The STC System for the CHiME-6 Challenge

Ivan Medennikov[1,2]*, Maxim Korenevsky*[1]*, Tatiana Prisyach*[1]*, Yuri Khokhlov*[1]*,*
Mariya Korenevskaya[1]*, Ivan Sorokin*[1]*, Tatiana Timofeeva*[1]*, Anton Mitrofanov*[1]*,*
Andrei Andrusenko[1,2]*, Ivan Podluzhny*[1]*, Aleksandr Laptev*[1,2]*, Aleksei Romanenko*[1,2]

[1]STC-innovations Ltd, St. Petersburg, Russia [2]ITMO University, St. Petersburg, Russia

{medennikov, korenevsky, knyazeva, khokhlov, korenevskaya, sorokin, timofeeva, mitrofanov-aa, andrusenko, podluzhnyi, laptev, romanenko}@speechpro.com

Abstract

This paper is a description of the Speech Technology Center (STC) systems for the CHiME-6 challenge aimed at multi-microphone multi-speaker speech recognition and diarization in a dinner party scenario. We participated in both Track 1 and Track 2 and submitted our results for Ranking A as well as Ranking B for each track.

The soft-activity based Guided Source Separation (GSS) as a front-end and a combination of advanced acoustic modeling techniques such as GSS-based training data augmentation, multi-stride and multi-stream self-attention layers, statistics layer and SpecAugment, as well as the lattice-level fusion of acoustic models were applied in the 1st track system. Our system for Track 1 was in the top three systems, achieving 30% relative WER reduction over the baseline. Additionally, lattice rescoring with a neural language model was applied for Ranking B. Overall, this led to 34% relative WER reduction over the baseline in Track 1.

For Track 2, we proposed a novel Target-Speaker Voice Activity Detection (TS-VAD) approach to solve the diarization problem. Good diarization results made it possible to perform GSS on the obtained segments. TS-VAD is based on i-vector speaker embeddings, which are initially estimated using a strong diarization system based on spectral clustering of x-vectors. The back-end from the Track 1 system was used in the second track. The system for Track 2 demonstrated state-of-the-art performance, outperforming the baseline by 39% DER, 45% JER, 43% WER (Ranking A) and 45% WER (Ranking B) relative.

Index Terms: automatic speech recognition, speaker diarization, guided source separation, target-speaker VAD, CHiME-6

1. Introduction

Significant progress in the Automatic Speech Recognition (ASR) area was made in recent years. However, there are still many problems, such as Distant Speech Recognition (DSR), which are far from being solved. DSR is a highly important problem for a wide range of real-world applications, and many initiatives have been organized in the last years to investigate it.

The CHiME-6 challenge [1] considers the problem of distant multi-microphone conversational speech recognition in everyday home environments. CHiME-6 is a replica of the previous CHiME-5 Challenge [2] in the sense of data used. The dataset consisting of the informal conversational communications between 4 persons in noisy real-life environments was recorded in a dinner party scenario. The speech was captured by the distant Microsoft Kinect microphone arrays that resulted in reverberated and low Signal-to-Noise Ratio (SNR) recordings. Another challenging feature was a large amount of overlapping speech in conversations. The main task of the CHIME-6 was to develop a speech recognition system with the best possible accuracy in the described conditions. The details on the challenge are described in [2].

There are numerous approaches developed for improving multi-microphone DSR. First of all, various dereverberation methods such as Weighted Prediction Error (WPE) [3, 4], and denoising approaches such as IRM masking [5] or denoising TasNet [6], can clean the acoustic signal from external distortion sources. Besides, beamforming approaches such as Minimum Variance Distortionless Response (MVDR) [7] and Generalized Sidelobe Canceller (GSC) [8], are really helpful in the case of multiple microphones. Various data augmentation techniques such as room acoustics simulation [9] and spectral augmentation [10], as well as different perturbations of a raw signal are also extremely useful in such a task. Finally, separation of overlapped speech is crucial for accurate speech recognition [11–16].

