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
Introduction

Track 1: Speech recognition only

Track 2: Diarization and ASR

Final results
Main challenges

- Conversational speech
- Noisy real-world environment
- Far-field conditions
- Large amount of overlapping speech
Introduction

Track 1: Speech recognition only

Track 2: Diarization and ASR

Final results
Array synchronization to generate the new CHiME-6 audio data from the CHiME-5 data
Data augmentation: room simulation, speed and volume perturbation
Data cleanup
LF-MMI TDNN-F with i-vectors based speaker adaptation
Speech enhancement: Weighted Prediction Error (WPE) + Guided Source Separation (GSS) + MVDR beamforming
2-stage decoding with i-vectors re-estimation

<table>
<thead>
<tr>
<th></th>
<th>Dev WER%</th>
<th>Eval WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHiME-6 baseline</td>
<td>51.76</td>
<td>51.29</td>
</tr>
<tr>
<td>CHiME-5 top system (USTC-iFlytek)</td>
<td>45.60</td>
<td>46.60</td>
</tr>
</tbody>
</table>
Front-end

- WPE dereverberation
- GSS
  - soft-activities obtained from TS-VAD
- Minimum Variance Distortionless Response (MVDR) beamforming
  - diagonal regularization of noise spatial covariance matrices
  - excluding one-third of all microphones with the worst Envelope Variance scores from beamforming

<table>
<thead>
<tr>
<th>Dev WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline TDNN-F</td>
</tr>
<tr>
<td>training on GSS-enhanced data</td>
</tr>
<tr>
<td>improved GSS (12 → 24 mic, 10 → 15 s context, 5 → 20 iterations)</td>
</tr>
<tr>
<td>+ MVDR: regularization</td>
</tr>
<tr>
<td>+ MVDR: excluding microphones by Envelope Variance</td>
</tr>
<tr>
<td>+ hard activity from ASR</td>
</tr>
<tr>
<td>+ soft activity from ASR</td>
</tr>
<tr>
<td>+ soft activity from TS-VAD</td>
</tr>
</tbody>
</table>

Back-end

ReLU + BatchNorm + Dropout

Multi-stream / multi-stride self-attention
5 streams
TDNN-F (7)

CNN (9)
SpecAugment

stats
fbank-80/gtf-80
MFCC-40

i-vectors

3-gram LM

Decoding

Fusion

MBR decoding

LSTM LM

Lattice rescoring

Ranking B

Ranking A

Fusion

MBR decoding
## Comparison of acoustic models

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>Dev WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDNN-F on MFCC</td>
<td>43.0</td>
</tr>
<tr>
<td>TDNN-F on fbank/gtf</td>
<td>42.5</td>
</tr>
<tr>
<td>+stats</td>
<td>41.9</td>
</tr>
<tr>
<td>+SpecAugment</td>
<td>41.0</td>
</tr>
<tr>
<td>CNN-TDNN-F+stats+SpecAugment</td>
<td>39.6</td>
</tr>
<tr>
<td>+multi-stride$^2$/multi-stream self-attention$^3$</td>
<td>37.7</td>
</tr>
<tr>
<td>+SMBR</td>
<td>36.8</td>
</tr>
</tbody>
</table>


- 3-layers LSTM with 2048 units per layer
- Regularization techniques from ASGD Weight Dropout (AWD) LSTM\(^4\) except Averaged Stochastic Gradient Descent (ASGD)
- Trained on Byte Pair Encoding (BPE) tokens
  - 3k BPE – best single model
  - 1k, 3k, 5k, 8k BPE were trained for using in ensemble

\(^4\)S. Merity, N. Keskar, and R. Socher, "Regularizing and Optimizing LSTM Language Models", in International Conference on Learning Representations, 2017
### ASR results for Track 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Dev WER%</th>
<th>Eval WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaldi baseline</td>
<td>51.76</td>
<td>51.29</td>
</tr>
<tr>
<td>Best single AM</td>
<td>36.82</td>
<td>38.59</td>
</tr>
<tr>
<td>Fusion (16 systems)</td>
<td>33.53</td>
<td>35.79</td>
</tr>
<tr>
<td>Lattice rescoring + Fusion</td>
<td>30.96</td>
<td>33.91</td>
</tr>
</tbody>
</table>
Speech recognition in real-world conditions is a challenging task

