AN INFORMATION FUSION APPROACH TO RECOGNIZING MICROPHONE ARRAY SPEECH IN THE CHIME-3 CHALLENGE BASED ON A DEEP LEARNING FRAMEWORK

Jun Du[1], Qing Wang[1], Yan-Hui Tu[1], Xiao Bao[1], Li-Rong Dai[1], Chin-Hui Lee[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA
{jundu,lrdai}@ustc.edu.cn, {xiaosong,tuyanhui,baox}@mail.ustc.edu.cn chl@ece.gatech.edu

ABSTRACT

We present an information fusion approach to robust recognition of microphone array speech for the recently launched 3rd CHiME Challenge. It is based on a deep learning framework with a large neural network consisting of subnets with different architectures. Multiple knowledge sources are integrated via an early fusion of normalized noisy features with different beamforming techniques, speech enhanced features, speaker related features, and other auxiliary features concatenated as the input to each subnet, and a late fusion by combining the outputs of all subnets to produce one single output set. Our experiments demonstrate that all information sources are complementary in our proposed framework. Our best system achieves an average word error rate reduction of 68% from the officially released baseline results on the test set of real data.

Index Terms— CHiME Challenge, deep learning, information fusion, microphone array, robust speech recognition

1. INTRODUCTION

With the emergence of tremendously speech-enabled applications using the techniques of automatic speech recognition (ASR) in mobile internet era, the environment robustness has been one of the most critical issues to be addressed to make the system more usable. For the past several decades, many techniques [1, 2] have been proposed to handle this difficult problem. In contrast, there were not many popular benchmarks for the noise robustness issues in the past due to the lack of the good solution to the strong demand in real application which led to a bad feedback loop. One remarkable benchmark was the Aurora series initiated by Nokia in 2000, including Aurora-2 [3], Aurora-3 [4, 5, 6, 7] and Aurora-4 [8] tasks. The Aurora-2 and Aurora-4 databases were designed with artificially generated noisy data for the recognition tasks of the small vocabulary and median vocabulary, respectively. Meanwhile, the Aurora-3 task aimed to recognize the digit strings in real automobile environments.

Evolving into the mobile era and the rising popularity of the deep learning technologies, the focus on noise robustness has been reactivated by a recent ASR series of CHiME challenges [9, 10, 11] in recent years. This series differs from the Aurora tasks in several aspects. First, the scenarios are extended to far-field speech recognition in the everyday listening conditions, e.g., the family living room. Second, the room impulse responses (RIRs) simulating speaker movements and reverberation have been convolved with the utterance to generate more realistic artificial noisy data. Third, research on the distant microphone arrays based ASR is more emphasized rather than the single-microphone techniques. One main difference of the CHiME-3 challenge launched this year from the past CHiME-1 and CHiME-2 challenges is a set of real-world data is collected from several typical scenes via mobile tablet devices equipped with microphone arrays. In this sense, the CHiME-3 challenge might start a drive of new research attempting to solve ASR problems in real-world applications. Furthermore, officially released results have also indicated that the traditional approaches working well for the simulation data could be totally failed for the real data.

Following the investigations of the techniques used in previous CHiME or REVERB [12] challenges and the specificity of the CHiME-3 challenge, we propose our solution via a large neural net consisting of subnets with different architectures, namely deep neural networks (DNNs) [13] and recurrent neural networks (RNNs) [14], to combine multiple knowledge sources by early feature fusion and late model fusion. In the early fusion stage, diversified features are concatenated to boost the recognition performance. First, the concatenation of multi-channel acoustic features is investigated with each channel corresponding to one beamforming result of a channel subset in a microphone array. This is quite different from the traditional approach that one single overall output, after beamforming combining all channels of the array, is fed to the recognizer. One reason such a proposed multi-channel concatenation approach can achieve a better performance might be that it reduces the risk caused by the imperfection of the existing beamforming approaches, especially for the microphone array with many highly diverse channels.

