# BUT System for CHiME-6 Challenge

K. Žmolíková, M. Kocour, F. Landini, K. Beneš, M. Karafiát,
H. K. Vydana, A. Lozano-Diez, O. Plchot, M. K. Baskar,
J. Švec, L. Mošner, V. Malenovský, L. Burget, B. Yusuf,
O. Novotný, F. Grézl, I. Szöke, J. Černocký

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

# Outline

# Dataset

- 20 real dinner parties
- each party has four speakers (friends), very natural conversations
- each party about 2 hours, 3 stages (kitchen, dining room, living room)
- about 20 % of overlap
- six 4-channel microphone arrays (Kinects) and binaural microphones
- 40 hours of training data
- fully transcribed

# Task

## Track1

- ASR on distant microphones
- oracle segmentation
- no external data allowed
- speaker-id provided for each segment, task is to transcribe this speaker (even when there are other overlapping speakers)

## Track2

- ASR + diarization on distant microphones (segmentation for test data not given)
- VoxCeleb data allowed
- task is to provide 4 long transcriptions of the entire session, these are then matched to the speakers in oracle way and scored

# Baseline system

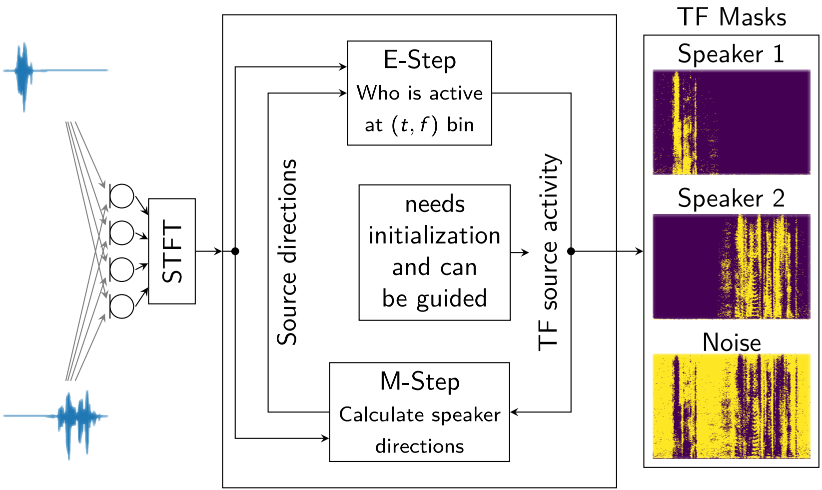
### Track1

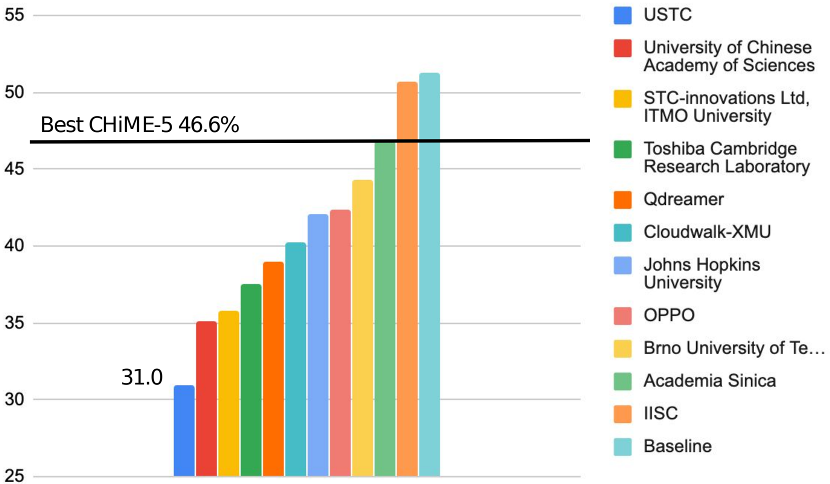
- WPE + Guided source separation as front-end
- Kaldi ASR: TDNNf + LF-MMI, data augmentation, data cleaning, i-vector speaker adaptation
- 3-gram Kneser-Ney language model
- WER: 51.8 % for dev, 51.3 % for eval