- Data augmentation techniques are quite effective for this type of data
- Separation of overlapping speech is extremely important
- Using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance
- Convolutional and multi-stream/multi-stride self-attention layers in AM provide a significant WER improvement
Speech recognition in real-world conditions is a challenging task
Data augmentation techniques are quite effective for this type of data
Separation of overlapping speech is extremely important
Using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance
Convolutional and multi-stream/multi-stride self-attention layers in AM provide a significant WER improvement
Track 1 conclusions

- Speech recognition in real-world conditions is a challenging task
- Data augmentation techniques are quite effective for this type of data
- Separation of overlapping speech is extremely important
- Using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance
- Convolutional and multi-stream/multi-stride self-attention layers in AM provide a significant WER improvement
Track 1 conclusions

- Speech recognition in real-world conditions is a challenging task
- Data augmentation techniques are quite effective for this type of data
- Separation of overlapping speech is extremely important
- Using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance
- Convolutional and multi-stream/multi-stride self-attention layers in AM provide a significant WER improvement
Track 1 conclusions

- Speech recognition in real-world conditions is a challenging task
- Data augmentation techniques are quite effective for this type of data
- Separation of overlapping speech is extremely important
- Using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance
- Convolutional and multi-stream/multi-stride self-attention layers in AM provide a significant WER improvement
1. Introduction

2. Track 1: Speech recognition only

3. Track 2: Diarization and ASR

4. Final results
Baseline

- ASR from Track 1
- SAD training (TDNN+LSTM)
- Diarization training (Kaldi x-vector extractor (VoxCeleb) + PLDA scores + Agglomerative hierarchical clustering)
- Decoding and scoring

<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>
</tbody>
</table>
Baseline diarization improvements

Unlabeled DEV/EVAL → 34-layer Wide ResNet → Cosine similarities

<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 (baseline)</td>
<td>63.42</td>
<td>70.83</td>
</tr>
<tr>
<td>WRN x-vectors⁵ + AHC</td>
<td>53.45</td>
<td>56.76</td>
</tr>
<tr>
<td>WRN x-vectors + SC⁶</td>
<td>47.29</td>
<td>49.03</td>
</tr>
</tbody>
</table>


The main problem is a large amount of overlapping speech

Lowest achievable DER on the development set for clustering-based systems is 25.6% due to Speaker Miss Errors

Approaches directly detecting each speaker (like EEND\textsuperscript{7}) can handle this

Our TS-VAD approach is inspired by EEND, TS-ASR\textsuperscript{8} and Personal VAD\textsuperscript{9}


\textsuperscript{8}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 model", in 2019 IEEE ASRU, 2019, pp.31-38

\textsuperscript{9}S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Moreno, "Personal VAD: Speaker-conditioned voice activity detection", ArXiv:1908.04284, 2019
Target-speaker VAD

General info
- Predicts presence/absence probabilities of each speaker in the current frame
- Takes MFCC + i-vectors for each speaker as inputs
- Requires an accurate initialization of i-vectors

Training details
- Kaldi ASR Toolkit
- i-vectors extractor training following the baseline recipe
- CHiME-6(worn+simu+u400k) + VoxCeleb(800h) data for Track 2
- only CHiME-6 data (0.5% DER worse) for Track 1
Single-channel target-speaker VAD scheme (TS-VAD-1C)

Combining 1-layer BLSTM

Speaker Detection 2-layer BLSTM (shared weights)

MFCC

4-layer CNN

SD1 SD2 SD3 SD4

output1 output2 output3 output4
Multi-channel processing

- WPE dereverberation reduces DER by 1%
- Averaging of per-channel probabilities provides 2% DER improvement
- Joint processing of channels is important

We introduce a multi-channel TS-VAD model (TS-VAD-MC)

- takes a set of TS-VAD-1C hidden representations from 10 random Kinect channels as input
- combines them by a simple attention mechanism
Multi-channel target-speaker VAD scheme (TS-VAD-MC)

Combining 1-layer BLSTMP

1D-CNN + Attention (shared weights)

Channel 1
- SD1
- SD2
- SD3
- SD4

Channel 2
- SD1
- SD2
- SD3
- SD4

... ...