A few issues need to be carefully considered in the proposed fusion approach. First, for multi-channel concatenation there is an increase of input layer size of DNN, which can be even larger than the size of the hidden layers and often leads to a performance degradation. To alleviate this problem, multiple frame expansion is applied to the main channel while only one central frame is used for the other channels. Second, appending multiple enhanced features is believed to be beneficial, motivated by an observation that the use of the enhanced features from the main channel alone could not improve over the noisy features on the real data possibly due to the large residual noise [11]. Furthermore, different feature normalization approaches, speaker related features, and auxiliary features are also studied in the early fusion. As for late fusion, the outputs of all subnets with different architectures are combined via a simple posterior average strategy [15] to generate one single output set for subsequent decoding. Based on our experiments, both early and late fusions are equally important and strongly complementary in terms of reducing the ASR word error rate (WER). Obviously the proposed two-stage fusion is superior to the purely early fusion or late fusion. If all the information are concatenated in the early fusion, then it is difficult to handle the issue of high-dimension in the input layer and dynamic range of different features. Similarly, if only the late fusion is used,

978-1-4799-7291-3/15/$31.00 ©2015 IEEE

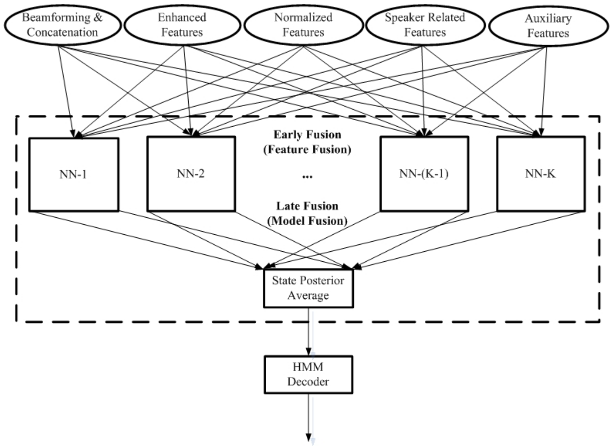

Fig. 1. System overview.

the poor performance of each subnet can be predictable.

2. SYSTEM DESCRIPTION

The overall flowchart of our proposed system is illustrated in Fig. 1. The dashed block conceptually denotes a large neural network, consisting of K subnets with different architectures. As for the input, multiple knowledge sources are exploited to generate different feature combinations. Each combination as an early fusion includes one type of multi-channel beamforming concatenations, with the enhanced features, feature normalization, speaker related features, and auxiliary features, to be elaborated in the following subsections. Each subnet is built independently with different architectures and learning methods. Finally, in recognition, the outputs of the large neural network for each frame are generated by a late fusion of all subnets in the output layer, which are then fed to a decoder with hidden Markov models (HMMs).

2.1. Early fusion

2.1.1. Beamforming and feature concatenation

Formulating a strategy to make a full use of multi-channel information of microphone array speech in the neural networks is critical to the recognition performance. The existing approaches can be divided into two broad classes, traditional beamforming to generate one single channel output for subsequent processing and channel concatenation. For example, in [16, 17], the concatenation of the noisy features in each channel of a microphone array outperforms the beamforming approach, especially for moving speech as it might preserve the signals from all directions. In [18], the beamformed features concatenated with the noisy features from the main channel of the microphone array yield a better recognition performance. In our current study, multiple sets of beamforming results are concatenated. Each beamformed result is generated on a subnet of channels in the microphone array. As for beamforming, two approaches are investigated. One approach is waveform averaging of specified channels, which is a special case of beamforming but robust to moving speech, denoted as **Avg** in Table 1 below. The other approach is generalized sidelobe canceller (GSC) [19, 20] based on a relative transfer function [21].

