ON REDUCING THE EFFECT OF SPEAKER OVERLAP FOR CHIME-5

Cătălin Zorilă and Rama Doddipatla

Toshiba Cambridge Research Laboratory, United Kingdom
{catalin.zorila,rama.doddipatla}@crl.toshiba.co.uk

ABSTRACT

The CHiME-5 speech separation and recognition challenge was recently shown to pose a difficult task for the current automatic speech recognition systems. Speaker overlap was one of the main difficulties of the challenge. The presence of noise, reverberation and the moving speakers have made the traditional source separation methods ineffective in improving the recognition accuracy. In this paper we have explored several enhancement strategies aimed to reduce the effect of speaker overlap for CHiME-5 without performing source separation. One is based on discarding the overlap segments using the speaker diarisation information from the challenge, another one is a neural network driven automatic gain control enhancement aimed to improve the previous speaker diarisation information, and the last one is based on optimal multi-array data selection. State-of-the-art acoustic models were used to perform the ASR experiments. Results have shown that proposed automatic gain control method yields word error rate (WER) reductions between 2% and 3% absolute on the development set of CHiME-5.

Index Terms— automatic speech recognition, CHiME-5, speaker overlap

1. INTRODUCTION

Modern automatic speech recognition (ASR) systems have already achieved recognition accuracies on clean data on a par with humans [1]. However, speaker overlap, noise and reverberation still pose serious challenges for ASR. This was recently proven in the latest CHiME speech separation and recognition challenge (CHiME-5) [2]. CHiME-5 challenge was primarily focused on the problem of distant multi-microphone conversational speech recognition in everyday home environments.

The corpus was made of 20 dinner party recordings lasting about two hours each. 16 out of the 20 recordings were used for training, two were used for development (dev) and the other two for evaluation. Each dinner party had four participants and it was divided into three phases depending on where the speakers were located: kitchen, dining or living room. Recordings were made using six Microsoft Kinect devices, which were referred to as arrays or distant sensors in our discussion below. There were two Kinects in each location and each device had four linearly distributed microphones. Additionally, each participant also had worn in-ear binaural microphones that were used to facilitate transcription. The worn and array sets were recorded asynchronously, therefore they had to be synchronized offline using a correlation based approach. Depending on how many arrays were available to decode the test data, the challenge had a single-array track and a multiple-array track. The baseline CHiME-5 system achieved 81.3% WER on the single-array track. More details about the challenge can be found in [2].

One of the main challenges of CHiME-5 is the speaker overlap. The average percentage of speech frames where two or more speakers are active at the same time is 23.8% (standard deviation 9.5%) for the training set, and 42.1% (standard deviation 10.7%) for the development set. Traditional source separation algorithms used to cope with the speaker overlap problem were proven ineffective due to the moving speakers, background noise, reverberation and lack of spatial resolution. The participants at the challenge managed to achieve WER improvements by combining data enhancement, advanced acoustic modeling and data augmentation, but the general consensus is that the overlap problem in CHiME-5 is not yet solved.

Du et al. [3] have trained a speech separation model using a two layer bi-directional long short-term memory (BLSTM) network using non-overlapping regions from each speaker, which were then mixed together to generate speaker-dependent training data. This approach had two stages and it was part of a pipeline consisting of other data enhancement, such as dereverberation, denoising or beamforming. Kanda et al. [4] have developed a minimum variance distortion-less response (MVDR) beamformer where the speaker adaptive masks were estimated using neural networks or complex Gaussian mixture models. Medennikov et al [5] have explored adaptation using a frame-level mask of a target speaker. Kitza et al [6] have proposed an algorithm based on the complex Angular Central Gaussian Mixture Model that exploited the time annotations to perform a guided source separation. Doddipatla et al. [7] have proposed a speaker dependent generalized eigenvalue (GEV) beamformer [8] to cope with the noise and the overlapped speakers. All speaker-dependent systems exploited the speaker diarisation information provided by the organizers for both train and test data.

