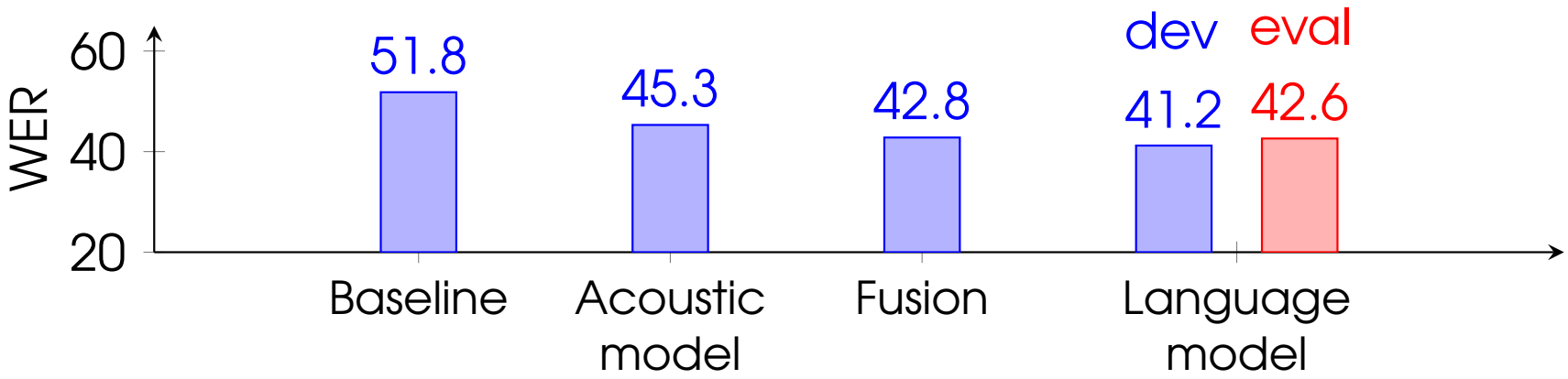
BUT System for CHiME-6 Challenge

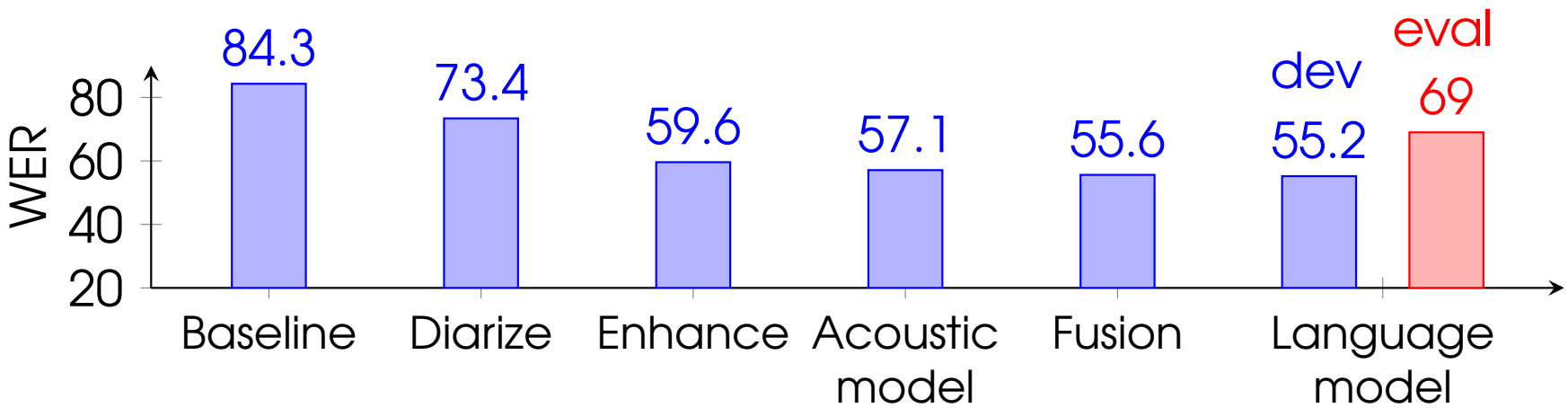
K. Žmolíková, M. Kocour, F. Landini, K. Beneš, M. Karafiát,
H. K. Vydana, A. Lozano-Diez, O. Plchot, M. K. Baskar,
J. Švec, L. Mošner, V. Malenovský, L. Burget, B. Yusuf,
O. Novotný, F. Grézl, I. Szöke, J. Černocký

