LETTER

Improve Multichannel Speech Recognition with Temporal and Spatial Information

Yu ZHANG[†a)], **Pengyuan ZHANG**[†], ***Nonmembers***, ***and*** **Qingwei ZHAO**[†], ***Member***

SUMMARY In this letter, we explored the usage of spatio-temporal information in one unified framework to improve the performance of multichannel speech recognition. Generalized cross correlation (GCC) is served as spatial feature compensation, and an attention mechanism across time is embedded within long short-term memory (LSTM) neural networks. Experiments on the AMI meeting corpus show that the proposed method provides a 8.2% relative improvement in word error rate (WER) over the model trained directly on the concatenation of multiple microphone outputs.

key words: *multichannel speech recognition, long short-term memory, attention mechanism, generalized cross correlation*

1. Introduction

Deep neural networks (DNNs) based acoustic models [1] have driven tremendous improvements in automatic speech recognition (ASR) in recent years. Further improvements are achieved by using more complex models such as long short-term memory based recurrent neural networks (LSTMs) [2]. However, it still remains challenging to perform recognition when the speaker is distant from the microphone, because of the presence of background noise, reverberation, and competing acoustic sources. In such cases, ASR systems often use signals from multiple microphones to enhance the speech signal and reduce the impact of noise and reverberation.

Multichannel ASR systems often adopt a two-part architecture, in which a beamforming algorithm is applied to enhance the speech, followed by conventional acoustic modeling approaches. And some mask estimation based beamforming techniques [3]–[5] have achieved good performance on the multichannel speech recognition task. However, the speech enhancement module is usually separate from the speech recognition module, which may lead to a suboptimal solution [6]. Therefore, joint training of speech enhancement and acoustic model was proposed to solve the problem. Sainath et al. [7]–[9] presented a multichannel neural network model trained directly from raw waveform input signal. Instead of filtering in the time domain, Xiao et al. [10] estimated the parameters of the frequency-domain beamformer from a generalized cross correlation (GCC) between microphones. In [11], a neural network which estimated masks for a statistically optimum beamformer was jointly trained with a neural network acoustic model. An LSTM adaptive beamformer was proposed in [12] to be jointly trained with an LSTM acoustic model. Ochiai et al. [13] presented a unified architecture for end-to-end multichannel speech recognition.

Some works [14], [15] have shown that DNNs can learn suitable representations for multichannel speech recognition by directly using multichannel outputs. These approaches, however, simply concatenated contextual acoustic features from multiple microphones without considering the temporal information within the contextual window of acoustic features and the spatial information across different channels. Over the past few years, attention-based recurrent neural networks have shown promising results in end-to-end speech recognition [16]–[18], in which the attention mechanism is used to learn the alignment between the input features and transcripts. [19] proposed a deep convolutional neural networks (CNNs) acoustic model which introduced location-based attention by weighting the contribution from each frame according to their distance to the current frame. And Kim et al. [20] proposed to embed an attention mechanism at inputs for distant speech recognition. Moreover, acoustic signals from microphone arrays can be used to improve the robustness in distant speech recognition due to the availability of additional spatial information. Therefore, exploiting the temporal and spatial information is essential for multichannel speech recognition.

In our previous work [21], an attention mechanism across time has been embedded successfully to improve the performance of speech recognition. Some studies [22], [23] have shown that performance can be improved by supplying complementary features as inputs to the network in parallel with the regular acoustic features for speech recognition. Motivated by the above work, we propose augmenting the acoustic features from microphone array with the spatial information to further improve the performance of multichannel speech recognition. Generalized cross correlation between microphones [24] is one of the representations that encode spatial information. It is considered as auxiliary features for acoustic models. Benefiting from the spatio-temporal information, significant improvements are obtained on the AMI [25] multichanel speech recognition task.

Manuscript received December 13, 2017.
Manuscript revised March 5, 2018.
Manuscript publicized April 6, 2018.

