CHiME WORKSHOP
The 4th International Workshop on
Speech Processing in Everyday Environments
formerly the International Workshop on Machine Listening in Multisource Environments

The USTC-iFlytek System For CHiME-4 Challenge

Jun Du

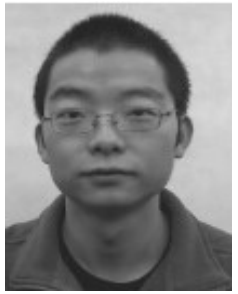
2016.09.13

Team

University of Science and Technology of China

(USTC)

Jun Du



Yan-Hui Tu

Lei Sun

iFlytek Research

Feng Ma

Hai-Kun Wang

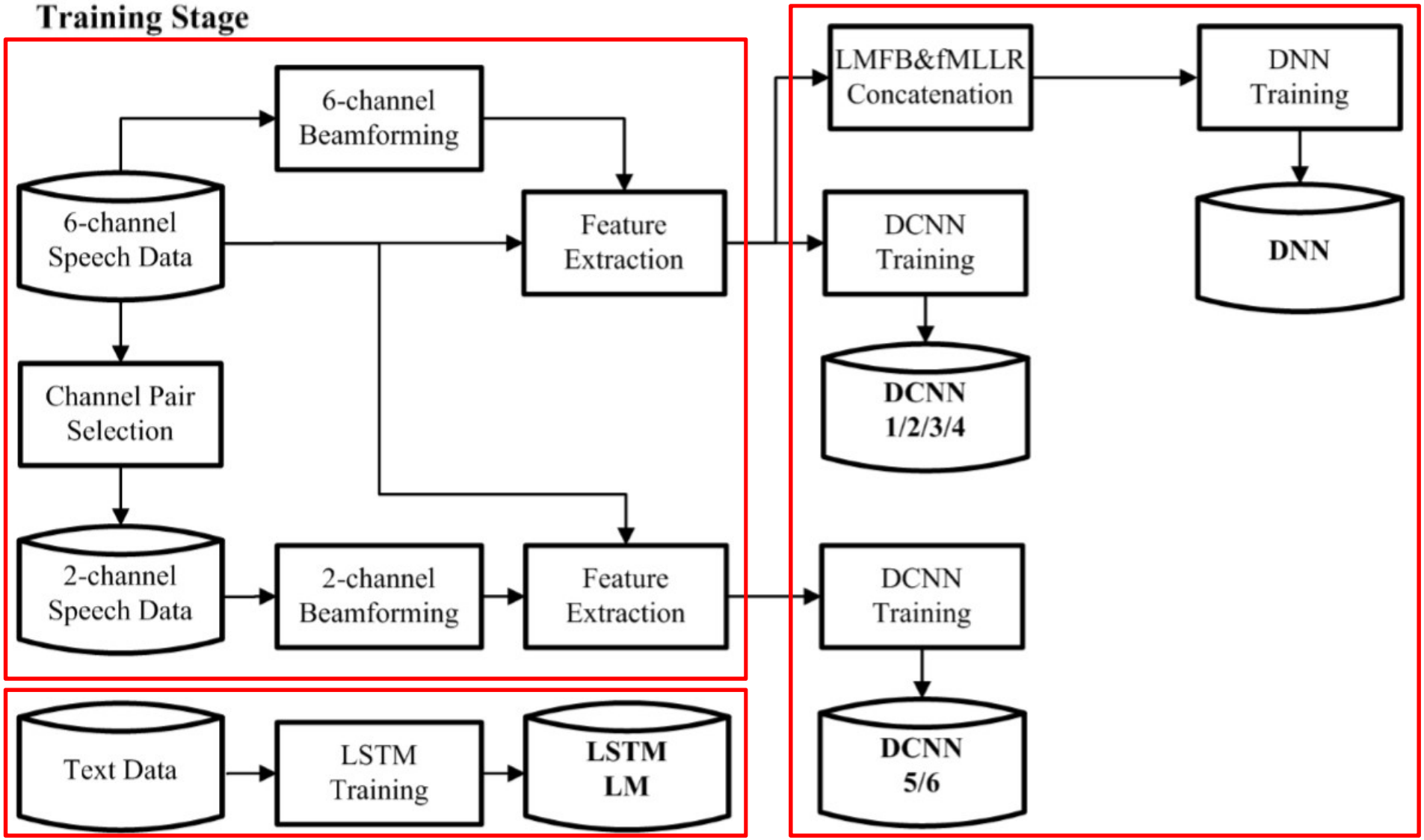
Jia Pan

Cong Liu

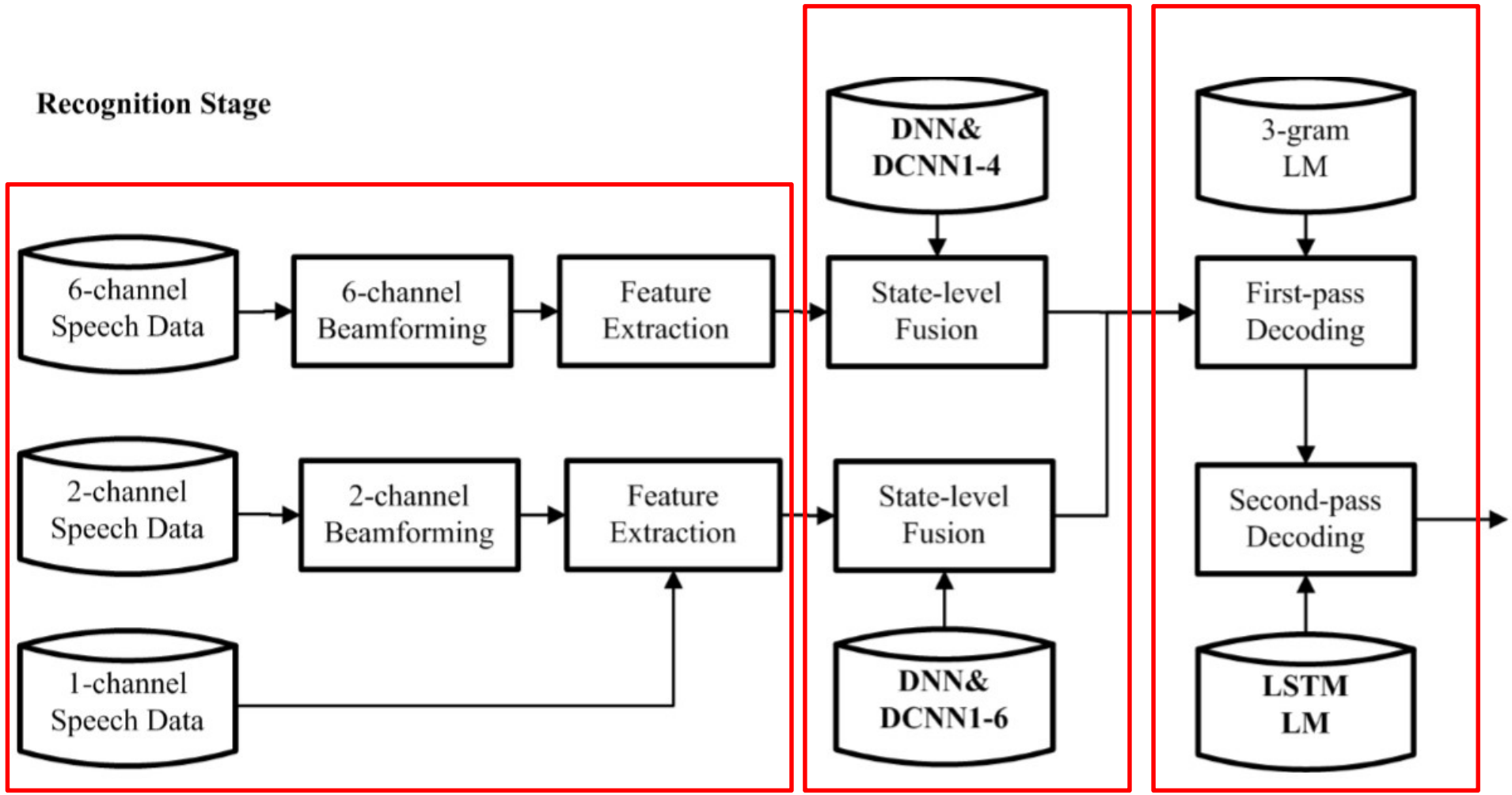
Georgia Institute of Technology

Chin-Hui Lee

Joint Framework For **X**-channel Tasks (I)



