{l-1} + \mathbf{b}_l, \tag{5}$$

$$\mathbf{o}_l = f(\mathbf{h}_l). \tag{6}$$

If $l = 1$, $\mathbf{o}_0$ is the input feature vector of DNN. The variables $\mathbf{W}_l$ and $\mathbf{b}_l$ are the weight matrix and the bias vector between the $(l-1)$-th hidden layer and the l-th hidden layer respectively. The function $f(a) = (1 + e^{-a})^{-1}$ is a non-linear activation function. Finally, the delay time is predicted in the last layer as:

$$\mathbf{y}(\tau) = \mathbf{W}_L\mathbf{o}_L + \mathbf{b}_L. \tag{7}$$

The DNN is then trained directly by minimizing the distance between the estimated delay time vector $\mathbf{y}(\tau)$ and the ground truth $\mathbf{t}(\tau)$. The cost function is computed with the total number of frames N for all speakers I and all microphones J based on a mean square error criterion:

$$L_{mse} = \frac{1}{NIJ}\sum_{\tau}\|\mathbf{y}(\tau) - \mathbf{t}(\tau)\|^2, \tag{8}$$

where i, j and τ denote a speaker index, a microphone index and a frame index respectively. The operation $\|*\|^2$ denotes the 2-norm square (Euclidean norm square). Our DNN is a forward deep neural network for a regression problem. It has two hidden layers with 1024 units. The size of its input layer is the sum of the dimension of a selected GCC feature vector and a estimated TDOA vector. The size of its output layer is obtained by multiplying the number of speakers by the number of microphones, e.g. $8 * I$ for 8 microphones. Thus, our DNN can be easily trained for the multi-speaker localization.

3.4 Delay-and-subtraction algorithm

Our delay-and-subtraction (DSB) beamformer aims to estimate each speaker' speech by using all speakers' locations. After each speaker' delay time is predicted from DNN, each speaker's speech can be separated from mixed speech by applying a delay-and-subtraction algorithm. For each speaker, mixed speech signals from all channels are firstly realigned by using the estimated delay time. Then, one of other speakers can be cancelled by using a subtraction process for the realigned signals from two channels. Given the i_1-th speaker's delay time, $0 < i_1 \le I$, the mixed speech from the jth microphone is realigned as follows by multiplying by the conjugate of the frequency response from the i_1-th speaker to the j-th microphone:

$$\begin{aligned} X_j(\tau, f)H_{i_1j}(\tau, f)^* &= \sum_i S_i(\tau, f)H_{ij}(f)H_{i_1j}(f)^*, \\ &= S_{i_1}(\tau, f) + \sum_{i\neq i_1} S_i(\tau, f)H_{ij}(f)H_{i_1j}(f)^*. \end{aligned} \tag{9}$$

© 2013 Information Processing Society of Japan

Here, $(.)^*$ is a conjugate operator. The speaker' speech spectrum $S_{i_1}(\tau, f)$ can be cancelled by using a subtraction process for realigned signals from random two channels. One channel is selected from all microphones as a reference channel. The other is selected from remaining microphones. After this subtraction process, $J-1$ signals including $I-1$ speakers remain. All signals from $J-1$ channels are realigned again. Then, one of $I-2$ speakers can be cancelled by using a subtraction process for two channels randomly chosen from $J-1$ channels. Then, $J-2$ signals with $I-2$ speakers are obtained. Each subtraction process aims to cancel one speaker. When the above steps are repeated until when only one speaker exists, the speaker' speech is recovered. Other speakers' speech is reconstructed in the same way. In this paper, let the number of speakers I be equal to 2. For two microphones with channel index j_1 and j_2 respectively, the speech spectrums for two speakers' $S_1(\tau, f)$ and $S_2(\tau, f)$ are separated from mixed speech spectrum by subtracting realigned speech and multiplying by a filter term as follows:

$$S_1(\tau, f) = \frac{X_{j_2}(\tau, f)H_{2j_2}(f)^* - X_{j_1}(\tau, f)H_{2j_1}(f)^*}{H_{1j_2}(f)H_{2j_2}(f)^* - H_{1j_1}(f)H_{2j_1}(f)^*}, \quad (10)$$

$$S_2(\tau, f) = \frac{X_{j_2}(\tau, f)H_{1j_2}(f)^* - X_{j_1}(\tau, f)H_{1j_1}(f)^*}{H_{2j_2}(\tau, f)H_{1j_2}(f)^* - H_{2j_1}(f)H_{1j_1}(f)^*}. \quad (11)$$

