The 6th International Workshop on Speech Processing in Everyday Environments

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

1. Introduction
2. Track 1: Speech recognition only
3. Track 2: Diarization and ASR
4. Final results

Main challenges

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

Baseline

- 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

	Dev WER%	Eval WER%
CHiME-6 baseline	51.76	51.29
CHiME-5 top system (USTC-iFlytek)	45.60	46.60

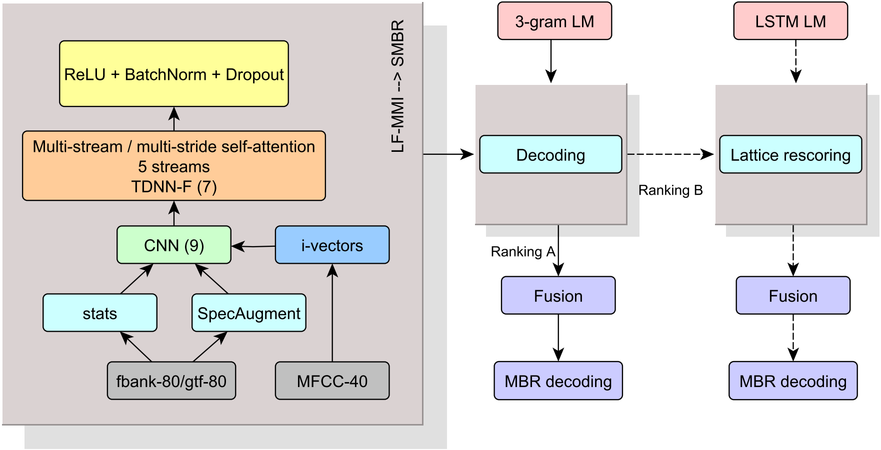
Front-end

- 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[1] scores from beamforming

	Dev WER%
baseline TDNN-F	51.8
training on GSS-enhanced data	47.6
improved GSS (12 $\rightarrow$ 24 mic, 10 $\rightarrow$ 15 s context, 5 $\rightarrow$ 20 iterations)	46.2
+ MVDR: regularization	46.0
+ MVDR: excluding microphones by Envelope Variance	45.8
+ hard activity from ASR	43.5
+ soft activity from ASR	43.3
+ soft activity from TS-VAD	43.0

[1] M. Wolf, C. Nadeu "Channel selection measures for multi-microphone speech recognition" in Speech Communication, 2014, v. 57, pp. 170—180.

Back-end



ReLU + BatchNorm + Dropout
Multi-stream / multi-stride self-attention
5 streams
TDNN-F (7)
CNN (9)
i-vectors
stats
SpecAugment
fbank-80/gtf-80
MFCC-40
LF-MMI --> SMBR
3-gram LM
LSTM LM
Decoding
Lattice rescoring
Ranking B
Ranking A
Fusion
Fusion
MBR decoding
MBR decoding

Comparison of acoustic models

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[2]/multi-stream self-attention[3]	37.7
+SMBR	36.8

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

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

Long Short-Term Memory (LSTM) Language Model (LM)

- 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

	Dev WER%	Eval WER%
Kaldi baseline	51.76	51.29
Best single AM	36.82	38.59
Fusion (16 systems)	33.53	35.79
Lattice rescoring + Fusion	30.96	33.91

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

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

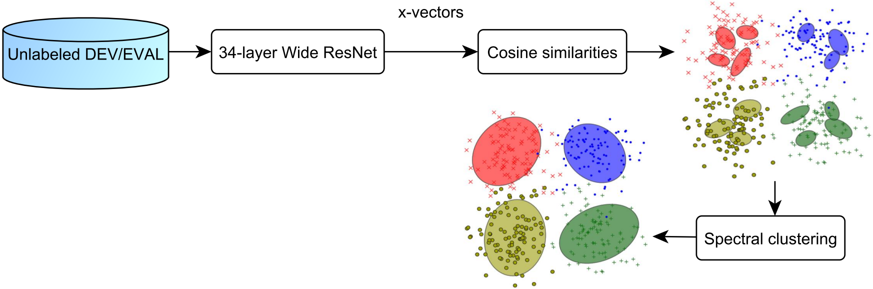
- Speech recognition in real-world conditions is a challenging task
- Data augmentation techniques are quite effective for this type of data
- Separation of overlapping speech is extremely important
- Using soft-activities from TS-VAD to initialize GSS instead of hard-activities improves system performance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

Baseline

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

	Dev WER%	Eval WER%
Kaldi baseline	84.25	77.94

Baseline diarization improvements



	DEV		EVAL	
	DER	JER	DER	JER
x-vectors + AHC (baseline)	63.42	70.83	68.20	72.54
WRN x-vectors[5] + AHC	53.45	56.76	63.79	62.02
WRN x-vectors + SC[6]	47.29	49.03	60.10	57.99

[5] A. Gusev et al., "Deep speaker embeddings for far-field speaker recognition on short utterances," arXiv preprint arXiv:2002.06033

[6] T. Park et al., "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," IEEE Signal Processing Letters, vol. 27, pp. 381–385, Dec 2019

