THE STC SYSTEM FOR THE CHIME 2018 CHALLENGE

Ivan Medennikov, Ivan Sorokin, Aleksei Romanenko, Dmitry Popov, Yuri Khokhlov, Tatiana Prisyach, Nikolay Malkovskii, Vladimir Bataev, Sergei Astapov, Maxim Korenevsky, Alexander Zatvornitskiy
STC- INNOVATIONS


2. Multi-disciplinary team with expertise in general machine learning, speech recognition, NLU, bi-modal (voice+face) identification

3. Close partnership with ITMO University
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Introduction

Unsuccess story

Success story

Conclusions

Final results on eval and future work
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Main challenges

- Conversational speech
- Noisy real-world environment
- Far-field conditions
- Great amount of overlapped speech
Beamforming and Enhancement: Unsuccess story
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

- MVDR + CGMM/Music/estnoiseg mask
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
Unsuccess story

- MVDR + CGMM/Music/estnoiseeg mask
- DeepBeam [Qian, 2018] *

*https://github.com/auspicious3000/deepbeam
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]*

*https://github.com/fgnt/nn-gev
Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoise seg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
Unsuccess story

- MVDR + CGMM/Music/estnoiseeg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
- Deep Clustering
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
- Deep Clustering
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
- Deep Clustering
- Permutation invariant training (PIT)
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
- Deep Clustering
- Permutation invariant training (PIT)
Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
- Deep Clustering
- Permutation invariant training (PIT)
- WPE
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask
- DeepBeam [Qian, 2018]
- GEV + BLSTM mask [Heymann, 2016]
- Denoising with CGMM mask
- Denoising Wavenet [Rethage, 2017]
- Deep Clustering
- Permutation invariant training (PIT)
- WPE

www.speechpro.com
Success story
Multi-channel speaker-aware model training: embeddings

- embedding training by triplet ranking loss [Ye and Guo, 2018]
Multi-channel speaker-aware model training: final model

- auxiliary inputs [Zmolikova, 2018]
- residual attention network [Wang, 2017]
- speaker-adapted classifier *
- sum and average all embeddings for speaker in utterance

*https://github.com/Microsoft/LightGBM
Success story

Speaker adaptation by frame-level mask: training

---

**Table:**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Mask Type</th>
<th>Phone</th>
<th>&lt;sil&gt;</th>
<th>word</th>
<th>&lt;sil&gt;</th>
<th>&lt;noise&gt;</th>
<th>&lt;sil&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>P01(id 1)</td>
<td>Ideal mask (general)</td>
<td>&lt;sil&gt;</td>
<td>0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P01(id 1)</td>
<td>Ideal targets (if P01)</td>
<td>&lt;sil&gt;</td>
<td>1 1 0 0 0 0 1 1 1 1 0 0 1 0 0 0 1 1 1 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P02(id 2)</td>
<td>&lt;sil&gt;</td>
<td>0 0 2 2 2 0 0 0 2 2 2 2 2 2 0 0 0 0 2 2 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P03(id 4)</td>
<td>&lt;sil&gt;</td>
<td>0 0 4 4 4 4 0 0 4 4 4 0 0 0 0 0 4 4 4 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P04(id 8)</td>
<td>&lt;sil&gt;</td>
<td>0 0 0 0 0 0 0 0 8 8 8 0 0 0 0 8 8 0 0 0 0 8 8 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121
### Speaker adaptation by frame-level mask: filtering

<table>
<thead>
<tr>
<th>Original acoustic feats</th>
<th>$x_{t,1}$</th>
<th>$x_{t+1,1}$</th>
<th>...</th>
<th>$x_{t+23,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>...</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$x_{t,n}$</td>
<td>$x_{t+1,n}$</td>
<td>...</td>
<td>$x_{t+23,n}$</td>
<td></td>
</tr>
<tr>
<td><strong>Speaker mask</strong></td>
<td>0.6</td>
<td>0.7</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Filtered acoustic feats</strong></td>
<td>$x_{t,1}$</td>
<td>$x_{t+1,1}$</td>
<td>$x_{t+2,1}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>Throw out</td>
</tr>
<tr>
<td>$x_{t,n}$</td>
<td>$x_{t+1,n}$</td>
<td>$x_{t+2,n}$</td>
<td></td>
<td>$x_{t+22,n}$</td>
</tr>
</tbody>
</table>
Success story

* [https://github.com/speechpro/mixup](https://github.com/speechpro/mixup) (for Kaldi)

Mixup [Medennikov, 2018] *

- virtual training examples by combining existing ones
- especially effective on mismatched test data

**Generation of new training data**

\[
\tilde{x} = \lambda x_i + (1 - \lambda) x_j \\
\tilde{y} = \lambda y_i + (1 - \lambda) y_j
\]
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Success story

System I

worn

WPE

half-sum

room simulation

kinects

WPE

CFFT features

CE speaker-aware model training

BN extraction

mixup augmentation

LF-MMI TDNN training
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

System II and III

Success story
System II and III

The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121
Success story

System II and III

The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

www.speechpro.com
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

System IV

- **worn**
  - WPE
  - half-sum
  - speed perturbation

- **kinects**
  - WPE
  - strict cleanup

- MFCC-40 + i-vectors
  - mixup augmentation
  - CE BLSTM training

Success story

www.speechpro.com
Decoding and models combination

- **Decoding**: application of softmax temperature to a prior distribution

- **Fusion**: posterior-level combination or two types of lattice-level combination
Fusion

The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

Success story

Fusion

- baseline
- system IV (BLSTM, MFCC)
- system III (TDNN-LSTM, MFCC)
- system I (CNN+TDNN, CFFT)
- system II (TDNN-LSTM, FBANK)
- fusion (single)
- fusion (single+dev)
WER (%) for the final system per session and location