### Track2

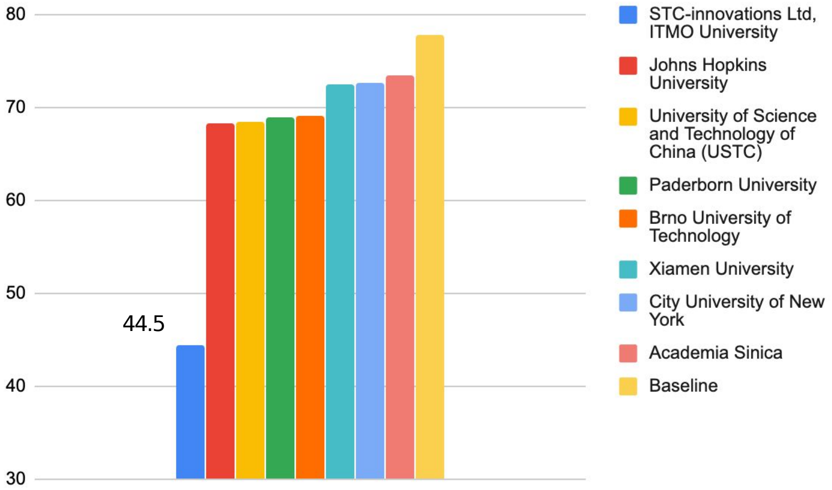
- WPE + BeamformIt as front-end
- SAD: 5-layer TDNN with statistics pooling
- Diarization: agglomerative hierarchical clustering on x-vectors
- ASR system same as Track1
- DER: 63.42 % for dev, 68.20 % for eval
- WER: 84.25 % for dev, 77.94 % for eval

# Baseline: GSS



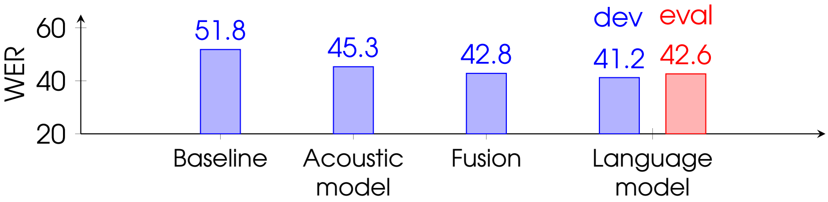

55
50
45
40
35
30
25
Best CHiME-5 46.6%
31.0
USTC
University of Chinese Academy of Sciences
STC-innovations Ltd, ITMO University
Toshiba Cambridge Research Laboratory
Qdreamer
Cloudwalk-XMU
Johns Hopkins University
OPPO
Brno University of Te…
Academia Sinica
IISC
Baseline

# Results: Track2

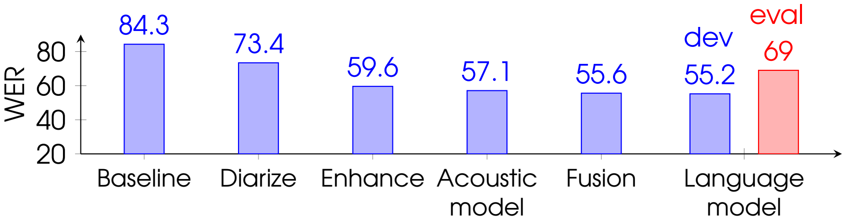


# Outline

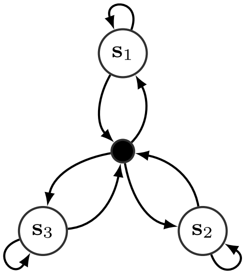
## Track 1




## Track 2



- x-vector clustering based on *Bayesian hidden Markov model* and *variational Bayes inference* (VBx)[1] (Diez et al. 2019)
- states corresponding to speakers, PLDA as state distribution
- x-vector extractor, SAD and PLDA from baseline
- x-vectors extracted every 0.25 seconds




| | Development | | Evaluation | |
|---|---|---|---|---|
| | DER | JER | DER | JER |
| Baseline | 63.42 | 70.83 | 68.20 | 72.54 |
| VBx | 51.67 | 53.20 | 75.81 | 69.46 |