2.1.2. Enhanced features

To demonstrate the effectiveness of the enhanced features (denoted as **Enh**) combined with the beamforming concatenation, we use the officially provided approaches [11]. The source localization technique in [22] is used to track the target speaker while the speech signal is estimated by time-varying minimum variance distortionless response (MVDR) beamforming with diagonal loading [23].

2.1.3. Normalized features

Utterance-based feature normalization is a widely used technique for ASR systems to eliminate the effect of the possible irrelevant variabilities, including speaker variability, background noises and channel distortions. Two normalization approaches, namely mean normalization (denoted as **MN** in Table 1) and mean variance normalization (denoted as **MVN** in Table 1) are applied to the acoustic features. MVN is more effective for the additive noises especially in low SNRs while MN is more stable for the high SNR cases.

2.1.4. Speaker related features

Similar to [24], the i-vectors (denoted as **iVec** in Table 1) to represent some speaker information are extracted via the standard procedure [25, 26] as the parallel features fed to the input layer of neural nets. The main advantage of this speaker adaptation approach is that the architecture of neural net remains unchanged and it is unnecessary to perform the first-pass decoding. Inspired by the beamforming concatenation, the multi-channel i-vectors are also extracted corresponding to each beamforming results, which is verified more effective than the single-channel i-vector. Note that for both training and testing, the i-vector is estimated based on the utterances of one single speaker and only changed across different speakers.

2.1.5. Auxiliary features

Besides the commonly used log Mel-filterbank (LMFB) features, other auxiliary features are also adopted. One feature set is the pitch and probability-of-voicing features proposed in [27], which are tuned for the ASR systems. It is believed that those features not only give large improvements for tonal language recognition but also yield remarkable gains for non-tonal languages which is also confirmed on our task. The other set is the cochleagram (CG) features well verified for ASR [28]. In our experiments, the pitch related features are always concatenated with the LMFB features while the CG features are optionally used.

As mentioned above, in early fusion, diversified features are concatenated together. One issue is to control the input feature dimension to avoid a possible performance degradation. Suppose the dimension of the basic acoustic features is D_1 and the size of acoustic context is τ frames. The number of channels after beamforming is M. The dimensions of the i-vector and auxiliary features are D_2 and D_3, respectively. Then the final dimension for the input feature vector is $D_1 * \tau + M * D_1 + M * D_2 + M * D_3$, which means the acoustic context expansion is only applied to the main channel of the basic acoustic features.

2.2. Neural network training

Three types of neural nets are adopted as subnets, namely DNN, long short-term memory (LSTM) based RNN [29], and bi-directional LSTM (BLSTM) based RNN [30]. Before neural network training, the state labels should be generated by the forced-alignment via a

Table 1. Description of 12 subsystems.

	Feature fusion	Feature dimension	NN type	NN training
S1	MN(Avg1+Avg2+C2+Enh)+iVec1+iVec2	1804	DNN	CE+ReFA+sMBR
S2	MN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	1804	DNN	CE+ReFA+sMBR
S3	MVN(Avg1+Avg2+C2+Enh+CG1+CG2)+iVec1+iVec2	1864	DNN	CE+ReFA+sMBR
S4	MVN(GSC1+GSC2+GSC3+Enh+CG1+CG2)+iVec1+iVec2	1864	DNN	CE+ReFA+sMBR
S5	MN(Avg1+Avg2+C2+Enh)+iVec1+iVec2	544	LSTM-RNN	CE+ReFA
S6	MN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	544	LSTM-RNN	CE
S7	MVN(Avg1+Avg2+C2+Enh)+iVec1+iVec2	544	LSTM-RNN	CE+ReFA
S8	MVN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	544	LSTM-RNN	CE
S9	MN(Avg1+Avg2+C2+Enh)+iVec1+iVec2	544	BLSTM-RNN	CE+ReFA
S10	MN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	544	BLSTM-RNN	CE
S11	MVN(Avg1+Avg2+C2+Enh)+iVec1+iVec2	544	BLSTM-RNN	CE+ReFA
S12	MVN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	544	BLSTM-RNN	CE

