INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

DNN-based Feature Enhancement using Joint Training Framework for Robust Multichannel Speech Recognition

Kang Hyun Lee, Tae Gyoon Kang, Woo Hyun Kang and Nam Soo Kim

Department of Electrical and Computer Engineering and INMC,
Seoul National University, Korea

khlee@hi.snu.ac.kr, tgkang@hi.snu.ac.kr, whkang@hi.snu.ac.kr and nkim@snu.ac.kr

Abstract

Ever since the deep neural network (DNN) appeared in the speech signal processing society, the recognition performance of automatic speech recognition (ASR) has been greatly improved. Due to this achievement, the demands on various applications in distant-talking environment also have been increased. However, ASR performance in such environments is still far from that in close-talking environments due to various problems. In this paper, we propose a novel multichannel-based feature mapping technique combining conventional beamformer, DNN and its joint training scheme. Through the experiments using multichannel wall street journal audio visual (MC-WSJ-AV) corpus, it has been shown that the proposed technique models the complicated relationship between the array inputs and clean speech features effectively via employing intermediate target. The proposed method outperformed the conventional DNN system.

Index Terms: deep neural networks (DNNs), multichannel speech recognition, distant speech recognition, feature compensation.

1. Introduction

In recent years, deep learning has been widely investigated in signal processing and become an opportunity for automatic speech recognition (ASR) to advance. In acoustic modeling, introduction of the deep neural network (DNN)-hidden Markov model (HMM) system which represents the relationship between the acoustic features and HMM states using a DNN instead of a Gaussian mixture model (GMM) is considered as a breakthrough. The DNN-HMM system has outperformed the conventional GMM-HMM system in various ASR tasks [1, 2, 3]. The remarkable performance of the DNN-HMM system is attributed to its capability in automatically learning the complicated nonlinear mapping between the input and target vectors.

Due to the progresses above, the ASR system has achieved great performance in close-talking environments. However, recent developments in speech and audio applications such as hearing aids and hands-free speech communication systems require speech acquisition in distant-talking environments. Unfortunately, as the distance from the speaker and the microphone increases, the recorded speech becomes more distorted due to the background noise and room reverberation. Although it may be possible to acquire the speech in close-talking environments by using a headset microphone, it is not a general solution because of the inefficiency in terms of cost and ease of use. Consequently, ASR performance in distant-talking environments is still far from that shown in close-talking environments.

In order to overcome this difficulty, various researches have focused on techniques for efficiently integrating the information obtained from multiple distant microphones to improve the ASR performance. One of the most conventional multichannel-based techniques is the beamformer method, which enhances the signals emanating from a particular location by individual microphone arrays. The simplest technique is the delay-and-sum (DS) beamformer [4], which compensates the delays of the microphone inputs so that only the target signal from a particular direction synchronizes with. In addition, there are many sophisticated beamforming methods [5, 6] which optimize the beamformers to produce a spatial pattern with a dominant response for the location of interest.

Feature mapping techniques based on DNN have been also investigated recently. DNN-based feature enhancement techniques [7, 8] have already been widely employed in robust ASR due to their advantage in directly representing the arbitrary unknown mapping between the noisy and clean features unlike the conventional techniques [9, 10, 11, 12] which usually require specific assumptions or formulations. Especially, [8] showed that the feature mapping technique combining beamformer and DNN improves the performance of the ASR system in multichannel distant speech recognition.

Meanwhile, recent researches on joint training technique of DNN [13, 14] have drawn attention. The joint training technique builds a DNN by concatenating two independently trained DNNs and jointly adjusting the parameters. Through this training technique, the synergy between two DNNs can be amplified. Traditionally, this joint training framework has been applied to incorporate two different tasks into one universal task, i.e., integrating speech separation and acoustic modeling [14]. In addition to the usage above, the joint training technique can be used for training a DNN in charge of a single task elaborately. In these circumstances, the performance of DNN depends on deciding which types of features are represented in the intermediate layer where junction between two DNNs occur. In [15], a performance of DNN was enhanced by giving appropriate intermediate concepts which the DNN should represent in the mid-level.

