The University of Cambridge System for the CHiME-7 DASR Task

Keqi Deng[*], *Xianrui Zheng*[*], *Philip C. Woodland*

Department of Engineering, University of Cambridge, Trumpington St., Cambridge, UK.
{kd502, xz396, pw117}@cam.ac.uk

Abstract

This paper summarises the Cambridge team's work in the DASR Task of the CHiME-7 Challenge for speaker diarisation and automatic speech recognition (ASR). For speaker diarisation, the combination of Pyannote and ECAPA-TDNN was explored. For ASR, a two-pass ASR system was built. The first-pass model was based on a CTC model fine-tuned from a pre-trained WavLM model, based on which test-time unsupervised adaptation was implemented before decoded with a 4-gram language model (LM). For the second-pass system, with WavLM-based encoders, forward and backward hybrid CTC/attention models, as well as a label-synchronous neural transducer model, were trained for re-scoring. As a result of these efforts, for the sub-track, the Cambridge system achieved 21.7% and 22.7% DA-WER on the overall Dev and Eval sets respectively, with 24.6% and 32.0% relative error-rate reductions over the challenge baselines. For the main track, an ECAPA-based system was used for diarisation. Using our diarisation, together with our proposed ASR, the submitted main-track system gave a DA-WER of 38.7% on the Eval set which is with a 30% relative reduction in error rate compared to the challenge baseline.

Index Terms: speech recognition, speaker diarisation

1. Introduction

The Cambridge entry for the DASR Task of the CHiME-7 Challenge [1] is presented in this paper. The DASR Task is divided into a sub-track and a main track. For the sub-track, the oracle diarisation/segmentation is provided, based on which an ASR system needs to be run. However, for the main track no pre-defined segmentation is provided and therefore diarisation needs to be performed. This paper describes our contributions in both ASR and diarisation to both tracks.

This paper describes a two-pass ASR system which was implemented based on end-to-end (E2E) trainable models. A first-pass ASR model was built using WavLM [2] and Connectionist Temporal Classification (CTC) [3] and was decoded with a 4-gram language model (LM). In addition, unsupervised test-time adaptation [4] was used to improve the recognition performance. For the second-pass re-scoring system, the model combination is employed, including forward and backward CTC/attention joint [5] models, as well as a label-synchronous neural transducer (LS-Transducer) [6]. A speaker diarisation system is also built for the main track, which employs an ECAPA-TDNN to extract speaker embeddings with automatic channel selections and combines multiple diarisation outputs using DOVER-Lap [7].

2. Sub-Track: Speech Recognition only

This section describes the ASR methods and models used in our submitted systems. For the sub-track, oracle diarisation is used. For front-end processing, including channel selection [8] and GPU-accelerated GSS [9], exactly the same methods used in the baseline system [1] were used.

2.1. ASR Model Training

The systems developed followed the baseline system to pre-process the data, including the division of training and Dev sets, data augmentation with room impulse responses and noises, and speed perturbation.

2.1.1. WavLM-based CTC model

A CTC model was fine-tuned from a WavLM Large model [2] for the first pass and English characters were employed as the modelling unit. Note the WavLM model enables a mask operation which functions as SpecAugment [10] during training.

2.1.2. Hybrid CTC/attention model with WavLM encoder

A hybrid CTC/attention model was built for the second-pass re-scoring. In contrast to the baseline model [1], which regards WavLM as a feature extractor, our model directly employed WavLM as the encoder whose parameters can be updated during training. Similar to the CTC model, the WavLM encoder has an implicit SpecAugment operation, therefore separate SpecAugment was not used.

In preliminary experiments, when using 50 BPE modelling units, it was found that the WER result on the Dev set from CTC greedy search or attention-based decoder with teacher forcing was much lower than when using 500 BPE units. Interestingly, when using CTC/attention joint decoding during inference, the model with 500 BPE units performed better, this is because the sequence is much longer when using 50 BPE units, and the error accumulates along the prediction. In our two-pass ASR system, the CTC/attention model was targeted at re-scoring, which already has complete hypotheses from the first-pass CTC model, therefore, 50 BPE units were chosen as the modelling units.