Channel N
- SD1
- SD2
- SD3
- SD4
Diarization system overview

Unlabeled DEV/EVAL → 34-layer Wide ResNet → Cosine similarities

x-vectors

Target-speaker VAD

Post-processing → i-vectors re-estimation → i-vectors

Spectral clustering

final segments
Post-processing

- Median filtering
- Thresholding
- Combining speech segments separated by short pauses
- Deleting too short speech segments
- Multi-speaker Viterbi Decoder
## Diarization results

<table>
<thead>
<tr>
<th>Method</th>
<th>DEV DER</th>
<th>DEV JER</th>
<th>EVAL DER</th>
<th>EVAL JER</th>
</tr>
</thead>
<tbody>
<tr>
<td>x-vectors + AHC (baseline)</td>
<td>63.42</td>
<td>70.83</td>
<td>68.20</td>
<td>72.54</td>
</tr>
<tr>
<td>WRN x-vectors + AHC</td>
<td>53.45</td>
<td>56.76</td>
<td>63.79</td>
<td>62.02</td>
</tr>
<tr>
<td>WRN x-vectors + SC</td>
<td>47.29</td>
<td>49.03</td>
<td>60.10</td>
<td>57.99</td>
</tr>
<tr>
<td>+ TS-VAD-1C (it1)</td>
<td>39.19</td>
<td>40.87</td>
<td>45.01</td>
<td>47.03</td>
</tr>
<tr>
<td>+ TS-VAD-1C (it2)</td>
<td>35.80</td>
<td>37.38</td>
<td>39.80</td>
<td>41.79</td>
</tr>
<tr>
<td>+ TS-VAD-MC</td>
<td>34.59</td>
<td>36.73</td>
<td>37.57</td>
<td>40.51</td>
</tr>
<tr>
<td><strong>Fusion (best DER)</strong></td>
<td>32.84</td>
<td>36.31</td>
<td>36.02</td>
<td>40.10</td>
</tr>
<tr>
<td><strong>Fusion (best WER)</strong></td>
<td>37.30</td>
<td>36.11</td>
<td>41.40</td>
<td>39.73</td>
</tr>
</tbody>
</table>
Diarization errors and their influence on GSS performance

<table>
<thead>
<tr>
<th>data</th>
<th>diarization</th>
<th>Spk Miss</th>
<th>False Alarm</th>
<th>Spk Error</th>
<th>DER</th>
<th>WER*</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>WRN xvec + SC</td>
<td>27.24</td>
<td>9.83</td>
<td>10.22</td>
<td>47.29</td>
<td>70.47</td>
</tr>
<tr>
<td></td>
<td>best DER</td>
<td>15.85</td>
<td>8.85</td>
<td>8.13</td>
<td>32.84</td>
<td>54.70</td>
</tr>
<tr>
<td></td>
<td>best WER</td>
<td>9.02</td>
<td>20.71</td>
<td>7.57</td>
<td>37.30</td>
<td>53.33</td>
</tr>
<tr>
<td>eval</td>
<td>WRN xvec + SC</td>
<td>25.58</td>
<td>16.09</td>
<td>18.43</td>
<td>60.10</td>
<td>72.86</td>
</tr>
<tr>
<td></td>
<td>best DER</td>
<td>16.31</td>
<td>10.15</td>
<td>9.56</td>
<td>36.02</td>
<td>55.56</td>
</tr>
<tr>
<td></td>
<td>best WER</td>
<td>9.32</td>
<td>23.27</td>
<td>8.81</td>
<td>41.40</td>
<td>54.85</td>
</tr>
</tbody>
</table>

GSS is able to suppress False Alarm errors!