<table>
<thead>
<tr>
<th>Track</th>
<th>Session</th>
<th>Kitchen</th>
<th>Dining</th>
<th>Living</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>S02</td>
<td>67.7</td>
<td>59.7</td>
<td>55.5</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>58.0</td>
<td>59.8</td>
<td>54.9</td>
<td></td>
</tr>
<tr>
<td>Single+Dev</td>
<td>S02</td>
<td>65.5</td>
<td>56.2</td>
<td>52.4</td>
<td>56.6</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>55.7</td>
<td>56.8</td>
<td>51.9</td>
<td></td>
</tr>
<tr>
<td>Multiple</td>
<td>S02</td>
<td>65.8</td>
<td>57.9</td>
<td>55.1</td>
<td>58.1</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>55.5</td>
<td>57.3</td>
<td>55.4</td>
<td></td>
</tr>
<tr>
<td>Multiple+Dev</td>
<td>S02</td>
<td>62.1</td>
<td>52.2</td>
<td>50.2</td>
<td>53.5</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>51.2</td>
<td>51.6</td>
<td>51.4</td>
<td></td>
</tr>
</tbody>
</table>

Success story

www.speechpro.com
### Summary

<table>
<thead>
<tr>
<th>Track</th>
<th>Features</th>
<th>Adaptation</th>
<th>Model</th>
<th>Loss</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>CFFT FBANK MFCC MFCC</td>
<td>Auxiliary soft-mask soft-mask ivec+mask</td>
<td>CNN+TDNN TDNN-LSTM TDNN-LSTM BLSTM</td>
<td>CE, LF-MMI LF-MMI LF-MMI CE</td>
<td>63.4 63.3 63.8 66.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fusion (4 systems)</td>
<td></td>
<td></td>
<td>59.4</td>
</tr>
<tr>
<td>Single+Dev</td>
<td>Fusion (4 systems)</td>
<td></td>
<td></td>
<td></td>
<td>56.6</td>
</tr>
<tr>
<td>Multiple</td>
<td>Fusion (4 systems)*</td>
<td></td>
<td></td>
<td></td>
<td>58.1</td>
</tr>
<tr>
<td>Multiple+Dev</td>
<td>Fusion (4 systems)*</td>
<td></td>
<td></td>
<td></td>
<td>53.5</td>
</tr>
</tbody>
</table>

The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121
Conclusions
Common speech processing approaches face great challenges in real-world conditions.
Common speech processing approaches face great challenges in real-world conditions.

Both speaker separation and speaker adaptation are extremely important.
Conclusions

- Common speech processing approaches face great challenges in real-world conditions
- Both speaker separation and speaker adaptation are extremely important
- Data augmentation and normalization are reasonably effective for this type of data
Conclusions

- Common speech processing approaches face great challenges in real-world conditions
- Both speaker separation and speaker adaptation are extremely important
- Data augmentation and normalization are reasonably effective for this type of data
- Fusion always gives a good performance improvement
The work was financially supported by the Ministry of Education and Science of the Russian Federation. Contract 14.579.21.0121, ID RFMEFI57915X0121

### Final results on eval and future work

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Our result</th>
<th>abs, %</th>
<th>rel, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73.3</td>
<td>55.5</td>
<td>-17.8</td>
<td>-24.3</td>
</tr>
</tbody>
</table>

- Joint training of all components (front-end and back-end)
- Diarization for unsegmented real-world data
Contributions of applied methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Abs WER improvement, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array synchronization</td>
<td>0.9</td>
</tr>
<tr>
<td>Room simulator</td>
<td>1.6</td>
</tr>
<tr>
<td>Alignment transfer (worn half-sum → kinect)</td>
<td>1.3</td>
</tr>
<tr>
<td>Speaker adaptation (gating/throw out)</td>
<td>7/5</td>
</tr>
<tr>
<td>Speaker adaptation (i-vector)</td>
<td>2.4</td>
</tr>
<tr>
<td>Speaker adaptation (auxiliary)</td>
<td>4.1</td>
</tr>
<tr>
<td>Multi-channel model</td>
<td>2.2</td>
</tr>
<tr>
<td>Strict cleanup</td>
<td>1.3</td>
</tr>
<tr>
<td>WPE</td>
<td>1.4</td>
</tr>
<tr>
<td>Mixup</td>
<td>1.1</td>
</tr>
<tr>
<td>Speed Perturbation</td>
<td>0.9</td>
</tr>
<tr>
<td>Backstitch training</td>
<td>0.5</td>
</tr>
<tr>
<td>Fusion</td>
<td>3.9</td>
</tr>
</tbody>
</table>
THANK YOU
ABOUT THE COMPANY

STC-Innovations is a leader in the multimodal biometric market. STC-Innovations develops multimodal biometric solutions based on person-identifying technologies via voice, face and other noncontact biometric features.

STC-Innovations is a spin-off company of the Speech Technologies Center, leading global provider of innovative systems in high-quality recording, audio and video processing and analysis, speech synthesis and recognition, and real-time, high-accuracy voice and facial biometrics solutions with over 20 years of research, development and implementation experience in Russia and internationally.

STC is ISO-9001: 2008 certified.

CONTACTS

Russia
4 Krasutskogo street, St. Petersburg, 196084
Tel.: +7 812 325-8848
Fax: +7 812 327 9297
Email: info@speechpro.com

USA
Suite 316, 369 Lexington ave
New York, NY, 10017
Tel.: +1 646 237 7895
Email: sales-usa@speechpro.com