2.1.3. Backward hybrid CTC/attention model

During re-scoring, to get richer information, we also built a backward CTC/attention model, which had a right-to-left decoder to utilise future information during re-scoring. Other details were the same as the normal (forward) CTC/attention model.

2.1.4. Label-synchronous neural transducer (LS-Transducer)

To further increase the model diversity used by the re-scoring system, an LS-Transducer [6] model, which is a streamable online model, was also implemented for re-scoring. In this case,

[*] Equal contribution.

Keqi Deng is funded by the Cambridge Trust. Xianrui Zheng is funded by an Amazon Studentship. This work used resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/T022159/1.

the offline WavLM encoder and 50 BPE units were employed, and other details followed [6]. Hence, our re-scoring system included different types of model, including forward and backward, offline and online.

2.2. ASR model decoding

Before ASR decoding, the CTC-based first-pass model first performed unsupervised test-time adaptation [4], which was an utterance-level adaptation and therefore was independent for each session. Then, the lexicon-based decoding was performed and N-best lists (N was set to 600) were saved for subsequent re-scoring.

2.2.1. Unsupervised test-time adaptation

The unsupervised test-time adaptation method proposed by [4] was implemented for the first-pass CTC model. Without access to labelled data during decoding, an unsupervised entropy-based loss function was used for adaptation, which sharpens class distributions via Entropy Minimisation (EM). Since the CTC blank token dominates the class distribution in L encoder output frames, the frames where the blank token yields the highest probability were excluded to mitigate the class-imbalance issue. Denote the vocabulary size as V, the objective $\mathcal{L}_{\text{em}}$ can be calculated as

$$\mathcal{L}_{\text{em}} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{V} P_{ij}\log P_{ij} \tag{1}$$

where P_{ij} is the probability of the j-th class in the i-th frame. Note that this test-time adaptation was at the utterance-level following [4], which was of course independent for each session.

2.2.2. First-pass decoding

During the first-pass decoding, a 4-gram LM that was generated from the training set transcription was used. In addition, a lexicon was also generated from the training set transcription, considering the 4-gram LM was word-level and the ASR modelling unit of CTC was English characters. Therefore, the lexicon mapped the word to the corresponding characters. The output N-best list was saved for the following re-scoring.

2.2.3. Second-pass decoding

After the first-pass decoding, the N-best lists were used for re-scoring. As mentioned, three different models were used for re-scoring, during which the hypotheses in the N-best list were fed into the three models to get their corresponding scores for each hypothesis. Suppose $\mathcal{S}_{\text{ctc}}$ is the score obtained from the original first-pass CTC model, $\mathcal{S}_{\text{aed}}$ is the score from the decoder of hybrid CTC/attention model, $\mathcal{S}_{\text{b-aed}}$ is the score from the decoder of backward CTC/attention model, and $\mathcal{S}_{\text{ls-t}}$ is the score from the LS-Transducer, the final score $\mathcal{S}_{\text{final}}$ of a hypothesis is computed as:

$$\mathcal{S}_{\text{final}} = \mathcal{S}_{\text{ctc}} + \alpha\mathcal{S}_{\text{aed}} + \beta\mathcal{S}_{\text{b-aed}} + \gamma\mathcal{S}_{\text{ls-t}} \tag{2}$$

where α, β, and γ are the coefficients for the three second-pass models. Inspired by [11, 12], the three interpolation coefficients were obtained by employing the covariance matrix adaptation evolution strategy (CMA-ES) [13, 14] as a black-box optimisation algorithm to minimise the WER on the Dev set.

3. Main Track: Diarisation and ASR

Our pipeline system for the main track consists of a diarisation system and a downstream ASR system, which has been described in the previous section. In this section, the details of our speaker diarisation system are presented.

3.1. Speaker Diarisation

Multi-channel weighted prediction error (WPE) dereverberation was applied to every channel of all corpora. No additional audio processing was performed.