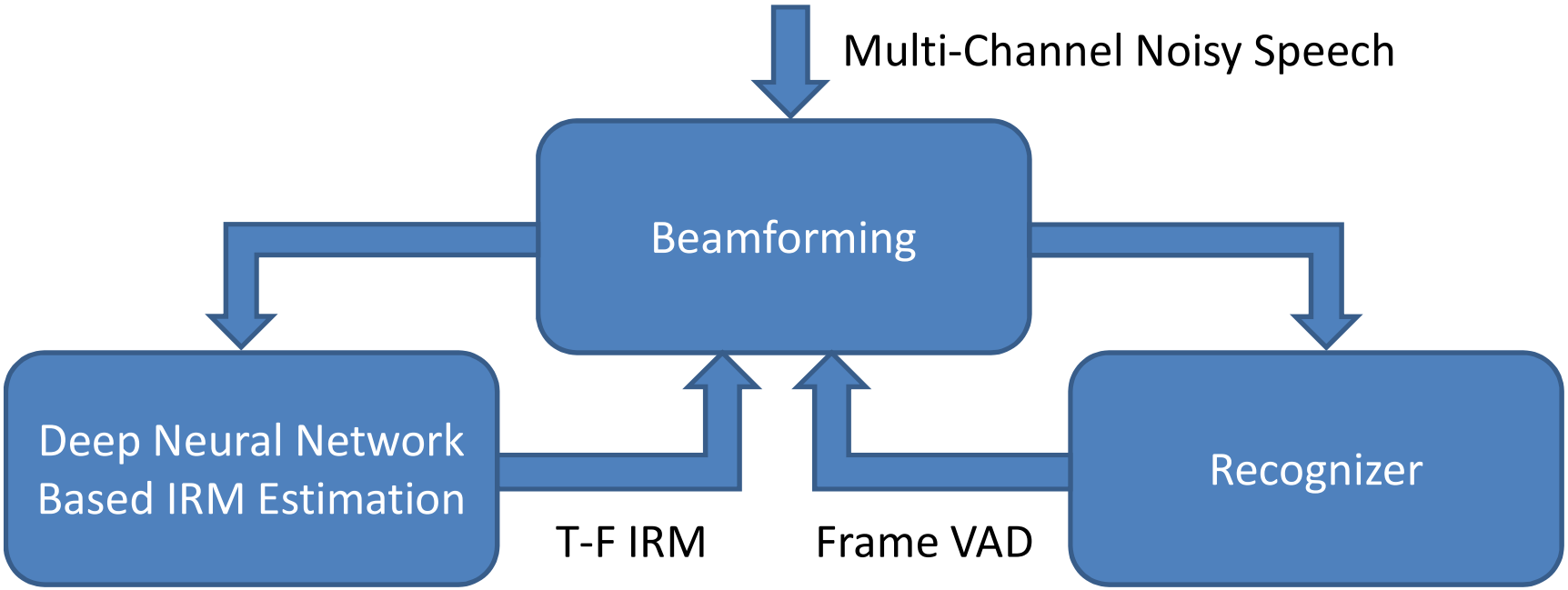
Joint Framework For **X**-channel Tasks (II)

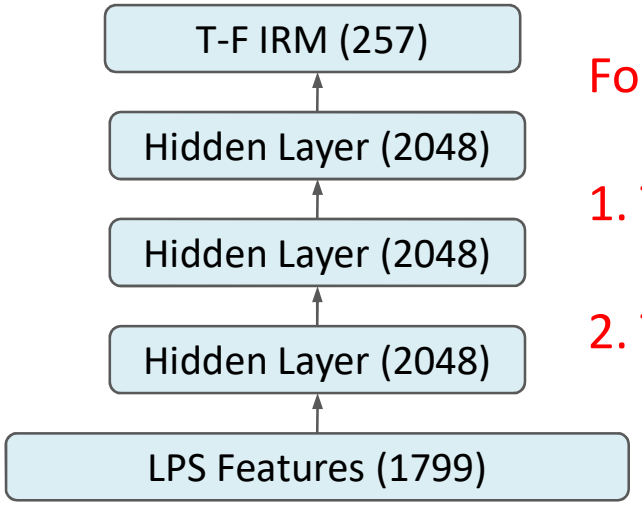


Implementation

- The official Kaldi recipe
  - Features: fMLLR and LMFB features
  - DNN-HMM acoustic model: concatenating fMLLR and LMFB
  - Model ensemble and two-pass decoding
- CNTK toolkit: IRM-DNN training
- Self-developed toolkit
  - Beamforming
  - DCNN-HMM acoustic model (only CE training)
  - LSTM language model

Feedback Loop Optimization



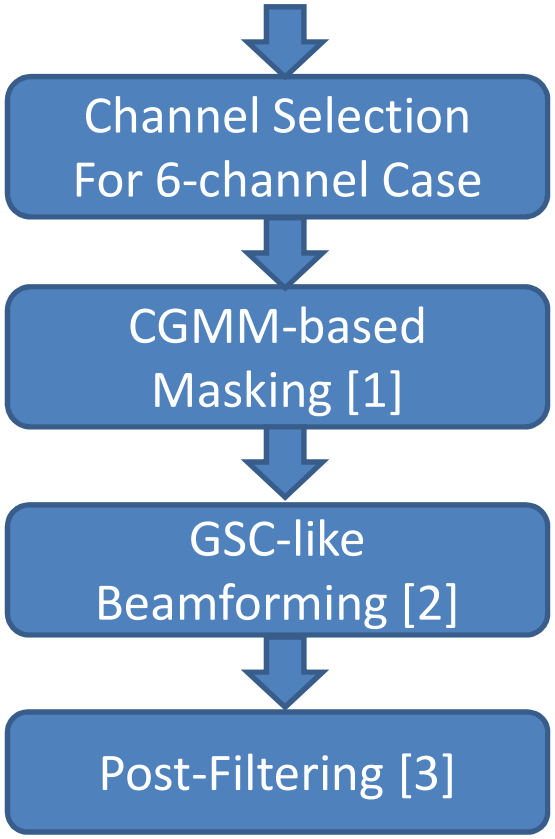


For the IRM-DNN training:

1. The output T-F IRM is defined between CH5 and clean

2. The input is the LPS feature of CH5

Beamforming



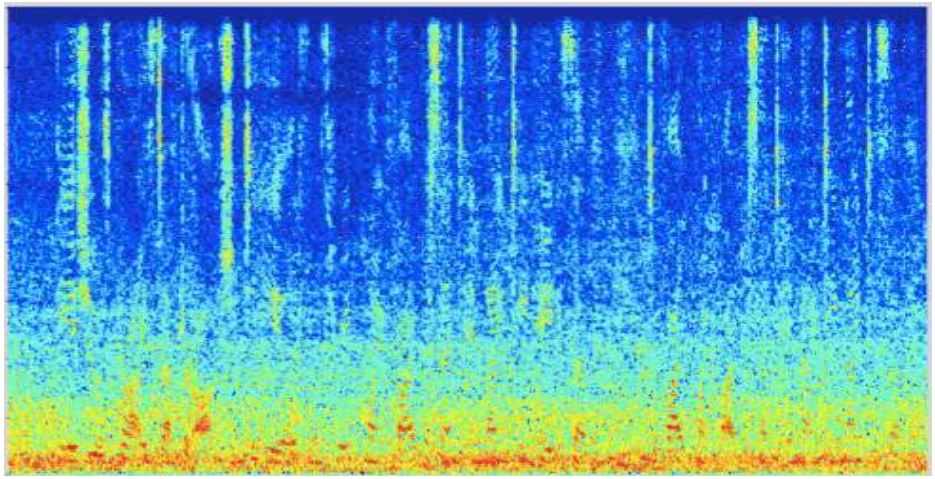
1. CGMM to estimate the noise/noisy covariance matrix
2. **Frame VAD and DNN-IRM to improve the masking**
3. Frame-level VAD to determine the noise segment
4. DNN-IRM to determine T-F units in speech segments

[1] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in ICASSP, 2016.

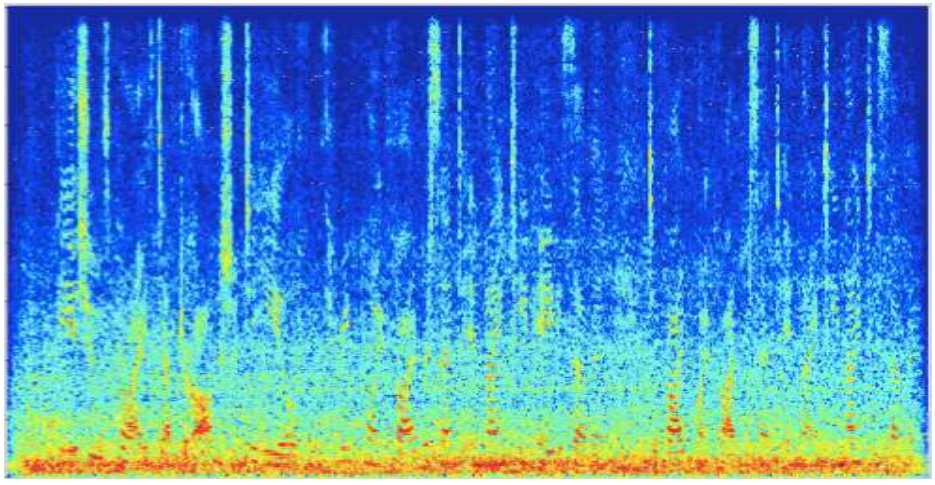
[2] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a gsc-like structure employing eigenvector-based transfer function ratios estimation," IEEE TASLP, vol. 19, no. 1, pp. 206–219, 2011.

[3] L. Wang, T. Gerkmann, and Simon Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," IEEE TASLP, vol. 23, no. 9, pp.1493-1508, 2015.

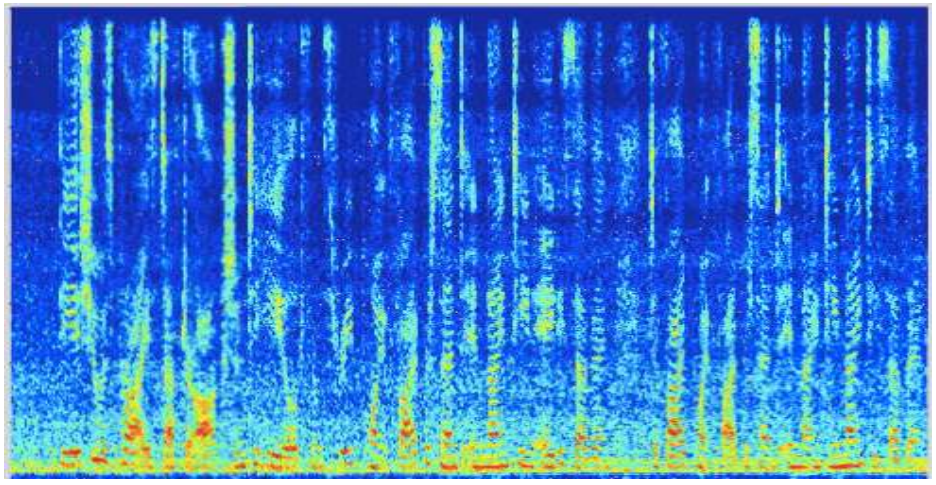
Spectrogram Comparison



CH5 (F06_446C020B_STR_REAL)