[1] `https://github.com/BUTSpeechFIT/VBx`

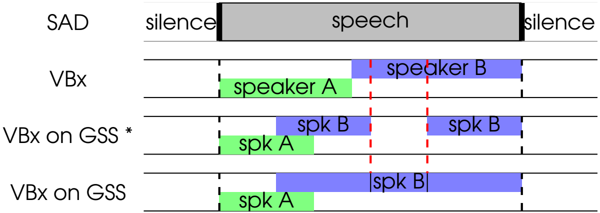
# Diarization

| | Development | | | | Evaluation | | | |
|---|---|---|---|---|---|---|---|---|
| | DER | Miss | FA | SpkErr | DER | Miss | FA | SpkErr |
| Baseline | 63.42 | – | – | – | 68.20 | – | – | – |
| AHC | 63.16 | 26.02 | 10.80 | 26.35 | 73.09 | 23.16 | 20.78 | 29.15 |
| VBx | 51.67 | 26.02 | 10.80 | 14.85 | 75.81 | 23.16 | 20.79 | 31.86 |
| Oracle VAD AHC | 52.40 | 25.60 | 0.01 | 26.80 | 51.91 | 21.75 | 0.02 | 30.14 |
| Oracle VAD VBx | 47.29 | 25.60 | 0.03 | 21.66 | 59.44 | 21.75 | 0.06 | 37.63 |

| | Silence | 1 Speaker | Overlap |
|---|---|---|---|
| Development | 1.07h (24.0%) | 2.04h (45.7%) | 0.97h (21.7%) |
| Evaluation | 1.74h (33.4%) | 2.52h (48.4%) | 0.78h (15.0%) |

# Diarization + Enhancement

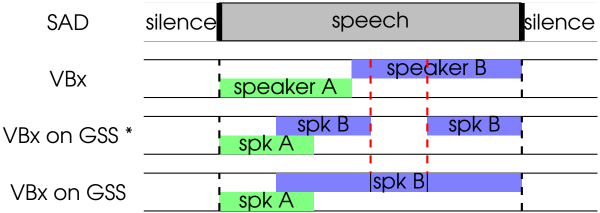
- enhancement by GSS with VBx diarization as guidance (Boeddeker et al. 2018)
- diarization reran on each of enhanced recordings and results combined




| | Development | |
|---|---|---|
| | DER | JER |
| Baseline | 63.42 | 70.83 |
| VBx | 51.67 | 53.20 |
| VBx on GSS | **51.44** | **48.45** |

# Diarization + Enhancement

- enhancement by GSS with VBx diarization as guidance (Boeddeker et al. 2018)
- diarization reran on each of enhanced recordings and results combined




| | Development | | Evaluation | |
|---|---|---|---|---|
| | DER | JER | DER | JER |
| Baseline | 63.42 | 70.83 | **68.20** | 72.54 |
| VBx | 51.67 | 53.20 | 75.11 | 71.77 |
| VBx on GSS | **51.44** | **48.45** | 80.57 | **66.33** |

# Acoustic model: Training data

| | |
|---|---|
| enhanced | training data after GSS |
| Worn (L) | left microphone from worn data |
| Worn (S) | both microphones (stereo) from worn data |
| WornRVB | reverberated worn data with artificial RIRs |
| 250k non-overlapped | 250k utterances from kinects,<br>only parts with 1 speaker |

| | | Size (h) | Track 1 | Track 2 |
|---|---|---|---|---|
| 1 | Worn (L) + enhanced | 200 | 48.94 | - |
| 2 | Worn (S) + enhanced | 300 | 47.85 | 59.29 |
| 3 | (2) + WornRVB | 1050 | 47.57 | 59.22 |
| 4 | (3) + 250k non-overlapped | 1330 | **47.31** | **59.02** |

similar conclusions in (Zorila et al. 2019)

# Acoustic model: Architecture and training

Improvements:

- CNN-TDNNf $>$ TDNNf
- sequence-discriminative training on top of LF-MMI

| | Track1 | Track2 |
|---|---|---|
| TDNNf | 49.37 | 60.64 |
| TDNN-LSTM | 49.95 | 61.62 |
| CNN-TDNNf | 47.85 | 59.29 |
| CNN-TDNNf + sMBR | **47.32** | **58.82** |

- trained on *Worn (S) + enhanced*
- Track 2 uses *VBx + GSS* diarization

# Acoustic model: Others

Improvements:

