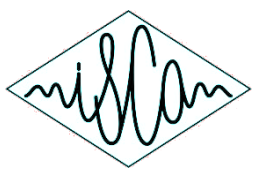
The 4th CHIME Speech Separation and Recognition Challenge

Emmanuel Vincent, Inria
Shinji Watanabe, Mitsubishi Electric Research Labs
Jon Barker, Ricard Marxer, University of Sheffield



Overview

1. From CHiME-1 to CHiME-3
2. Environment, simulation, and mic mismatches in CHiME-3
3. CHiME-4 tracks and baselines
4. CHiME-4 results
5. Discussion

CHiME-1 and 2

Initial challenges in the CHiME series (2011, 2013).

Focus: real noise backgrounds composed of multiple competing sources.

Scenario:

- domestic noise, single room
- clean speech from Grid/WSJ 5k
- mixed using fixed/time-varying impulse responses recorded with a binaural mic pair at 2 m distance.

Top CHiME-1 systems came close to human performance.

CHiME-2 stepped in the right direction but doubts remained about using simulated data (i.e. artificially mixed speech + noise).

CHiME-3 objectives

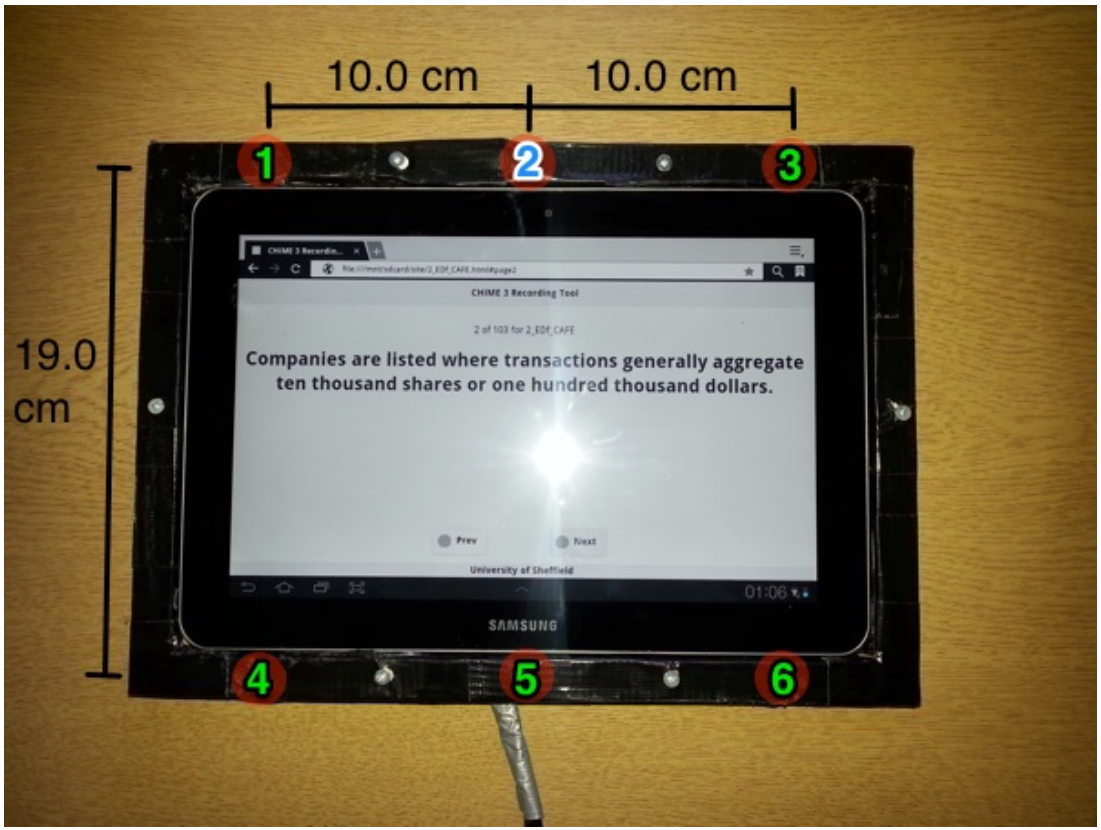
CHiME-3 challenge held in 2015 with the following objectives:

- commercially relevant scenario and hardware,
- more varied noise environments,
- real data, i.e. utterances spoken and recorded live in noise,
- use of close-talking mic to create matched simulated data.

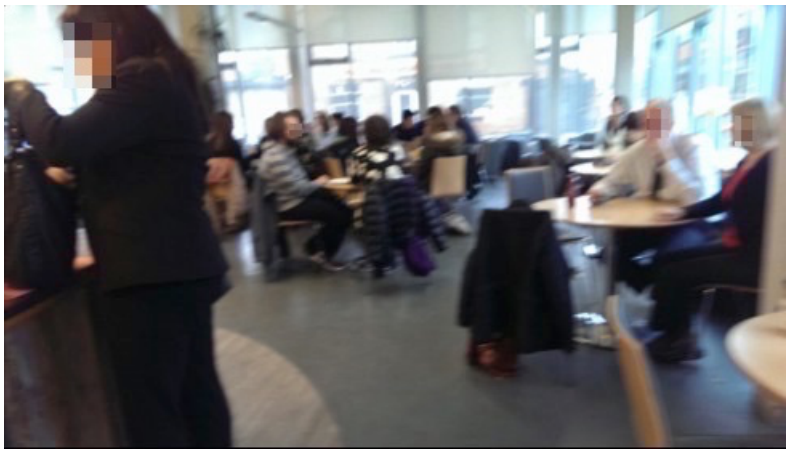
CHiME-3 scenario and hardware

ASR running on a tablet device being used in noisy everyday environments.

WSJ 5k sentences respoken live and recorded by a custom array of 6 mics (5 forward, 1 backward, occasional mic failures), plus a close-talking mic.



CHiME-3 environments



Sitting in a cafe (**CAF**)

Standing at a street junction (**STR**)

Travelling on a bus (**BUS**)

In a pedestrian area (**PED**)

CHiME-3 datasets

Real data recorded from 12 native US talkers.

Simulated data created by:

- estimating speaker movements, SNR, and noise signal from real data,
- remixing clean speech with corresponding time-varying delay and same noise signal or other noise signal with same SNR.

Speaker ID known, but environment ID supposed unknown.

Dataset		# speakers	# utterances
Training	real	4	1600
	simu	83	7138
Devel	real	4	410
	simu	4	410
Test	real	4	330
	simu	4	330

CHiME-3 results

WER on real data decreased from 33.4% (baseline) to 5.8% (best) by

- training data augmentation,
- robust multichannel speech enhancement,
- feature normalization,
- advanced DNN-HMM acoustic modeling,
- RNN language modeling.

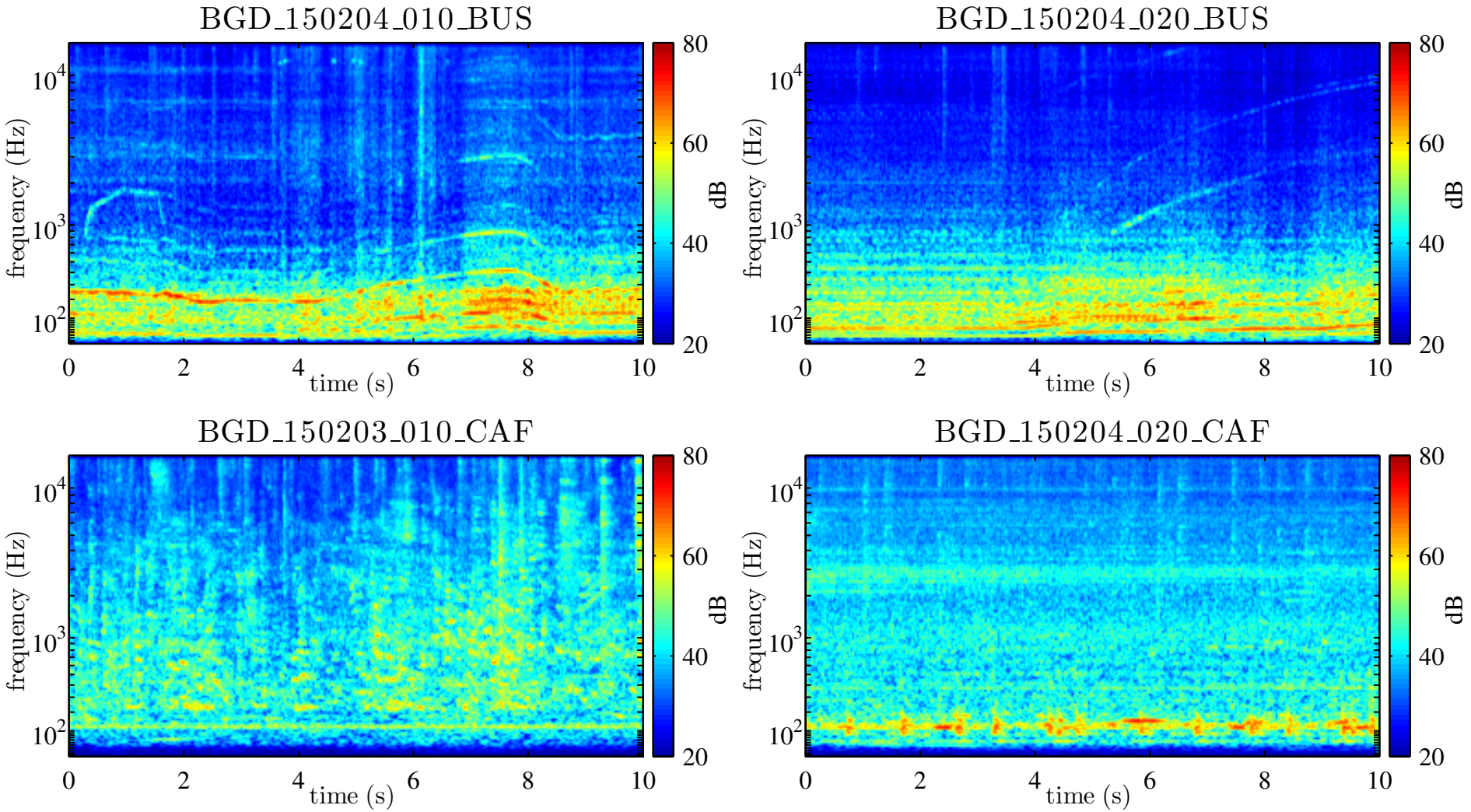
Best WER close to clean speech performance, possibly due to matched hardware and "all-inclusive" training data.