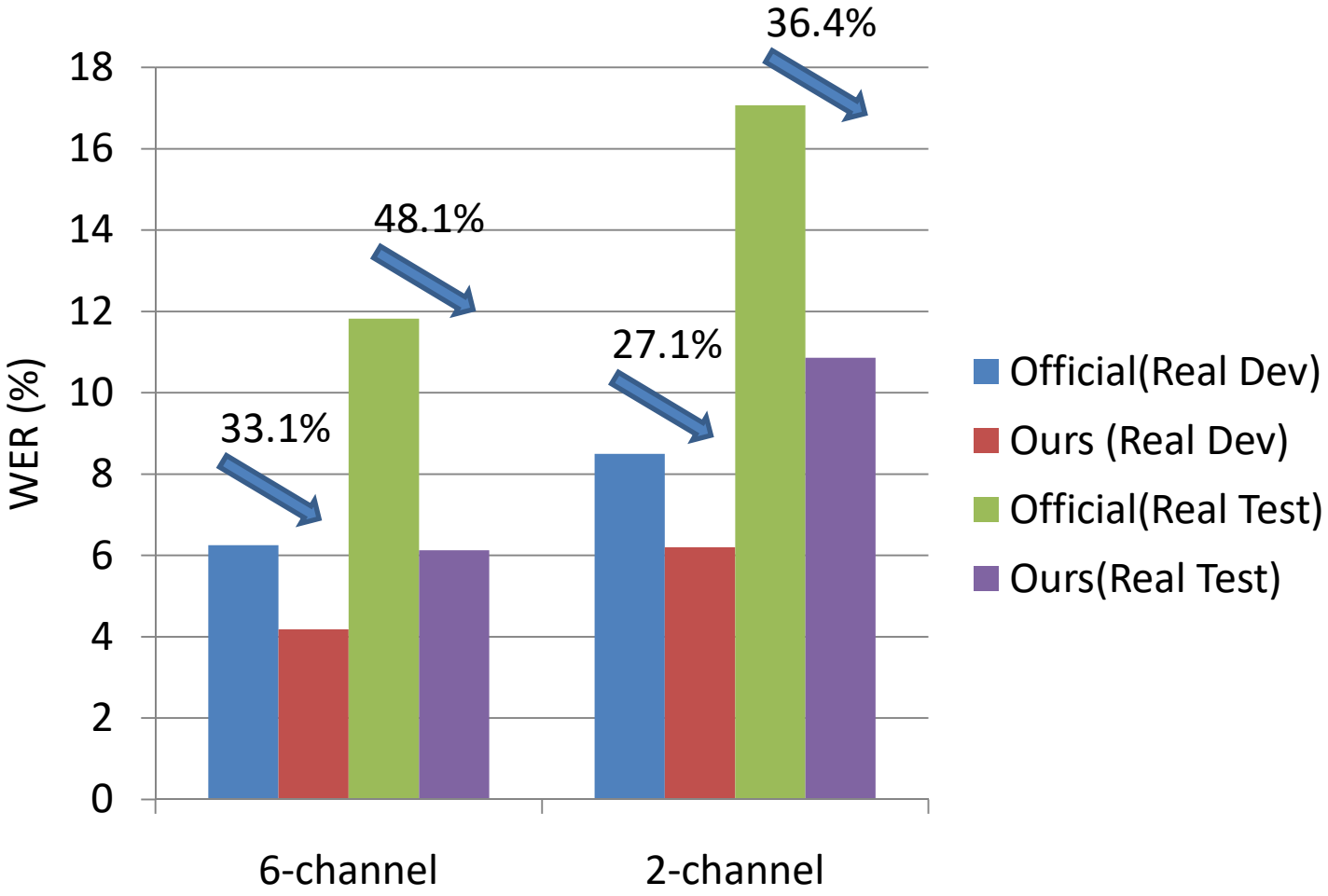


The official 6-channel beamforming



The proposed 6-channel beamforming

Beamforming (Official vs. Ours)



Evaluation on the official baseline DNN system

More effective for more adverse environments and more microphones!

