論文 / 著書情報
Article / Book Information

Title	A Unified Network for Multi-Speaker Speech Recognition with Multi-Channel Recordings
Authors	Conggui Liu, Nakamasa Inoue, Koichi Shinoda
Citation	Proc. APSIPA, , , pp. 1304-1307
Pub. date	2017, 12
DOI	https://doi.org/10.1109/APSIPA.2017.8282233
URL	http://www.ieee.org/index.html
Copyright	(c)2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

A unified network for multi-speaker speech recognition with multi-channel recordings

Conggui Liu*, Nakamasa Inoue*, and Koichi Shinoda*
* Tokyo Institute of Technology, Tokyo, Japan
E-mail: conggui@ks.cs.titech.ac.jp, inoue@ks.c.titech.ac.jp, shinoda@c.titech.ac.jp

***Abstract*—Despite the recent progress in speech recognition, meeting speech recognition is still a challenging task, since it is often difficult to separate one speaker's voice from the others in meetings. In this paper, we propose a joint training framework of speaker separation and speech recognition with multi-channel recordings for this purpose. The location of each speaker is first estimated and then used to recover her/his original speech in a delay-and-subtraction (DAS) algorithm. The two components, speaker separation and speech recognition, are represented by one deep net, which is optimized as a whole using training data. We evaluated our method using simulated data generated from WSJCAM0 database. Compared with the independent training of the two components, our proposed method improved word accuracy by 15.2% when the locations of speakers are known, and by 53.6% when the locations of speakers are unknown.**

I. Introduction

Many people spend many hours in meetings, but their contents of meetings tend to be forgotten, cannot be rechecked later, while it is costly to manually transcribe them. Accurate automatic speech recognition has been strongly demanded for this purpose. Deep learning has significantly improved speech recognition accuracy for clean speech [1][2], but it is still challenging for recognizing speech in meetings. The major reasons include 1) ambient noise [3], 2) the time-varying number of speakers [4], and 3) overlapped speech where multiple speakers talk simultaneously [5]. In this paper, we focus on the problem of overlapped speech in the meeting speech recognition. Many methods using a single microphone have been developed to solve this problem, but their performance has not yet been sufficiently high. On the other hand, the cost of microphones has become much cheaper than before, and many people now can use their mobile devices to collect voices. We can use many microphones to detect speakers' locations and use them to obtain higher speech recognition accuracies.

The conventional meeting speech recognition methods using multi-channel recordings [6][7] first separate multiple speakers' speech from recordings and then recognize each speaker's speech. Several approaches for multi-channel speech separation have been investigated, including blind source separation, beamforming, and deep neural network (DNN). Blind source separation [8] tries to estimate the signals from each source by maximizing the statistical independence between the estimated sources. It cannot perform well when the number of active sources changes over time, and cannot give the label for each source which may be important for meeting speech recognition.

Beamforming can solve these problems by using source locations. It first identifies the locations of sources, and then use them to design spatial filters which are used to separate the sources. Delay-and-sum (DS) beamforming [9] uses the time difference of arrival (TDOA) to estimate the spatial filters. Minimum Variance Distortionless Response (MVDR) beamforming [10] further suppresses the sounds from other sources, where the spatial filters are estimated by minimizing the variance of recorded signals between channels.

Although these beamforming techniques have significantly improved the accuracy of meeting speech recognition, it is still far lower than that for clean speech. One possible reason is that the front-end separation and the back-end acoustic modeling are optimized independently. A criterion based on word error rates (WER) is proposed in [11][12] to optimize the front-end separation, but the parameters of back-end acoustic modeling are unchanged.

Recently, DNN-based beamforming is explored for separating speech from noise and jointly trained with an acoustic model [13]. This approach successfully improved the performance of speech recognition over the conventional method without joint training. However, it is to separate a single speaker's speech from noise; it cannot be directly used for our application where multiple speakers exist and their speech should be separated.