In [7], we have explored a neural network supported automatic gain control (AGC) mechanism for improving the baseline speaker diarisation by suppressing the interfering speakers. There, the enhancement was applied only during the test phase. In this paper, we expand on this idea and provide a detailed study on this method in the context of CHiME-5 task. We have explored three enhancement strategies aimed to reduce the effect of speaker overlap without performing source separation. First approach has exploited the baseline CHiME-5 diarisation to detect the segments where only one speaker is active, the second one has aimed to improve the former method by using the AGC technique recently proposed in [7], and the last one has aimed to reduce further the effect of interfering speakers by algorithmically selecting the Kinect device from where the data should be extracted. In-depth ASR evaluations on the individual and combined enhancements were performed using state-of-the-art acoustic models (AMs). Both matched and mismatch scenarios were considered. The results presented in this paper are based on the development set of CHiME-5 only.

The reminder of this paper is organized as follows. Section 2 details the enhancement methods used for speaker overlap suppression, the experimental setup is described in Section 3, and the results and

discussion are presented in Section 4. Finally, Section 5 concludes the paper.

2. ENHANCEMENT FOR SPEAKER OVERLAP SUPPRESSION

Speaker overlap has a negative impact on the ASR performance during both training and testing. Three approaches were explored to reduce the effect of speaker overlap in the context of CHiME-5 without performing source separation. One was a hard speaker overlap suppression using the baseline speaker diarisation from the challenge, another one was based on a soft overlap suppression using frame-wise masks estimated from a speaker-dependent neural network model, and the last approach was based on multi-array data selection for reducing the interfering speech.

2.1. Hard speaker overlap suppression

Speaker diarisation information for both train and test data are available for CHiME-5. Based on the speaker label time stamps, binary masks were estimated, where the unity gain was used when there was only one active speaker and the zero gain was applied otherwise. The time resolution for the binary mask was 16-ms. Short-term speech frames were extracted at the same rate, re-scaled according to the mask value and then overlap-and-added to synthesize the modified waveform. This enhancement is referred to as hard overlap suppression (HOS).

2.2. Soft speaker overlap suppression using AGC

The accuracy of the baseline speaker diarisation is limited, therefore HOS may discard useful information. For instance, an audio segment may be marked as overlapped, although the interfering and target speakers take turns. To address this limitation, we have proposed a soft overlap suppression that would highlight the target speaker and attenuate the interfering ones. An automatic gain control mechanism is driven by neural network predictions to identify the frames where the target speaker is dominant. The block diagram of the proposed algorithm is depicted in Fig. 1.

The main component of proposed system is a deep neural network (DNN) designed to perform frame-wise speaker classification. The DNN was trained on HOS (non-overlapped) data. Input features were 24-dimensions MFCCs with delta and delta-delta, extracted using a Hann window of length 32-ms at every 16-ms. The DNN had three hidden layers with 1024, 512 and 256 hidden nodes, respectively, and sigmoid activations. The output layer had 4 nodes, same as the number of speakers, and the labels were one-hot vectors. The training of the DNN was performed in TensorFlow using the softmax cross entropy cost function and ADAM optimizer (learning rate 1e-4). The batch size was 50 and the training was stopped after 15 iterations. For the proposed AGC enhancement, a separate DNN was trained for each dinner party.

After DNN training, predictions were made on all speech frames from each recoding. The maximum likelihood criterion was used to decide the dominant speaker in each frame. Frames where the dominant speaker was the target speaker had gain one, the other frames had gain 0.001. Postfiltering was applied on the prediction vector to reduce the noise and to smooth the transitions between frames. The noise reduction was performed with a 11 taps median filter, and the smoothing effect was achieved with a double exponential moving average filter whose attack constant was 0.1 and release constant was 0.98. Finally, corrections were applied on top of the previous gains to avoid the suppression of frames where there is no speaker overlap.

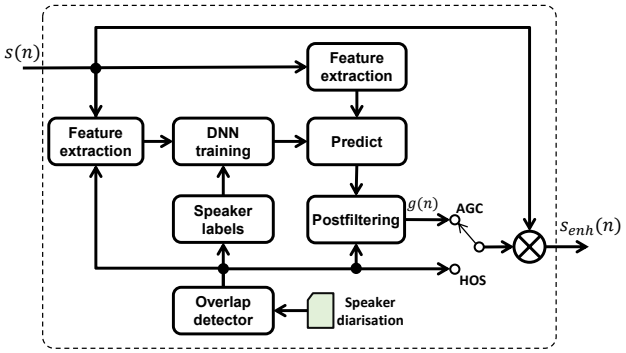

Fig. 1. Block diagram of the DNN-based automatic gain control system for soft speaker overlap suppression.