* 12-microphone GSS enhancement, baseline TDNN-F
Recognition of diarized segments

<table>
<thead>
<tr>
<th></th>
<th>Dev WER%</th>
<th>Eval WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaldi baseline for Track 1</td>
<td>51.76</td>
<td>51.29</td>
</tr>
<tr>
<td>Kaldi baseline for Track 2</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>
<tr>
<td>Our best Track 1 result</td>
<td>30.96</td>
<td>33.91</td>
</tr>
</tbody>
</table>
Track 2 conclusions

- Multi-microphone multi-speaker conversational speech recognition for unsegmented recordings is an extremely challenging problem
- TS-VAD approach directly solves the diarization problem and allows performing GSS
- Iterative re-estimation of i-vectors significantly reduces DER
- Best ASR results are obtained when using diarization with larger False Alarm rate instead of the best DER diarization
- More details on TS-VAD approach in the upcoming INTERSPEECH paper

---

Track 2 conclusions

- Multi-microphone multi-speaker conversational speech recognition for unsegmented recordings is an extremely challenging problem
- TS-VAD approach directly solves the diarization problem and allows performing GSS
- Iterative re-estimation of i-vectors significantly reduces DER
- Best ASR results are obtained when using diarization with larger False Alarm rate instead of the best DER diarization
- More details on TS-VAD approach in the upcoming INTERSPEECH paper

---

Track 2 conclusions

- Multi-microphone multi-speaker conversational speech recognition for unsegmented recordings is an extremely challenging problem
- TS-VAD approach directly solves the diarization problem and allows performing GSS
- Iterative re-estimation of i-vectors significantly reduces DER
- Best ASR results are obtained when using diarization with larger False Alarm rate instead of the best DER diarization
- More details on TS-VAD approach in the upcoming INTERSPEECH paper

---

Track 2 conclusions

- Multi-microphone multi-speaker conversational speech recognition for unsegmented recordings is an extremely challenging problem
- TS-VAD approach directly solves the diarization problem and allows performing GSS
- Iterative re-estimation of i-vectors significantly reduces DER
- Best ASR results are obtained when using diarization with larger False Alarm rate instead of the best DER diarization
- More details on TS-VAD approach in the upcoming INTERSPEECH paper

---

Multi-microphone multi-speaker conversational speech recognition for unsegmented recordings is an extremely challenging problem.

TS-VAD approach directly solves the diarization problem and allows performing GSS.

Iterative re-estimation of i-vectors significantly reduces DER.

Best ASR results are obtained when using diarization with larger False Alarm rate instead of the best DER diarization.

More details on TS-VAD approach in the upcoming INTERSPEECH paper.\textsuperscript{10}

Introduction

Track 1: Speech recognition only

Track 2: Diarization and ASR

Final results
### Track 1 results

<table>
<thead>
<tr>
<th></th>
<th>Dev WER%</th>
<th>Eval WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>51.76</td>
<td>51.29</td>
</tr>
<tr>
<td>Ranking A</td>
<td>33.53</td>
<td>35.79</td>
</tr>
<tr>
<td>Ranking B</td>
<td>30.96</td>
<td>33.91</td>
</tr>
</tbody>
</table>

### Track 2 results

<table>
<thead>
<tr>
<th></th>
<th>DEV DER</th>
<th>DEV JER</th>
<th>DEV WER</th>
<th>EVAL DER</th>
<th>EVAL JER</th>
<th>EVAL WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>63.42</td>
<td>70.83</td>
<td>84.25</td>
<td>68.20</td>
<td>72.54</td>
<td>77.94</td>
</tr>
<tr>
<td>Ranking A</td>
<td>37.30</td>
<td>36.11</td>
<td>41.56</td>
<td>41.40</td>
<td>39.73</td>
<td>44.49</td>
</tr>
<tr>
<td>Ranking B</td>
<td>37.30</td>
<td>36.11</td>
<td>39.56</td>
<td>41.40</td>
<td>39.73</td>
<td>42.67</td>
</tr>
</tbody>
</table>
THANK YOU!