[†]The authors are with Key Laboratory of Speech Acoustics and Content Understanding, University of Chinese Academy of Sciences, Beijng, China.
a) E-mail: zhangyu@hccl.ioa.ac.cn

DOI: 10.1587/transinf.2017EDL8268

2. Acoustic Model for Multichannel Speech Recognition

The proposed acoustic model for multichannel speech recognition is shown in Fig. 1. The attention mechanism proposed in [21] is adopted to utilize the temporal information at input layer. And GCC between microphones is supplied as spatial feature compensation. In this section, the attention mechanism and feature extraction of GCC are described respectively.

2.1 Temporal Information: Attention Mechanism

At each time step, a concatenation of features from each microphone in a microphone array is considered as one input frame for acoustic modeling. Traditionally, the input of acoustic model is comprised of L input frames in a contextual window, which results in that the temporal information within L frames is not well considered. However, the contribution from each frame to the state prediction may be diffferent. Therefore, an attention mechanism is used here, which makes the model iteratively select relevant input across a long time-scale.

As shown in Fig. 1, by scaling input x_t with attention weights α_t, a weighted representation $\hat{x}_t$ is generated for the LSTM acoustic model, which estimates the probability for context-dependent hidden Markov model (HMM) state $p(s|x_t)$. The attention weights $\boldsymbol{\alpha}_t$ enable the LSTM model to tune its attention to the input frames at each time step. The attention-based LSTM model can be described by the following equations:

$$e_t = \text{Attention}(x_t, s_{t-1}, \alpha_{t-1}) \tag{1}$$

$$\alpha_{tl} = \frac{\exp(e_{tl})}{\sum_{l=1}^{L} \exp(e_{tl})} \tag{2}$$

$$\hat{x_{tl}} = \alpha_{tl} x_{tl} \tag{3}$$

$$p(s|x_t) = \text{LSTM}(\hat{x}_t) \tag{4}$$

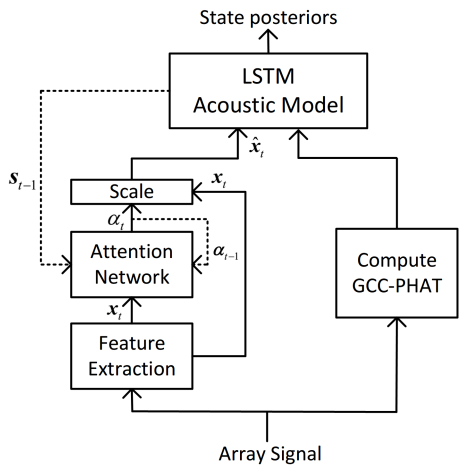

Fig. 1 Structure of acoustic model for multichannel speech recognition

where Attention() is a deep neural network that computes the attention scores e_t, and LSTM() stands for LSTM acoustic model that predicts state labels. As in Eq. (1), the attention score e_t depends on the three inputs: the input feature x_t, the prediction from previous frame s_{t-1}, and the attention weights history α_{t-1}. Equation (2) shows that the attention weights α_{tl} are normalizations of the attention scores e_{tl}. In Eq. (3), the weighted representation $\hat{x}_t$ is generated by scaling x_{tl} with the attention weight α_{tl}. Finally, the weighted representation $\hat{x}_t$ is served as input of the following LSTM acoustic model, instead of the conventional raw input x_t.

2.2 Spatial Information: Generalized Cross Correlation

Generalized cross correlation has been successfully used to determine the time delay of arrival (TDOA) of propagating waves between two spatially separated microphones. And TDOA estimated from multiple microphone pairs can be used to parameterize the source location. Hence, GCC encodes the spatial information. In this work, a generalized cross correlation with phase transform (GCC-PHAT) algorithm [26] is adopt to compute GCC between each pair of microphones due to its robustness to reverberation.

Given two channel signals $x_i(n)$ and $x_j(n)$, and their Fourier transforms $X_i(f)$ and $X_j(f)$, GCC is computed as follows:

$$\text{gcc}_{ij}(n) = \text{IFFT}\left(\frac{X_i(f) X_j^*(f)}{|X_i(f) X_j^*(f)|}\right) \tag{5}$$

where $*$ denotes the complex conjugate, and IFFT means inverse fast Fourier transform. Ideally, there exist phase differences across channel signal $x_i(n)$ and $x_j(n)$, and $\text{gcc}_{ij}(n)$ should exhibit a peak over a restricted range, which corresponds to the TDOA between microphone i and j. The separation distance of the microphones physically limits the range of valid time delays. The acoustic path length of each signal differs according to the location of the microphone, and these differences in arrival time are even greater when the space between microphones is larger. This finite range is determined by the distance between the microphones divided by the speed of sound.