2.3. Data selection using correlation analysis

The baseline acoustic model of CHiME-5 was trained using data from the worn microphones and data from randomly chosen arrays. However, randomly selecting the arrays from where to extract speech is not optimal since, at a given time, a particular array may be able to pick up a cleaner signal for the target speaker than the other arrays. Using standard beamforming for this task is not trivial in the context of CHiME-5 due to audio synchronization errors, speaker overlap, background noise and reverberation.

An alternative approach was followed inspired by the correlation analysis used to synchronize the initial recordings [2]. The method aimed to find the Kinect having the strongest correlation peak with the worn microphone of target speaker. The normalized correlation coefficient was used as a metric. The search interval was the length of the segment plus an additional time guard of one second on either ends. The left and right channels of the worn microphone recordings were mixed together for the correlation analysis, and only the first channel from each Kinect device was used. This method is referred to as data selection (DTS).

3. EVALUATION SETUP

A conventional (not end-to-end) ASR architecture was used to perform the experiments. ASR training and decoding were performed in Kaldi [9].

3.1. Data

The baseline AM was trained using worn data and 100k randomly chosen segments from the Kinects. The worn data were unprocessed and the same for all experiments. The array data were either unprocessed (baseline), or enhanced using individual HOS, AGC or DTS modifications, or combinations of enhancements (e.g., DTS+AGC). Less than 80k segments were generated using DTS.

Regarding the test data, as mentioned in the Introduction, the evaluation was performed on the development set of CHiME-5. The Kinect data of the development set were all previously enhanced using a weighted delay-and-sum beamformer (BF, BeamformIt [10]). On top of BF, additional enhancements were applied (e.g., BF+AGC)

3.2. Front-end

13-dimensions MFCCs and the standard HMM/GMM recipe of CHiME-5 were used to train the alignment model. The acoustic features for the ASR acoustic model (AM) were 40-dimensions MFCCs. All features were normalized in mean at the segment level.

3.3. Acoustic model

Several acoustic models were tested. To reduce the turn around time for the results, the initial experiments were performed with a simpler lattice-free maximum mutual information (LF-MMI) time-delay neural network (TDNN) AM [11]. Two experimental configurations were used for this model. TDNN-A did not use data cleaning for training, neither i-vectors nor speed perturbation [12]. TDNN-B did use data cleaning and i-vectors, but no speed perturbation. The dimension of the i-vectors was 100.

Two more advanced AMs recently used for CHiME-5 [7] were also tested. One was based on convolutional neural networks (CNNs) [13] in combination with uni-directional long short-term memory networks (LSTMs) [14]. The CNN-LSTM AM consisted of two CNN layers followed by three LSTM layers. The CNN layers were 2D with 256 and 128 filters, respectively (3x3 filter kernels). Each LSTM layer had a cell dimension of 1024. The i-vectors were bypassed from the CNNs directly to the LSTMs. The other advanced AM was similar with the previous one, but it used bi-directional LSTM (BLSTM) layers instead of uni-directional ones. Both CNN-LSTM and CNN-BLSTM AMs used data cleaning, i-vectors and speed perturbation. In all cases, the standard 3-gram language model of CHiME-5 was used for decoding. A summary of the AMs used for evaluation is shown in Table 1.

Table 1. Acoustic models used for the evaluation.

AM	Configuration		
	cleaning	i-vectors	SP
TDNN-A	-	-	-
TDNN-B	+	+	-
CNN-LSTM	+	+	+
CNN-BLSTM	+	+	+

4. RESULTS & DISCUSSION

4.1. Single array track

The WER results of the experiments using TDNN-A AM are shown in Table 2. Three separate TDNN-A AMs were trained using either unprocessed data (baseline), or data enhanced with AGC or HOS processing. The baseline training set had the most amount of speaker overlap, while the HOS train set had the least amount of speaker overlap.