How much do environment, data simulation, or mic mismatch affect learning-based enhancement and ASR performance?

E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition", *Computer Speech and Language*, to appear

Environment mismatch

Noise characteristics vary within and across environments.



Environment mismatch — Impact on enhancement

WER achieved by multichannel DNN-based enhancement + DNN-HMM backend trained on the full training set enhanced by the enhancement system to be evaluated

Training (real)	Test (real) BUS	CAF	PED	STR	Avg.
BUS	21.03	13.06	17.92	9.28	15.32
CAF	31.48	13.15	16.95	8.78	17.59
PED	27.89	12.20	17.04	8.93	16.51
STR	24.30	11.80	16.42	8.48	15.25
1/4 of all	20.83	11.65	15.94	8.72	14.28
all but BUS	22.62	10.72	15.47	7.55	14.09
all but CAF	18.90	10.59	16.07	7.53	13.27
all but PED	18.56	10.76	14.93	8.09	13.08
all but STR	18.19	10.03	15.08	7.94	12.81
3/4 of all	18.84	10.98	15.41	7.79	13.26

1 training environment:
Multicondition: 14.28%
Matched: 14.93%
Mismatched: 16.58%

3 training environments:
Multicondition: 13.26%
Mismatched: 14.02%

⇒ multicondition training preferable to matched training

⇒ on average, performs well on environments not seen in training

⇒ a few large differences for certain pairs of environments

Environment mismatch — Impact on ASR

WER achieved by no enhancement + speaker-independent GMM-HMM backend

Training (real + simu)	Test (real) BUS	CAF	PED	STR	Avg.
BUS	66.33	56.03	51.18	34.37	51.97
CAF	71.99	44.98	40.99	34.12	48.02
PED	70.46	46.51	38.86	36.03	47.96
STR	68.84	52.56	47.80	30.93	50.03
1/4 of all	65.00	47.63	42.66	31.75	46.76
all but BUS	62.07	43.72	36.02	28.52	42.58
all but CAF	61.43	44.51	38.60	27.14	42.92
all but PED	63.40	44.75	40.34	28.82	44.32
all but STR	61.48	41.47	36.00	27.20	41.46
3/4 of all	62.09	43.07	37.18	27.29	42.40

1 training environment:
Multicondition: 46.76%
Matched: 45.28%
Mismatched: 50.91%

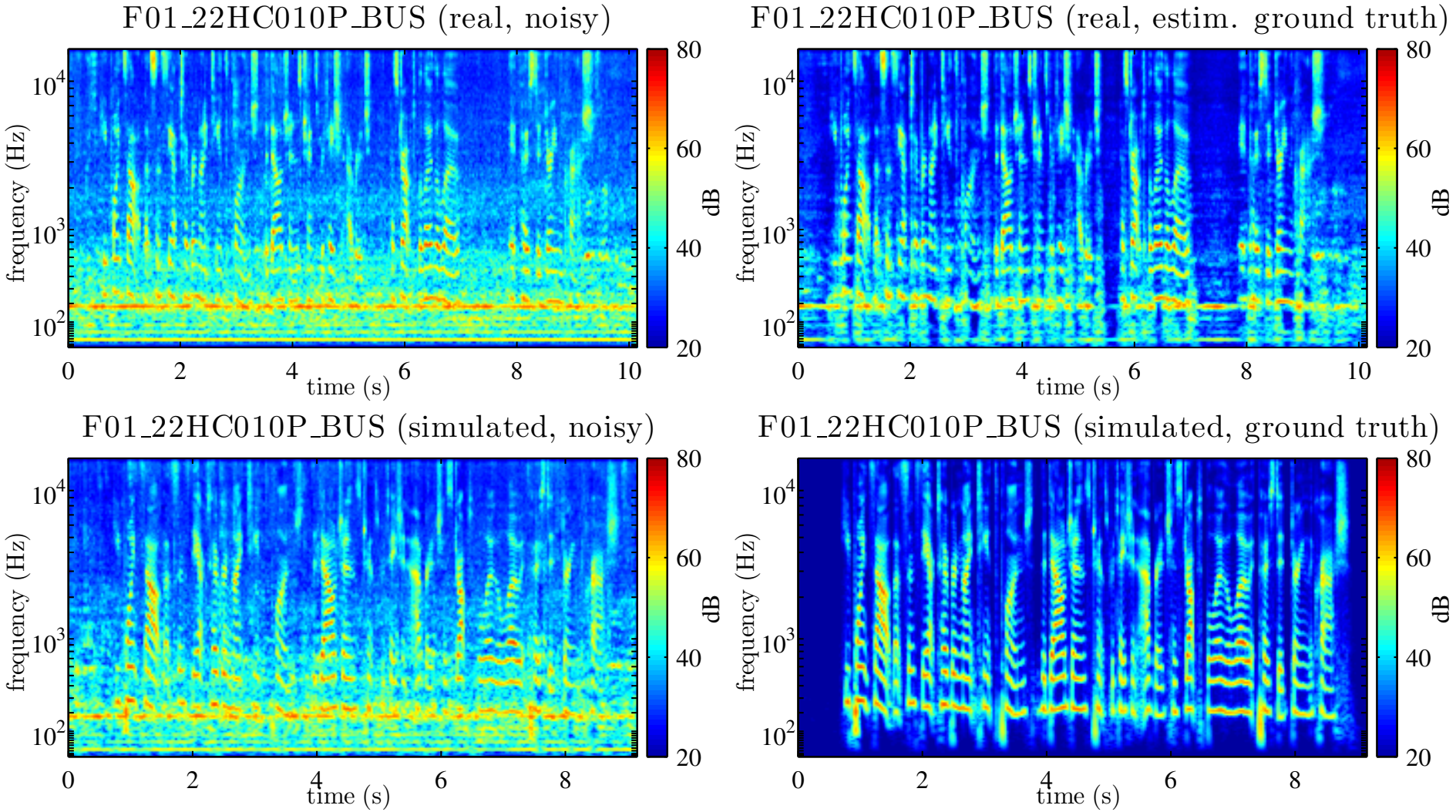
3 training environments:
Multicondition: 42.40%
Mismatched: 43.53%

⇒ similarly small impact on ASR as on enhancement

⇒ matched slighly better, probably due to GMM-HMM instead of DNN

Data simulation

Real and simulated signals similar in terms of speech and noise characteristics and SNR at each frequency. Ground truths more different.



Data simulation — Impact on enhancement

WER achieved by various beamformers/postfilters + GMM-HMM backend
reproduced from Prudnikov et al. (2015)

Enhancement	Dev		Difference
	real	simu	
none	18.70	18.71	+0.01
MVDR	18.20	10.78	−7.42
DS	12.43	14.52	+2.09
DS + Zelinski	14.29	15.25	+0.96
DS + Simmer	12.75	14.14	+1.39
MCA	10.72	12.50	+1.78

⇒ MVDR affected by data simulation mismatch (well known issue)

⇒ other learning-free enhancement techniques much more robust

Also holds for learning-based enhancement techniques, however real vs simulated training data can make a difference. See paper for details.

Data simulation — Impact on ASR

WER achieved by various DNN-HMM backends
reproduced from Yoshioka et al. (2015)

Acoustic model	Dev		Difference
	real	simu	
DNN (4 hidden)	13.64	13.51	−0.07
DNN (10 hidden)	12.27	11.97	−0.30
CNN (2 hidden)	11.94	11.70	−0.24
CNN (3 hidden)	11.52	11.25	−0.27
NIN	11.21	10.64	−0.57

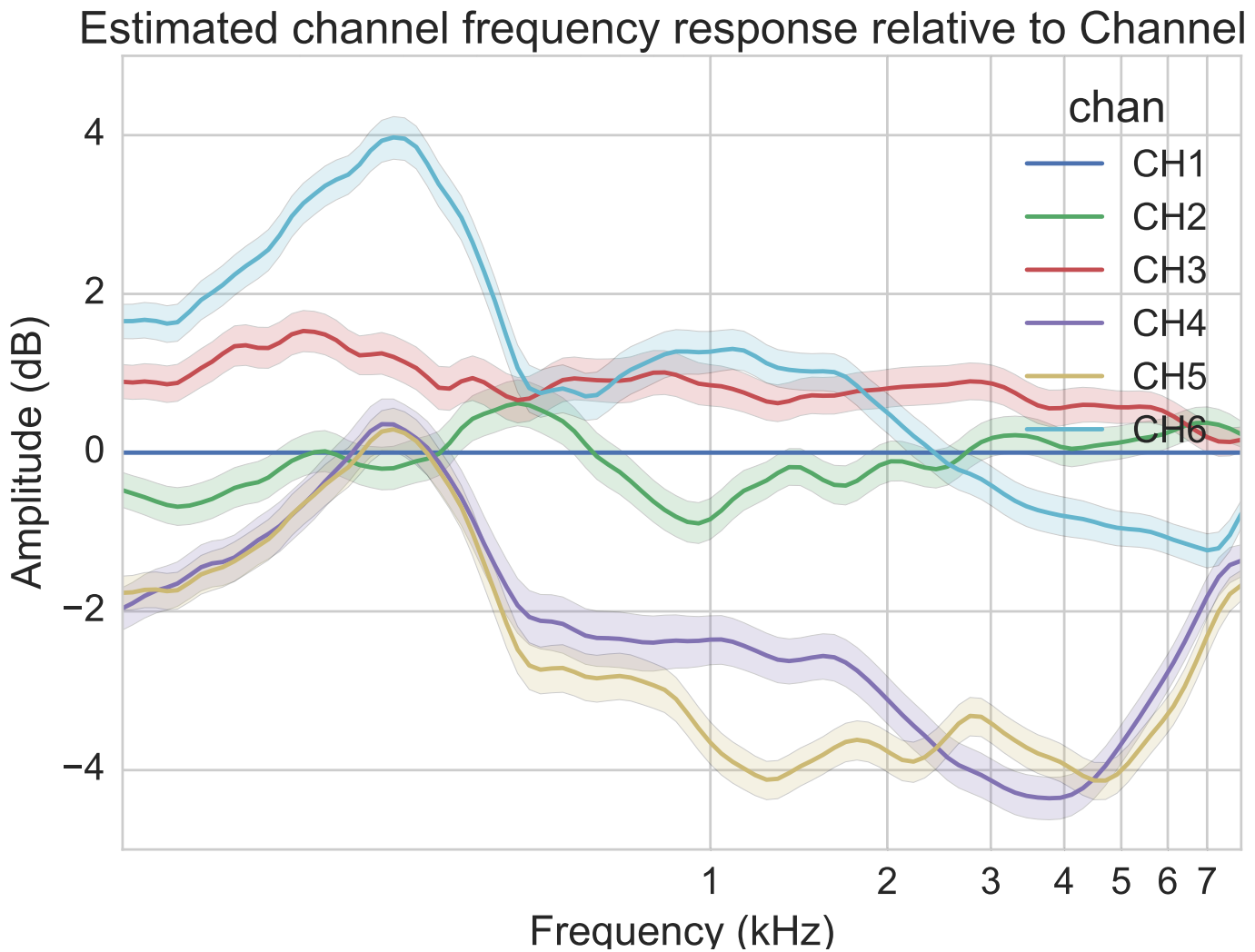
$\Rightarrow$ ASR backend little affected by data simulation mismatch

Training on simulated data alone: 4% relative WER increase only, essentially due to fewer data.