In this work, our models are trained and evaluated on the AMI meeting corpus, in which an 8-microphone 10cm radius uniform circular array is used. As mentioned in [10], we also use 588-dimensional GCC vectors as auxiliary features for the neural network acoustic model. They are computed as follows. The maximum distance between any pair of microphones is 20cm. So the maximum delay between two microphones is $\tau = \frac{0.2m}{340m/s} = 0.588$ms. It corresponds to a less than 10 sample delay at a sample rate of 16kHZ. Therefore, the center 21 correlation coefficients for each microphone pair are sufficient to encode the location of the speaker. There are totally 28 microphone pairs in the 8-microphone array. On the whole, the dimension of GCC features is $21 \times 28 = 588$. It encapsulates the relevant spatial information in this vector representation.

As shown in Fig. 1, for both training and testing, the GCC features are concatenated to the weighted acoustic features $\hat{x}_t$ at each time step. Thus the neural network acoustic model is informed which speech segment comes from which location. It enables the neural network acoustic model to make better use of acoustic signals from different channels.

3. Experimental Setup

The AMI corpus contains around 100 hours of meetings speech, in which acoustic signal is captured by individual headset microphones, lapel microphones, and one or more microphone arrays. The primary microphone array data, referred to as MDM, is used in the experiments. Our models are trained and tested using this split: a training set of 80 hours, a development set and a evaluation set, each of 9 hours. Kaldi [27] is exploited for building speech recognition systems. Three LSTM layers of 1024 memory cells are used as acoustic models. 40-dimensional log-Mel filterbank features are extracted from every recording. The network is unrolled for 20 time steps for training with truncated back-propagation through time (BPTT) and acoustic models are trained with cross-entropy (CE) criterion.

For comparison, the results of single distant microphone (SDM) and traditional beamforming are also shown. Experiments with SDM make use of the first microphone of the microphone array. For the beamforming experiments, the BeamformIt toolkit [28] is adopted to implement a weighted delay-and-sum beamforming, in which GCC-PHAT is also used to compute the TDOA to create a single enhanced signal. Meeting speech recognition is challenging by speech overlapp. The overlapping segments not excluded during training stage. We show results on the full set as well as the subset that only contains the non-overlapping speech segments.

4. Results

To find the best setup for the proposed model, we explore the effect of input frame number on the attention mechanism across time first and then the size of analysis window to compute GCC. Lastly, performance comparison with baseline models is shown.

4.1 Number of Input Frames

The effect of input contexts on the attention mechanism across time is explored in this subsection. And GCC features are not used in this group of experiments. The LSTM model with attention mechanism is denoted as ALSTM. For the baseline model, the concatenation of multiple microphone outputs is used as input. The word error rate (WER) results with different input contexts are shown in Table 1.

The configuration in the second column of Table 1 stands for spliced context. For instance, splicing together frames from $t-3$ to $t+3$ at the input layer is written compactly as $[-3, 3]$. From Table 1, it can be observed that

Table 1 WER with different input contexts

Data	Input Context	LSTM		ALSTM	
		dev	*eval*	*dev*	*eval*
MDM	$[-3, 3]$	37.5%	42.4%	36.7%	41.7%
	$[-5, 5]$	37.8%	42.7%	36.0%	41.4%
	$[-7, 7]$	38.0%	43.3%	36.4%	41.5%

Table 2 WER with different window sizes of GCC

Window size (ms)	25	55	75	105	155
dev	36.6%	36.4%	35.9%	35.8%	36.5%
eval	41.7%	41.5%	41.0%	40.8%	41.5%

$[-5, 5]$ is the optimal temporal context for the attention-based LSTM model. This indicates that 11 frames at input layer are sufficient for the attention mechanism. And it achieves more than 1% absolute reduction in WER compared with the LSTM baseline model.