This paper provides a description of the STC system for the CHiME-6 challenge. The system for Track 1 uses a strong front-end provided by organizers of the challenge, but with several modifications, providing a noticeable WER improvement. The back-end is trained on GSS-based augmented data and utilizes different modern techniques such as SpecAugment [10] and multi-stride [17] / multi-stream [18] self-attention layers. For Track 2, we proposed a novel Target-Speaker Voice Activity Detection (TS-VAD) approach, which is described in detail in [16]. TS-VAD allowed us to perform a good diarization and apply GSS on the obtained segments. Additionally, lattice rescoring with a strong Language Model (LM) was applied prior to fusion for Ranking B in both Track 1 and Track 2.

The rest of the paper is organized as follows. Section 2 describes the conditions and our contributions corresponding to the first track of the challenge, Section 3 relates to the second track. Finally, Section 4 concludes the paper.

2. Track 1: Speech recognition only

2.1. Front-end

Track 1 conditions allow the participants to use the information about the speakers boundaries for each utterance. So it is possible to use Guided Source Separation (GSS) [12, 13], which was developed during the CHiME-5 Challenge [2] and later allowed to improve the recognition accuracy significantly [19, 20]. The STC system uses the combination of the WPE dereverberation method [3, 4], GSS, and the MVDR beamforming [7] adopted from the baseline system. As in [19], this enhancement was also applied for training data augmentation and provided a significant WER reduction.

As noted in [20], the use of the refined utterance boundaries

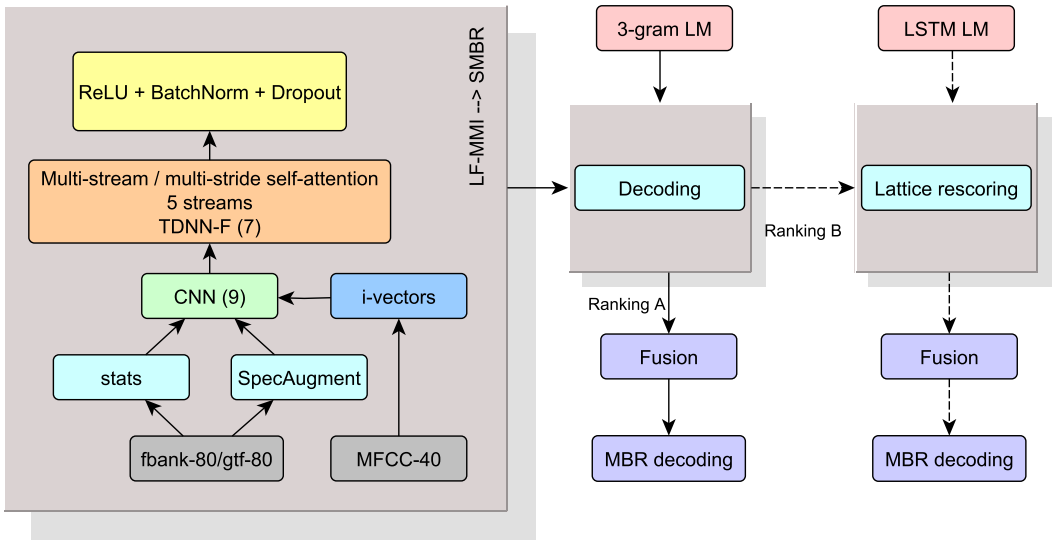

Figure 1: *Back-end scheme*

obtained after the first-pass decoding can provide an additional WER improvement. By default, per-frame speaker activities induced from hard label information are multiplied by the spectral masks after each iteration of GSS. We supposed that using soft-activity labels can improve the masks estimates. Soft-activities can be extracted from the first-pass decoding lattices. However, we found that better results can be obtained using speaker activity probabilities estimated by a special TS-VAD model. A more detailed description of such models is given in Section 3.2 and in [16].

The basic MVDR-beamforming procedure included in the *pb_chime*[1] package uses spectral masks obtained from GSS. After a thorough analysis of this procedure, we found a couple of ways to improve the accuracy slightly. The first one is a diagonal regularization of noise spatial covariance matrices. The second one is excluding one-third of all microphones with worst Envelope Variance [21] scores from the beamforming.

The contributions of applied approaches are presented in Table 1.