In this paper, we propose a joint training framework of multi-speaker separation and speech recognition with multi-channel recordings. A localization network predicts the delay time from each speaker to each microphone. The delay times are used to recover each speaker's speech in a delay-and-subtraction (DAS) algorithm. The parameters of the localization network are then jointly trained with that of an acoustic model network. Different from the conventional beamforming techniques, our method localizes speakers using a deep neural network and utilizes all speakers' locations simultaneously to estimate each speaker's speech. We evaluate our method using an 8-channel microphone array and simulated data of two-speaker and four-speaker meetings generated from WSJCAM0 database.

The rest of this paper is organized as follows. Section 2 describes the details of the proposed method. Section 3 shows and discusses experiment results. Section 4 concludes this paper.

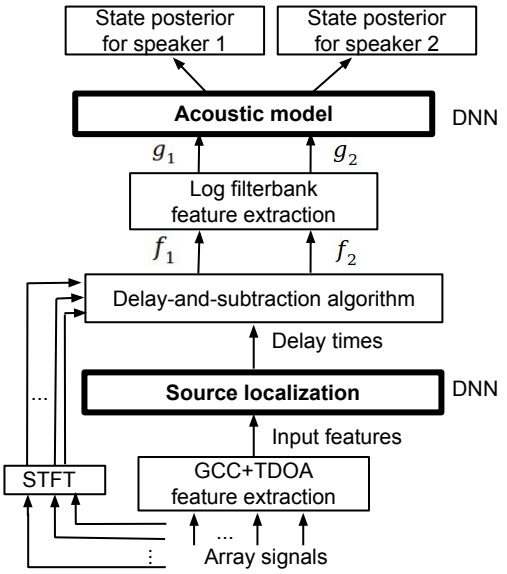

Fig. 1. The proposed joint training framework for recognizing meeting speech of two speakers. Here, f_1 and f_2 denote speech spectrum for Speaker 1 and Speaker 2 respectively, g_1 and g_2 denote log mel-filter bank feature for Speaker 1 and Speaker 2 respectively.

II. Speaker separation

A. Outline

Fig. 1 shows our joint training framework. It consists of two parts: speaker separation and speech recognition. For speaker separation, we apply deep learning-based beamforming with a delay-and-subtraction algorithm. For speech recognition, we employ a DNN-based acoustic model [1]. While Xiao et al. [13] utilized deep learning for beamforming, we use it for source localization. Let N and M be the number of speakers and microphones respectively ($N \leq M$). The short-time Fourier transform (STFT) $X_m(\tau, f)$ of the mixed signal in the m-th channel can be written as:

$$X_m(\tau, f) = \sum_{n=1}^{N} S_n(\tau, f) H_{nm}(f), \tag{1}$$

where $\tau = 1, ..., T$, $n = 1, .., N$, $m = 1, .., M$ and $f = 1, .., F$ are frame, speaker, microphone, and frequency bin indices respectively. $S_n(\tau, f)$ denotes STFT of the n-th speaker' speech. The frequency response $H_{nm}(f) = e^{-i2\pi f t_{nm}}$ can be computed using delay time t_{nm}. We assumes speakers' locations are unchanged. We first estimate delay time t_{nm} to obtain $H_{nm}(f)$, and then use $H_{nm}(f)$ to obtain $S_n(\tau, f)$ from $X_m(\tau, f)$.

B. Localization DNN

In the feature extraction for speaker separation, we calculate two kinds of features: 1) the generalized cross-correlation (GCC) [14] and 2) the time difference of arrival (TDOA). We first apply voice activity detection (VAD) to detect an utterance from one channel. Then we apply the same onset and offset times for the corresponding signals in all other channels. The GCC features are extracted from the voiced parts of signals.