state-of-the-art system with Gaussian mixture continuous density HMMs (GMM-HMMs) [31]. The only difference is the use of multi-channel concatenation of acoustic features after waveform average beamforming. This set of state labels is used for the training of all subnets. For the DNN training, the Kaldi recipe for CHiME-2 challenge [32] is adopted with the standard procedure, namely the pre-training using restricted Boltzmann machines plus the cross entropy (CE) training. And the DNN can be refined by re-alignment (ReFA) and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion [13]. As for the training of LSTM-RNN or BLSTM-RNN, the CE training [29] and ReFA are adopted with the truncated backpropagation through time (BPTT) learning algorithm to update the model parameters.

2.3. Late fusion

For all K subnets, the outputs share the same tied state set from the HMM topology of the GMM-HMM or DNN-HMM system. So late fusion can be implemented by a simple strategy of state posterior averaging in the output layers [15]. This approach has been verified to be more effective than lattice fusion or ROVER [33] in our experiment, which is reasonable as fusion at the frame level (state level) is with a higher resolution than fusion at the text-level and not affected by the language model. Back to Fig. 1, if we treat the early fusion and late fusion as the internal operations of the large neural net in dashed box, then the input might be a high-dimensional vector with diversified features from multiple knowledge sources while the output is the normal state posterior representation.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

The CHiME-3 challenge was designed to focus on a real-world and commercially motivated scenarios that a person talking to a mobile tablet device in a variety of real and challengingly public noisy conditions [11]. Four environments were selected, namely café (CAF), street junction (STR), public transport (BUS) and pedestrian area (PED). For each environment, both the real and simulated noisy speech data were provided. The real data consists of 6-channel recordings of sentences from the WSJ0 corpus [34] spoken in four environments. The simulated data was constructed by mixing clean utterances with the background noises [35]. The training set contained 1600 real noisy utterances from the combinations of four speakers and four environments with 100 utterances for each combination, and 7138 simulated utterances. The development set and test set included the same 410 and 330 utterances as in the WSJ0 task. In each environment the set was split into four random partitions and each was assigned to a different talker. This resulted in 1640 (410×4) and 1320 (330×4) real utterances for the development set and test set, respectively. Similarly, the simulated data could be generated for the development and test sets. Other details can be found in [11].

For the GMM-HMM systems, the 182-dimensional feature vector consisted of 13-dimensional Mel-frequency cepstral coefficients (MFCCs) with 7-frame context expansion and 2-channel concatenation. The two sets were formed by waveform averaging of the original channels (4,5,6) and channels (1,3). The number of tied states was 1965 which was also the size of the output layer for all subnets. And a total of 15019 Gaussians were used. Other settings were the same as in the Kaldi recipe [31, 11].

As for the input features to the neural network, 40-dimensional LMFB features and 2-dimensional pitch features with their first-order and second-order derivatives were formed as the basic acoustic features (D_1=126). The dimension of i-vector for each channel was set to 20 (D_2=20). And the dimension of CG features for each channel was 30 (D_3=30). The acoustic context size τ was 11. For the DNN architecture, 7 hidden layers with 2048 neurons for each layer were used. The other parameters of DNN can refer to the Kaldi recipe [11]. For both LSTM-RNN and BLSTM-RNN, 3 recurrent layers were used. For LSTM-RNN, the number of memory cells was 2048 while the size of the recurrent projection layer was 512. For BLSTM-RNN, the number of memory cells was 1024 while the size of the recurrent projection layer was 800 for both forward and backward layers.