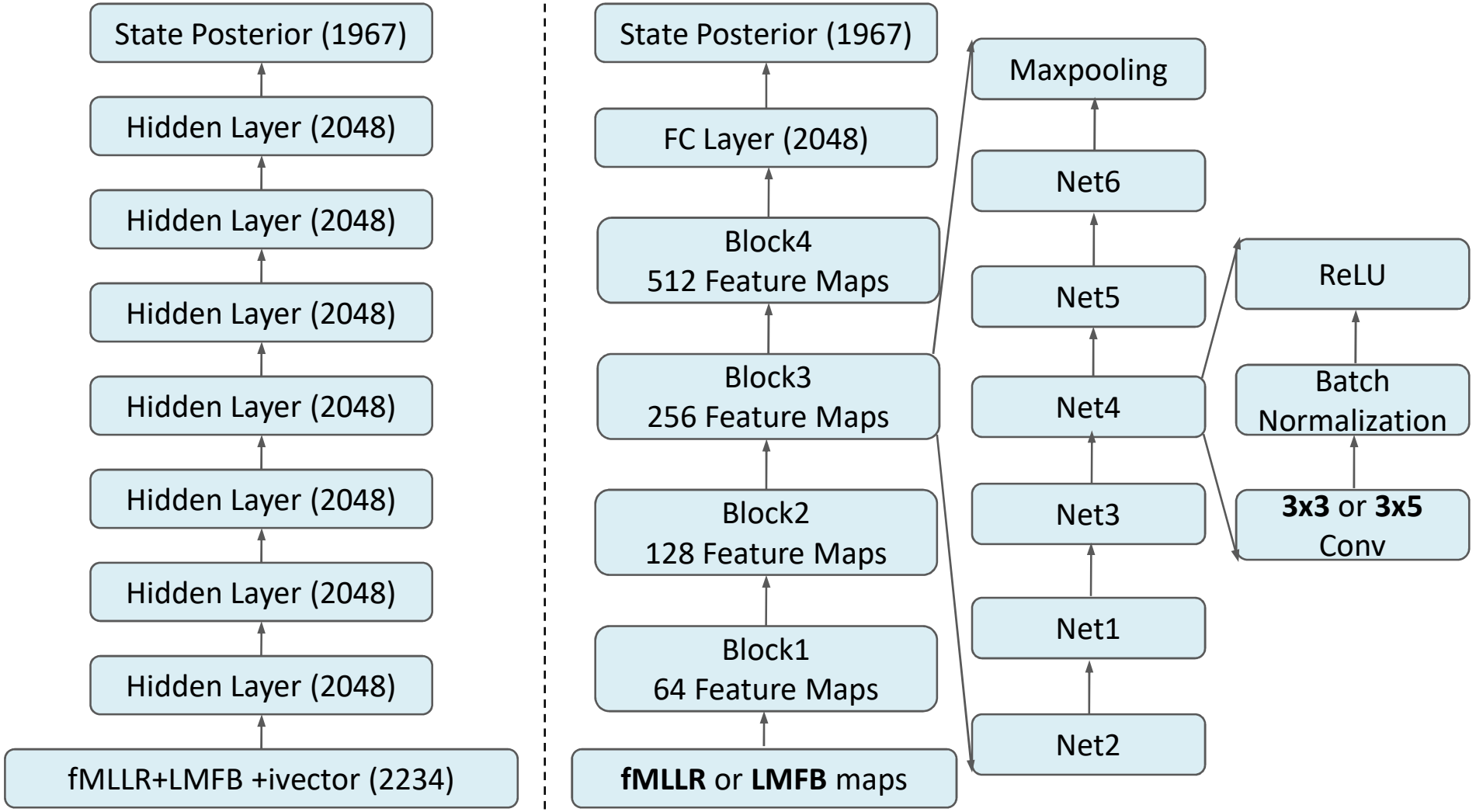
Training Data Augmentation

- Multi-style training
  - A: 1-channel (1,3,4,5,6) noisy speech simulating 1-channel case
  - B: 2-channel beamformed speech simulating 2-channel case
  - C: 6-channel beamformed speech simulating 6-channel case

- Training for 6-channel case
  - A+C for 1 DNN and 4 DCNNs

- Training for 2-channel and 1-channel cases
  - A+C for 1 DNN and 4 DCNNs, A+B for 2 DCNNs