3.1.1. System Description

We use the provided Pyannote [15] baseline diarisation, ECAPA-TDNN [16] and DOVER-Lap [7] in our diarisation system. The baseline diarisation system was utilised to identify the channel exhibiting the most speech activity within a fixed-length window (5 seconds). This information is then used to give a best channel prediction for each final speech segment of the baseline system. Since each speech segment can contain multiple 5-second windows, and the baseline system typically chooses different channels for each window, we sample a channel based on the distribution of the best channels across all included windows. We sample three times for each dataset and employ ECAPA-TDNN to extract speaker embeddings and then perform clustering-based diarisation on each of these three groups of chosen channels. These diarisaton results, along with supplied baseline diarisation output are then combined using DOVER-Lap.

4. Experimental Results

The following subsections first give the results on the sub-track and then the main track.

Table 1: *Sub-track: %DA-WER results with oracle diarisation on CHiME7 Dev sets, which contains Chime6 Dev, Dipco Dev, and Mixer6 Dev sets.*

ASR Model	Chime6	Dipco	Mixer6	Overall Dev
Baseline	32.6	33.5	20.2	28.8
Whisper	30.9	34.5	21.2	28.8
Our ASR	**22.8**	**28.5**	**13.8**	**21.7**
w/o Re-score	23.3	29.0	14.2	22.2
w/o TTA	24.1	31.4	14.8	23.4

Table 2: *Sub-track: %DA-WER results with oracle diarisation on CHiME7 Eval sets, which contains Chime6 Eval, Dipco Eval, and Mixer6 Eval sets.*

ASR Model	Chime6	Dipco	Mixer6	Overall Eval
Baseline	35.5	36.3	28.6	33.4
Whisper	36.6	35.7	25.2	32.5
Our ASR	**26.2**	**25.1**	**16.8**	**22.7**
w/o Re-score	26.8	25.4	17.0	23.1
w/o TTA	27.4	28.0	21.0	25.5

4.1. Results on the Sub-track

The results of our two-pass ASR system on Chime7 Dev and Eval sets for the sub-track are shown in Tables 1 and 2, which

Table 3: *Main track: %DER results with the submitted diarisation system on CHiME7 Dev sets, Chime6 (F) uses forced-aligned as the scoring reference.*

Model	Chime6	Chime6 (F)	Dipco	Mixer6
Baseline	40.0	39.0	29.9	16.6
Ours	39.0	30.5	29.4	17.5

Table 4: *Main track: %DER results with the submitted diarisation system on CHiME7 Eval sets.*

Model	Chime6	Dipco	Mixer6
Baseline	56.3	27.9	9.3
Ours	48.2	25.6	10.3

Table 5: *Main track: %DA-WER results with speaker diarisation on Chime7 Dev sets. SD denotes speaker diarisation.*

ASR	SD	Chime6	Dipco	Mixer6	Overall Dev
Baseline	Baseline	62.4	56.6	22.5	47.2
Our ASR	Baseline	52.5	48.3	17.7	39.5
Our ASR	Our Diar	**44.4**	**45.8**	**20.5**	**36.9**

Table 6: *Main track: %DA-WER results with speaker diarisation on Chime7 Eval sets. SD denotes speaker diarisation.*

ASR	SD	Chime6	Dipco	Mixer6	Overall Eval
Baseline	Baseline	77.4	54.7	33.7	55.3
Our ASR	Baseline	70.1	40.8	19.4	43.4
Our ASR	Our Diar	**56.1**	**36.8**	**23.1**	**38.7**

shows that our two-pass ASR system greatly outperformed the baseline and Whisper on the sub-track, with 24.6% and 32.0% relative reduction in DA-WER. In addition, the test-time adaptation (TTA) technique was shown to be very effective for ASR, especially on the Eval set with 9.4% relative reduction in DA-WER. However, TTA made the CTC model overconfident, resulting in only about 0.5 % absolute DA-WER reduction even if a strong re-scoring system was built.