4.2 Analysis Window Size of GCC

The computation of GCC between each microphone pair is repeated along the recording in order to respond to changes in the location of the speaker. During this computation, a big analysis window leads to a reduction in the resolution of changes in the location of the speaker. However, using a very small analysis window reduces the robustness of the cross-correlation estimation, as less acoustic frames are used to compute it. To match the time-scale of acoustic features, GCC is also computed every 10ms. The attention mechanism across time is not considered. The LSTM acoustic models are trained directly on the concatenation of multiple microphone outputs and the corresponding GCC features. With different GCC window sizes, the WER results are summarised in Table 2. It shows that more improvements could be obtained with larger windows until 105ms.

4.3 Comparison to Baseline Models

Three baseline models are prepared: (1) training the LSTM acoustic model on the SDM data; (2) beamforming the multichannel signals into a single channel and following the standard acoustic modeling approach used for the SDM case; (3) training the LSTM acoustic model directly on the concatenation of features from each microphone in the array.

Firstly, the frame accuracies on training set and validation set of these models are shown in Fig. 2. Three pair results are given: the red is the baseline model trained on the beamformed signal, the blue is the baseline model trained on the concatenation of microphone array, and the green is the proposed model with attention mechanism and GCC features. And the overlapping segments are not excluded from the validation set and training set. It can be observed that the proposed model achieves a significantly higher frame accuracy than the baseline models, which shows that the proposed model improves the ability to model the acoustic signal from the microphone array.

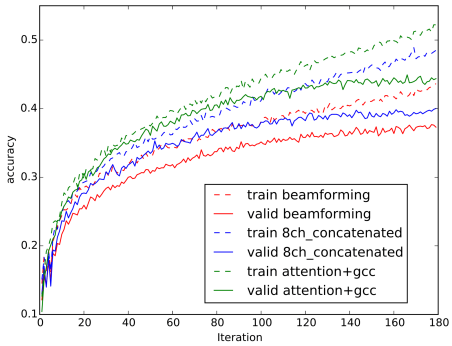

Fig. 2 Frame accuracy on the validation set and training set.

Table 3 Performance comparison on the AMI corpus

Data	Model	*dev*	*dev**	*eval*	*eval**
SDM	-	42.8%	34.3%	47.2%	38.3%
MDM	Beamforming	39.5%	30.1%	43.3%	34.0%
	Concatenated	37.8%	30.7%	42.7%	34.5%
	+ attention	36.0%	29.7%	41.4%	33.6%
	+ GCC	35.8%	29.5%	40.8%	32.9%
	+ attention + GCC	34.4%	28.7%	39.5%	31.5%

Then, recognition performances of these models are evaluated. To evaluate the effectness of spatial and temporal information, experiments are conducted with attention mechanism or GCC respectively. Table 3 shows the WER results. In the first row of Tabel 3, ones without star mean full test data, while those with star means only non-overlapping segments.

Compared MDM with SDM experiments, significant improvements are achieved by using multichannel data. It shows the benefit of additional spatial information in improving the performance of distant speech recognition. Although the beamformed model shows slight better results than raw 8-channel concatenated model on non-overlapping speech, it performs worse on the overlapping segments. That is probably because the competing acoustic source results in less accurate TDOA estimations for beamforming.

From the fifth row, the attention mechanism across time provides about a 4% relative improvement. It suggests that the temporal information is well utilized by the attention mechanism. From the next row, the GCC vectors are supplied as auxiliary features for acoustic modeling. On average 5% relative improvements are observed on the two testsets. This indicates that additional spatial information is beneficial for the neural network acoustic model and could be utilized directly by the neural network.

Finally, we combine the attention mechanism across time and the spatial feature compensation. The results are shown in the last row of Table 3. It shows that the improvements from the temporal and spatial information are combined. Compared with the 8-channel baseline in the forth row, the proposed model achieves 8.2% and 7.5% relative improvements in WER for all segments and non-overlapping segments. It performs better than the two baseline models on both the overlapping and non-overlapping segments.