In Table 1, the K (K=12) subsystems corresponding to the K subnets before late fusion in Fig. 1 are listed with the detailed configurations. For example, system S1 adopted the waveform average beamforming where Avg1 and Avg2 denoted the average of channels (4,5,6) and (1,3), respectively. And the channel 2 (C2) was also concatenated. This design was inspired by the positions of the 6 channels on the tablet. Mean normalization (MN) was applied to Avg1, Avg2, C2 and enhancement (Enh) features. iVec1 and iVec2 are the i-vectors corresponding to Avg1 and Avg2, respectively. The dimension of the input feature vectors for S1 was 1804. For S2, GSC1, GSC2, GSC3 represented the GSC-based beamforming of channel combinations (4,5), (1,3), (2,5), respectively.

Table 2. WER (%) comparison of different stages of the early fusion for S1 system on the development and test sets of real data.

System	BUS	CAF	PED	STR	Avg.
Development set					
Baseline(C5)	20.93	12.89	9.18	13.58	14.15
Avg1+Avg2	16.74	11.83	7.79	11.58	11.98
+C2	15.74	11.52	7.92	11.47	11.66
+Enh	14.35	10.06	7.76	10.50	10.67
+iVec1+iVec2	12.33	9.45	6.83	10.37	9.75
+ReFA	11.7	9.00	6.86	9.76	9.33
+sMBR	10.87	7.92	6.14	8.88	8.45
Test set					
Baseline(C5)	34.77	26.24	20.76	16.23	24.50
Avg1+Avg2	28.04	22.23	17.30	13.54	20.28
+C2	27.28	21.22	17.28	12.87	19.66
+Enh	25.31	21.26	15.88	12.27	18.68
+iVec1+iVec2	22.79	20.02	15.13	11.60	17.39
+ReFA	21.33	18.88	14.78	11.32	16.58
+sMBR	19.09	16.74	13.19	10.53	14.89

Table 3. WER (%) comparison of different stages of the early fusion for S2 system on the development and test sets of real data..

System	BUS	CAF	PED	STR	Avg.
Development set					
GSC1	16.92	10.44	7.57	11.28	11.55
+GSC2	16.37	9.73	7.23	10.57	10.98
+GSC3	15.96	9.97	7.26	10.6	10.95
+Enh	14.62	9.29	7.23	9.97	10.28
+iVec1+iVec2	13.57	8.83	6.9	10.09	9.85
+ReFA	13.39	8.45	6.9	9.95	9.68
+sMBR	12.16	8.14	6.12	8.64	8.77
Test set					
GSC1	27.16	21.78	17.12	13.34	19.85
+GSC2	25.82	20.94	15.98	12.83	18.89
+GSC3	25.69	20.86	15.21	12.63	18.59
+Enh	24.27	21.22	15.56	12.29	18.33
+iVec1+iVec2	22.36	20.30	14.52	11.58	17.19
+ReFA	22.56	19.91	14.87	11.6	17.24
+sMBR	20.02	17.26	12.91	10.4	15.15

3.2. Experimental Results

3.2.1. Early fusion

First, the experiments on early fusion are shown in Table 2 and 3. Table 2 gives a WER comparison at different stages of early fusion for the S1 system on the development and test sets of real data. "Baseline(C5)" denoted the baseline DNN system using the speech data of channel 5 and CE training. Our proposed beamforming and concatenation system "Avg1+Avg2" consistently outperformed the baseline system for all testing cases, e.g., relative WER reductions of 15.3% and 17.2% were achieved for the development and test sets in average. Then by appending of the C2 features, the recognition performance was slightly improved. More interestingly, the concatenation with the enhanced features brought about an absolute 1% WER reduction for both the development and test sets. In contrast, according to the officially released preliminary results [11], no gain was observed by the use of the enhanced features only. This indicated the necessity of the parallel beamformed and enhanced features, which might be strongly complementary. Furthermore, the additional i-vector features gave remarkable gains which demonstrated the effectiveness of this speaker adapted features. As for DNN training, ReFA and sMBR could consistently reduce the WER. Overall, the relative WER reductions of 40.3% and 39.2% were yielded from the baseline system for the development and test sets, respectively. By considering that the test set was more difficult than the development set, these similarly relative improvements showed a generalization ability of our proposed early fusion.