4.2. Results on the Main-track

Table 3 shows the DERs on different sub-sets of the Dev set. With the manual reference, our diarisation approach only gave less than 1% absolute reduction on the Chime6 and Dipco subsets. However, with the force-aligned reference, there is more than 8% absolute reduction on the Chime6 Dev sub-set. The DA-WER results are listed in Tables 5 and 6, in which our overall system, including speaker diarisation and ASR, gave 10.3% and 16.6% absolute DA-WER reductions over the baseline on the Dev and Eval sets respectively. In addition, with our ASR, our diarisation method gave lower DA-WERs on the than the baseline diarisation by 2.6% (Dev) and 4.7% (Eval) absolute.

4.3. DER investigation

From the DER results in both Tables 3 and 4, there was an improvement relative to the baseline only in the Chime6 and Dipco subsets only, whereas the DERs for Mixer6 show an increase in both the Dev and Eval sets. Therefore, we looked at breakdown of DERs to gain a better understanding of the underlying reasons for this phenomenon. From Table 7 it can be seen that our improved diarisation was mostly due to the reduction in SER. Specifically, the Chime6 SER decreases from 14.5% to 4.5% with the manual scoring reference and from 18.2% to 6.7% with the force-aligned reference.

Table 7: *%DER breakdowns on CHiME7 Dev sub-sets. Chime6 (F) uses the forced-aligned reference. The three components of DER are % missed speech (MS), % false alarm (FA) and % speaker error rate (SER).*

Model	Data	MS	FA	SER
Baseline	Chime6	22.3	3.2	14.5
	Chime6 (F)	9.5	11.3	18.2
	Dipco	12.0	4.8	13.0
	Mixer6	13.8	1.0	1.7
Ours	Chime6	32.8	1.7	4.5
	Chime6 (F)	16.4	7.3	6.7
	Dipco	17.9	2.8	8.8
	Mixer6	15.0	0.7	1.8

However, although the reduction in SER indicates the quality of speaker embeddings improves after using ECAPA, there is a significant increase in missed speech (MS). This is because the ECAPA pipeline assumes there is no overlap in the data. Even though we have used DOVER-Lap to combine the various ECAPA outputs results with the baseline that does considers overlaps, the issue of MS remains a significant challenge.

5. Future Work

This section discusses potential future improvements for our current setup. One improvement would involve developing a more effective method for selecting channels, and the other one is to take into account overlapping data. In order to improve the channel selection method, we developed a way to find the "ideal" single-best channel selection in order to measure the performance of channel selection techniques. The aim is to establish a lower bound for DER in future work by utilising perfect VAD and knowing the ideal channel selection for our ECAPA-based system.

5.1. Channel Selection

Apart from the Chime6 Dev subset, where the reference device is available for all sessions, the reference channel is not provided for other subsets. To assess the impact of optimal channel selection on our diarisation system, we employ oracle utterance segmentation with a simplified ASR system and measure the WER of each segment. This simplified one-pass ASR uses greedy decoding and is applied for all channels for all utterances and chooses the channel for each utterance yields the lowest WER. The WERs from different channel selection methods are presented in Table 8 where it can be seen that random selection yields the poorest results, as anticipated. The GSS method gives the the same input to the ASR as the provided CHiME7 baseline.[1] The final two lines demonstrate that using just one top-performing channel, without any beamforming, can achieve better results than GSS.

[1] With the simplified ASR, the error rate is, as expected, lower than the baseline. The comparison of the simplified ASR system and the submitted one can be found by comparing the error rates from the GSS lines of Table 8 and the sub-track values in Tables 1 and 2.

Table 8: *%WER with different channel selection method on Dev sets, using the simplified ASR. 'Random' selects a single channel randomly; 'GSS' is the same as the input from the baseline. 'Best' selects the single channel that gives the lowest WER for an utterance.*

Method	Partition	Chime6	Dipco	Mixer6
Random	Dev	52.3	65.2	36.1
	Eval	51.3	68.2	36.3
GSS	Dev	31.9	44.6	19.7
	Eval	33.5	47.0	29.7
Best	Dev	30.5	40.4	18.6
	Eval	27.3	43.1	13.2

Table 9: *%DER with different pre-processing merging methods with oracle speech regions and best channel assignment*