The proposed delay-and-subtraction algorithm, strictly speaking, is a filter-and-subtraction method, because a filter is applied before each subtraction process. Each speaker' speech can be recovered only by using locations of other speakers. The recovered speech usually suffers from attenuation which may affect the quality of speech perception. In order to improve speech quality, a filter, which is consisted of speakers' transform functions, is applied to further enhance the recovered speech signals. If there are more than two speakers, mixed speech should be selected from more microphones. Speakers' speech can be obtained by iteratively cancelling other speakers' speech in the same way as above. A filter should be add in the each subtraction process.

4. Experiments

4.1 Conditions

Experiments are implemented on WSJCAM0 corpus [27] which is recorded by native speakers of British English. In this paper, only two speakers, a female and a male, are used and their order is fixed: the first speaker is female, the second speaker is male. Since DNN training needs the ground truth of training data, we generated 80 hours simulated data by randomly selecting clean speech from 7861 clean training sentences. Speaker locations are limited on a circle with 1 meter radius. Angle of all speakers is set to $[0, 360)$. When simulating the training data, speaker locations are also randomly chosen from 360 degrees for each sentence and the difference of two speakers' angle is fixed to be 90 degree.

For development data, two data sets are generated: 1) a data set with known speakers' locations and 2) a data set with unknown speakers' locations. The data set with known speakers' locations is synthesized by using clean speech from the development data with 368 clean utterances and known locations which are present in the training data. Both clean speech and speakers' locations are randomly chosen. The data set with unknown speakers' locations is synthesized by using a same process but with unknown speakers' locations which are not present in the training data. In the training step, the data set with known speakers' locations is applied for validation constraints. Test data is also simulated in a similar way as above simulated development data.

Table 1 Performance of source localization

Method	Precision (%)	Angle error (degree)
GCC-PHAT	98.0	1.29
GCC-NON	98.0	1.56
DNN	100.0	0.40

The performance of our method is evaluated in terms of source localization and speech recognition. For the source localization, the evaluation is usually done by computing three values (precision, recall and F-measure) through the estimated TDOA and its ground truth. Since the estimated number of speakers is assumed to be same as the true number of speakers, they should be equal to each other. Only the precision of the TDOA measuring how many sentences are estimated accurately is computed. In addition, angle (or direction of arrival) error measuring how far the estimated TDOA is from the ground truth is also used. The precision P is computed using the number N_c of correctly estimated TDOA and the number N of true TDOA :

$$P = \frac{N_c}{N}. \quad (12)$$

In this paper, a far-field assumption is used. Then, the TDOA denoting by ∇t can be computed using a following equation:

$$\nabla t = \frac{d\cos(\theta)}{c}. \quad (13)$$

Here, d is the distance between a speaker and a microphone in meter, θ is direction of arrival (DOA) and source speed is c (meter per second). Angle error E_a is measured by computing mean absolute error (MAE) for all speakers:

$$E_a = \frac{1}{I}\sum_i |\widehat{\theta_i} - \theta_i|, \quad (14)$$

where $\widehat{\theta_i}$ denotes the estimated direction of arrival for the ith speaker.

In order to evaluate speaker separation, three experiments are implemented in this section: 1) source localization, 2) speech recognition results of test data with known speakers' locations and 3) speech recognition results of test data with unknown speakers' locations.

4.2 Source localization

The delay times between speakers and microphones are predicted by applying the DNN. We first discuss the performance of each source localization method. We use the simulated development data with known speakers' locations for the evaluation. Both the GCC-PHAT method and the GCC-NON method are used as benchmarks.

All results shown in Table 1 are the mean value of all sentences. Our DNN-based method achieves 100% precision, while benchmarks obtain same precision, 98%. Although speakers' locations are not close to each other, the false detections of the speakers'

© 2013 Information Processing Society of Japan

Table 2 Comparing WER (%) for the proposed method and delay-and-sum beamformer.