Let $X_1(\tau, f)$ and $X_2(\tau, f)$ be the STFTs of two signals recorded by two microphones, e.g., Microphone 1 and Microphone 2, respectively. The cross-correlation of these two signals for the τ-th frame and the f-th frequency bin is computed by $\widehat{R}(\tau, f) = X_1(\tau, f) X_2(\tau, f)^*$. Then, the GCC for $l \in -K, .., 0, .., K$ is computed by:

$$v_{\mathrm{gcc}}(\tau, l) = \Psi^{-1}\left(\frac{\widehat{R}(\tau, f)}{|\widehat{R}(\tau, f)|}\right), \tag{2}$$

where an operator $\Psi^{-1}(.)$ denotes inverse STFT. Only the $(2L+1)$ central elements of the GCC are selected to form a vector, $\mathbf{v}_{\mathrm{gcc}}(\tau) = [v_{\mathrm{gcc}}(\tau, -L), .., v_{\mathrm{gcc}}(\tau, 0).., v_{\mathrm{gcc}}(\tau, L)]^T$, assuming the other elements don't contain the information about speakers' locations. The parameter L is estimated in the same way as [13], $L = \lceil df_s/c \rceil$, where f_s is the sampling rate of signals, d is the distance between two microphones, and c is the speed of sound.

In addition to the GCC features, we use the TDOA features which are estimated for each utterance, not for each frame, to identify the locations precisely. For each possible value of TDOA $t_{\mathrm{tdoa}} \in [-d/c, d/c]$, the phase spectrum $\phi(t_{\mathrm{tdoa}})$ [15] is calculated as:

$$\phi(t_{\mathrm{tdoa}}) = \max_{\tau} \sum_{f=1}^{F} \Re\left(\frac{\widehat{R}(\tau, f)}{|\widehat{R}(\tau, f)|} e^{-i2\pi f t_{\mathrm{tdoa}}}\right), \tag{3}$$

where $\Re(Z)$ denotes the real part of a complex value Z. The TDOAs for two speakers are estimated by finding the locations of the largest two peaks of the phase spectrum $\phi(t_{\mathrm{tdoa}})$, and form a 2-dimensional TDOA vector $\mathbf{v}_{\mathrm{tdoa}}$.

We calculate the GCC vector $\mathbf{v}_{\mathrm{gcc}}(\tau)$ and the TDOA vector $\mathbf{v}_{\mathrm{tdoa}}$ for each microphone pair. The dimension of input features to source localization DNN is $M(M-1)(2L+3)/2$ for M microphones. For N speakers, it should be $M(M-1)(2L+1+N)/2$.

The DNN for source localization has two hidden layers and each hidden layer has 1024 sigmoidal units. The output of the DNN in the τ-th frame is a vector $\mathbf{y}(\tau) = [y_{11}(\tau), .., y_{nm}(\tau), .., y_{NM}(\tau)]^T$, where $y_{nm}(\tau)$ is the estimation of the delay time t_{nm} from the n-th speaker to the m-th microphone. Its dimension should be $N \times M$. The DNN is trained by minimizing the mean square error (MSE) between the predicted delay time $y_{nm}(\tau)$ and its ground truth t_{nm}. Again, we average the predicted delay times over all the frames in one utterance for each speaker, $\overline{y}_{nm} = \frac{1}{T}\sum_{\tau=1}^{T} y_{nm}(\tau)$.

C. Delay-and-subtraction algorithm

When we separate multiple sources, the locations of the other sources may be effectively used for identifying the location of each source. Instead of the conventional methods, such as DS beamforming [9], we apply a delay-and-subtraction (DAS) algorithm using all speakers' locations to estimate each speaker's speech. The inputs of delay-and-subtraction (DAS) algorithm are STFT coefficients $X_m(\tau, f)$ of multi-channel array signals and averaged delay time $\overline{y}_{nm}$.