- semi-supervised training on VoxCeleb
  (system trained on CHiME used as teacher)
- i-vectors clean-up
  *speaker vector:* i-vector extracted from entire session
  *non-overlapped vector:* i-vector extracted from
  non-overlapped parts of the session

| | Track1 | Track2 |
|---|---|---|
| CNN-TDNNf + sMBR | 47.32 | 58.82 |
| (1) + VoxCeleb | 46.80 | **57.92** |
| (1) + speaker + online i-vector | 46.63 | 58.46 |
| (1) + non-overlapped + online i-vector | **46.47** | - |

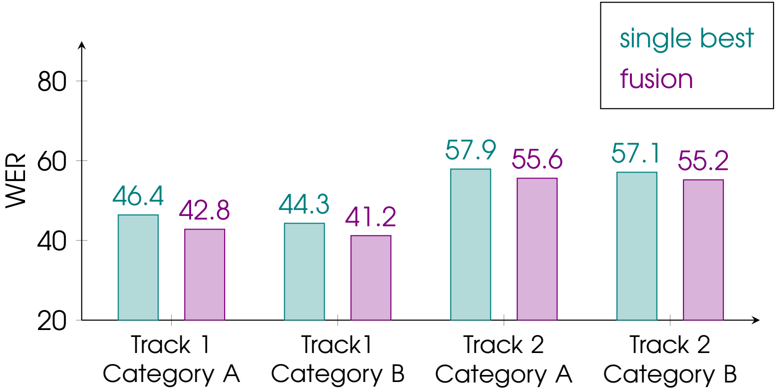
- LSTM language model, BrnoLM toolkit[2]
- rescoring of 3000-best hypothesis
- hidden state of LSTM carried over segments to include context
- regularization:
  - dropout 0.5
  - randomly replacing input tokens with rate 0.3

| | Perplexity | WER (%) |
|---|---|---|
| baseline | 157.7 | 48.24 |
| + LSTM | 152.1 | 46.94 |
| + across-segment | 136.5 | 46.61 |
| + input corruption | **131.1** | **46.08** |

[2] https://github.com/BUTSpeechFIT/BrnoLM

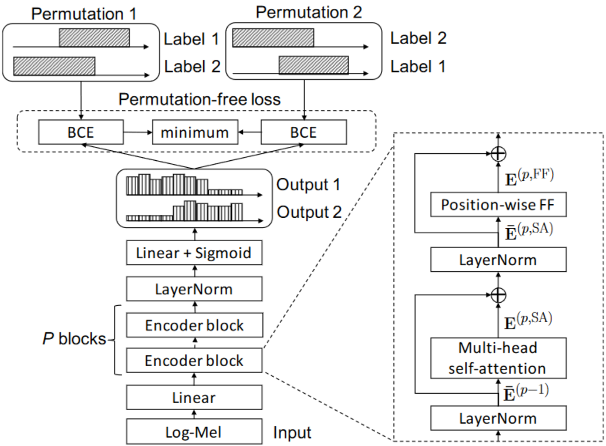
# Fusion

- ROVER fusion over different acoustic models (enhancement and diarization the same in all)
- max(scorea, scoreb) is chosen as the confidence score
- Conf. values were calibrated using simple logistic regression but no gains reflected in %WER
- 7 systems fused for Track1, 8 systems fused in Track2



# Towards end-to-end diarization

- transformer-based system (encoder part), with PIT objective (Fujita et al. 2019)
- overlaps allowed



# Towards end-to-end diarization

- transformer-based system (encoder part), with PIT objective (Fujita et al. 2019)
- overlaps allowed
- mismatch between training annotations and “new RTTMs”

| Method | Data | Del 1min | VoxCeleb pretrain | DER (%) Old RTTMs | DER (%) New RTTMs |
|---|---|---|---|---|---|
| Baseline | – | – | – | 59.87 | 63.25 |
| VBx | – | – | – | 50.83 | 51.67 |
| E2E | CH1 | ✗ | ✗ | 70.3 | 80.6 |
| E2E | CH1 | ✗ | ✓ | 64.6 | 73.9 |
| E2E | CH1 | ✓ | ✓ | 63.6 | 71.7 |
| E2E | mix | ✓ | ✓ | 63.5 | 71.7 |
| E2E | WPE+mix | ✓ | ✓ | **62.4** | **70.9** |