In this paper, we propose a novel DNN-based feature enhancement technique for multichannel distant speech recognition in modern multichannel environments where various types of microphone data are given as training data. The main contribution of the proposed approach is to construct a multichannel-based feature mapping DNN algorithm by properly combining a conventional beamformer, DNN and its joint training technique with lapel microphone data which has an intermediate level of acoustic information between DNN input and the target. To implement the technique making use of various micro-

http://dx.doi.org/10.21437/Interspeech.2016-105

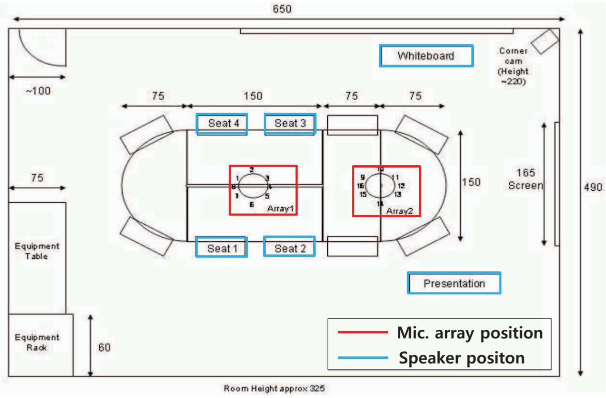

Figure 1: The layout of the UEDIN Instrumented Meeting Room (measurements in cm). Array microphones are numbered 1-16. Cameras are mounted under Array 1 to give closeup views of participants in the seated locations. The six reading locations are indicated as Seat 1-4, Presentation and Whiteboard.

phone types and evaluate the performance, we used a data set of single speaker scenario from multichannel wall street journal audio visual (MC-WSJ-AV) corpus [16] which is a re-recorded version of WSJCAM0 [17] in a meeting room environment.

The remainder of this paper is organized as follows. In section 2, we describe the MC-WSJ-AV corpus. In section 3, we present the proposed DNN-based multichannel feature enhancement technique. In section 4, we provide results from the ASR experiments performed to evaluate the proposed methods on evaluation set of single speaker scenario in MC-WSJ-AV corpus, and in section 5, we provide a summary and final conclusions.

2. Database description on single speaker stationary scenario of MC-WSJ-AV corpus

MC-WSJ-AV corpus can be categorized into three scenarios: single speaker stationary, single speaker moving and overlapping speakers scenarios. Since we are dealing with only the audio data in the single speaker stationary scenario, this section overviews the recording of the single speaker stationary scenario in MC-WSJ-AV database.

For the recording of the single speaker stationary scenario data, the data is recorded in three sites: The Centre for Speech Technology Research, Edinburgh (UEDIN), The IDIAP Research Institute, Switzerland (IDIAP) and TNO Human Factors, The Netherlands (TNO). Instrumented meeting rooms installed at the three sites allow the audio to be fully synchronized. The layout of the UEDIN room with the positions of the microphone arrays and the six reading positions, is shown in Figure 1. The room contains two eight-element circular microphone arrays, one mounted at the center and one at the end of the meeting room table.

In addition, the speakers are provided with close-talking radio headset and lapel microphones. The TNO and IDIAP rooms contain the similar recording equipments, but differ in their physical layout and acoustic conditions. In the single speaker stationary condition, the speaker was asked to read sentences from six positions within the meeting room: four seated around the table, one standing at the whiteboard and one standing at the presentation screen. For each speaker, one sixth of the sentences are read from each position.

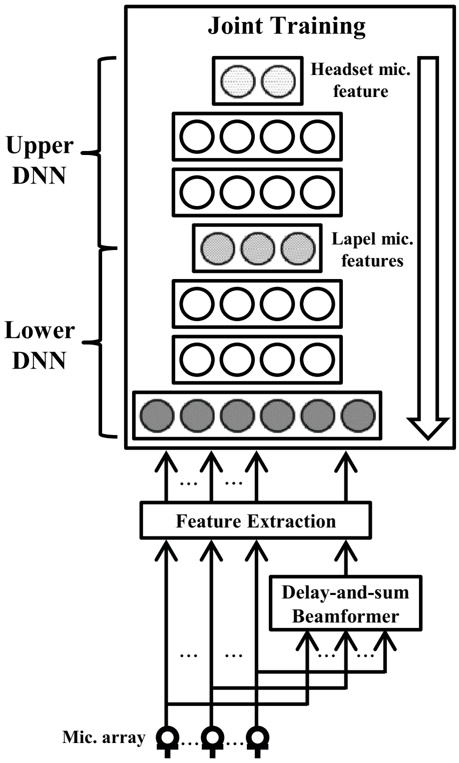

Figure 2: The schematic diagram of proposed technique.