Acoustic Model

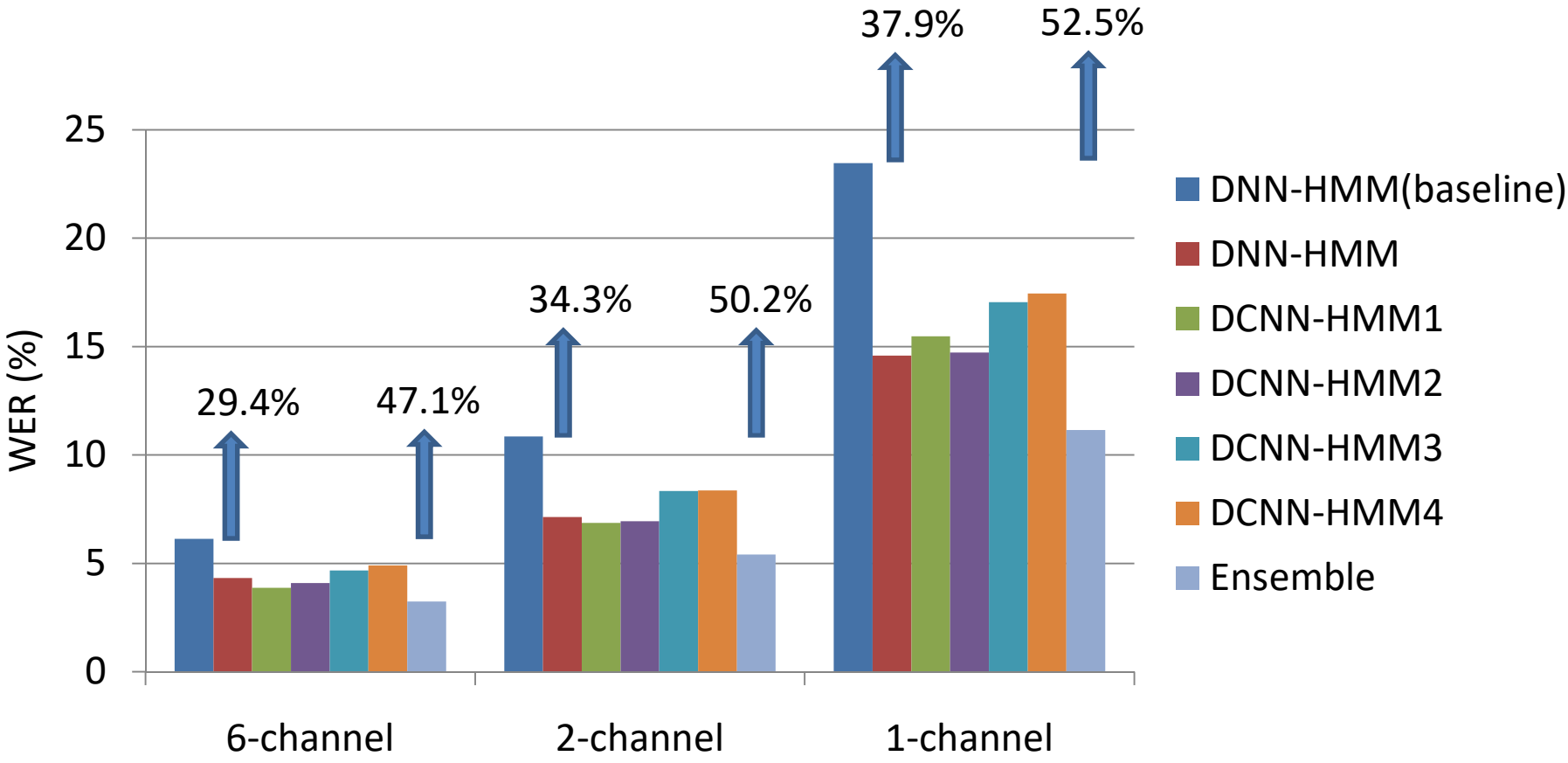


DNN-HMM (Sequence Training)

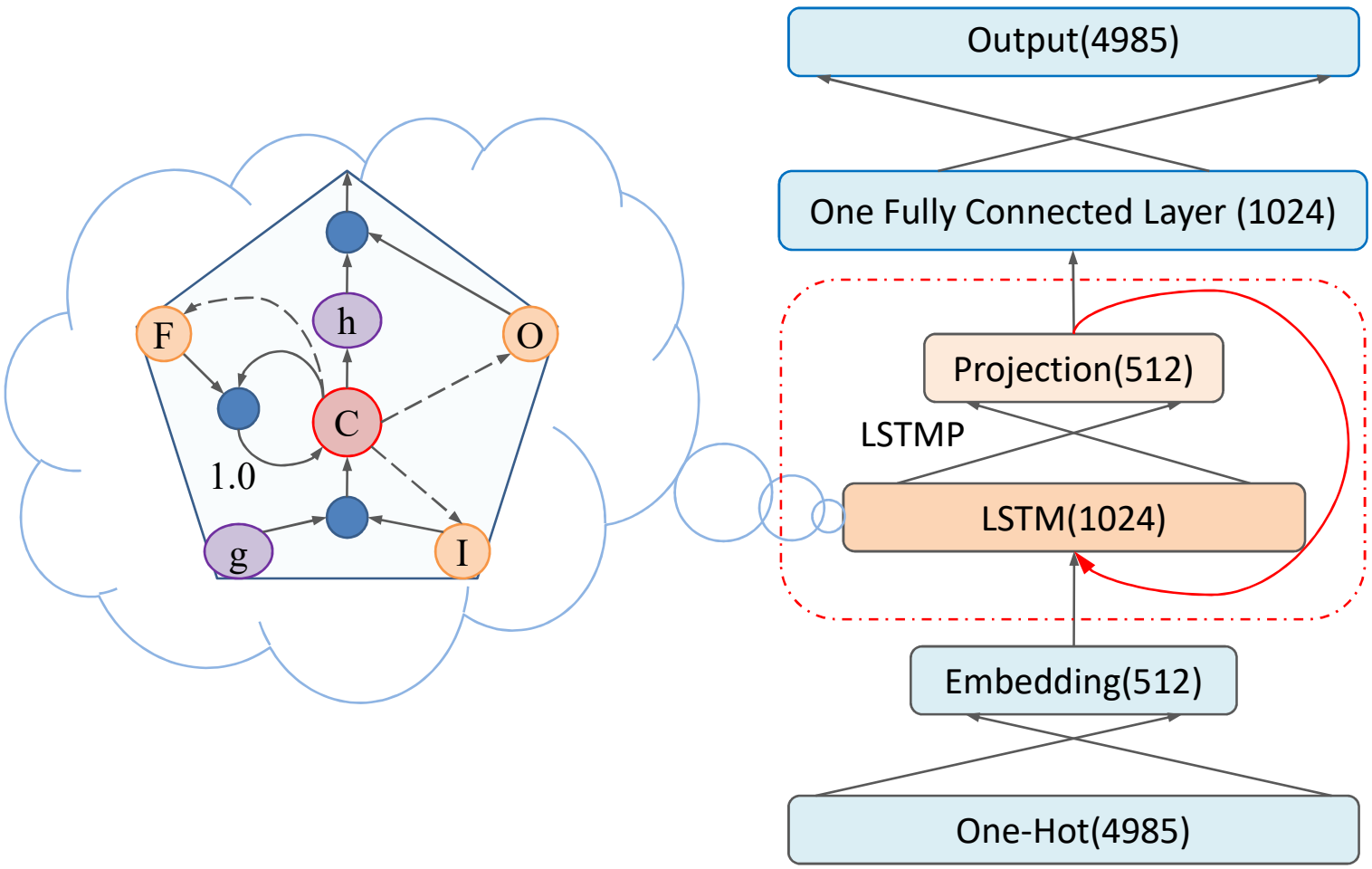
DCNN-HMM (Cross-Entropy Training)