Training set on $\times 3$ simulated data: 10% relative WER improvement.

Mic response mismatch

Relative responses obtained by averaging over 1 min noise segments.



Mic response mismatch — Impact on ASR

WER achieved by no enhancement + GMM-HMM backend

Training (real + simu)	Test (real) ch1	ch2	ch3	ch4	ch5	ch6
ch1	36.78	81.12	41.10	35.45	31.50	34.95
ch2	37.69	79.92	41.57	36.62	32.73	36.57
ch3	37.51	81.29	41.73	35.87	31.58	35.62
ch4	39.09	83.69	43.31	36.83	32.64	37.16
ch5	39.64	83.82	43.59	37.30	32.73	37.68
ch6	36.63	81.72	40.54	35.09	30.75	34.51

Matched: 43.75%

Mismatched: 44.46%

$\Rightarrow$ ASR backend little affected by mic response mismatch

To sum up:

- all tested mismatches have little impact on enhancement or ASR,
- the number of mics has a much greater impact on performance

CHiME-4 tracks

CHiME-4:

- revisits the datasets originally recorded for CHiME-3
- increases the level of difficulty by constraining the number of mics available for testing.

Three tracks:

- 6ch
- 2ch
- 1ch

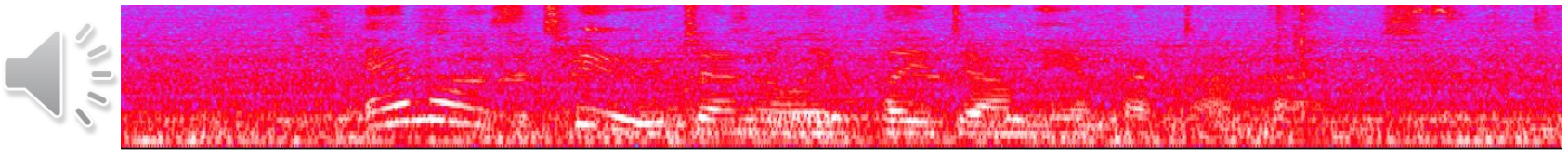
Test channels randomly selected, avoiding mic failures.

All 6 channels can still be used for training.

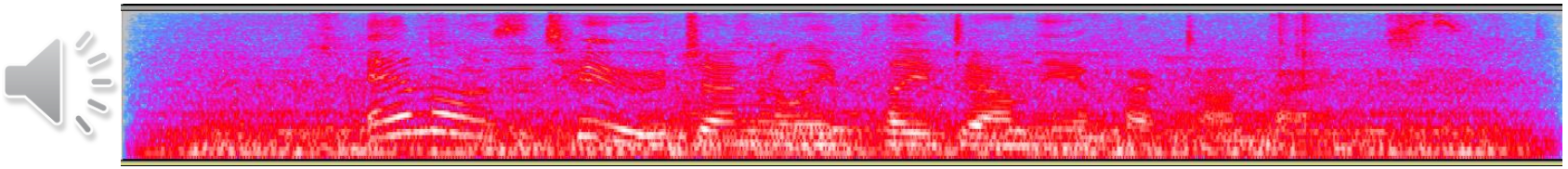
Baseline enhancement

- BeamformIt
  - https://github.com/xanguera/BeamformIt
  - Developed for NIST RT project by X. Anguera (ICSI)
  - Weighted delay and sum beamformer based on GCC

Original 5th channel signal



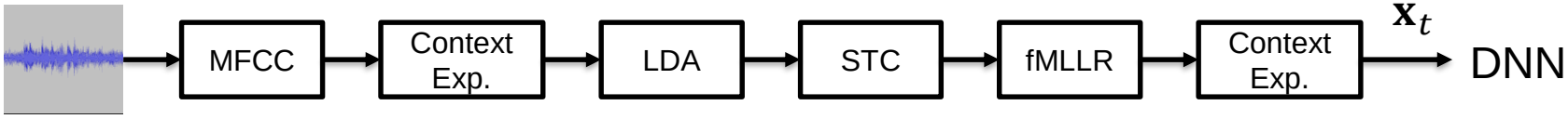
BeamformIt



Baseline features, AM, and LM

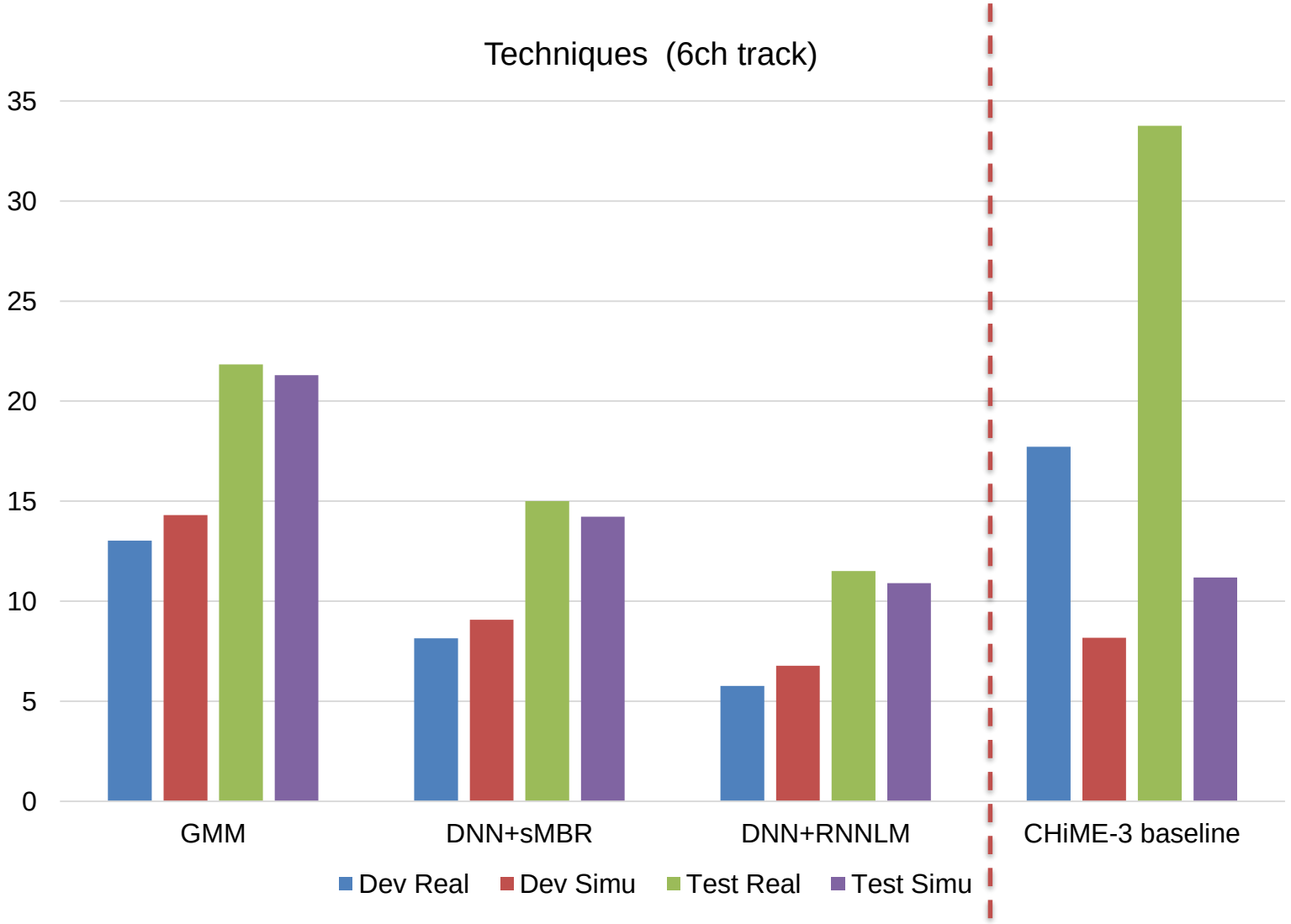
Red bold indicates update from CHiME-3 baseline

- Kaldi baseline (based on [Hori'15])
  - CHiME-4 official baseline package
  - Kaldi github https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4
- Features
  - So called **fMLLR** features

