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, [2]ITMO University, Saint Petersburg, Russia

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

Abstract

This paper describes the STC system for the CHiME-6 Challenge aimed at multi-microphone multi-speaker speech recognition and diarization in a dinner party scenario. The system for Track 1 utilizes soft-activity based Guided Source Separation (GSS) front-end and a combination of advanced acoustic modeling techniques, including GSS-based training data augmentation, multi-stride and multi-stream self-attention layers, statistics layer and spectral augmentation, as well as lattice-level fusion of acoustic models. The system showed WER of 33.53%/35.79% on the development/evaluation data.

For Track 2, we proposed a novel Target-Speaker Voice Activity Detection (TS-VAD) approach, which directly solves the diarization problem and allows performing GSS on top of the diarized segments. Our TS-VAD is based on i-vector speaker embeddings, which are initially estimated using a strong x-vector diarization system with spectral clustering. This approach allowed to achieve DER of 37.30%/41.40%, JER of 36.11%/39.73%, and WER of 41.56%/44.49% using acoustic models from the Track 1 system.

Additionally, lattice rescoring with a neural language model was applied for Ranking B and provided WER reduction to 30.96%/33.91% in Track 1 and 39.56%/42.67% in Track 2.

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

1. Track 1: Speech recognition only

1.1. Front-end

Track 1 conditions allow the participants to use the information about the speaker boundaries for each utterance. So it is possible to use Guided Source Separation (GSS) [1, 2], which was developed during the CHiME-5 Challenge [3] and later [4, 5] allowed to improve the recognition accuracy significantly. The STC system uses the combination of the Weighted Prediction Error (WPE) dereverberation method [6], GSS and the Minimum Variance Distortionless Response (MVDR) beamforming [7] adopted from the baseline system.

As noted in [5], the use of the refined utterance boundaries 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 model. A more detailed description of such models is given in Section 2.2.

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 several 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 [8] scores from the beamforming.

1.2. Back-end

As demonstrated in [4], 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 (microphone set, context length, number of iterations). We also used the room simulation, speed and volume perturbation included in the baseline recipe.

Our basic AM consists of 9-layer Convolutional Neural Network (CNN) [9] with residual connections, followed by 8-layer Factorized Time-Delay Neural Network (TDNN-F) [10]. The network takes an 80-dimensional log Mel filterbank or Gammatone filterbank [11] feature vectors as an input. Mean and standard deviation statistics computed by the "stats" layer are used as additional input channels, and a SpecAugment [12] layer is applied for spectral perturbation. Speaker embeddings are also used to provide a speaker-aware training. We obtained the best results when using i-vectors [13] as speaker embeddings, however, models with x-vectors [14, 15]) were also included in an ensemble. We also observed a noticeable improvement after adding multi-stride and multi-stream self-attention layers [16, 17] into the model. All the models were trained according to the Lattice Free Maximum Mutual Information (LF-MMI) [18] criterion and fine-tuned for one more epoch of state-level Minimum Bayes Risk (sMBR) [19] training.

Finally, we performed lattice fusion followed by MBR decoding [20] to combine recognition results from different models and different versions of GSS.

As part of Ranking B, the regularized Long Short-Term Memory (LSTM) LM [21] on Byte Pair Encoding (BPE) [22] text decomposition was applied for lattices rescoring [23] prior to fusion. This provided an additional WER reduction.

Recognition results are presented in Table 1.

	Dev WER%	Eval WER%
Kaldi baseline	51.76	51.29
Best single AM	36.82	38.59
Fusion	33.53	35.79
Lattice rescoring + Fusion	30.96	33.91

Table 1: *ASR results for Track 1*

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

2. 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 over time. In order to tackle 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 [24, 25], Target-Speaker ASR [26] and Personal VAD [27]. TS-VAD takes standard acoustic features (MFCC) along with the embeddings of each speaker as its inputs and gives 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.

2.1. Baseline diarization improving

Firstly, Track 2 conditions allow the participants to use the VoxCeleb [28] data for the diarization models training. So we used the improved 34-layer Wide ResNet (WRN) x-vector extractor [29] trained on the VoxCeleb data. Basic AHC clustering of these WRN x-vectors computed on the same VAD segments by PLDA scores improved DER 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 [30] instead of AHC, which reduced DER by another 5-7% abs. Such diarization accuracy is already sufficient to provide a good start for TS-VAD.