Overall results

Track 1

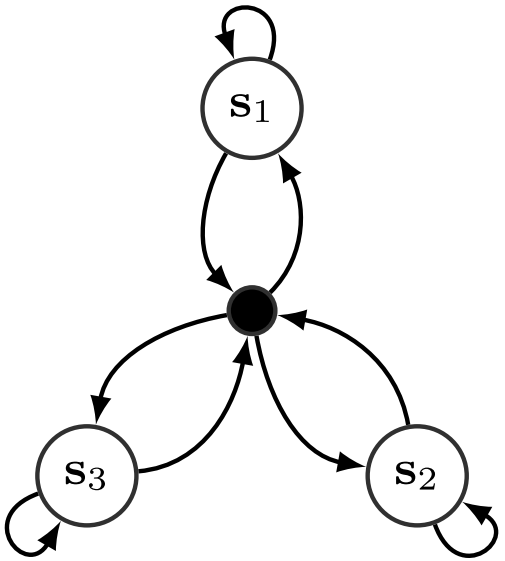


Track 2



Diarization

- 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

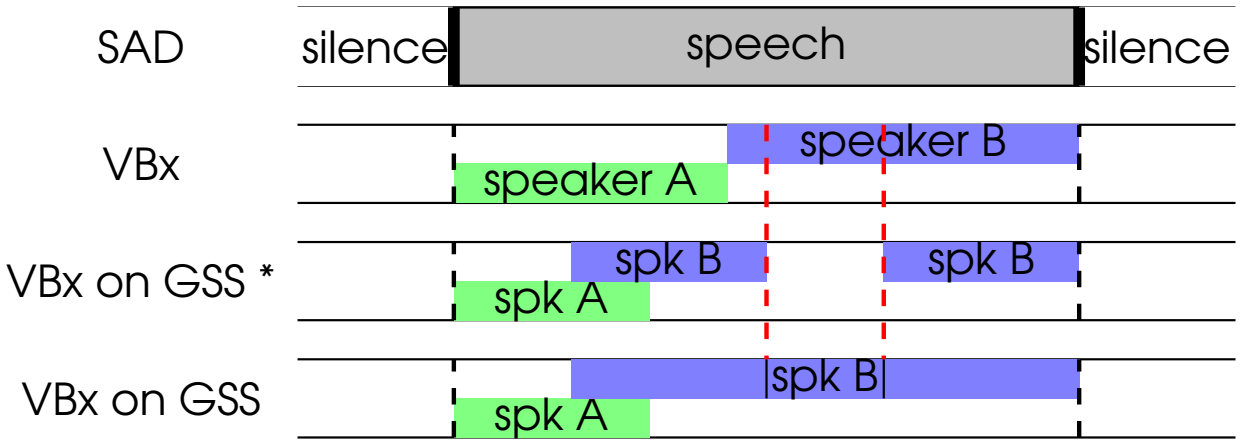


	Development	
	DER	JER
Baseline	63.42	70.83
VBx	**51.67**	**53.20**

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

Diarization + Enhancement

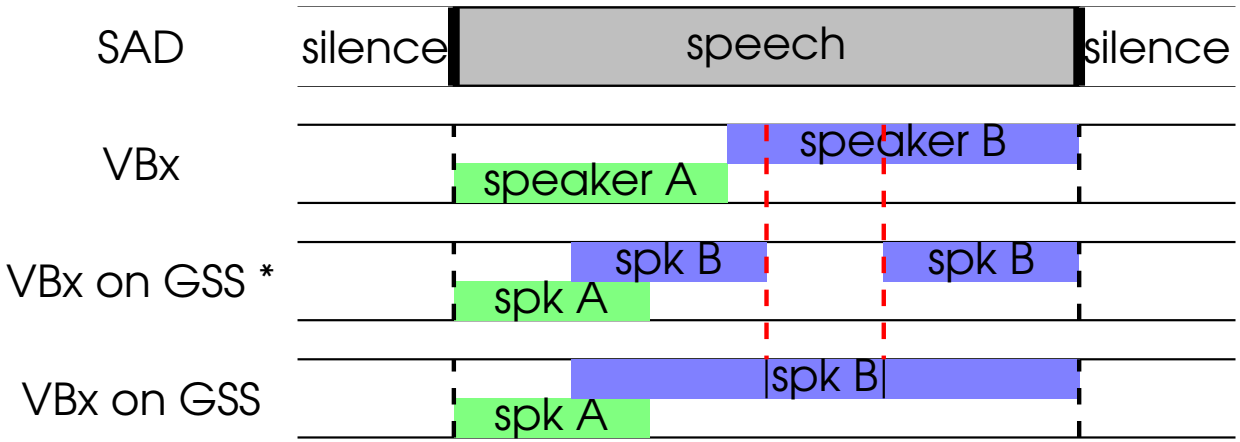
- enhancement by GSS with VBx diarization as guidance (Boeddeker et al. 2018)
- diarization reran on each of enhanced recordings and results combined