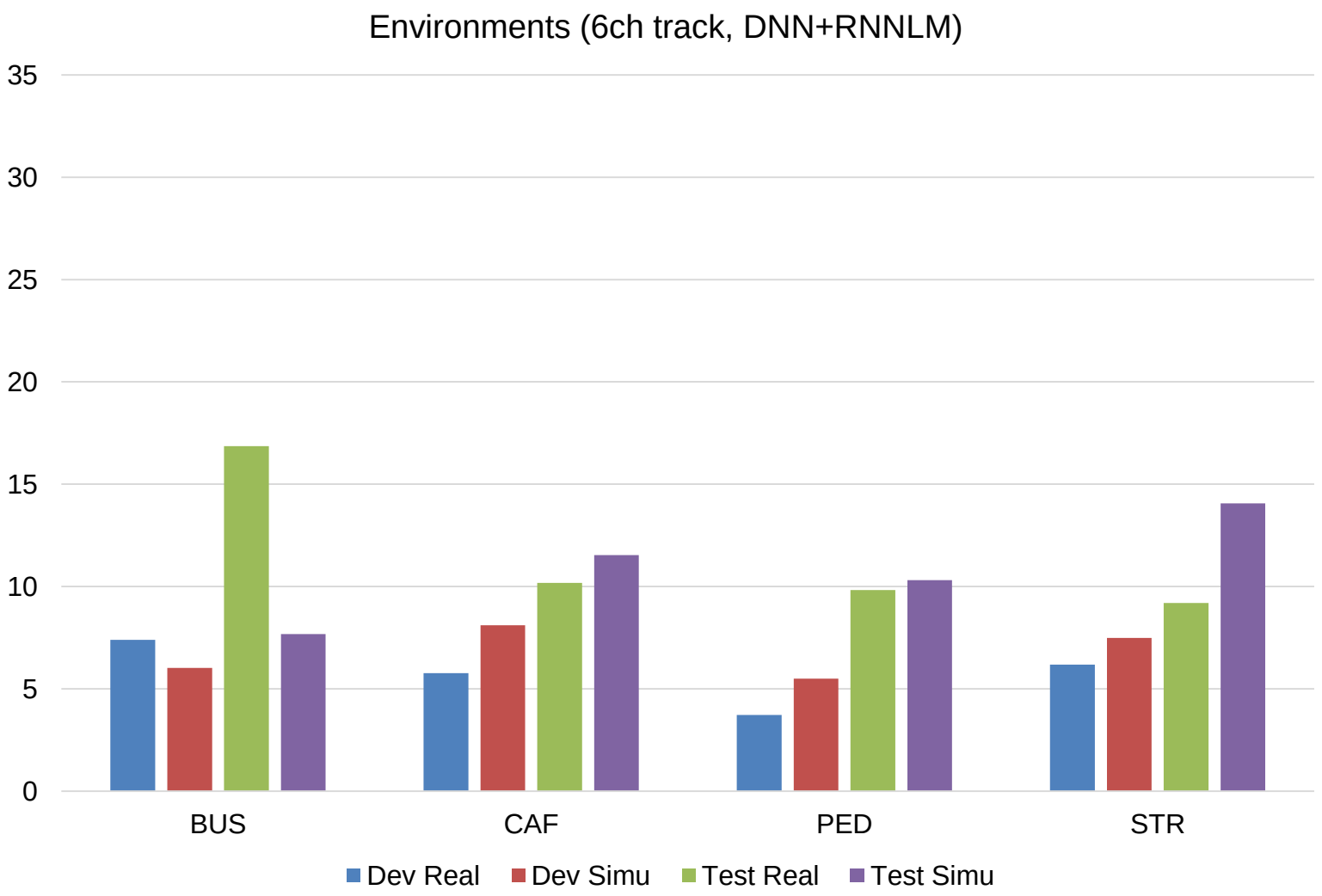


  - Needs GMM construction step
- Acoustic model (nnet1 in Kaldi)
  - **Noisy data training** using 5th channel (Not enhanced data training)
  - 7layer-2048neuron-DNN with sequence discriminative training (sMBR)
- Language model
  - 3-gram LM (provided by WSJ0)
  - **5-gram** LM (SRILM) and **RNNLM** rescoring (Mikolov's RNNLM)

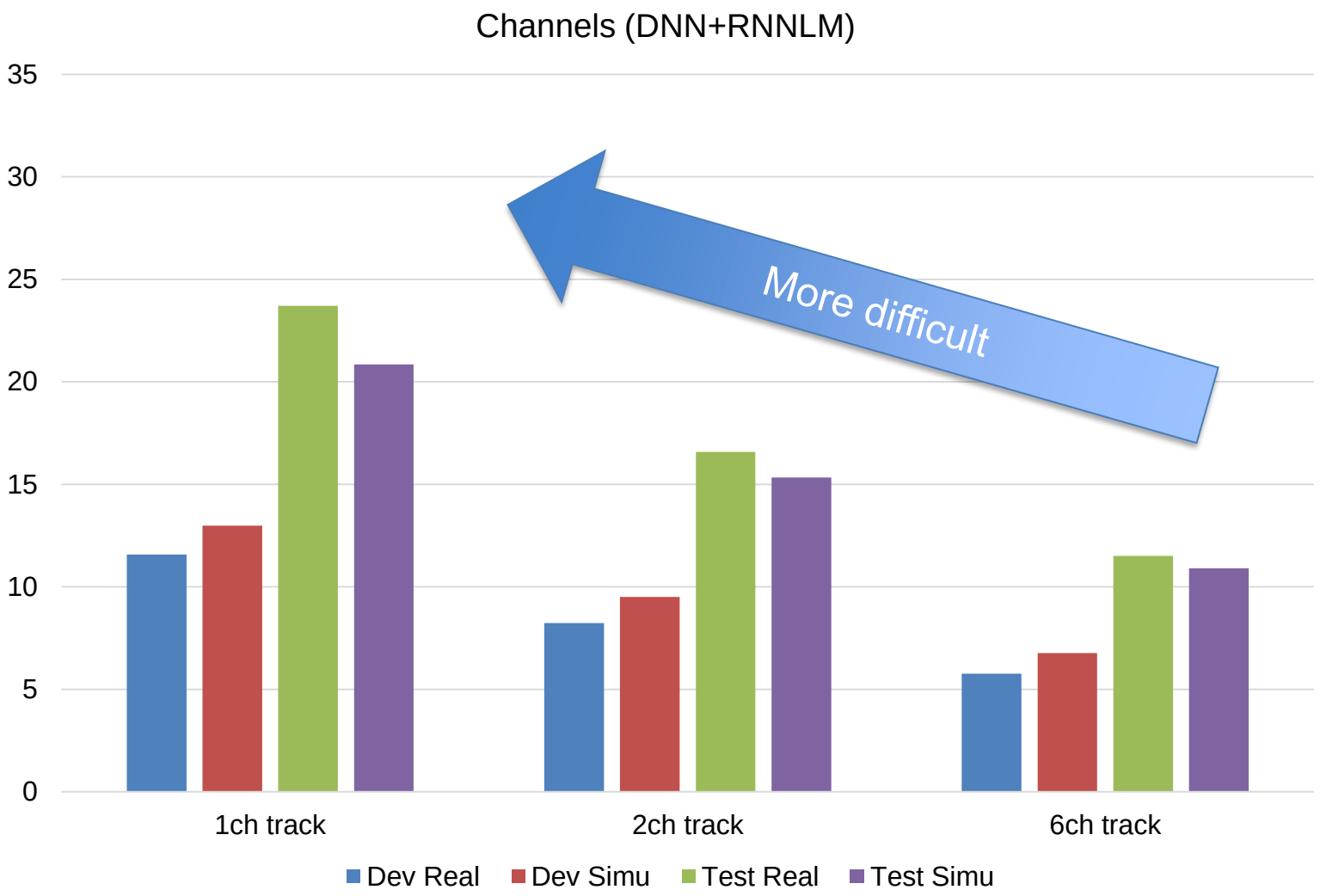
Baseline results



Baseline results



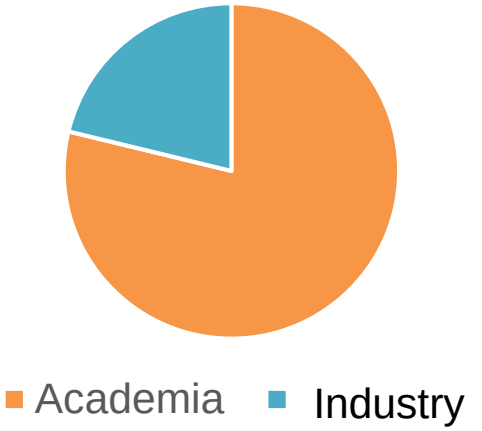
Baseline results



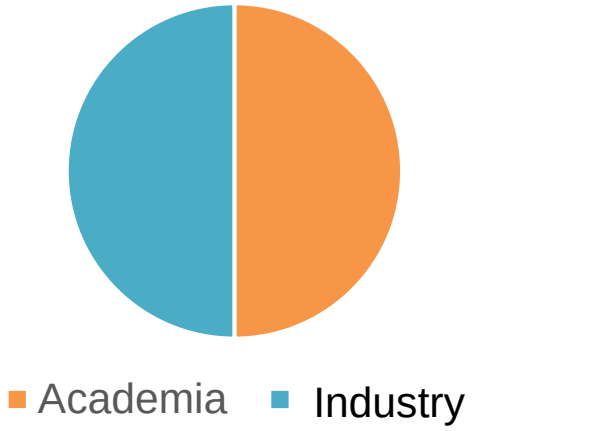
CHiME-4 submissions

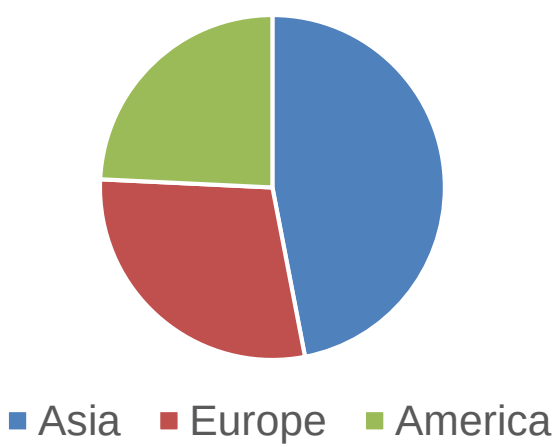
- Dataset distributed to 66 (CHiME-3) + 34 (CHiME-4) = **100 groups** (!)