Table 3 lists a WER comparison at different stages of early fusion for the S2 system on the development and test sets of real data. Similar observations to those for S1 as in Table 2 could be made. The main difference from S1 to S2 was the use of GSC-based beamforming. It was interesting that in the stage of pure beamforming concatenation, GSC-based approach outperformed the waveform average approach, e.g., average WER from 19.66% to 18.59% on the test set. However, by the performance comparison of the final systems, S1 and S2, we could make an opposite observation that the waveform average was slightly better than GSC which implied that the simple average operation in the time domain was a more robust beamforming approach. Finally, both S1 and S2 gave significant gains over the baseline system and each feature set in the early fusion stage made a contribution in reducing the WER.

3.2.2. Late fusion

Before late fusion, the recognition performances of the 12 subsystems are shown in Table 4. Clearly, no single subsystem could achieve the best performance for all environments, even in one subset, e.g. the development set with real data. For the test set with real data, S11 achieved the best performance in average but still not the best for each of four environments. And there were 8 subsystems with one best performance case at least. Those observations delivered important messages. On one hand, the noise statistics should be quite different in four environments. Each subsystem with one feature combination could not well handle all the noise conditions. On the other hand, all the subsystems might be complementary, which was one key motivation of our late fusion strategy.

Table 5 illustrates a WER comparison of different combinations in late fusion on the development and test sets of the real and simulated data. We designed the fusion experiments from two aspects, namely fusion of different neural networks with the fixed input feature combination and fusion of different inputs with the fixed type of neural networks. From the results of F(1,5,9), F(2,6,10), F(3,7,11), F(4,8,12), significant improvements were achieved by fusing different architectures (DNN, LSTM-RNN, BLSTM-RNN), e.g., WER on the real test data was reduced from 14.82% in the best single subsystem to 12.1% in F(3,7,11) in average, indicating that learning different architectures could help each other in predicting the state posteriors at the output layer. With the fixed neural network type, the improvements by fusing different feature inputs were also significant on the real data. What's interesting was the small dynamic range of the WERs for the first 7 fusion systems which could be potentially boosted with further fusion. So we fused all 12 subsystems in F(1-12), and a relative WER reduction of 28.8% was obtained from the best single subsystem on the real test data. Finally, the F(1-12) system consistently achieves the best results for both development and test set of real data and this observation could be also applied for the most of best systems on the simulation data.

Table 4. WER (%) comparison of 12 subsystems on the development and test sets of real and simulated data.