Suppose there exist two speakers and two microphones, i.e. $N = 2$ and $M = 2$, for cancelling the first speaker, the spectrum of the mixed speech signal $X_m(\tau, f)$ for $m \in 1, 2$ is first multiplied by a weight $H_{1m}(f)^*$ as follows:

$$\begin{aligned} Y_m(\tau, f) &= X_m(\tau, f) H_{1m}(f)^* \\ &= S_1(\tau, f) + S_2(\tau, f) H_{2m}(f) H_{1m}(f)^*, \end{aligned} \tag{4}$$

where $(.)^*$ denotes a conjugate operator. The speech spectrum $S_1(\tau, f)$ can be canceled by using a subtraction process, $Y_2(\tau, f) - Y_1(\tau, f)$. Then, the second speaker's speech spectrum $S_2(\tau, f)$ is estimated as follows:

$$\widehat{S}_2(\tau, f) = \frac{X_2(\tau, f)\widehat{H}_{12}(f)^* - X_1(\tau, f)\widehat{H}_{11}(f)^*}{\widehat{H}_{22}(f)\widehat{H}_{12}(f)^* - \widehat{H}_{21}(f)\widehat{H}_{11}(f)^*}. \tag{5}$$

Here, $\widehat{H}_{nm}(f) = e^{-i2\pi f \overline{y}_{nm}}$ is the estimation of $H_{nm}(f)$, where $n = 1, 2$ and $m = 1, 2$. The speech spectrum $S_1(\tau, f)$ for the first speaker can be estimated in the same way.

For the case with more than two speakers $S_1, ..., S_N$, i.e. $M \geq N > 2$, each speaker' speech can be recovered one by one. For example, first S_1 is estimated by subtracting $(N-1)$ speakers' speech spectrum $S_2, .., S_N$, then, S_2 is estimated by subtracting $(N-2)$ speakers' speech spectrum $S_3, .., S_N$. This process continues to estimate the speech spectrum of all the speakers $\widehat{S}_1, ..., \widehat{S}_N$.

III. Joint Training

Speaker separation uses the feedback of speech recognition to adjust its parameters. At the same time, the module of speech recognition is adapted to reduce the effect of the mismatch between separated speech and its original speech. We apply a multi-task framework to recognize multiple speakers' speech simultaneously as shown in Fig. 1, where each task is to classify the context-dependent states obtained through forced alignment and estimate state posterior probabilities, and all tasks share layers. The training steps are as follows:

1) Train the localization DNN from simulated data.
2) Train the DNN acoustic model from clean data
3) Retrain the DNN acoustic model using separated speech.
4) Train the localization DNN jointly with the DNN acoustic model using back-propagation with a cross-entropy objective function and simulated data.

IV. Experiments

A. Setting

We generated training data by randomly selecting clean speech from 7861 training sentences in the WSJCAM0 corpus [16]: 80 hours for DNN-based source localization, and about 15 hours for re-training and joint training. Test data is created in the same way as the training data using the test set in the WSJCAM0 corpus. Two-speaker and four-speaker meetings are simulated. The simulation uses an eight equally spaced circle microphone array with 0.1-meter radius. Speakers are located on a circle with the 1-meter radius.

For two-speaker meetings, the 1st speaker is female, and the 2nd speaker is male. The 1st speaker's location α_1 is randomly selected from a total of 270 directions (1^o interval), where $\alpha_1 \in [0^o, 270^o)$. The 2nd speaker's location α_2 is fixed to $\alpha_1 + 90^o$. To automatically and correctly assign estimated locations to speakers, the patterns of speakers' locations in the test sets should be same as that in the training data, e.g., the angle of the 2nd speaker is always 90 degrees larger than that of the 1st speaker. Two test sets, *Known* set and *Unknown* set, are generated. The 1st speaker's location α_1 for the *Known* and *Unknown* sets are selected from a total of 270 directions (1^o interval) within $[0^o, 270^o)$ and a total of 90 directions (1^o interval) within $[270.5^o, 360.5^o)$, respectively. In both cases, the 2nd speaker's location α_2 is $\alpha_1 + 90^o$.

TABLE I
Angle error (degree) for known test set and unknown test set. "PHAT" is GCC-PHAT algorithm for localizing speakers, and "DNN" is our localization method.