Column-wise, the WERs are progressively decreasing from top to bottom, confirming that the AMs trained with data having less speaker overlap is better. Remarkably, decoding the HOS test data on the first two AMs has yielded a sharp increase of the WER, but not for the last AM (HOS). This indicates that the HOS processing has introduced a large mismatch with the first two train sets, while in the latter case the train and test sets are matched, therefore the accuracy is improved. Since we are interested in demonstrating the effect of AGC with the same AM, the results on the HOS dev data have been omitted for the rest of the paper.

Table 2. Recognition accuracy in WER(%) using the TDNN-A acoustic model (single array track).

Train data	Dev data enh.		
	BF	BF+HOS	BF+AGC
Baseline	88.3	98.5	88.2
AGC	87.9	97.1	86.6
HOS	87.2	85.0	85.6

The results in Table 2 show that AGC improves the robustness of the baseline AM when the enhancement is applied on the train data, and it yields lower WERs when applied on the test data. One may notice the relatively modest improvement for the baseline TDNN-A AM with AGC enhancement. However, since AGC is aiming to isolate the target speaker from the interfering background, it is able to provide complementary information that can improve the overall WER. To test this hypothesis, we have combined the lattices of the systems with and without AGC (Table 3). As shown in Table 3, the performance of the combined system (A+B) was significantly better than the performance of the baseline system (A), confirming the hypothesis above.

Table 3. System combination with the TDNN-A acoustic model (single array track).

Train data	Dev data enh.		
	BF (A)	BF+AGC (B)	A+B
Baseline	88.3	88.2	86.0
AGC	87.9	86.6	84.8
HOS	87.2	85.6	83.3

A similar analysis was performed using more advanced acoustic models (Table 4). In this case the AMs were trained with unprocessed data only. Noticeably, the absolute WERs improved by more than 14% with the CNN-LSTM compared with the TDNN-A model, and the AGC enhancement was shown to consistently improve the performance in all cases. Note that the acoustic models did not share the same training settings (see Table 1), however, we have shown in [7] that the CNN-LSTM clearly outperforms the TDNN for the same training configuration.

Table 4. Recognition accuracy in WER(%) using different acoustic models trained with unprocessed data (single array track).

AM	Dev data enh.		
	BF (A)	BF+AGC (B)	A+B
TDNN-A	88.3	88.2	86.0
TDNN-B	80.9	80.1	77.2
CNN-LSTM	74.0	74.3	71.8

4.2. Multiple-array track

Similar results were found for the multiple array case where there was no restriction on the Kinect device used to decode the test data.

Table 5. Recognition accuracy in WER(%) using the TDNN-A acoustic model (multiple-array track). Second left-hand column is the single array case.

Train data	Dev data enh.		
	BF	BF+DTS	BF+DTS+AGC
Baseline	88.3	86.3	85.8
AGC	87.9	85.7	84.5
HOS	87.2	84.6	83.6
DTS	85.0	83.1	82.5

Therefore, the data selection (DTS) approach described in the previous section was used to choose the best array candidate for each speech segment. In a first round of experiments, DTS was also utilized for selecting the training data for AMs. The WER results are depicted in Table 5.

Compared with the baseline case in Table 2, the DTS TDNN-A model has achieved more than 3% absolute WER reduction on all test sets. Regarding the impact of AGC enhancement on the evaluation data, that is improving the recognition accuracy in all cases (last column in Table 2). Same as for the single array case, combining the lattices of the decoded evaluation data with and without AGC has yielded between 2 and 3% absolute WER reduction, confirming the complementary nature of the information contained in the AGC signal (Table 6).

Table 6. System combination with the TDNN-A acoustic model (multiple-array track).

Train data	Dev data enh.		
	BF+DTS (C)	BF+DTS+AGC (D)	C+D
Baseline	86.3	85.8	83.6
AGC	85.7	84.5	82.5
HOS	84.6	83.6	80.5
DTS	83.1	82.5	80.1

In Table 7, the performance of DTS and AGC was assessed with the advanced acoustic models. Again, the WER improvements by using AGC were consistent with the previous experiments.

Table 7. Recognition accuracy in WER(%) using different acoustic models trained with unprocessed data (multiple-array track).

AM	Dev data enh.		
	BF+DTS (C)	BF+DTS+AGC (D)	C+D
TDNN-A	86.3	85.8	83.6
TDNN-B	78.6	77.7	74.7
CNN-LSTM	71.6	71.1	68.9

4.3. Speaker-dependent GEV (single-array track)

A last set of experiments was performed using the speaker-dependent GEV (SDGEV) enhancement proposed in [7]. Only the single-array track results were available at the time this manuscript was written.