	Dev WER%
baseline TDNN-F	51.8
training on GSS-enhanced data	47.6
improved GSS[2]	46.2
+ MVDR: regularization	46.0
+ MVDR: EV	45.8
+ hard activity from ASR	43.5
+ soft activity from ASR	43.3
+ soft activity from TS-VAD	43.0

Table 1: *Contributions of various front-end approaches. EV stands for excluding microphones by Envelope Variance*

2.2. Back-end

As demonstrated in [19], using GSS-enhanced data in training improves ASR results significantly. Following this, we trained AM on a dataset consisting of worn microphones recordings and data obtained using four versions of GSS with various settings (microphones set, context length, number of iterations). We also used the room simulation [9] and both speed and volume perturbations included in the baseline recipe.

The scheme of our system back-end is presented in Figure 1. Our basic AM consists of 9-layer Convolutional Neural Network (CNN) [22] with residual connections, followed by 8-layer Factorized Time-Delay Neural Network (TDNN-F) [23]. The network takes 80-dimensional log Mel filterbank (fbank-80) or Gammatone filterbank (gtf-80) [24] feature vectors as an input.

Mean and standard deviation statistics computed by the Kaldi *stats-layer* are used as additional input channels, and the Kaldi *spec-augment-layer* (SpecAugment) [10] is applied for spectral perturbation. We tried both local (per-chunk) and global (per-utterance) statistics. The best result was obtained with local statistics computed in parallel to the SpecAugment layer.

Speaker embeddings are also used to provide speaker-aware training. We obtained the best results when using i-vectors [25] as speaker embeddings; however, models with x-vectors [26, 27]) were also included in an ensemble. It should be noted here that x-vectors were good enough only when they were evaluated on the segments with no speakers overlaps. Such segments were obtained from the JSON files provided by organizers in Track 1 and from the TS-VAD-based diarization segments in Track 2.

We also observed a noticeable improvement after adding multi-stride [17] and multi-stream [18] self-attention layers into the model. To perform multi-stream self-attention layers, we replaced 8-layer TDNN-F from the CNN-TDNN-F described above, with 5 streams of 7 TDNN-F layers and 12 self-attention heads per stream. For multi-stride self-attention layers, we extended CNN-TDNN-F with 3 self-attention blocks with different strides and 12 self-attention heads each. All the models were trained according to the Lattice Free Maximum Mutual Information (LF-MMI) [28] criterion for 4 epochs and fine-tuned for one more epoch of state-level Minimum Bayes Risk (sMBR) [29] training. Different acoustic models are compared in Table 2.

[1] `https://github.com/fgnt/pb_chime5`

[2] Microphones: 12 $\rightarrow$ 24, context: 10 s $\rightarrow$ 15 s, iterations: 5 $\rightarrow$ 20

In the baseline recipe, the decoding is carried out in two stages. According to the results of the first stage, the weights extracted from lattices are used for i-vectors recalculation. We found, however, that using speech segments from a simple neural VAD intersected with manual utterance boundaries is much faster and provides comparable accuracy in a single decoding stage.

Finally, we performed lattice fusion followed by MBR decoding [30] to combine recognition results from different models and different versions of GSS.

Acoustic model	Dev WER%
TDNN-F on MFCC	43.0
TDNN-F on fbank/gtf	42.5
+stats	41.9
+SpecAugment	41.0
CNN-TDNN-F+stats+SpecAugment	39.6
+multi-stride self-attention	38.3
+multi-stream self-attention	37.8
+sMBR	36.8

Table 2: *Comparison of acoustic models*

2.3. Advanced language modeling and rescoring

As part of Ranking B, the regularized Long Short-Term Memory (LSTM) LM [31] was applied for lattices rescoring [32] prior to fusion. The LM consisted of 3 layers with 2048 units per layer. The network was trained for 800 epochs with the Adam optimizer and learning rate of 3e-3. The training data was obtained with the Byte Pair Encoding (BPE) text decomposition provided by YouTokenToMe[3]. The model averaging over 10 best epochs was applied to obtain the final LM. The best single LM was trained on 3k BPE text decomposition. BPE-Dropout [33] also provided an additional improvement. Initializing the hidden states of LM during rescoring procedure with the hidden states obtained from the previous lattice also improved WER slightly. The combination of models trained on 1k, 3k, 5k, and 8k BPE was used in the final rescoring. The contributions of the applied methods are presented in Table 3.