Target-speaker VAD: Motivation

- 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[7]) can handle this
- Our TS-VAD approach is inspired by EEND, TS-ASR[8] and Personal VAD[9]

[7] 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 ASRU, 2019, pp. 296-303

[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

[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

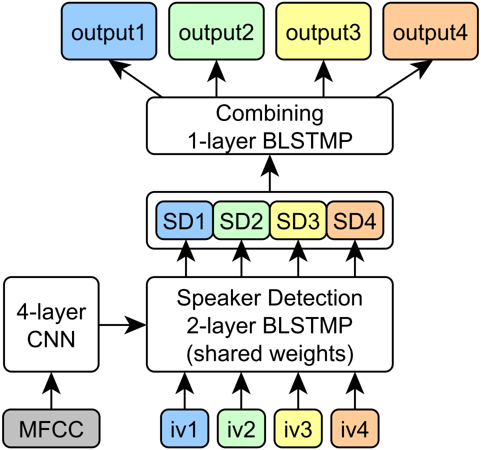
General info

- 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)



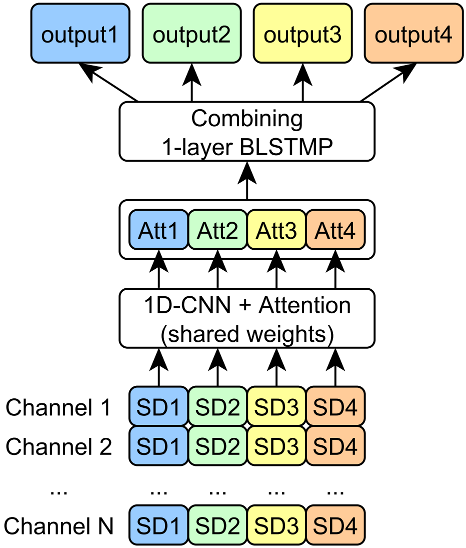
Multi-channel processing

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

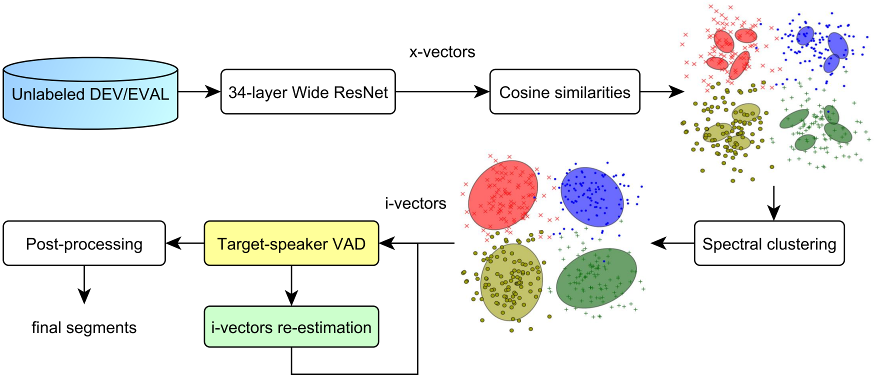
We introduce a multi-channel TS-VAD model (TS-VAD-MC)

- 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)



Diarization system overview



Post-processing

Post-processing

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

Diarization results

	DEV		EVAL	
	DER	JER	DER	JER
x-vectors + AHC (baseline)	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**

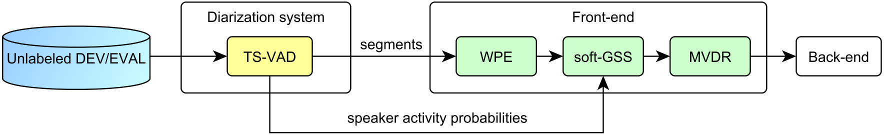
Diarization errors and their influence on GSS performance

data	diarization	Spk Miss	False Alarm	Spk Error	DER	WER*
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

GSS is able to suppress False Alarm errors!

* 12-microphone GSS enhancement, baseline TDNN-F

Recognition of diarized segments



	Dev WER%	Eval WER%
Kaldi baseline for Track 1	51.76	51.29
Kaldi baseline for Track 2	84.25	77.94
Best single AM	44.89	47.67
Fusion	41.56	44.49
Lattice rescoring + Fusion	39.56	42.67
Our best Track 1 result	30.96	33.91

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[10]

[10] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario", submitted to INTERSPEECH 2020

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[10]

[10] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario", submitted to INTERSPEECH 2020

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[10]

[10] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario", submitted to INTERSPEECH 2020

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[10]

[10] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario", submitted to INTERSPEECH 2020

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[10]

[10] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario", submitted to INTERSPEECH 2020

Submission results

Track 1 results

	Dev WER%	Eval WER%
Baseline	51.76	51.29
Ranking A	33.53	35.79
Ranking B	30.96	33.91

Track 2 results

	DEV			EVAL		
	DER	JER	WER	DER	JER	WER
Baseline	63.42	70.83	84.25	68.20	72.54	77.94
Ranking A	37.30	36.11	41.56	41.40	39.73	44.49
Ranking B			39.56			42.67

THANK YOU!