The STC System for the CHiME-6 Challenge

Ivan Medennikov\textsuperscript{1,2}, Maxim Korenevsky\textsuperscript{1}, Tatiana Prisyach\textsuperscript{1}, Yuri Khokhlov\textsuperscript{1}, Mariya Korenevskaya\textsuperscript{1}, Ivan Sorokin\textsuperscript{1}, Tatiana Timofeeva\textsuperscript{1}, Anton Mitrofanov\textsuperscript{1}, Andrei Andrusenko\textsuperscript{1,2}, Ivan Podluzhny\textsuperscript{1}, Aleksandr Laptev\textsuperscript{1,2}, Aleksei Romanenko\textsuperscript{1,2}

\textsuperscript{1}STC-innovations Ltd, \textsuperscript{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)\textsuperscript{1,2}, which was developed during the CHiME-5 Challenge\textsuperscript{3} and later\textsuperscript{4,5} to improve the recognition accuracy significantly. The STC system uses the combination of the Weighted Prediction Error (WPE) dereverberation method\textsuperscript{6}, GSS and the Minimum Variance Distortionless Response (MVDR) beamforming\textsuperscript{7} adopted from the baseline system.

As noted in\textsuperscript{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\textsuperscript{2}. The basic MVDR-beamforming procedure included in the \pkg{pb_chime} 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\textsuperscript{8} scores from the beamforming.

1.2. Back-end

As demonstrated in\textsuperscript{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)\textsuperscript{9} with residual connections, followed by 8-layer Factorized Time-Delay Neural Network (TDNN-F)\textsuperscript{10}. The network takes an 80-dimensional log Mel filterbank or Gammatone filterbank\textsuperscript{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\textsuperscript{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\textsuperscript{13} as speaker embeddings, however, models with x-vectors\textsuperscript{14,15} were also included in an ensemble. We also observed a noticeable improvement after adding multi-stride and multi-stream self-attention layers\textsuperscript{16,17} into the model. All the models were trained according to the Lattice Free Maximum Mutual Information (LF-MMI)\textsuperscript{18} criterion and fine-tuned for one more epoch of state-level Minimum Bayes Risk (sMBR)\textsuperscript{19} training.

Finally, we performed lattice fusion followed by MBR decoding\textsuperscript{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\textsuperscript{21} on Byte Pair Encoding (BPE)\textsuperscript{22} text decomposition was applied for lattices rescoring\textsuperscript{23} prior to fusion. This provided an additional WER reduction.

Recognition results are presented in Table\textsuperscript{1}.

\begin{table}[h]
\begin{tabular}{|c|c|c|}
\hline
 & Dev WER\% & Eval WER\% \\
\hline
Kaldi baseline & 51.76 & 51.29 \\
Best single AM & 36.82 & 38.59 \\
Fusion & 33.53 & 35.79 \\
\hline
Lattice rescoring + Fusion & 30.96 & 33.91 \\
\hline
\end{tabular}
\caption{ASR results for Track 1}
\end{table}

\textsuperscript{1}https://github.com/fgnt/pb_chime5

\section*{References}

\begin{thebibliography}{9}
\bibitem{1} I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko, \textit{The STC System for the CHiME-6 Challenge}, 2020.
\end{thebibliography}
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 invented 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.

<table>
<thead>
<tr>
<th></th>
<th>DEV</th>
<th>EVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DER</td>
<td>JER</td>
</tr>
<tr>
<td>x-vectors + AHC</td>
<td>63.42</td>
<td>70.83</td>
</tr>
<tr>
<td>WRN x-vectors + AHC</td>
<td>53.45</td>
<td>56.76</td>
</tr>
<tr>
<td>WRN x-vectors + SC</td>
<td>47.29</td>
<td>49.03</td>
</tr>
<tr>
<td>+ TS-VAD-1C (it1)</td>
<td>39.19</td>
<td>40.87</td>
</tr>
<tr>
<td>+ TS-VAD-1C (it2)</td>
<td>35.80</td>
<td>37.38</td>
</tr>
<tr>
<td>+ TS-VAD-MC</td>
<td>34.59</td>
<td>36.73</td>
</tr>
<tr>
<td>Fusion (best DER)</td>
<td>32.84</td>
<td>36.31</td>
</tr>
<tr>
<td>Fusion (best WER)</td>
<td>37.30</td>
<td>36.11</td>
</tr>
</tbody>
</table>

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 2.1) 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.

Table 3: ASR results for Track 2

<table>
<thead>
<tr>
<th></th>
<th>Dev WER%</th>
<th>Eval WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaldi baseline</td>
<td>84.25</td>
<td>77.94</td>
</tr>
<tr>
<td>Best single AM</td>
<td>44.89</td>
<td>47.67</td>
</tr>
<tr>
<td>Fusion</td>
<td>41.56</td>
<td>44.49</td>
</tr>
<tr>
<td>Lattice rescoring + Fusion</td>
<td>39.56</td>
<td>42.67</td>
</tr>
</tbody>
</table>

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.

3. Acknowledgments

This research was financially supported by the Foundation NTI (contract 20/18gr) ID 000000007415QR20002.

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