3. Proposed approach

In this work, the (m)-th array microphone feature, the DS-beamformed feature from the array, lapel microphone feature and headset microphone feature being extracted at the t-th frame are denoted as $\mathbf{a}_t^{(m)}$, $\mathbf{b}_t$, $\mathbf{l}_t$ and $\mathbf{h}_t$, respectively. Additionally, we denote a subsequence of vectors $[\mathbf{x}'_{n_1-n_2}\ \mathbf{x}'_{n_1-n_2+1} \cdots \mathbf{x}'_{n_1+n_2-1}\ \mathbf{x}'_{n_1+n_2}]$ with the prime representing matrix or vector transpose from frame index $n_1 - n_2$ to $n_1 + n_2$ as $\mathbf{x}_{n_1,n_2}$.

We propose a novel DNN-based feature enhancement approach for multichannel distant speech recognition. The purpose of our technique is to estimate the clean features from the distant array features. However, there exists two problems for enabling the DNN to achieve this adverse task. The first problem is the phase differences among each signal of array microphones originated from the distances between the speaker and each microphone. And the second problem, which is more serious, is the lack of acoustic information of the array. Due to the distances between each of the array microphones and the speaker, the microphones have low ratio of direct-to-reverberant speech energy which becomes a huge limitation on reconstructing the clean speech entirely. To compensate for these problems, we propose the DNN as shown in Figure 2.

The proposed DNN is constructed by concatenating two individually fine-tuned DNNs and training the unified DNN jointly. We call the first DNN as lower DNN since it is placed in the lower part of the DNN in Figure 2. The second DNN which is called the upper DNN, deals with modeling the relationship between the output vector generated by the lower DNN and the headset microphone feature.

3.1. Lower DNN

For training the lower DNN, DS beamforming [4] is employed to the microphone array to align the phases of microphone inputs. Once the beamforming has been applied, the input vector of the lower DNN $\mathbf{v}_t$ is formed by concatenating a window of several adjacent frames of feature from the beamformed source and additional windows covering each array microphone features, i.e.,

$$\mathbf{v}_t = [\mathbf{a}_{t,\tau}^{(1)}, \mathbf{a}_{t,\tau}^{(2)}, \cdots, \mathbf{a}_{t,\tau}^{(M-1)}, \mathbf{a}_{t,\tau}^{(M)}, \mathbf{b}_{t,\tau}] \quad (1)$$

where τ represents the temporal coverage required for figuring out the clean feature of t-th frame and M represents the number of the array elements. This input structure helps the lower DNN to learn the correlations among features of array microphones.

As the target vector of the network, we used a window of several frames of feature obtained from lapel microphone which has a much higher ratio of direct-to-reverberant speech energy than those of the array microphones but lower than those of the headset microphones. Therefore, the lower DNN output can be represented as follows:

$$\widehat{\mathbf{o}}_t^L = [\,\widehat{\mathbf{l}}_{t,\tau}]. \quad (2)$$

3.2. Upper DNN and joint training

In the training stage of the upper DNN training, the network learns the mapping between the output vector of the lower DNN and the corresponding headset microphone feature which can be interpreted as a ideal clean feature. The mapping can be represented as follows:

$$\widehat{\mathbf{o}}_t^U = [\widehat{\mathbf{h}}_t] \cong f(\widehat{\mathbf{l}}_{t,\tau}). \quad (3)$$

Here, function f is a function which deals with the mapping from the reconstructed lapel microphone features to the headset microphone feature. Since the clean features are estimated from the reconstructed lapel features which have more abundant acoustic information than the array features, we can expect more accurate reconstruction of clean features.

After training the upper DNN, two different networks are cascaded to form a single larger DNN and the unified DNN jointly adjusts the weights using the backpropagation algorithm. In detail, the error signal between the clean target and the output of unified DNN flows back to the lapel microphone feature layer and the lower DNN, and consequently training all the parameters. With this series of processes, learning the relationship between the array features and the headset features can be enhanced by guiding the DNN through the intermediate level features. For training all the DNNs in the proposed method, the stochastic gradient descent algorithm is used to minimize the mean squared error (MSE) function which is given by

$$C_{MSE} = \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{O}_t - \widehat{\mathbf{O}}_t||^2 \quad (4)$$

where $\mathbf{O}_t$, $\widehat{\mathbf{O}}_t$, and T denote the target, output vector of network and number of training samples, respectively.