	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, 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
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**	-

Language model

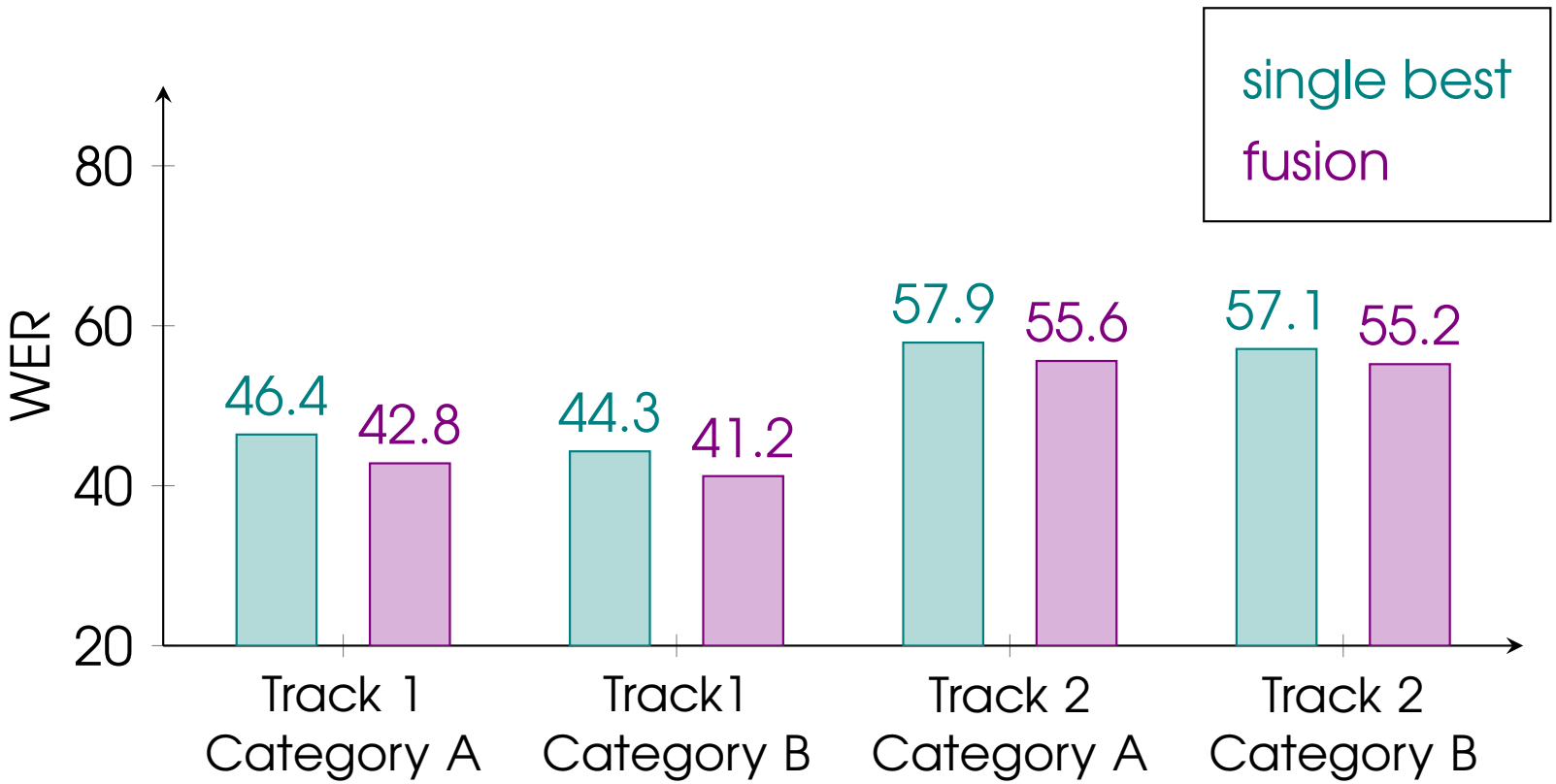
- LSTM language model, BrnoLM toolkit[2]
- rescoring of 3000-best hypothesis
- hidden state of LSTM carried over segments to include context
- regularization:
  - dropout 0.5
  - randomly replacing input tokens with rate 0.3