	Dev WER%
baselines (best single AM):	
no rescoring	36.8
rescoring with regularized LSTM	34.2
+ model averaging	34.0
+ BPE-Dropout	33.9
+ hidden states initialization	33.8
fusion of LMs (1k+3k+5k+8k BPE)	33.7

Table 3: *Contributions of language modeling and rescoring approaches*

2.4. Track 1 results

Final recognition results for Track 1 are presented in Table 4. The fusion of acoustic models provided 3.3% absolute WER improvement on development set, and lattice rescoring for Ranking B reduced WER of more 2.6%. These results were scored 3rd in Ranking A and 2nd in Ranking B of Track 1 of the challenge.

[3] https://github.com/VKCOM/YouTokenToMe

	Dev WER%	Eval WER%
Kaldi baseline	51.76	51.29
Best single AM	36.82	38.59
Fusion (Ranking A)	33.53	35.79
+ LSTM-LM (Ranking B)	30.96	33.91

Table 4: *ASR results for Track 1*

3. Track 2: Diarization and ASR

In Track 2, participants are not allowed to use the information about the speakers boundaries for utterances. Detection of such boundaries is one of the goals of Track 2.

Baseline recipe uses the agglomerative hierarchical clustering (AHC) of x-vectors on VAD segments. However, this approach does not allow one to take into account the regions where speakers overlap. Since the baseline diarization is not accurate enough, GSS cannot be applied directly, which results in a very high word error rate. Therefore, our main intention was to achieve a substantial diarization improvement, which would be sufficient for applying GSS effectively. In order to perform this, we investigated a novel approach referred to as Target-Speaker Voice Activity Detection (TS-VAD), which was inspired by End-to-End Neural Diarization [34], Target-Speaker ASR [35, 36], and Personal VAD [37]. TS-VAD takes standard acoustic features (MFCC) along with the i-vector of each speaker as its inputs. The model produces the probability of each speaker activity on each frame. However, TS-VAD requires a sufficiently accurate initial diarization to estimate i-vectors for each speaker. To obtain such a diarization, we improved the baseline procedure in two main directions.

3.1. Baseline diarization improving

Firstly, Track 2 conditions allow the participants to use the VoxCeleb [38] data for the diarization models training. So we used the improved 34-layer Wide ResNet (WRN) x-vector extractor [39] trained on the VoxCeleb data. Basic AHC clustering of these WRN x-vectors computed on the same VAD segments by PLDA scores improved DER on the development set by about 12% abs. compared to the baseline extractor. Secondly, we replaced PLDA scores with cosine similarities and applied Spectral Clustering (SC) with automatic selection of the binarization threshold [40] instead of AHC, which reduced DER by another 5-7% abs. Such diarization accuracy was already sufficient to provide a good start for TS-VAD. The proposed diarization scheme is presented in Figure 2.

3.2. Target-speaker VAD

The STC system includes two types of TS-VAD models. The first one (TS-VAD-1C) is single-channel; it is presented in Figure 3 and can be described as follows. Input MFCC features are transformed by a 4-layer CNN and then fed to four parallel Speaker Detection (SD) blocks. Each SD block is a 2-layer Bidirectional LSTM (BLSTM) with projections [41], which takes an i-vector corresponding to the speaker as an additional input. It is important to note that the parameters of four SD blocks are shared. Then, combined outputs of four SD blocks are passed to one more BLSTM layer followed by four parallel fully connected layers and 2-class softmax layers on top of them. Four pairs of outputs produced by the TS-VAD model represent the probabilities of the presence/absence