VoxCeleb pretrain — “conversations” of 2 speakers simulated from VoxCeleb data

Del 1min — omitting first minute with introductions

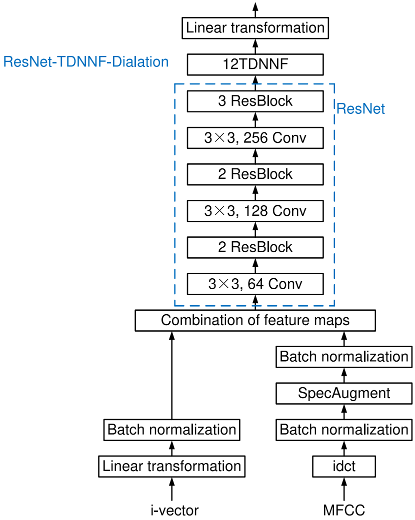
# Towards end-to-end ASR

| Acoustic model<br>(Training data)<br>(Architecture) (Target units) | Dev-worn | Dev-enhanced |
|---|---|---|
| LSTM<br>(worn+enhanced)<br>(5enc-1dec-320H)(char) | 60.19 | 66.51 |
| Transformer<br>(worn)<br>(6enc-6dec-256H-4heads)(char) | 66.06 | 73.39 |
| Transformer<br>(worn-data+enhanced)<br>(12enc-6dec-256H-4heads)(char) | 64.66 | 68.70 |
| Transformer<br>APC-Pre-training(voxcelb)+(worn+enhanced)<br>(12enc-6dec-256H-4heads)(char) | 61.60 | 66.7 |

# Outline

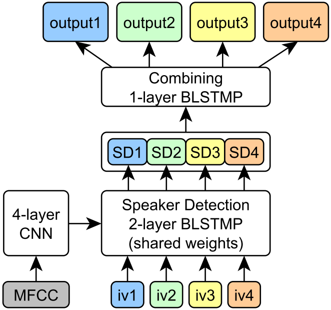
# USTC system (Track1 winners)

- a bit unclear where improvements are coming from ☹
- post-processing of GSS outputs with a neural network (+2%)
- data augmentation (worn + GSS + volume and speed pertrubation + SpecAugment)
- acoustic model architectures (ResNet-TDNNF-Dialation)



# STC system (Track2 winners)

- TS-VAD!!!
- Wide residual networks for x-vectors + 10% DER/JER
- Spectral clustering of x-vectors +5-10% DER/JER
- GSS
  - increasing context, iterations, of mics, +1%
  - Using soft activity from TS-VAD, +3%
- SpecAugment +1%
- AM with multi-stride and multi-stream attention +2%
- Quite nice improvements from LM (+2%)



# Other highlights

- SpecAugment in many systems, seems to give a good improvement
- generally many different AM architectures (Resnets, Attention), also bigger
- no real break-throughs on separation
- STC also had a paper with end-to-end ASR with reasonable results

# Outline

# Retrospective

## What went good

- many people interested and experimenting
- connected different parts of the group
- beaten (strong) baseline and good diarization
- interesting research threads started
- managed to co-operate distantly

## What to improve

- earlier start?
- avoid the rushed ending
- have more tools ready to use
- keep better track of current outputs / to-dos

# Future

- research opportunities
  - Igi - denoising / RIR estimation
  - Alicia - end-to-end diarization
  - Hari - end-to-end ASR
  - Katka - speech separation
  - KarelB - ASR-aware language model
  - Fede - multichannel VBx
- implementing good stuff from other teams
  - different acoustic model architectures
  - TS-VAD
  - x-vector / diarization improvements (Resnet, angular margin softmax)
  - SpecAugment
- making of nice simple unified recipe

# References

Christoph Boeddeker et al. "Front-end processing for the CHiME-5 dinner party scenario". In: CHiME5 Workshop, Hyderabad, India. 2018.

Mireia Sánchez Diez et al. "Analysis of Speaker Diarization based on Bayesian HMM with Eigenvoice Priors". In: IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING 28.1 (2019), pp. 355–368. ISSN: 2329-9290. DOI: 10.1109/TASLP.2019.2955293. URL: https://www.fit.vut.cz/research/publication/12139.

Yusuke Fujita et al. End-to-End Neural Speaker Diarization with Permutation-Free Objectives. 2019. arXiv: 1909.05952 [eess.AS].

Catalin Zorila et al. "An Investigation into the Effectiveness of Enhancement in ASR Training and Test for CHiME-5 Dinner Party Transcription". In: arXiv preprint arXiv:1909.12208 (2019).