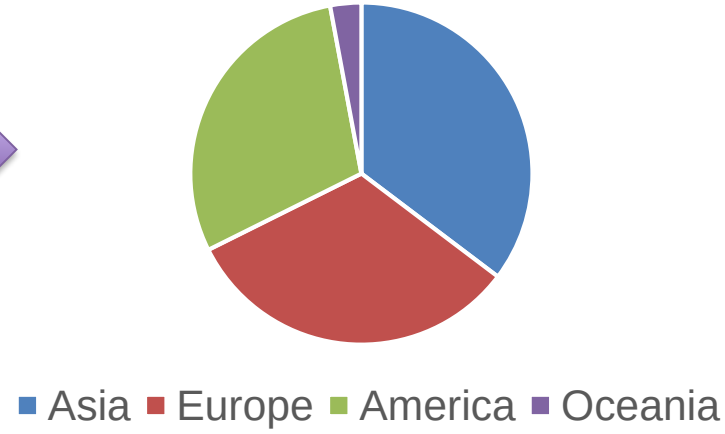
CHiME-3



CHiME-4



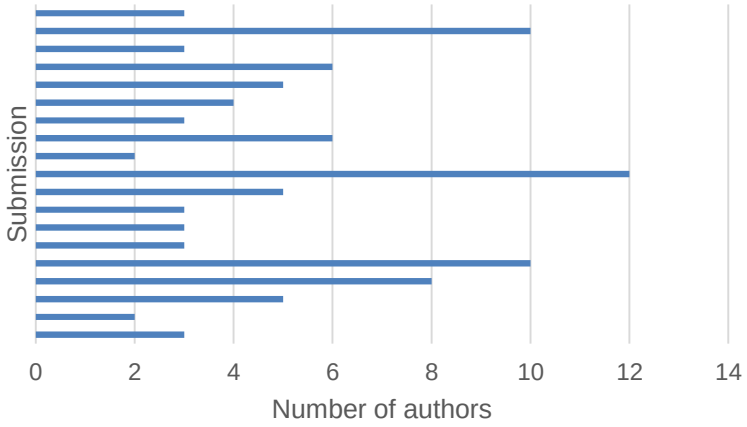




CHiME-4 submissions

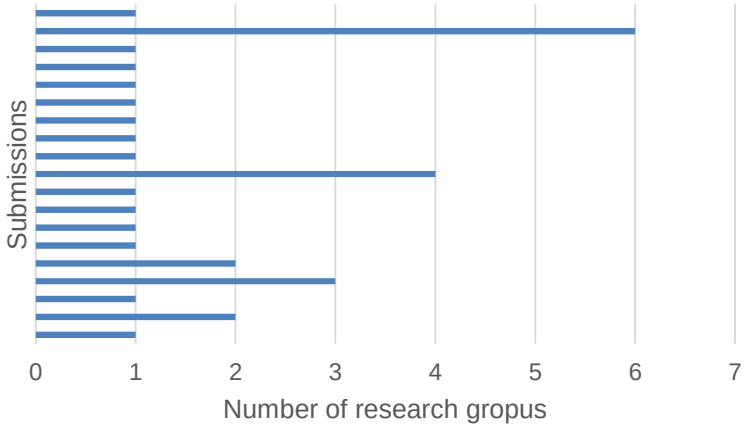
- Totally **43** submissions by **19** teams (cf 26 teams in CHiME-3)
  - 1ch track: 13, 2ch track: 14, 6ch track: 16
  - Most teams submitted multiple tracks
- Number of participants (author base): **96**

max 12, min 2, median 3, average 5



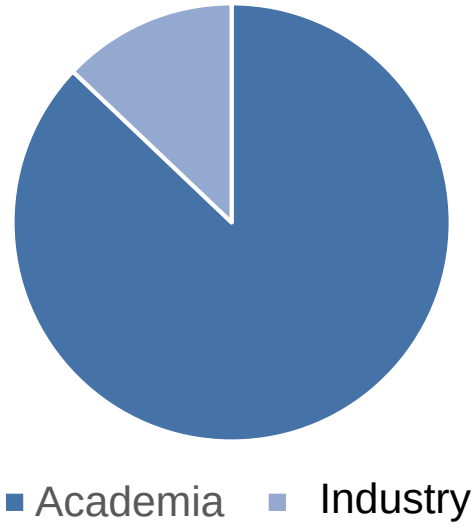
- Number of research groups: **31**

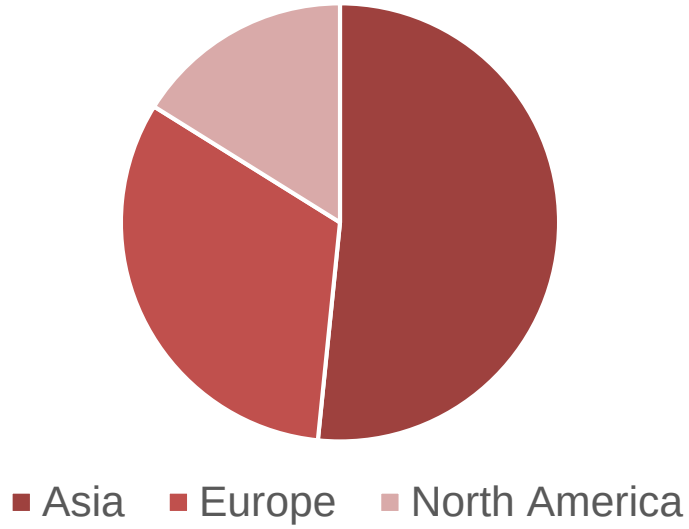
max 6, min 1, median 1, average 1.6

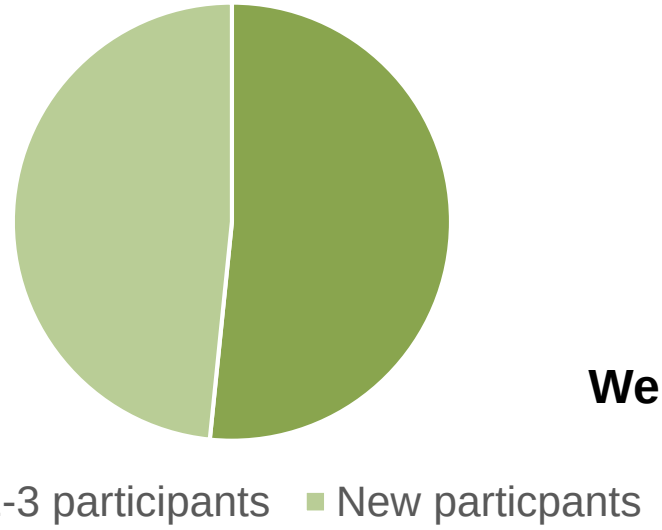


CHiME-4 submissions

- More info about challenge participants (research group base)