Table 8. Recognition accuracy in WER(%) using SDGEV data enhancement [7] (single array track).

	Dev data enh.		
	SDGEV (E)	SDGEV+AGC (F)	E+F
CNN-BLSTM (SDGEV enh.)	64.9	65.0	63.7

In this case, the AM was CNN-BLSTM and both train and test data were enhanced using SDGEV, while AGC was applied only on the test set. Results are depicted in Table 8. Noticeably, the performance for system F (SDGEV+AGC) was slighly worse than for the baseline system E (SDGEV), the reason being that there is a mismatch between the training set and the test set of system F, while there is no mismatch for system E. Nevertheless, by fusing system E and system F, the WER still drops significantly (last column in Table 8), indicating the complementary nature of SDGEV and AGC enhancements.

A detailed analysis of the WER has shown that the AGC is effective in reducing the insertion error, which suggests a decrease of the speaker overlap and therefore a better speaker diarisation. The effect of varying the amount of worn microphone ('clean') training data has not been assessed, however, increasing the amount of clean training is likely to increase the mismatch between the train and the test sets (the decoding is performed on the array data), and thus worsening the recognition accuracy.

In the future we will explore combining the AGC with source separation. AGC will be used to predict the audio frames where only one speaker is active, while the frames with overlapped speakers will be processed using source separation algorithms.

5. CONCLUSIONS

In this paper we have performed a thorough analysis of a time-domain enhancement method aimed to reduce the effect of speaker overlap for the CHiME-5 task. The proposed method was an automatic gain controller driven by a DNN-based speaker classifier that would improve the speaker diarisation. ASR experiments with state-of-the-art acoustic models have shown that the proposed approach yields WER reductions between 2 and 3% absolute on the development set of CHiME-5.

6. REFERENCES

[1] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," 2017.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[3] J. Du, T. Gao, L. Sun, F. Ma, Y. Fang, D.Y. Liu, Q. Zhang, X. Zhang, H.K. Wang, J. Pan, J.Q. Gao, C.H. Lee, and J.D. Chen, "The USTC-iFlytek systems for CHiME-5 challenge," in *Proc. CHiME-5 Workshop*, 2018.

[4] N. Kanda, S. Ikeshita, R. Horiguchi, Y. Fujita, K. Nagamatsu, X. Wang, V. Manohar, N. Soplin, M. Maciejewski, S.J. Chen,

A. Subramanian, R. Li, Z. Wang, J. Naradowsky, P. Garcia-Perera, and G. Sell, "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *Proc. CHiME-5 Workshop*, 2018.

[5] I. Medennikov, I. Sorokin, A. Romanenko, D. Popov, Y. Khokhlov, T. Prisyach, N. Malkovskii, V. Bataev, S. Astapov, M. Korenevsky, and A. Zatvornitskiy, "The STC system for the CHiME 2018 challenge," in *Proc. CHiME-5 Workshop*, 2018.

[6] M. Kitza, W. Michel, Boeddeker C., J. Heitkaemper, T. Menne, R. Schluter, H. Ney, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "The RWTH/UPB system combination for the CHiME 2018 workshop," in *Proc. CHiME-5 Workshop*, 2018.

[7] R. Doddipatla, T. Kagoshima, C.T. Do, P. Petkov, C. Zorila, E. Kim, H. Hayakawa, H. Fujimura, and Y. Stylianou, "The Toshiba entry to the CHiME 2018 challenge," in *Proc. CHiME-5 Workshop*, 2018.

[8] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[9] 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 *Proc. ASRU*, 2011, pp. 3586–3589.

[10] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 15, no. 7, pp. 2011–2023, 2007.

[11] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, and V. Manohar, "Purely sequence-trained neural networks for asr based on lattice-free MMI," in *Proc. Interspeech*, 2016, p. 2751–2755.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[13] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, and G. Penn, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, vol. 22, no. 10, pp. 1533–1545, 2014.

[14] S. Hochreiter and J. Schmidhuber, "Acoustic beamforming for speaker diarization of meetings," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.