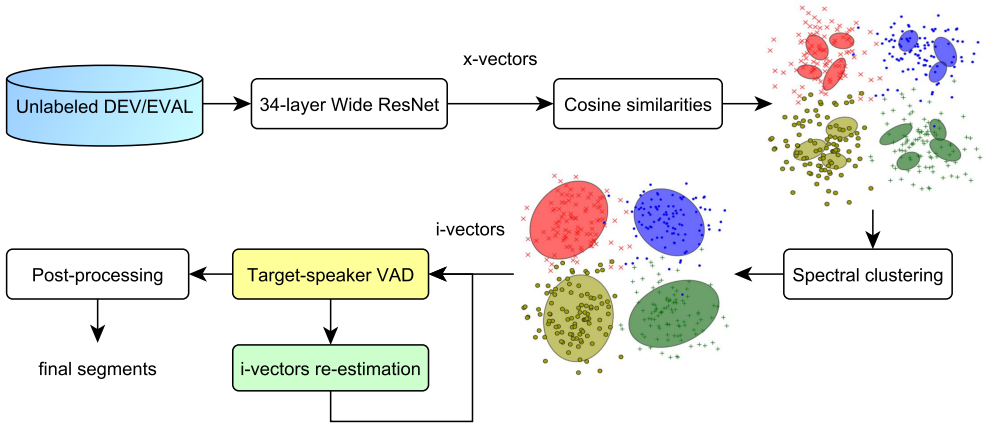

Figure 2: *Diarization scheme for Track 2*

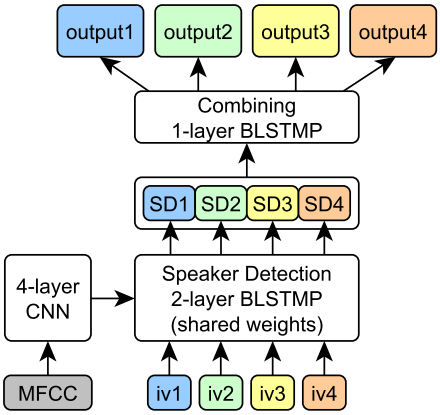

Figure 3: *Single-channel TS-VAD*

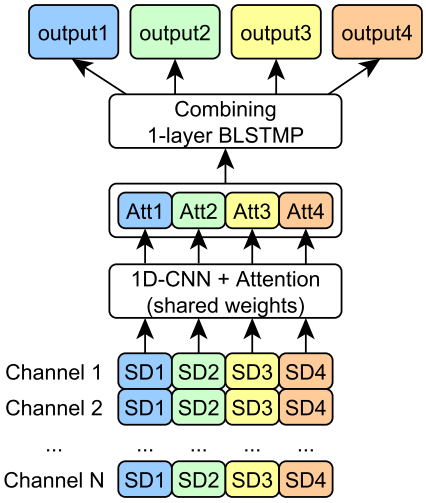

Figure 4: *Multi-channel TS-VAD*

of each speaker on the current frame. The training loss is a sum of 4 cross-entropies computed from speaker alignment. The described TS-VAD model is applied to each of the Kinect channels separately. After that, the output probabilities are averaged over the channels for each speaker. After simple post-processing (thresholding, median filtering, combining speech segments separated by short pauses, deleting too short speech segments) of these probabilities, one can obtain an improved speaker segmentation with significantly reduced DER. These probabilities can be used as weights for recalculating the i-vectors. We used the obtained embeddings in the second iteration of the described approach, which provides an additional DER improvement. The third iteration, however, did not improve DER anymore.

The second TS-VAD model (TS-VAD-MC) is multichannel and takes a combination of TS-VAD-1C model SD blocks outputs from a set of 10 Kinect recordings as an input. The channels of input Kinect recordings are chosen randomly for training, and the 1st and 4th channels are taken at test-time. This way of combining information from different channels is more effective than a simple averaging of probabilities, as in the TS-VAD-1C model. All the SD vectors for each speaker are passed through a convolutional layer and then combined using a simple attention mechanism. Combined outputs of the attention for all speakers are passed through a single BLSTM layer and converted into a set of per-frame probabilities of each speaker's presence/absence. The scheme of TS-VAD-MC is presented in Figure 4. Both types of TS-VAD models are described in detail in [16].

We used both CHiME-6 and an 800h subset of the VoxCeleb data for training the TS-VAD models for Track 2. Besides, we used the probabilities obtained from the TS-VAD model trained only on CHiME-6 data in Track 1 as soft-activities (see Section 2.1) to improve GSS performance. We also found that

- TS-VAD works better (1% abs. DER reduction) on top of blockwise WPE dereverberation;
- Fusion of probabilities from several TS-VAD models further improves diarization;
- Best ASR results (up to 2.5% abs. WER improvement) are obtained when using diarization with a larger False