4. Experiments

The proposed technique was trained on development set (DEV) and its performance was evaluated on evaluation set (EVAL1). The selection of read sentences for these sets was based on the development and evaluation sets of the WSJCAM0 British English corpus [17]. Each speaker prompt contained 17 adaptation sentences, 40 sentences from the 5000-word sub-corpus, respectively.

In this section, some basic experimental results obtained from DS-beamformed source (*DS*) of microphone array, headset microphone (*Headset*), lapel microphone (*Lapel*) and single distant microphone (*SDM*) recordings were presented. Here, the microphone array refers to Array 1 which is the left one among the two arrays in Figure 1 and single distant microphone is the no. 1 microphone of the Array 1. Also, the comparison of performances with conventional DNN-based feature mapping methods were included.

4.1. Recognition system and feature extraction

A baseline DNN-HMM system was trained on the WSJCAM0 database. The training set consisted of 53 male and 39 female speakers. We used the Kaldi speech recognition toolkit [18] for feature extraction, acoustic modeling of ASR and ASR decoding. For feature extraction, 13-dimensional MFCCs (including C_0) with their first and second derivatives were extracted and the cepstral mean normalization algorithm was applied for each speaker. In order to provide the target alignment information for the DNN-based acoustic model, we built a GMM-HMM system with 2047 senones and 15026 Gaussian mixtures in total. The target senone labels of the DNN-HMM system were obtained over the training data. As for the language model, we applied the standard 5k WSJ trigram language models.

For the DNN training of the acoustic model, we applied five hidden layers with 2048 nodes. As for the input of the DNNs, input features consisted of consecutive 11-frame (5 frames on each side of the current frame) context window of 13 dimensional MFCC features with their first and second order derivatives, resulting with the input dimension of 429. The input features of the DNNs were normalized to have zero mean and unit variance. The output dimension of the DNN was 2047. Generative pre-training algorithm for the restricted Boltzmann machines was carried out to initialize the DNN parameters as described in [19]. The errors between the DNN output and the target senone labels were calculated according to the cross-entropy criterion [2]. In order to speed up the training, we applied the learning rate scheduling scheme and the stop criteria presented in [19].

4.2. Training and structures of DNN-based techniques

The performance of the proposed method was compared with four different versions of DNN-based feature enhancement approaches. The compared techniques are

- *FE-SDM*: mapping single array microphone into a clean target source,
- *FE-DS*: mapping DS-beamformed source of the array into a clean target source,
- *FE-Array*: mapping multiple sources from microphone array into a clean target source,
- *FE-DS&Array*: mapping multiple sources including the sources from the microphone array and DS-beamformed source of the array into a clean target source.

For training all the DNN-based feature enhancement techniques, we used cepstral mean normalized MFCC feature of 13 dimension with their first and second derivatives as an input.

Table 1: WERs (%) on EVAL1 according to various source types

Channel	WER (%)
SDM	58.00
DS	41.97
Lapel	13.18
Headset	7.49

Table 2: Input and output dimensions of the DNN-based techniques.

Method	Input dim.	Output dim.
FE-SDM	429	13
FE-DS	429	13
FE-Array	3432	13
FE-DS&Array	3861	13
Proposed	3861	13

All the techniques used one or more windows depending on the number of sources and each window consists of 11 consecutive MFCCs. Meanwhile, the feature mapping DNNs commonly estimated 13-dimensional static MFCC of current frame and the outputs of DNNs were fed into the recognizer after extraction of their dynamic component. Table 2 shows the input and output dimensions of each DNN-based techniques. The networks had 5 hidden layers with 1024 rectified linear units (ReLUs) [20] are applied except for the proposed technique which contains the intermediate layer because of its unique structure. The parameters of the DNN-based techniques are randomly initialized and fine-tuned using stochastic gradient descent algorithm with minimum MSE objective function like those of the proposed method.

Mini-batch size for the stochastic gradient descent algorithm was set to be 256 for all of the DNN-based feature enhancement techniques. The momentum was set to be 0.5 at the first epoch and increased to 0.9 afterward. The learning rate was initially set to be 0.01 and exponentially decayed over each epoch with decaying factor of 0.9 except for the cases of the lower DNN and joint training of the proposed method. For lower DNN and the joint training, learning rate was initially set to be 0.001 and exponentially decayed over each epoch with a decaying factor of 0.95. All the training of DNN-based techniques were stopped after 50 epochs.