	Known		**Unknown**	
Spacing(cm)	PHAT	DNN	PHAT	DNN
7.65	2.60	0.41	38.14	2.93
14.14	0.35	0.36	27.53	5.54
18.48	0.31	0.39	21.50	4.81
20.00	0.30	0.42	20.12	3.60
Mean	0.89	**0.40**	26.82	**4.22**

For four-speaker meetings, the 1st, and 3rd speakers are females, and the others are males. A test set, *Known* set, is generated. For both training and evaluation, the 1st speaker's location α_1 is randomly selected from the first quadrant. The i-th speaker's location α_i $(1 < i < 5)$ is fixed to $\alpha_{i-1} + 90^o$.

To evaluate source localization performance, we use the angle error between the estimated direction of arrival (DOA) and the ground truth of DOA. For U sentences and N speakers, the angle error E_a is calculated as:

$$E_a = \frac{1}{NU} \sum_{n=1}^{N} \sum_{u=1}^{U} |\widehat{\theta}_{nu} - \theta_{nu}|, \tag{6}$$

where $\widehat{\theta}_{nu}$ is the estimation of true DOA θ_{nu} for the n-th speaker and the location of the u-th utterance. We assume speakers' sound waves arrive at individual microphones in a parallel way (the far-field assumption). Then, the angle $\widehat{\theta}_{nu}$ is computed by $\widehat{\theta}_{nu} = c\Delta\widehat{t}/d$, where c (m/sec) is the speed of sound, $\Delta\widehat{t}$ is the estimated TDOA for two microphones, $\Delta\widehat{t} = \overline{y}_{nm_2} - \overline{y}_{nm_1}$, for the m_1-th microphone and the m_2-th microphone, and d (m) is the distance between these two microphones. With the circle microphone array, there are four possible choices for the microphone distance (spacing) d: 1) 7.65 cm, 2) 14.14 cm, 3) 18.48 cm, and 4) 20.00 cm. For speech recognition, all results are evaluated for a microphone pair with the smallest microphone distance in terms of angle error and word error rate (WER).

B. Angle error of source localization

Table I shows the performance of DNN-based source localization for two speaker meetings. Compared with PHAT [14], DNN (our method) achieves smaller angle errors for both the *Known* set and the *Unknown* set. The improvement

TABLE II
WER (%) FOR KNOWN TEST SET AND UNKNOWN TEST SET. "DNN+DAS" DENOTES OUR SPEAKER SEPARATION METHOD, "DNN+DAS+RE" IS RETRAINING ACOUSTIC MODEL USING SEPARATED DATA, AND "DNN+DAS+RE+JOINT" IS JOINT TRAINING OF SPEAKER SEPARATION AND ACOUSTIC MODEL.

No.	Method	Known	Unknown
(1)	PHAT+MVDR	22.1	65.0
(2)	DNN+MVDR	20.2	21.7
(3)	PHAT+DAS	16.8	63.5
(4)	DNN+DAS	12.0	23.6
(5)	DNN+DAS+Re	8.2	13.7
(6)	DNN+DAS+Re+JOINT	**6.9**	**11.4**

TABLE III
WER (%) FOR KNOWN TEST SET.

Method	Known
PHAT+MVDR	37.8
DNN+DAS	12.7
DNN+DAS+Re	7.8
DNN+DAS+Re+JOINT	**7.4**

is significant when the distance between two microphones is small and in the *Unknown* case. The results confirm the effective use of the DNN-based localization method.

C. Performance of two speakers' speech recognition

In Table II, we compare (4) DNN+DAS (our method) with (1) PHAT+MVDR [17]: MVDR beamforming with PHAT, (2) DNN+MVDR: the combination of our DNN-based source localization with MVDR, and (3) PHAT+DAS: the combination of PHAT with our DAS algorithm. Both the DNN-based localization and the DAS algorithm are effective to improve speech recognition performance. For example, DNN+DAS achieves 10% improvement for the *Known* set, and 41.4% improvement for the *Unknown* set over PHAT+MVDR.