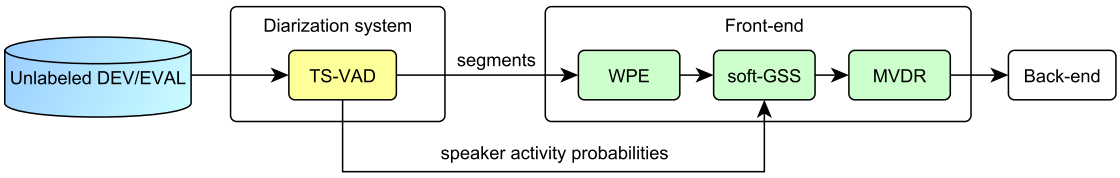

Figure 5: *Scheme of STC system for Track 2*

data	diarization	Spk Miss	False Alarm	Spk Error	DER	WER [4]
dev	WRN xvec + SC	27.24	9.83	10.22	47.29	70.47
	best DER	15.85	8.85	8.13	32.84	54.70
	best WER	9.02	20.71	7.57	37.30	53.33
eval	WRN xvec + SC	25.58	16.09	18.43	60.10	72.86
	best DER	16.31	10.15	9.56	36.02	55.56
	best WER	9.32	23.27	8.81	41.40	54.85

Table 5: *Diarization errors and their influence on GSS performance*

Alarm rate instead of the best DER diarization.

The results of the successive application of the approaches described above are presented in Table 6. It is interesting to note that the best ASR results were obtained on the segmentation not optimal with respect to DER but better in terms of JER.

Table 5 shows the breakout of the total DER into three types of errors, namely speaker miss, false alarm, and speaker error, obtained on the best clustering-based system and TS-VAD. The last column shows the Word Error Rates of the baseline acoustic model after applying the GSS-based front-end. Our best clustering-based segmentation is not good enough for GSS. On the other hand, TS-VAD segmentation allows one to perform GSS effectively. Besides, the segmentation which provides the best speech recognition accuracy has a larger False Alarm rate but significantly smaller speaker error and speaker miss rates. So the False Alarms are not very harmful to GSS, unlike speaker errors and misses. Note that the sum of Speaker Error and Miss rates is significantly less than the lowest possible error of clustering based diarization on the challenge data, which is 25%, as mentioned in [16].

	DEV		EVAL	
	DER	JER	DER	JER
x-vectors + AHC	63.42	70.83	68.20	72.54
WRN x-vectors + AHC	53.45	56.76	63.79	62.02
WRN x-vectors + SC	47.29	49.03	60.10	57.99
+ TS-VAD-1C (it1)	39.19	40.87	45.01	47.03
+ TS-VAD-1C (it2)	35.80	37.38	39.80	41.79
+ TS-VAD-MC	34.59	36.73	37.57	40.51
Fusion (best DER)	32.84	36.31	36.02	40.10
Fusion (best WER)	37.30	36.11	41.40	39.73

Table 6: *Diarization results for Track 2*

3.3. Recognition of diarized segments

The good diarization results obtained with TS-VAD made it possible to apply front-end technologies that we used successfully in Track 1, namely WPE + GSS + MVDR, for Track 2 as well. As in Track 1, this led to a substantial improvement of WER. Moreover, the ASR performance gap between TS-VAD and manual segmentation is rather small. The recognition results on the TS-VAD segments are compared in Table 7, while the scheme of the STC system provided for the second Track is presented in Figure 5.

	Dev WER%	Eval WER%
Kaldi baseline	84.25	77.94
Best single AM	44.89	47.67
Fusion (Ranking A)	41.56	44.49
+ LSTM-LM (Ranking B)	39.56	42.67

Table 7: *ASR results for Track 2*

4. Conclusion

In this paper we presented the STC system for the CHiME-6 challenge. The data augmentation approaches turned out to be extremely useful in such a task. Convolutional, statistics, and multi-stream/multi-stride self-attention layers in AM also provided a significant WER improvement.

We also presented TS-VAD, a novel approach for the diarization of conversations with multiple speakers and overlapping speech, which provided state-of-the-art results in a complex multi-channel dinner party scenario. It directly solves the diarization problem and allows performing GSS in the 2nd track of the challenge. It should be noted here, that using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance, so TS-VAD was useful in both tracks of the challenge. And re-estimation of i-vectors utilized by TS-VAD reduced DER in the 2nd track of the challenge significantly.

The described approaches allowed us to develop a competitive system for the first track of the CHiME-6 challenge and to achieve state-of-the-art results in Track 2.

5. Acknowledgments

This research was financially supported by the Foundation NTI (contract 20/18gr) ID 0000000007418QR20002.