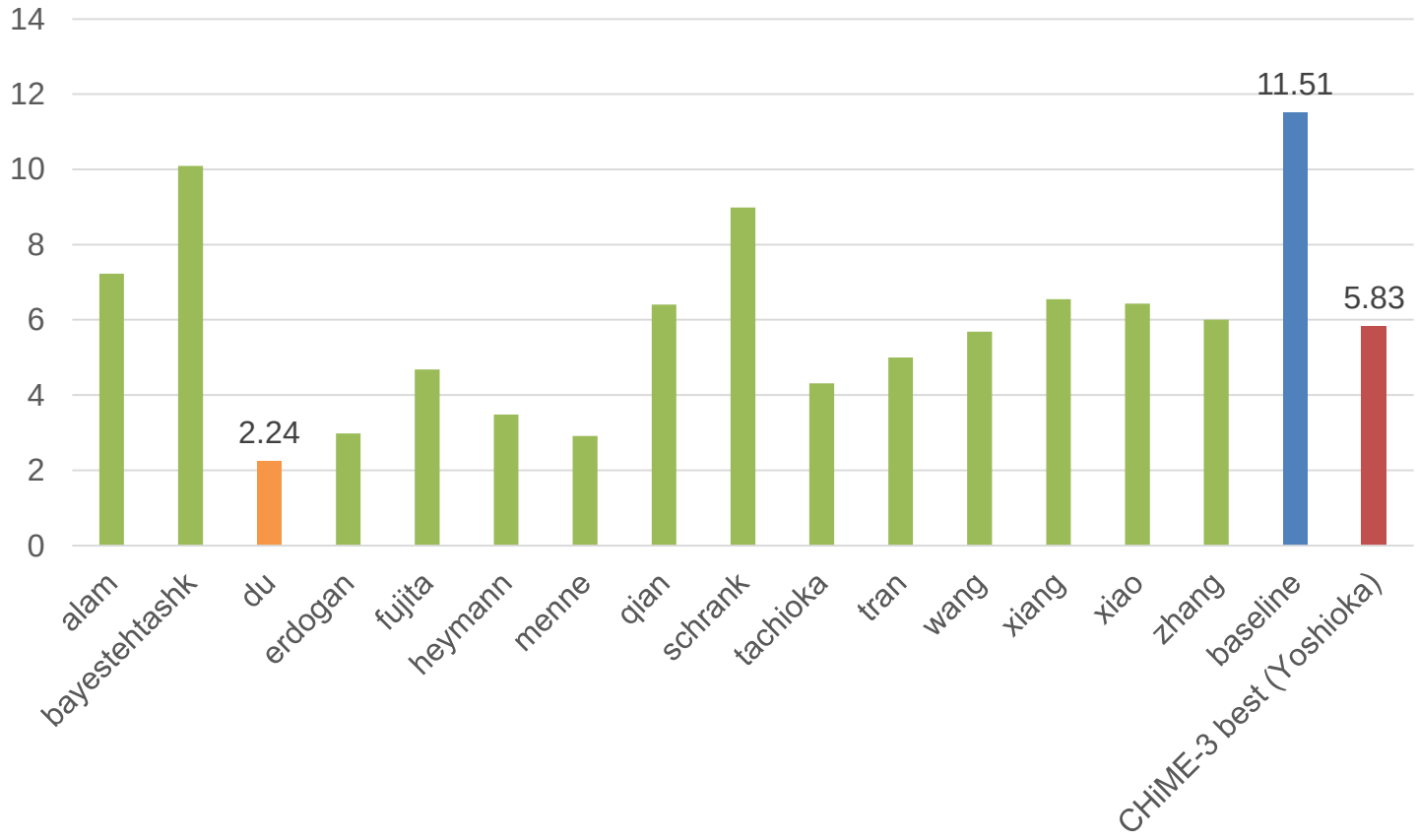




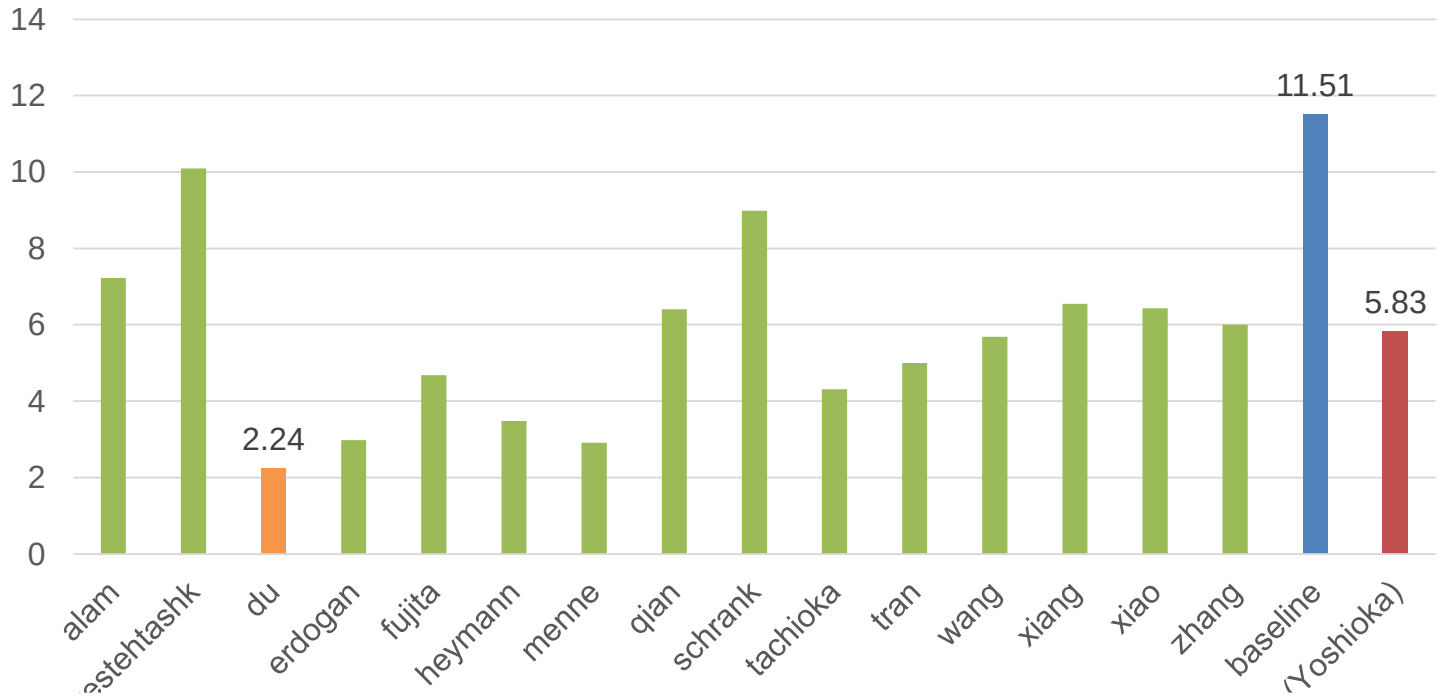


Welcome to CHiME!

6ch WER (Test Real)



6ch WER (Test Real)

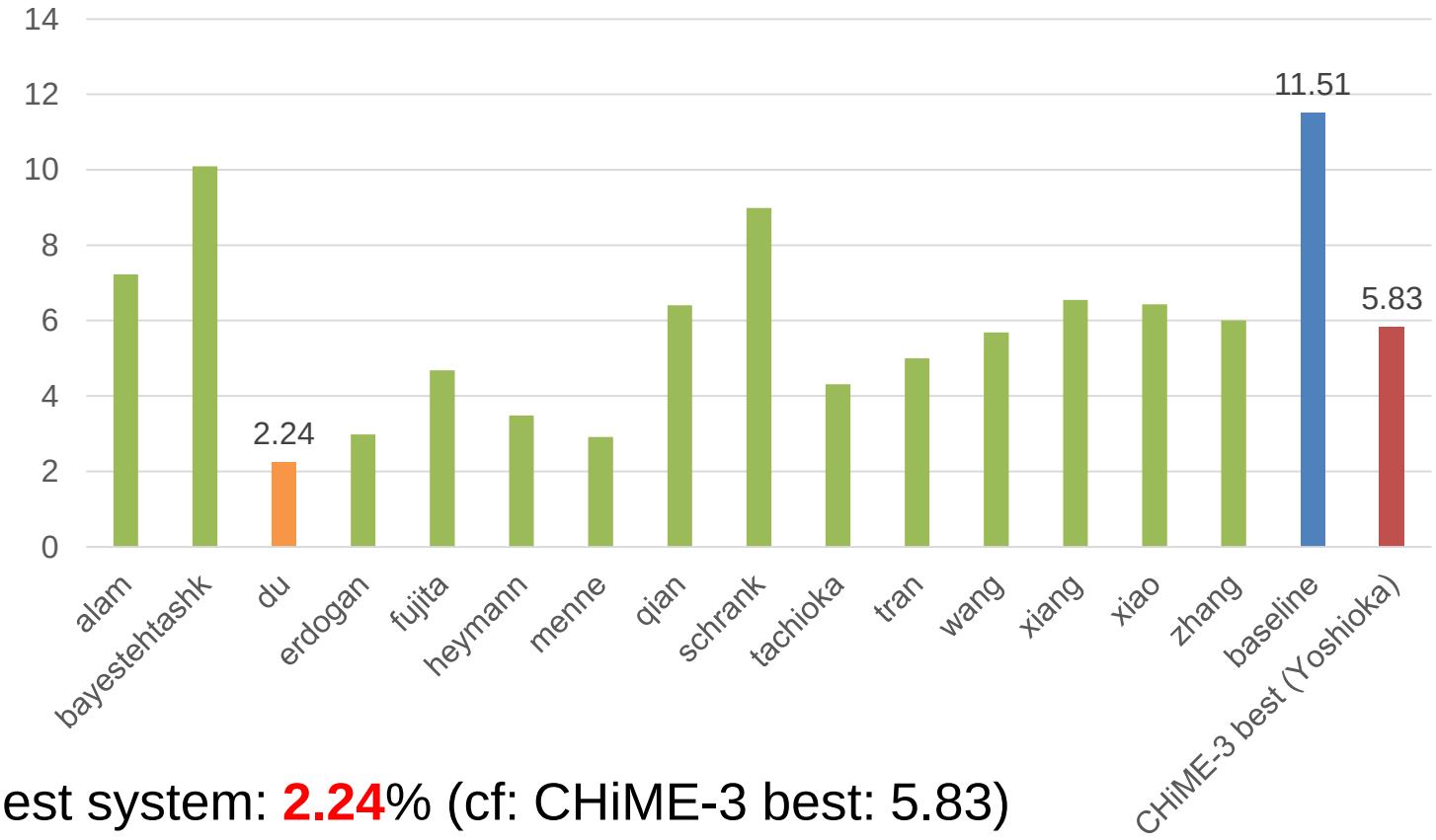


The USTC-iFlytek System for CHiME-4 Challenge

Jun Du[1], *Yan-Hui Tu*[1], *Lei Sun*[1], *Feng Ma*[2], *Hai-Kun Wang*[2], *Jia Pan*[2], *Cong Liu*[2], *Chin-Hui Lee*[3]

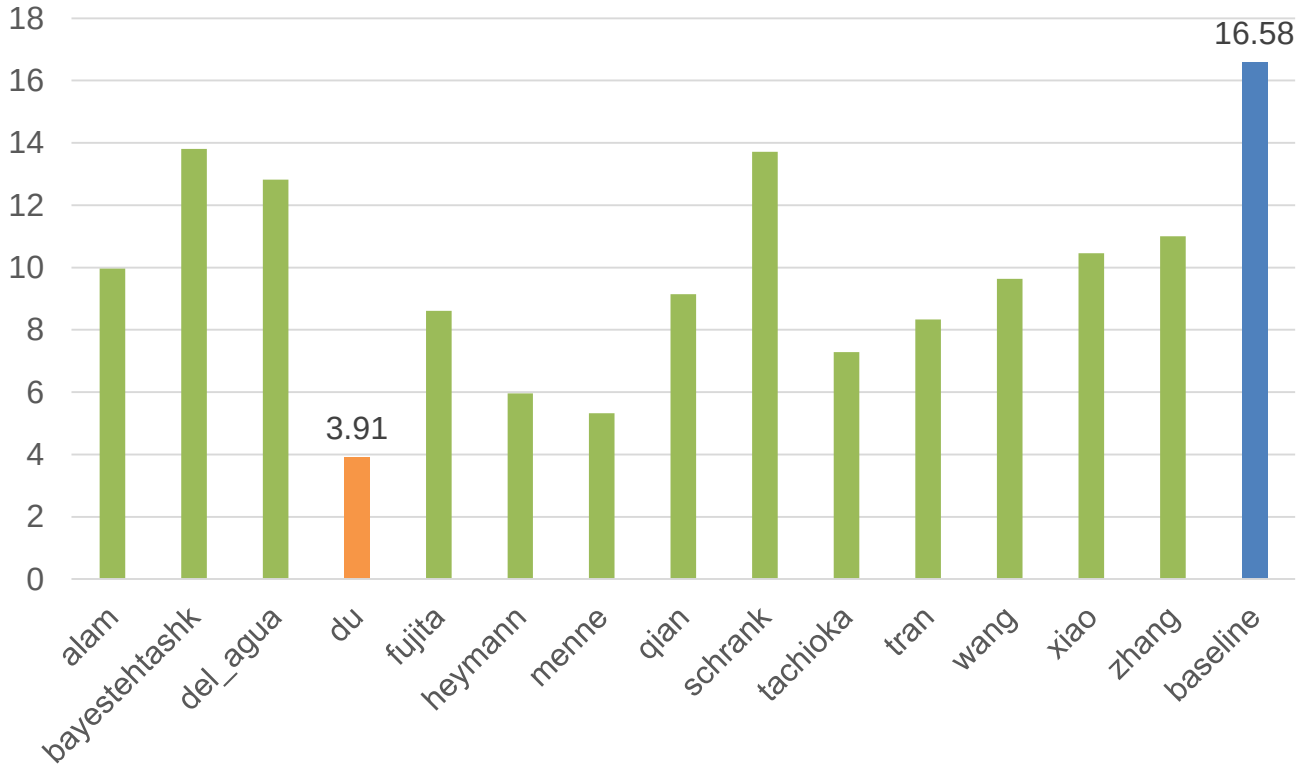
[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
jundu@ustc.edu.cn, {tuyanhui,sunlei17}@mail.ustc.edu.cn
[2]iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, P. R. China
{fengma,hkwang,jiapan,congliu2}@iflytek.com
[3]Georgia Institute of Technology, Atlanta, Georgia, USA
chl@ece.gatech.edu