We also show the result of (6) DNN+DAS+Re+JOINT which represents the joint training of speaker separation and the acoustic model in Table II. DNN+DAS+Re+JOINT decreases WER by 15.2 points for the *Known* set and 53.6 points for the *Unknown* set compared to PHAT+MVDR. This result shows the effectiveness of our method in improving the performance of meeting speech recognition.

D. Performance of four speakers' speech recognition

Table III shows the performance of speech recognition in four speaker meetings. Comparing with PHAT+MVDR, DNN+DAS+Re+JOINT decreases WER by 30.4 points for test data. The improvement comes from using the DNN-based localization, the DAS algorithm, the acoustic model retraining, and the joint training. Though the improvement derived by the joint training step is slightly smaller than that in two speaker meetings, our method still performs well for the case with more than two speakers.

V. CONCLUSIONS

We have proposed a joint training framework of multi-channel speaker separation and acoustic model using a deep neural network. For speaker separation, a localization DNN is used to estimate speakers' locations, and all estimated locations are used to recover each speaker' speech from multi-channel recordings. The speaker separation is then jointly trained with acoustic model. Different from conventional localization methods, the localization DNN can automatically assign estimated locations to speakers because it maintains the order of speakers through four-quadrant angles. Our method can effectively localize speakers, separate multiple speakers' speech, and decrease WER by 15.2 points for the *Known* set and 53.6 points for the *Unknown* set.

In future, we plan to do meeting speech recognition in a real environment, e.g., where more speakers are recorded by simple microphone arrays, such as smartphones, in a noisy environment, and the number of speakers is unknown.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI 16H02845 and by JST CREST Grant Number JPMJCR1687, Japan.

REFERENCES

[1] K. Veselỳ, *et al.* "Sequence-discriminative training of deep neural networks." *INTERSPEECH*, pp. 2345–2349, 2013.
[2] G. Hinton, *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.
[3] T. Yoshioka, *et al.* "Impact of single-microphone dereverberation on dnn-based meeting transcription systems," *ICASSP*, pp. 5527–5531, 2014.
[4] A. Stolcke, *et al.* "Leveraging speaker diarization for meeting recognition from distant microphones," *ICASSP*, pp. 4390–4393, 2010.
[5] E. Shriberg, *et al.* "Observations on overlap: findings and implications for automatic processing of multi-party conversation." *INTERSPEECH*, pp. 1359–1362, 2001.
[6] A. Stolcke, *et al.* "Making themost from multiple microphones in meeting recognition," *ICASSP*, pp. 4992–4995, 2011.
[7] S. Araki, *et al.* "Blind speech separation in a meeting situation with maximum snr beamformers," *ICASSP*, vol. 1, pp. 41–44, 2007.
[8] H. Sawada, *et al.* "Frequency-domain blind source separation," *Speech enhancement*, pp. 299–327, 2005.
[9] J. Benesty, *et al.* "Microphone array signal processing," *Springer Science & Business Media*, 2008.
[10] B. D., *et al.* "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, 1988.
[11] M. L., *et al.* "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 489–498, 2004.
[12] M. L., *et al.* "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," *Hands-Free Speech Communication and Microphone Arrays*, pp. 104–107, 2008.
[13] X. Xiao, *et al.* "Deep beamforming networks for multi-channel speech recognition," *ICASSP*, pp. 5745–5749, 2016.
[14] C. Knapp, *et al.* "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
[15] Z. El Chami, *et al.* "A phase-based dual microphone method to count and locate audio sources in reverberant rooms," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 209–212, 2009.
[16] T. Robinson, *et al.* "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," *ICASSP*, vol. 1, pp. 81–84, 1995.
[17] X. Mestre, *et al.* "On diagonal loading for minimum variance beamformers," *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 459–462, 2003.