	Dev.	Test
Clean	5.4	5.6
Idea+DS	54.3	59.1
GCC-PHAT+DS	71.1	75.3
GCC-NON+DS	69.7	72.9
GCC-PHAT+DSB	11.4	10.7
GCC-NON+DSB	10.8	10.4
DNN+DSB	9.18	9.7

locations happen in the conventional methods. The DNN-based method improves the precision of localization. The angle error (its unit is degree) is the error of direction of arrival (DOA). The DNN-based method obtains the smaller error of the DOA than benchmarks. Our DNN-based method can reduce the false detections of the speakers' locations from the benchmarks.

4.3 Speech recognition results when speakers' locations are known

After the evaluation of source localization, we evaluate the speech separation using the word error rate (WER). In this subsection, we evaluate the performance of speech separation for the test data with known speakers' locations by comparing with benchmark methods. The speech recognition system is built on a triphone system with the Linear Discriminant Analysis (LDA) and the Maximum Likelihood Linear Transform (MLLT). The acoustic model is based on the Deep Neural Network Hidden Markov Models (DNN-HMMs) and trained by using the WSJ-CAM0 database.

After multiple speakers are localized by applying traditional methods, it is difficult for them to mark the estimated locations for the speakers. In order to compare with our method, this problem is ignored and the estimated speakers' locations are realigned by using the true speakers' locations. Since there is no such problem in the our proposed method, we don't do the same process. To evaluate the performance of the beamformer, the delay-and-sum (DS) beamformer is used as our benchmark.

Speech recognition results are shown in Table 2. The proposed method (DNN+DSB) can obtain low WER. With the same source localization methods (GCC-PHAT and GCC-NON), our method reduces WER by about 60% for the development data, and about 63% for the test data compared with the DS beamformer. For the ideal case where speakers' locations are the ground truths, the DS beamformer still obtains a high WER. To recover each speaker's speech, it is difficult for the DS beamformer to cancel the other speakers' speech, because only one speaker's location is used. On the contrary, our method can cancel them by using all speakers' locations. Compared with different source localization techniques, DNN-based method improves WER about 2.3% over that with GCC-PATH, about 1.7% over that with GCC-NON for the development data. For the test data, our method also achieves better results than source localization benchmarks. Compared with the DSB algorithm, the DNN-based source localization is slightly improve the performance of the whole speech separation system. We don't discuss speaker order problem which tends to be present in the traditional localization methods, but there is no such problem in our DNN-based method. Since the proposed beamformer aims to recover speech by using speakers' locations, it needs the high-precision multi-talker localization.

Table 3 Comparing WER (%) for the case where each speaker's location is known and that for the case where each speaker' location is unknown

Speakers' locations	Dev	Test
Known	9.2	9.7
Unknown	9.3	10.3

4.4 Speech recognition results when speakers' locations are unknown

In the above subsection, the performance of speech recognition is evaluated for the test data with known speakers' locations. In this subsection, we further evaluate the performance of our method for the test data with unknown speakers' locations. We use the same speech recognition system as that in the above subsection.

Compared with the case where speakers' locations are known, the performance of our method is slight worse for the test data with unknown speakers' locations. The results are shown in Table 3. Thus, the proposed method can work even for unknown speakers' locations. In order to achieve this, the test data should be recorded by the microphone array with the same structure as that used in the training data. In addition, the DNN should be trained with a large amount of data.

5. Conclusion and Future work

We have explored the deep neural network for multi-talker speech separation with multi-channel inputs by applying a beamforming technique. Its performance is evaluated for source localization and speech recognition for a known case and an unknown case. Compared with the delay-and-sum beamformer, our delay-and-subtraction beamformer works much better. This improvement mainly comes from the delay-and-subtraction algorithm, which aims to separate each speaker's speech by using all speakers' locations and works even when the activation of speakers changes over time. The performance is further improved by using a DNN-based source localization which can reduce the location detection error. Experiments also demonstrate it works even for the situation when speakers' locations in the test data are different from that in the training data.

However, a noise-free situation including two speakers is tested in our experiments. It may be difficult for speaker separation in a real environment, e.g., where more speakers are present in a noisy environment. In such a environment, the clues of speakers' locations are unclear from input features so that it is difficult for the DNN to localize speakers. In addition, speakers' location ambiguities may happen with growth of the number of speakers. In future, this research will be extended to cover these problems.

References

[1] M.N.Schmidt, and R.K. Olsson: Single-channel speech separation using sparse non-negative matrix factorization, *In Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*,(2006).

[2] M.H. Radfar, and R.M. Dansereau: Single-channel speech separation using soft mask filtering, *IEEE Transactions on Audio, Speech, and Language Processing*, pp: 2299-2310,(2007).