6ch WER (Test Real)

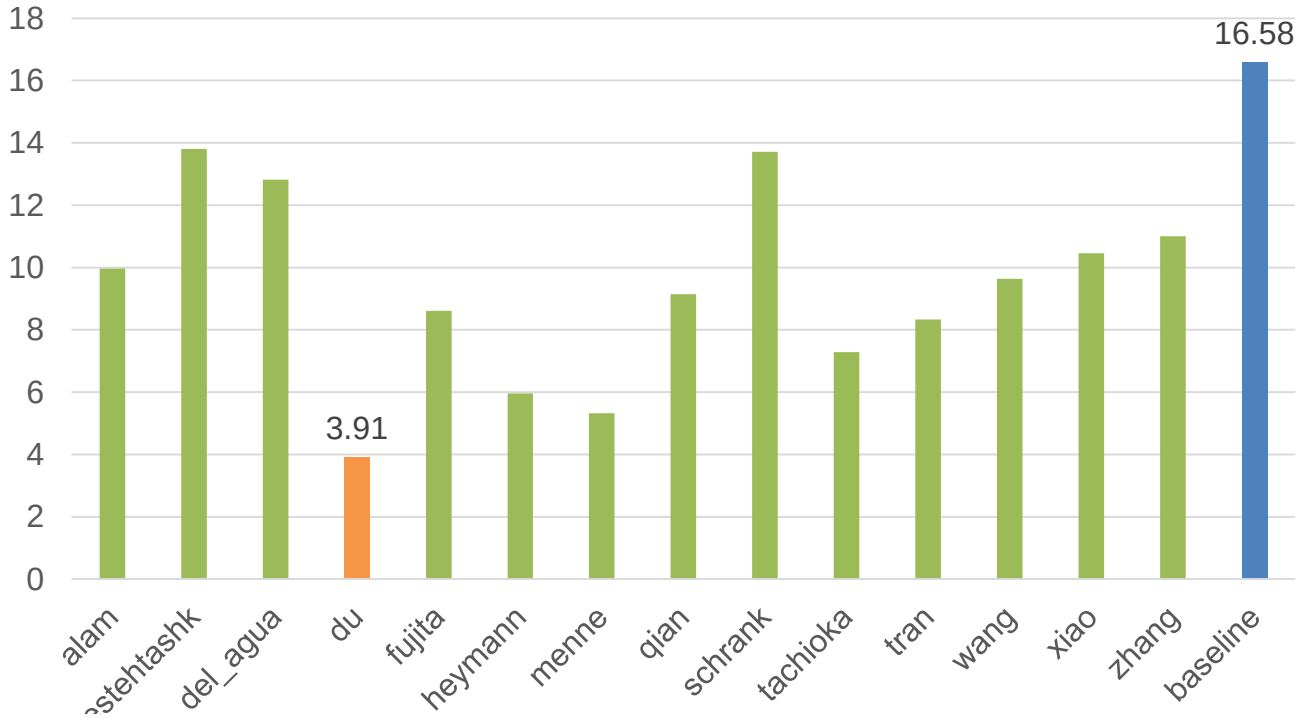


- Best system: **2.24**% (cf: CHiME-3 best: 5.83)
- 8 among 15 systems outperform CHiME-3 best system

2ch WER


18
16
14
12
10
8
6
4
2
0
16.58
3.91
alam
bayestehtashk
del_agua
du
fujita
heymann
menne
qian
schrank
tachioka
tran
wang
xiao
zhang
baseline

2ch WER

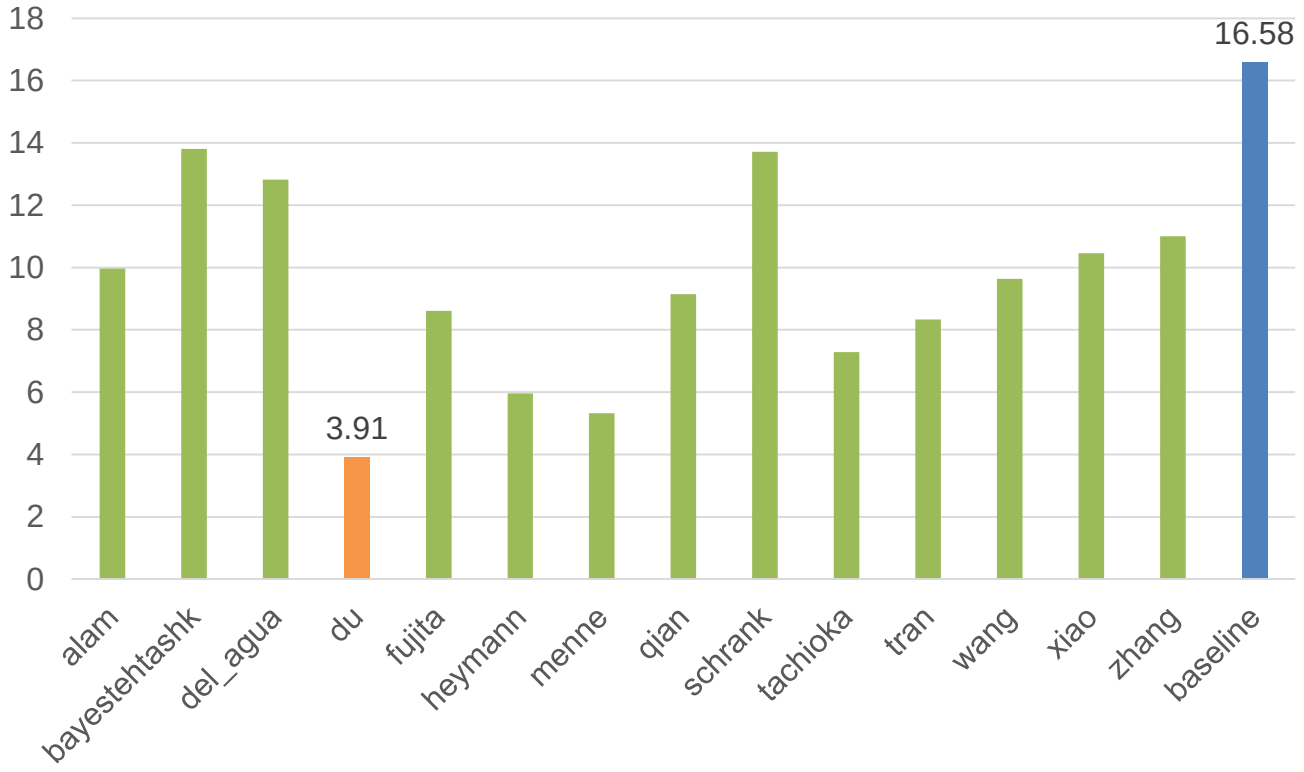


The USTC-iFlytek System for CHiME-4 Challenge

Jun Du[1], *Yan-Hui Tu*[1], *Lei Sun*[1], *Feng Ma*[2], *Hai-Kun Wang*[2], *Jia Pan*[2], *Cong Liu*[2], *Chin-Hui Lee*[3]

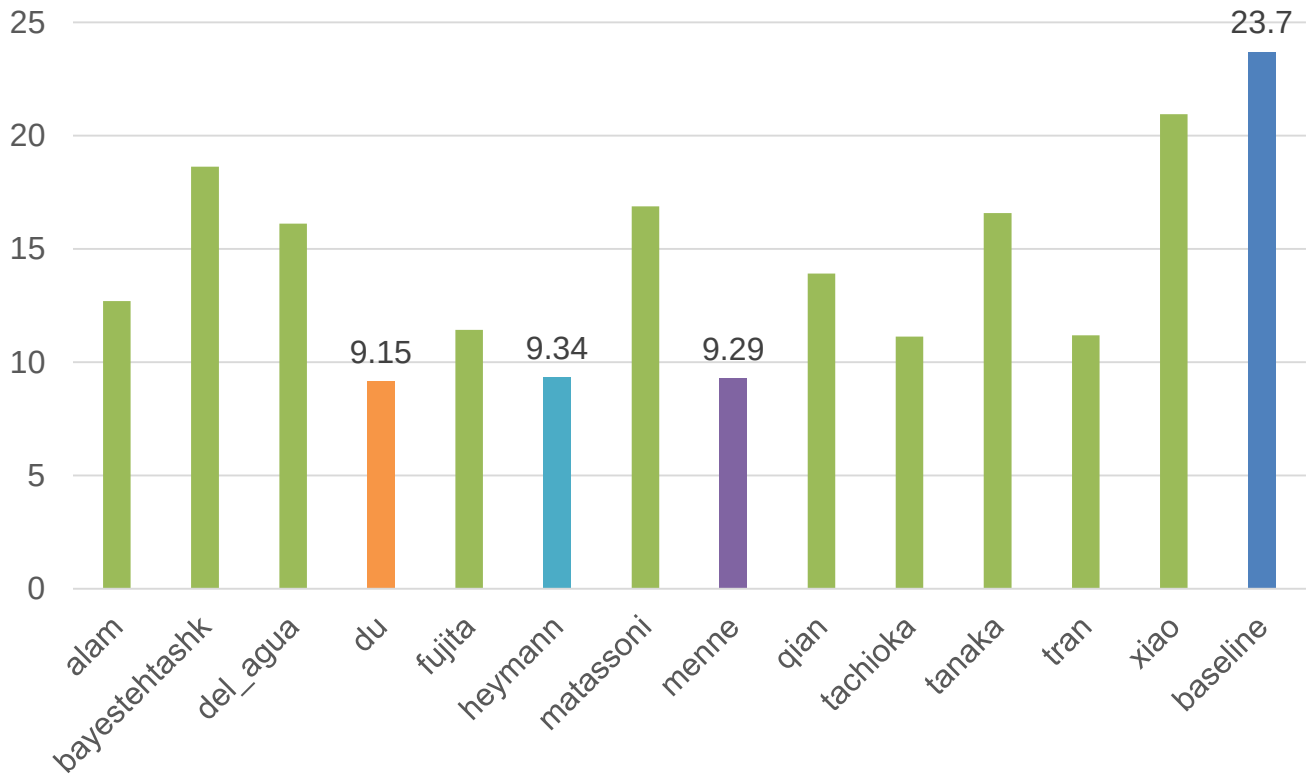
[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
jundu@ustc.edu.cn, {tuyanhui,sunlei17}@mail.ustc.edu.cn
[2]iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, P. R. China
{fengma,hkwang,jiapan,congliu2}@iflytek.com
[3]Georgia Institute of Technology, Atlanta, Georgia, USA
chl@ece.gatech.edu