Model Ensemble

- Ensemble via the state posterior average of NN output
- For 6-channel, 5-model ensemble (DNN,DCNN1-4)
- For 2-channel and 1-channel, 7-model ensemble (DNN,DCNN1-6)

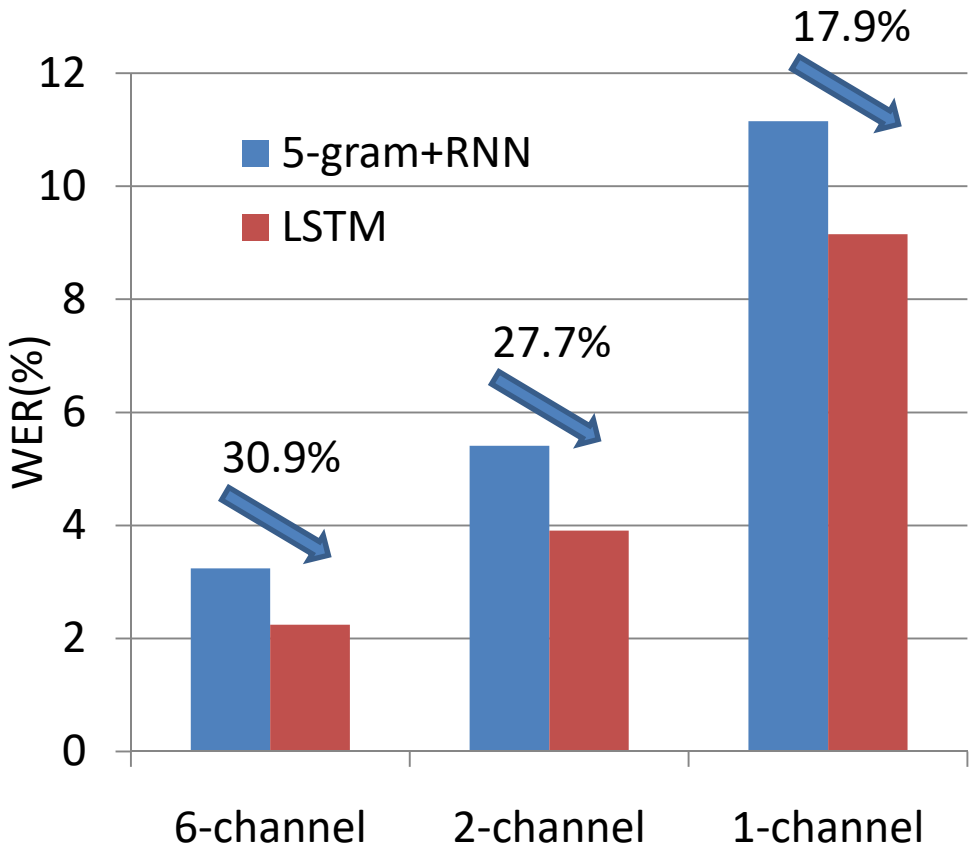


Language Model



Two LSTM-LMs (Forward and Backward) are combined

5-gram+RNN vs. LSTM



Evaluation on real test set for the **best configured** system

Better front-end and acoustic models, more effective LSTM-LM!

Summary

Table 2: WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	5.84	4.90	14.10	7.58
	CAF	5.09	9.84	9.64	14.98
	PED	2.66	4.84	6.89	11.58
	STR	4.63	6.86	5.98	13.09
2ch	BUS	2.74	2.83	5.16	3.83
	CAF	2.18	4.29	3.83	5.66
	PED	1.73	2.94	3.18	6.14
	STR	2.65	3.79	3.49	7.32
6ch	BUS	2.05	1.64	2.65	1.36
	CAF	1.50	1.99	2.09	1.87
	PED	1.50	1.55	1.74	2.35
	STR	1.71	1.93	2.48	2.91

The best system for all tasks (1ch: 9.15%, 2ch:3.91%, 6ch:2.24%)

Thanks!

Q&A