	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 2: *Diarization results for Track 2*

2.2. Target-speaker VAD

The STC system includes two types of TS-VAD models. The first one (TS-VAD-1C) 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 [31] taking 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 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, and then the 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 provide any improvement.

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 by means of a simple attention mechanism. Combined outputs of attention for all speakers are passed through a single BLSTM layer and converted into a set of per-frame probabilities of each speaker presence/absence.

We used both CHiME-6 and a 800h subset of the VoxCeleb data for training the TS-VAD model 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-activity (see section 1.1) to improve GSS performance. We also found that

- TS-VAD works better (1% abs. DER reduction) on top of 2-minute long block WPE dereverberation;
- Fusion of probabilities from several TS-VAD model further improves diarization;
- Best ASR results (up to 2.5% abs. WER improvement) are obtained when using diarization with larger False Alarm rate instead of the best DER diarization.

The results of the successive application of the approaches described above are presented in Table 2.

2.3. ASR over diarization 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 leads to a substantial improvement of WER. Moreover, the ASR performance gap between TS-VAD and manual segmentation is rather small. The recognition results over the TS-VAD segments are presented in Table 3.

	Dev WER%	Eval WER%
Kaldi baseline	84.25	77.94
Best single AM	44.89	47.67
Fusion	41.56	44.49
Lattice rescoring + Fusion	39.56	42.67

Table 3: *ASR results for Track 2*

3. 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. References

[1] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-End Processing for the CHiME-5 Dinner Party Scenario," in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 35–40.

[2] M. Kitza, W. Michel, C. Boeddecker, J. Heitkaemper, T. Menne, R. Schlüter, 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 2018 Workshop on Speech Processing in Everyday Environments*, Sep 2018, pp. 53–57.

[3] 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*, Mar 2018, pp. 1561–1565.

[4] 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," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 47–53, Dec 2019.

[5] N. Kanda, C. Boeddecker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr," in *Proc. Interspeech 2019*, Sep 2019, pp. 1248–1252.

[6] 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, Dec 2012.

[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, Mar. 2010.

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

[9] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using cnns," in *Proc. Interspeech 2016*, Sep 2016, pp. 3434–3438.

[10] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech 2018*, Sep 2018, pp. 3743–3747.

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

[12] 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 *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[13] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, Dec 2013, pp. 55–59.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018, pp. 5329–5333.

[15] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin, "On deep speaker embeddings for text-independent speaker recognition," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, Jun 2018, pp. 378–385.

[16] K. Han, J. Huang, Y. Tang, X. He, and B. Zhou, "Multi-stride self-attention for speech recognition," in *Proc. Interspeech 2019*, Sep 2019, pp. 2788–2792.

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

[18] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Interspeech 2016*, Sep 2016, pp. 2751–2755.

[19] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2345–2349, Jan 2013.

[20] 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, Oct 2011.

[21] S. Merity, N. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," in *International Conference on Learning Representations*, Aug 2017.

[22] I. Provilkov, D. Emelianenko, and E. Voita, "Bpe-dropout: Simple and effective subword regularization," *ArXiv*, vol. abs/1910.13267, Oct 2019.

[23] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018, pp. 5929–5933.

[24] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, and S. Watanabe, "Auxiliary interference speaker loss for target-speaker speech recognition," in *Proc. Interspeech 2019*, Sep 2019, pp. 236–240.

[25] N. Kanda, S. Horiguchi, Y. Fujita, Y. Xue, K. Nagamatsu, and S. Watanabe, "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2019, pp. 31–38.

[26] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2019, pp. 296–303.

[27] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. Moreno, "Personal vad: Speaker-conditioned voice activity detection," *ArXiv*, vol. abs/1908.04284, Aug 2019.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, Jun 2017, pp. 2616–2620.

[29] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva, A. Ivanov, T. Pekhovsky, and Y. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," *ArXiv*, vol. abs/2002.06033, Feb 2020.

[30] 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, Dec 2019.

[31] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 338–342, Jan 2014.