5. Conclusion

In this letter, we add the spatio-temporal information into neural network acoustic modeling for multichannel speech recognition. The attention mechanism utilizes the temporal information at input layer, which makes the acoustic model focus more attention to more relevant frames. As for the spatial information, GCC vectors are supplied to acoustic models. It is used as spatial feature compensation to improve the performance of multichannel speech recognition. Experiments on the AMI corpus show that the proposed model outperforms the multiple input baseline model and beamforming baseline model.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 11590770-4, U1536117) and the National Key Research and Development Plan (Nos. 2016YFB0801203, 2016YFB0801200).

References

[1] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Trans. Audio, Speech, Language Process., vol.20, no.1, pp.30–42, 2012.
[2] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," Proc. of Interspeech, pp.338–342, 2014.
[3] H. Erdogan, T. Hayashi, J.R. Hershey, et al., "Multi-channel speech recognition: LSTMs all the way through," Proc. of CHiME 2016 Workshop, 2016.
[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," Proc. of ICASSP, pp.196–200, 2016.
[5] Y.-H. Tu, J. Du, L. Sun, F. Ma, and C.-H. Lee, "On design of robust deep models for CHiME-4 multi-channel speech recognition with multiple configurations of array microphones," Proc. of Interspeech, pp.394–398, 2017.
[6] M.L. Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," Proc. of HSCMA, pp.104–107, 2008.
[7] T.N. Sainath, R.J. Weiss, K.W. Wilson, A. Narayanan, M. Bacchiani, and Andrew, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," Proc. of ASRU, pp.30–36, 2015.
[8] T.N. Sainath, R.J. Weiss, K.W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," Proc. of ICASSP, pp.5075–5079, 2016.
[9] B. Li, T.N. Sainath, R.J. Weiss, K.W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," Proc. of Interspeech, pp.1976–1980, 2016.
[10] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M.L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," Proc. of ICASSP, pp.5745–5749, 2016.
[11] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: end-to-end training of a beamformer-supported multi-channel asr system," Proc. of ICASSP, pp.5325–5329, 2017.
[12] Z. Meng, S. Watanabe, J.R. Hershey, and H. Erdogan, "Deep long

short-term memory adaptive beamforming networks for multichannel robust speech recognition," Proc. of ICASSP, pp.271–275, 2017.
[13] T. Ochiai, S. Watanabe, T. Hori, J.R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," IEEE J. Sel. Topics Signal Process., vol.11, no.8, pp.1274–1288, 2017.
[14] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," Proc. of ASRU, pp.285–290, 2013.
[15] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," Proc. of HSCMA, pp.172–176, 2014.
[16] J. Chorowski, D. Bahdanau, D. Serdyuk, et al., "Attention-based models for speech recognition," Proc. of NIPS, pp.577–585, 2015.
[17] L. Lu, X. Zhang, K. Cho, et al., "A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition," Proc. of Interspeech, pp.3249–3253, 2015.
[18] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," Proc. of ICASSP, pp.4945–4949, 2016.
[19] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention," Proc. of Interspeech, pp.17–21, 2016.
[20] S. Kim and I. Lane, "Recurrent models for auditory attention in multi-microphone distant speech recognition," Proc. of Interspeech, pp.3838–3842, 2016.
[21] Y. Zhang, P. Zhang, and Y. Yan, "Attention-based lstm with multi-task learning for distant speech recognition," Proc. of Interspeech, pp.3857–3861, 2017.
[22] M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," Proc. of ICASSP, pp.7398–7402, 2013.
[23] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," Proc. of ASRU, pp.55–59, 2013.
[24] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Audio, Speech, Language Process., vol.24, no.4, pp.320–327, 1976.
[25] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," Language Resources and Evaluation, vol.41, no.2, pp.181–190, 2007.
[26] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," Proc. of ICASSP, pp.375–378, 1997.
[27] D. Povey, A. Ghoshal, G. Boulianne, et al., "The kaldi speech recognition toolkit," Proc. of ASRU, no. EPFL-CONF-192584, 2011.
[28] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," IEEE Trans. Audio, Speech, Language Process., vol.15, no.7, pp.2011–2022, 2007.