			S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Dev	Simu	BUS	6.64	5.91	7.24	6.52	7.73	6.80	7.18	6.50	6.78	**5.80**	6.73	6.27
		CAF	9.09	9.23	10.27	10.04	10.15	10.41	9.31	10.03	9.40	9.20	**8.97**	9.31
		PED	6.17	**5.72**	6.90	6.39	6.90	6.42	6.68	6.65	6.45	6.09	6.80	6.25
		STR	8.33	6.96	9.01	7.29	9.07	8.16	8.45	7.57	7.82	**6.90**	7.74	7.52
		Avg.	7.56	**6.96**	8.36	7.56	8.46	7.95	7.91	7.69	7.61	7.00	7.56	7.34
	Real	BUS	10.87	12.16	12.16	**7.08**	13.26	14.43	13.69	13.69	11.37	12.45	11.77	12.17
		CAF	**7.92**	8.14	8.83	10.12	10.38	9.75	10.31	9.73	8.75	8.63	8.88	8.81
		PED	6.14	6.12	**5.94**	8.48	7.57	7.80	7.60	8.16	6.73	7.17	7.15	6.80
		STR	8.88	8.64	**7.47**	10.31	10.34	10.50	9.92	10.29	8.94	9.05	9.03	8.98
		Avg.	**8.45**	8.77	9.10	9.00	10.39	10.62	10.38	10.47	8.95	9.33	9.21	9.19
Test	Simu	BUS	7.68	6.74	7.45	13.32	7.55	7.71	7.13	6.71	7.10	7.27	**6.35**	6.37
		CAF	11.60	9.51	11.00	**9.17**	11.80	12.03	9.92	9.88	11.09	10.96	9.3	9.56
		PED	11.58	8.37	10.96	**6.37**	11.21	10.70	9.58	8.87	10.24	9.77	8.59	8.74
		STR	11.52	9.21	11.94	**8.98**	11.75	10.89	11.15	10.09	11.04	9.97	11.11	9.60
		Avg.	10.59	**8.46**	10.34	9.46	10.58	10.33	9.45	8.89	9.87	9.49	8.84	8.57
	Real	BUS	**19.09**	20.02	23.35	25.24	21.84	23.72	22.06	22.43	19.48	22.99	19.41	21.07
		CAF	16.74	17.26	19.78	20.58	18.57	19.74	19.05	19.44	**16.14**	17.59	16.64	17.69
		PED	13.19	12.91	14.69	14.07	16.11	16.42	15.13	15.53	14.52	14.41	13.19	**12.65**
		STR	10.53	10.40	11.90	12.57	11.06	12.18	11.28	11.09	**9.92**	10.96	10.05	10.16
		Avg.	14.89	15.15	17.43	18.11	16.89	18.02	16.99	17.12	15.01	16.49	**14.82**	15.39

4. CONCLUSION

For the recently launched CHiME-3 challenge, we propose to integrate multiple knowledge sources denoted by multiple feature sets into the neural nets with different architectures. The early fusion is adopted as a local feature concatenation while the late fusion acts as the model average. The use of both fusions can reduce about two-thirds of WER over the officially released baseline results [11].

5. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants No. 61305002. The authors would like to thank other colleagues, Tian Gao, Yong Xu, Na-Na Fan and Shi-Xue Wen for their contributions on brainstorming and experiments.

6. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, Vol. 16, No. 3, pp. 261-291, 1995.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, Vol. 22, No. 4, pp. 745-777, 2014.

[3] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, 2000, pp.181-188.

[4] Aurora document AU/217/99, "Availability of Finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation," Nokia, Nov. 1999.

[5] Aurora document AU/271/00, "Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: description and baseline results," UPC, Nov. 2000.

[6] Aurora document AU/273/00, "Description and baseline results for the subset of the SpeechDat-Car German database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation," Texas Instruments, Dec. 2001.

[7] Aurora document AU/378/01, "Danish SpeechDat-Car digits database for ETSI STQ-Aurora advanced DSR," Aalborg University, Jan. 2001.

[8] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, version 2.0," ETSI STQ-Aurora DSR Working Group, Nov. 2002.

[9] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, Vol. 27, No. 3, pp. 621-633, 2013.

[10] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: datasets, tasks and baselines," *Proc. ICASSP*, 2013, pp. 126-130.

[11] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," *Submitted to IEEE Workshop on ASRU*, 2015.

[12] K. Kinoshita, "Summary of the REVERB challenge," *The workshop on REVERB challenge*, 2014.

[13] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *Proc. INTERSPEECH*, 2013, pp. 2345-2349.

[14] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. Thesis, University of Toronto, 2012.

[15] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," *IEEE Wordshop on ASRU*, 2013, pp. 279-284.

Table 5. WER (%) comparison of different combinations in the late fusion on the development and test sets of real and simulated data.