4.3. Dropout

As one of the most well-known regularization techniques, dropout was also applied. Dropout is a method that improves the generalization ability of the DNN. It can be easily implemented by randomly dropping the input and hidden neuron units. As pointed out by Hinton et al. [21], dropout can be considered as a bagging technique that averages over a large amount of models with shared parameters of the DNN. A dropout percentage of 20% was applied to every DNN-based feature enhancement technique.

4.4. Performance Evaluation

Table 3: WERs (%) on EVAL1 according to variety of DNN-based feature enhancement techniques.

Method	WER (%)	
Dropout Percentage	0%	20%
FE-SDM	27.36	25.88
FE-DS	24.25	21.91
FE-Array	21.54	20.44
FE-DS&Array	19.77	19.63
Proposed	**18.84**	**17.70**

Table 1 and Table 3 show the results according to various source types and DNN-based techniques, respectively. Comparison among the DNN-based approaches shows that high variety of input structure of the DNN guarantees better performance. We can see that the proposed method outperformed other DNN-based techniques including *FE-DS&Array* which has the same input structure but more parameters than the proposed approach.

Especially, the performance gap between the proposed method and *FE-DS&Array* became larger especially when dropout was applied. With dropout training, the average relative error rate reductions (RERRs) of the proposed method over *FE-DS&Array* was 9.8%. This confirms that our proposed approach which intervenes the DNN through information of reconstructed lapel microphone data can be effective in making the network to learn the complicated relationship between features from the distant microphone array, DS-beamformer and headset microphone sources.

5. Conclusion

In this paper, we have proposed a novel DNN-based feature enhancement approach for multichannel distant speech recognition. The proposed approach constructs a multichannel-based feature mapping DNN using conventional beamformer, DNN and its joint training technique with lapel microphone data. Through a series of experiments on MC-WSJ-AV corpus, we have found that the proposed technique clarifies the relationship between the features obtained from distant microphone array and clean speech. Future study will deal with techniques considering other speaking scenarios such as overlapping speakers.

6. Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2015R1A2A1A15054343), and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP (Institute for Information & communications Technology Promotion).

7. References

[1] A. Mohamed, G. E. Dahl, and G. E. Hinton, “Acoustic modeling using deep beliefs networks,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pretrained deep neural networks for large vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[3] B. Kingsbury, T. N. Sainath, and H. Soltau, “Scalable minimum

Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Proc. Interspeech*, 2012, pp. 10–13.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, Springer Science & Business Media, 2008.

[5] O. L. Frost, "An algorithm for linearly constrained adaptive array processing, *Proc. IEEE*, vol. 60, no. 8, pp. 926-935, Aug. 1972.

[6] L. J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. Antennas Propag.*, vol. AP-30, no. 1, pp. 27-34, Jan. 1982.

[7] A. Narayanan, and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 826–835, Apr. 2014.

[8] W. Li, L. Wang, Y. Zhou, J. Dines, M. Magimai.-Doss, H. Bourlard, and Q. Liao, "Feature mapping of multiple beamformed sources for robust overlapping speech recognition using a microphone array ," *IEEE/ACM Trans. Audio, Speech, Language Process*., vol. 22, no. 12, pp. 2244–2255, Dec. 2014.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process*., vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust speech recognition using a cepstral minimum-meansquare-error-motivated noise suppressor," *IEEE Trans. Audio, Speech, Language Process*., vol. 16, no. 5, pp. 1061–1070, Jul. 2008.

[11] N. S. Kim, "IMM-based estimation for slowly evolving environments," *IEEE Signal Process. Lett*., vol. 5, no. 6, pp. 146–149, Jun. 1998.

[12] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech, Audio, Process*., vol. 11, no. 6, pp. 568–580, Nov. 2003.

[13] A. Narayanan, and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 92–101, Jan. 2015.

[14] Z. Wang, and D. Wang, "A joint training framework for robust autiomatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 796–806, Apr. 2016.

[15] T. Kowaliw, N. Bredeche, and R. Doursat, *Growing adaptive machines: combining development and learning in artificial neural networks*, Springer, 2014.

[16] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proc. ASRU*, 2005, pp. 357–362.

[17] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: a british english speech corpus for large vocabulary continuous speech recognition, in *Proc. ICASSP*, 1995, pp. 81-84.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[19] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," Proc. in *Proc. Interspeech*, 2013, pp. 2345–2349,.

[20] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.,* 2011, pp. 315–323.

[21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.