Merging Method	Data	Dev	Eval
All	Chime6	36.8	36.9
	Chime6 (F)	27.8	-
	Dipco	22.4	20.6
	Mixer6	17.1	11.9
Channel-mediated	Chime6	27.7	31.4
	Chime6 (F)	24.2	-
	Dipco	18.5	17.1
	Mixer6	11.9	6.0

5.2. Overlap Considerations

As demonstrated above, most errors when using ECAPA result from the assumption that there are no overlap in the data. The pipeline merges all overlapped regions into a single stream during the pre-processing step. In the previous experiments, when overlapped speech regions are assigned with different channels, we just use the first channel and merge those regions. Now instead we only combine overlapped regions with the same channel, and treat them as a separate stream when they have different channel assignments. We refer this merging method as channel-mediated method. Table 9 shows that when using the oracle speech regions from the reference, together with the best (using the simplified ASR WER) channel information obtained from the previous steps, the DER is much better when using the channel-mediated method, which has a relative reduction of 24% and 21% in overall DER on the Dev and Eval sets respectively. It is worth mentioning that since the oracle speech regions are used so there is no false alarm in both cases, the DER reductions only come from MS and SER. The DER values for the overlap can be considered as the lower bound for the DER that the ECAPA system can achieve without applying any beamforming techniques on the audio. It is expected that further improvements may be achieved by combining multiple channels or using a diarisation system that does not solely rely on ECAPA.

6. Conclusion

This paper summarises the Cambridge entry for the CHiME-7 Challenge DASR Task. For the sub-track, a two-pass ASR system was implemented while exploring the use of self-supervised pre-trained models (i.e. WavLM) in both frame-synchronous (i.e. CTC) and label-synchronous models. In addition, model combination was explored to boost the performance of second-pass re-scoring. For the main track, we used combination of Pyannote and ECAPA TDNN with DOVER-Lap to perform speaker diarisation on overlapped speech. Experiments showed that our system gave reduced DA-WER compared to the challenge baseline, with 32% and 30% relative reduction on the overall Eval set for the sub and main tracks. For future work, we establish a performance goal for a single-channel ECAPA-based diarisation system without using beamforming.

7. References

[1] S. Cornell, M. Wiesner, S. Watanabe, D. Raj, X. Chang, P. Garcia, M. Maciejewski, Y. Masuyama, Z.-Q. Wang, S. Squartini, and S. Khudanpur, "The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios," 2023.

[2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[3] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[4] G.-T. Lin, S.-W. Li, and H. yi Lee, "Listen, Adapt, Better WER: Source-free single-utterance test-time adaptation for automatic speech recognition," in *Proc. Interspeech*, 2022, pp. 2198–2202.

[5] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[6] K. Deng and P. C. Woodland, "Label-synchronous neural transducer for end-to-end ASR," *arXiv*, vol. abs/2307.03088, 2023.

[7] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "DOVER-Lap: A method for combining overlap-aware diarization outputs," in *Proc. SLT*, 2021.

[8] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, 2014.

[9] D. Raj, D. Povey, and S. Khudanpur, "GPU-accelerated guided source separation for meeting transcription," *arXiv*, vol. abs/2212.05271, 2022.

[10] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.

[11] P. C. Woodland, X. Liu, Y. Qian, C. Zhang, M. J. F. Gales, P. Karanasou, P. Lanchantin, and L. Wang, "Cambridge University transcription systems for the multi-genre broadcast challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 639–646.

[12] Q. Li, C. Zhang, and P. C. Woodland, "Combining hybrid DNN-HMM ASR systems with attention-based models using lattice rescoring," *Speech Communication*, vol. 147, pp. 12–21, 2023.

[13] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, pp. 159–195, 2001.

[14] X. Zheng, C. Zhang, and P. C. Woodland, "Adapting GPT, GPT-2 and BERT language models for speech recognition," in *Proc. ASRU*, 2021.

[15] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *Proc. ICASSP*, 2020.

[16] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN embeddings for speaker diarization," *arXiv preprint arXiv:2104.01466*, 2021.