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

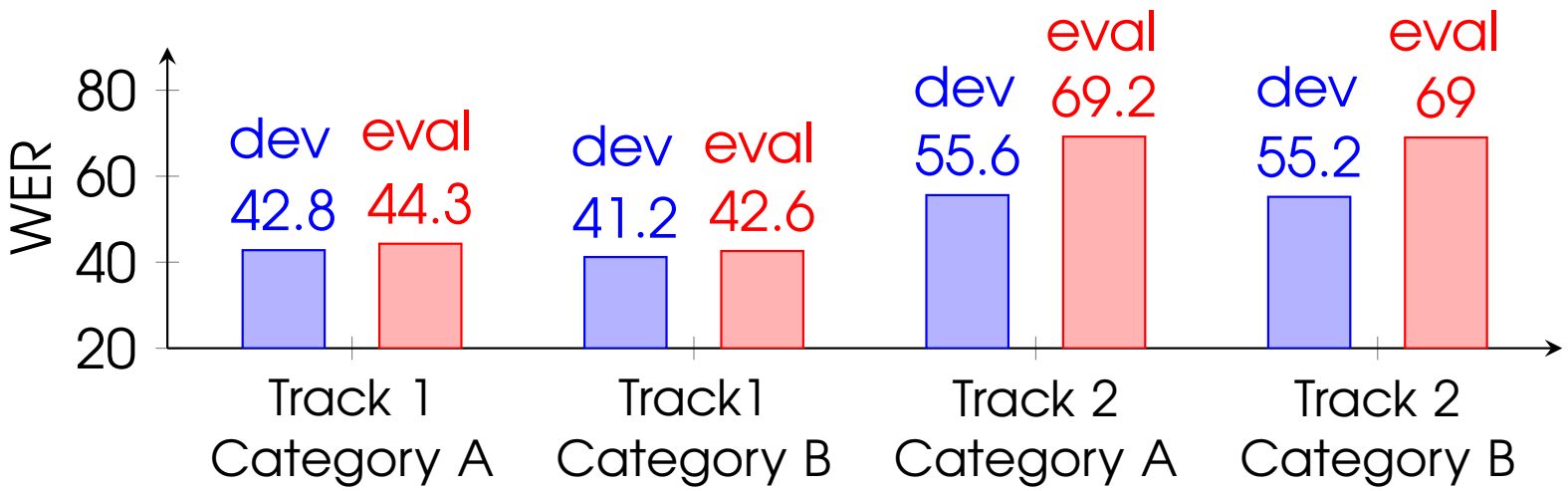
[2] https://github.com/BUTSpeechFIT/BrnoLM

Fusion

- ROVER fusion over different acoustic models (enhancement and diarization the same in all)
- 7 systems fused for Track1, 8 systems fused in Track2



Conclusion



Improvements from:

- **Diarization** VBx, so far not effective on evaluation data
- **Acoustic model** data, architecture, training
- **Language model** LSTM-LM, context, regularization

Thank you to the organizers of the challenge!

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).

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"

Data	Del 1min	VoxCeleb pretrain	DER (%) Old RTTMs	DER (%) New RTTMs
CH1	✗	✗	70.3	80.6
CH1	✗	✓	64.6	73.9
CH1	✓	✓	63.6	71.7
mix	✓	✓	63.5	71.7
WPE+mix	✓	✓	**62.4**	**70.9**

VoxCeleb pretrain — "conversations" of 2 speakers simulated from VoxCeleb data

Del 1min — omitting first minute with introductions from training data

Table: End-to-End ASR models

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