			F(1,5,9)	F(2,6,10)	F(3,7,11)	F(4,8,12)	F(1-4)	F(5-8)	F(9-12)	F(1-12)
Dev	Simu	BUS	5.53	4.97	5.38	**4.82**	5.50	5.16	5.18	5.07
		CAF	7.98	7.96	7.61	8.05	8.10	8.17	8.01	**6.95**
		PED	5.41	5.22	5.44	5.18	4.99	5.49	5.49	**4.73**
		STR	6.61	5.80	6.36	5.97	6.11	5.87	6.05	**5.62**
		Avg.	6.38	5.99	6.2	6.01	6.17	6.17	6.18	**5.59**
	Real	BUS	9.2	10.33	9.81	10.86	10.05	10.31	10.08	**8.76**
		CAF	7.26	7.27	7.02	7.17	7.39	7.37	7.01	**6.37**
		PED	5.66	6.02	5.68	5.99	5.4	5.74	6	**5.03**
		STR	7.46	8.23	7.27	7.92	7.71	7.26	7.34	**6.44**
		Avg.	7.4	7.96	7.44	7.89	7.64	7.67	7.61	**6.65**
Test	Simu	BUS	5.88	5.83	5.49	5.70	6.33	6.11	5.79	**5.30**
		CAF	9.25	9.11	7.86	8.03	8.52	8.91	8.46	**7.71**
		PED	8.69	8.07	7.38	6.99	8.11	7.94	7.6	**6.82**
		STR	8.91	**7.77**	9.02	8.03	9.00	8.31	8.74	7.96
		Avg.	8.19	7.70	7.44	7.20	7.99	7.82	7.65	**6.95**
	Real	BUS	15.87	17.74	16.04	17.46	16.79	17.76	16.3	**13.78**
		CAF	13	13.52	13.35	14.21	14.34	14.51	13.99	**11.36**
		PED	11.53	11.53	10.69	10.63	10.71	12.31	11.15	**9.30**
		STR	8.5	9.1	8.33	9.02	9.17	9.34	9.02	**7.77**
		Avg.	12.22	12.97	12.1	12.83	12.75	13.48	12.62	**10.55**

[16] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," *Proc. ICASSP*, 2014, pp. 5542-5546.

[17] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," *The Joint Workshop on HSCMA*, 2014.

[18] 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. on Audio, Speech, and Language Processing*, Vol. 22, No. 12, pp. 2244-2255, 2014.

[19] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagation*, Vol. 30, No. 1, pp. 27-34, 1982.

[20] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 12, No. 6, pp. 561-571, 2004.

[21] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 17, No. 4, pp. 546-555, 2009.

[22] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," *Proc. LVA/ICA*, 2010, pp. 41-48.

[23] X. Mestre and M. Lagunas, "On diagonal loading for minimum variance beamformers," *Proc. ISSPIT*, 2003, pp. 459-462.

[24] G. Saon, H. Soltau, D. Nahamoon, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," *IEEE Workshop on ASRU*, 2013, pp. 55-59.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 19, No. 4, pp. 788-798, 2011.

[26] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "Simplification and optimization of i-vector extraction," *Proc. ICASSP*, 2011, pp. 4516-4519.

[27] P. Ghahremani, B. BabaAli, and D. Povey, "A pitch extraction algorithm tuned for automatic speech recognition," *Proc. ICASSP*, 2014, pp. 2513-2517.

[28] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," *Proc. ICASSP*, 2014, pp. 7089-7093.

[29] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Proc. INTERSPEECH*, 2014, pp. 338-342.

[30] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *Proc. ICASSP*, 2013, pp. 6645-6649.

[31] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME Challenge Benchmark," *The 2nd International Workshop on Machine Listening in Multisource Environments*, 2013, pp. 19-24.

[32] C. Weng, D. Yu, S. Watanabe, and B.-W. Juang, "Recurrent deep neural networks for robust speech recognition," *Proc. ICASSP*, 2014, pp. 5569-5572.

[33] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," *IEEE Workshop on ASRU*, 1997, pp. 347-352.

[34] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSRI (WSJ0) Complete," Linguistic Data Consortium, Philadelphia, 2007.

[35] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, Vol. 87, No. 8, pp. 1933-1950, 2007.