We are grateful to STC Voice Biometrics Team for the awesome speaker embeddings extractor and valuable discussions.

[4] 12-microphone GSS enhancement, baseline TDNN-F

6. References

[1] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv:2004.09249*, 2020.

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

[3] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, pp. 2707–2720, 2012.

[4] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *ITG*, 2018, pp. 1–5.

[5] Z. Du, X. Zhang, and J. Han, "Investigation of monaural front-end processing for robust asr without retraining or joint-training," *ArXiv:1810.09067*, 2018.

[6] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," *ArXiv:2003.03998*, 2020.

[7] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 260–276, 2010.

[8] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[9] C. Kim, A. Misra, K. K. Chin *et al.*, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *INTERSPEECH*, 2017.

[10] D. S. Park, W. Chan, Y. Zhang *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019, pp. 2613–2617.

[11] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, pp. 332–353, 2008.

[12] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer *et al.*, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME Workshop*, 2018, pp. 35–40.

[13] M. Kitza, W. Michel, C. Boeddecker *et al.*, "The RWTH/UPB system combination for the CHiME 2018 workshop," in *CHiME Workshop*, 2018, pp. 53–57.

[14] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, 2018.

[15] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," *ArXiv:1910.14104*, 2019.

[16] I. Medennikov, M. Korenevsky, T. Prisyach *et al.*, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," in *INTERSPEECH*, (sumbitted) 2020.

[17] K. Han, J. Huang, Y. Tang *et al.*, "Multi-stride self-attention for speech recognition," in *INTERSPEECH*, 2019, pp. 2788–2792.

[18] K. Han, R. Prieto, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions," in *IEEE ASRU Workshop*, 2019, pp. 54–61.

[19] C. Zorila, C. Boeddecker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for CHiME-5 dinner party transcription," in *IEEE ASRU Workshop*, 2019, pp. 47–53.

[20] N. Kanda, C. Boeddecker, J. Heitkaemper *et al.*, "Guided source separation meets a strong ASR backend: Hitachi/paderborn university joint investigation for dinner party ASR," in *INTERSPEECH*, 2019, pp. 1248–1252.

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

[22] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *INTERSPEECH*, 2016, pp. 3434–3438.

[23] D. Povey, G. Cheng, Y. Wang *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *INTERSPEECH*, 2018, pp. 3743–3747.

[24] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *INTERSPEECH*, 2010, pp. 570–573.

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

[26] D. Snyder, D. Garcia-Romero, G. Sell *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE ICASSP*, 2018, pp. 5329–5333.

[27] S. Novoselov, A. Shulipa, I. Kremnev *et al.*, "On deep speaker embeddings for text-independent speaker recognition," in *Odyssey*, 2018, pp. 378–385.

[28] D. Povey, V. Peddinti, D. Galvez *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH*, 2016, pp. 2751–2755.

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

[30] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, pp. 802–828, 2011.

[31] S. Merity, N. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," in *ICLR*, 2017.

[32] H. Xu, T. Chen, D. Gao *et al.*, "A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition," in *IEEE ICASSP*, 2018, pp. 5929–5933.

[33] I. Provilkov, D. Emelianenko, and E. Voita, "BPE-dropout: Simple and effective subword regularization," *arXiv:1910.13267*, 2019.

[34] Y. Fujita, N. Kanda, S. Horiguchi *et al.*, "End-to-end neural speaker diarization with self-attention," in *IEEE ASRU Workshop*, 2019, pp. 296–303.

[35] N. Kanda, S. Horiguchi, R. Takashima *et al.*, "Auxiliary interference speaker loss for target-speaker speech recognition," in *INTERSPEECH*, 2019, pp. 236–240.

[36] N. Kanda, S. Horiguchi, Y. Fujita *et al.*, "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models," in *IEEE ASRU Workshop*, 2019, pp. 31–38.

[37] S. Ding, Q. Wang, S.-y. Chang *et al.*, "Personal VAD: Speaker-conditioned voice activity detection," *arXiv:1908.04284*, 2019.

[38] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017, pp. 2616–2620.

[39] A. Gusev, V. Volokhov, T. Andzhukaev *et al.*, "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv:2002.06033*, 2020.

[40] T. Park, K. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.

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