[3] E.M. Grais, and H. Erdogan: Single channel speech music separation

© 2013 Information Processing Society of Japan

using nonnegative matrix factorization and spectral masks, *In 2011 17th International Conference on Digital Signal Processing (DSP)*, pp. 1-6, (2011).

[4] E. Weinstein, M. Feder, and A.V. Oppenheim: Multi-channel signal separation by decorrelation, *IEEE transactions on Speech and Audio Processing*, pp: 405-413,(1993).

[5] M. Handa, T. Nagai, and A. Kurematsu: Frequency domain multi-channel speech separation and its applications, *In Acoustics, Speech, and Signal Processing(ICASSP)*, pp. 2761-2764, (2001).

[6] M.J. Reyes-Gomez, B. Raj, and D. R. W. Ellis: Multi-channel source separation by factorial hmms, *In Acoustics, Speech, and Signal Processing(ICASSP)*, (2003).

[7] R.H. Lambert: Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures, *PhD diss., University of Southern California*, (1996).

[8] P. Smaragdis: Blind separation of convolved mixtures in the frequency domain, *Neurocomputing 22*, pp: 21-34,(1998).

[9] R. M. Toroghi, F. Faubel, D. Klakow: Multi-channel speech separation with soft time-frequency masking, *In Proc. SAPA-SCALE Conference, Portland, Oregon*, (2012).

[10] A. Hyvrinen, and E. Oja: Independent component analysis: algorithms and applications, *Neural networks 13*, pp: 411-430,(2000).

[11] H. Oja, and K. Nordhausen: Independent component analysis, *Encyclopedia of Environmetrics* (2001).

[12] M.Z. Ikram, and D.R. Morgan: A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation, *In Acoustics, Speech, and Signal Processing (ICASSP)*, (2002).

[13] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari: Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming, *In Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop*, pp. 445-454, (2002).

[14] C. Blandin, A. Ozerov, and E. Vincent: Multi-source TDOA estimation in reverberant audio using angular spectra and clustering, *Signal Processing*, pp: 1950-1960,(2012).

[15] R. Takeda, and K. Komatani: Sound source localization based on deep neural networks with directional activate function exploiting phase information, *In Acoustics, Speech and Signal Processing (ICASSP)*, pp. 405-409, (2016).

[16] X. Xiao, S. Zhao, X. Zhong, D.L. Jones, E.S. Chng, and H. Li: A learning-based approach to direction of arrival estimation in noisy and reverberant environments, *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2814-2818, (2015).

[17] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, D. Yu: Deep beamforming networks for multi-channel speech recognition, *In Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5745-5749 (2016).

[18] X. Xiao, S. Watanabe, E.S. Chng, H. Li: Beamforming Networks Using Spatial Covariance Features for Far-field Speech Recognition, *in APSIPA* (2016).

[19] B. Rafaely: Microphone array signal processing, *The Journal of the Acoustical Society of America 125*, pp: 4097-4098,(2009).

[20] Van Veen, D. Barry, and K.M. Buckley: Beamforming: A versatile approach to spatial filtering, *IEEE assp magazine 5*, pp: 4-24,(1988).

[21] M. Brandstein, and D. Ward: Microphone arrays: signal processing techniques and applications, *Springer Science and Business Media*, (2013).

[22] H.L. Van Trees: Detection, estimation, and modulation theory, optimum array processing, *John Wiley and Sons*, (2004).

[23] S. Araki, H. Sawada, and S. Makino: Blind speech separation in a meeting situation with maximum SNR beamformers, *In Acoustics, Speech and Signal Processing(ICASSP)*, pp. I-41, (2007).

[24] C. Knapp, G. Carter: The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 320-327 (1976).

[25] B. Loesch, B. Yang: Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions, *In International Conference on Latent Variable Analysis and Signal Separation, Springer Berlin Heidelberg*, pp. 41-48 (2010).

[26] R. Takeda and K. Komatani: Discriminative Multiple Sound Source Localization based on Deep Neural Networks using Independent Location Model, *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, pp.603-609, (2016).

[27] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals: WSJ-CAMO: a British English speech corpus for large vocabulary continuous speech recognition, *In Acoustics, Speech, and Signal Processing(ICASSP)*, pp. 81-84,(1995).

© 2013 Information Processing Society of Japan