2ch WER

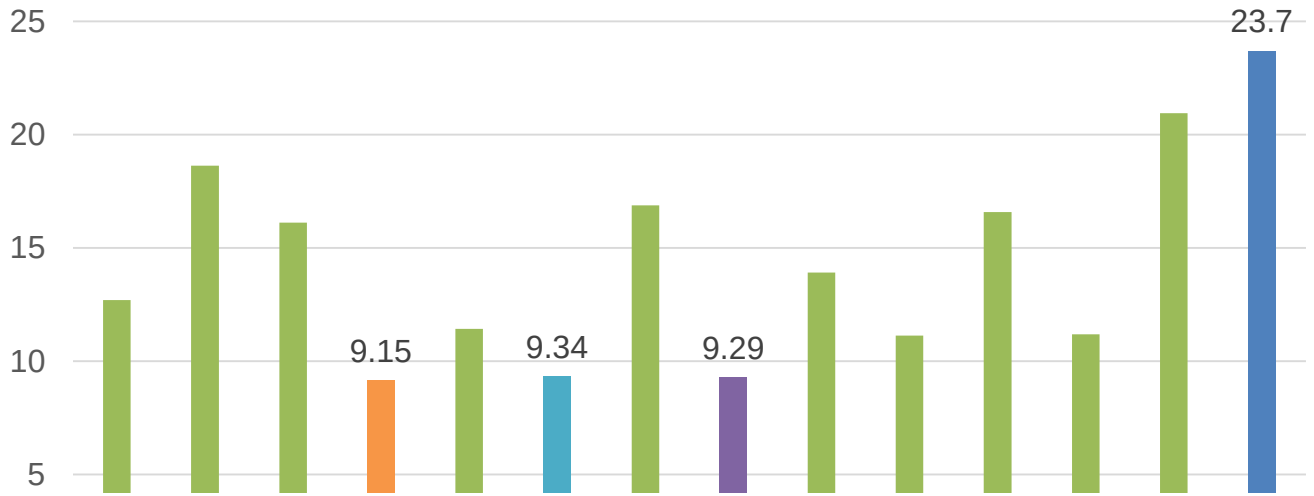


- Best system: **3.91**% (6ch track best: 2.24%)
- Two systems outperform CHiME-3 6 channel best performance

1ch WER


25
20
15
10
5
0
9.15
9.34
9.29
23.7
alam
bayestehtashk
del_agua
du
fujita
heymann
matassoni
menne
qian
tachioka
tanaka
tran
xiao
baseline

1ch WER



The USTC-iFlytek System for CHiME-4 Challenge

Jun Du[1], Yan-Hui Tu[1], Lei Sun[1], Feng Ma[2], Hai-Kun Wang[2], Jia Pan[2], Cong Liu[2], Chin-Hui Lee[3]

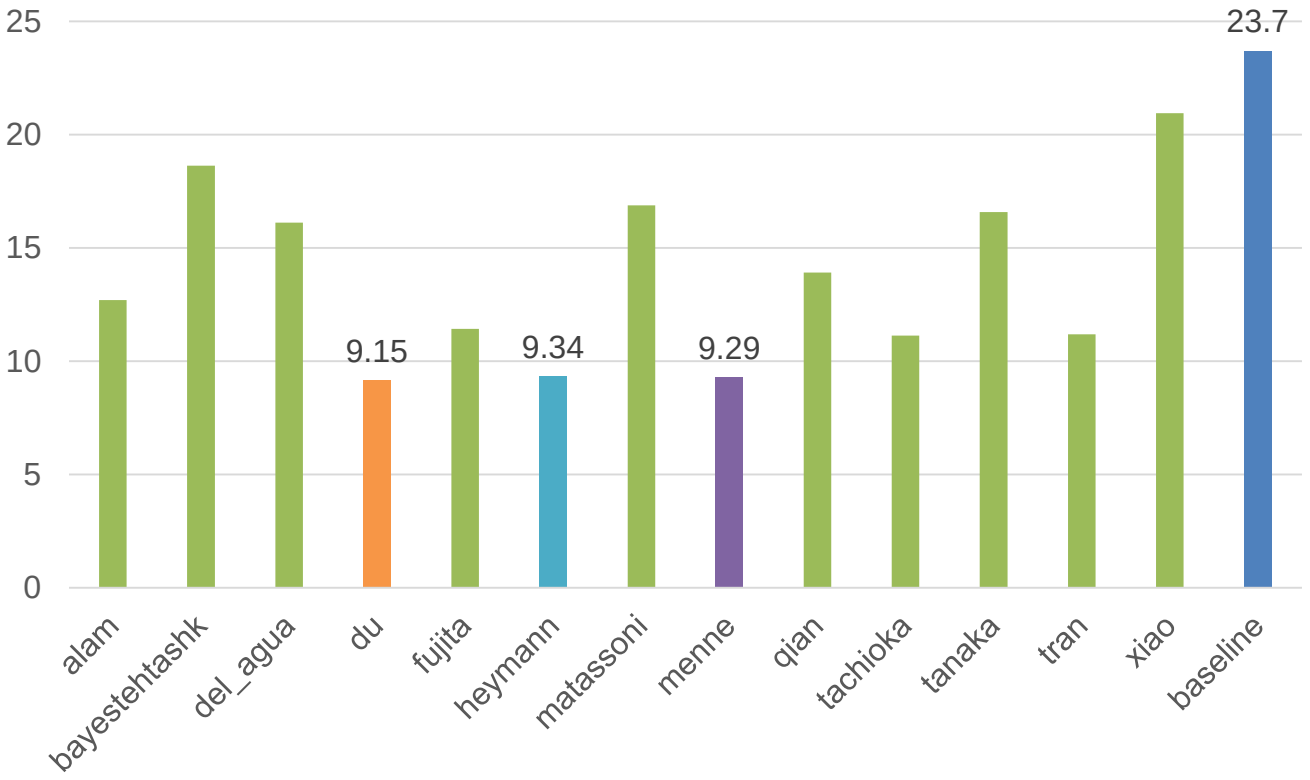
The RWTH/UPB/FORTH System Combination for the 4th CHiME Challenge Evaluation

Tobias Menne[1], Jahn Heymann[2], Anastasios Alexandridis[3,4], Kazuki Irie[1], Albert Zeyer[1], Markus Kitza[1], Pavel Golik[1], Lukas Drude[2], Ralf Schlüter[1], Hermann Ney[1], Reinhold Haeb-Umbach[2], Athanasios Mouchtaris[3,4]

Wide Residual BLSTM Network with Discriminative Speaker Adaptation for Robust Speech Recognition

Jahn Heymann, Lukas Drude, Reinhold Haeb-Umbach

1ch WER



- Best systems: **~9.2**% (6ch track best: 2.24%, 2ch track best: 3.91%)
- Still large gap between single and multi channel systems

Problem solved?

- Multi-channel, single device, and constrained vocabulary
  (Tablet or smart phone scenarios)

Problem solved?

- Multi-channel, single device, and constrained vocabulary
  (Tablet or smart phone scenarios)
  - Yes
- Single-channel scenario
  - No
- Multiple devices and/or spontaneous speech
  - ???

Successful front-ends

- Mask based beamforming (all top 5 systems use these techniques)

$$\mathbf{R} = \mathbb{E}[\mathbf{y}(t,f)\mathbf{y}^H(t,f)] \approx \frac{\sum_t M(t,f)\mathbf{y}(t,f)\mathbf{y}^H(t,f)}{\sum_t M(t,f)}$$

  - Spatial clustering (Complex GMM)
  - DNN/LSTM masking
- Variants of beamforming
  - MVDR
  - Max SNR
  - GSC
- System combination with different beamforming systems
- Single-channel speech enhancement: not so successful

Successful back-ends

- Data augmentation
  - All 6 channel data
  - Enhanced data
- Acoustic modeling
  - CNN
  - BLSTM
  - Model adaptation
- Language modeling
  - LSTM
- Joint training
  - Integrates beamforming and acoustic models with a single deep network

Future scientific directions (toward CHiME-5?)

- More real data (but not too much) >50h
- Speaking styles
  - Read speech → Spontaneous speech
- Acoustic environments
  - Multiple microphone devices
  - Multiple rooms/places with different speaker and microphone locations
- Speaker constraints
  - More speakers, and unbalanced amount of data per speaker
  - Speaker movement
  - Speaker overlaps
- Speaker diarization, speech activity detection
- Restrict computational resources at test time
  - Off-line → on-line, real time scenarios

Towards a sustainable challenge series

- What is the measure criterion of the challenge success?
  A. Produce novel and effective techniques
  B. Establish standard techniques for the problem
  C. More participants
  D. Attract community, and gain visibility

These are closely related each other

- Publicly available data (B, C, D)
- Complete set of state-of-the-art baselines (B, C)
  - People focus on developing a new technique rather than building a system ☺ (A)
- Design the task appropriately considering scientific findings and real scenarios (C, D)
- Place (satellite workshop of major conference) and timing (C, D)
- Clear evaluation metric (e.g., WER) (B)

Towards a sustainable challenge series

- What is the measure criterion of the challenge success?
  A. Produce novel and effective techniques
  B. Establish standard techniques for the problem
  C. More participants
  D. Attract community, and gain visibility

These are closely related each other

- Force participants to report the improvement of each modification to the baseline (A, B)
  - Multiple tracks increase the risk of loosing participants per track
- Increase the difficulty while improving baseline performance (CHiME3 -> CHiME4) (A, B)
- Fairness (B)
  - Provide a web-based scoring server etc. (blind test set to avoid over-tuning of test data)

Toward CHiME-5!!

Thanks!!

Questionnaire

https://goo.gl/forms/wROvDZgA0j0PhBJ12