Channel Selection from DNN Posterior Probability for Speech Recognition with Distributed Microphone Arrays in Everyday Environments

— CHiME-5 Challenge

Feifei Xiong, Jisi Zhang*, *Bernd T. Meyer*, Heidi Christensen, **Jon Barker**

Speech and Hearing Group (SPandH), University of Sheffield, Sheffield, UK
Medical Physics and Cluster of Excellence Hearing4All, University of Oldenburg, Germany

(supported by a PhD Scholarship support from Toshiba Ltd)*

Outline

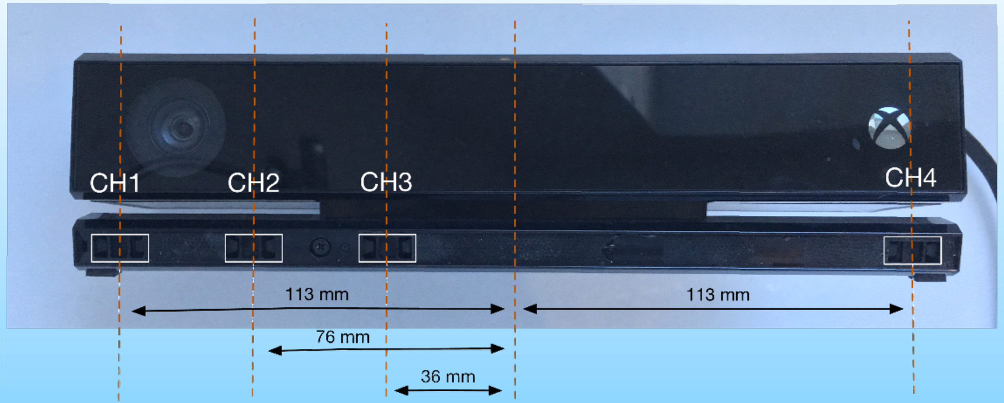
- Introduction
  - Motivation
- Proposed System
  - Channel selection
- Results
  - Separate evaluation
- Conclusions

Introduction

- CHiME-5: conversational speech recognition in everyday home environments with distributed microphones/arrays
- Challenges:
  - Natural conversational speech in a dinner party scenario
  - To recognize speech from each speaker
- Tasks:
  - Single-array track: one given reference Kinect (coarsely) based on Video
  - Multiple-array track: all 6 Kinects can be exploited
  - Ranking A: frame-level tied phonetic targets and official language modeling
  - Ranking B: all other systems

Microphone Configuration

- Kinect's microphone configuration
- Binaural 'worn' microphones
  - Potential 'best' channel
- BeamformIt using CH1 ~ CH4
  - Slight improvement over raw channel
  - Alternatively: CH2



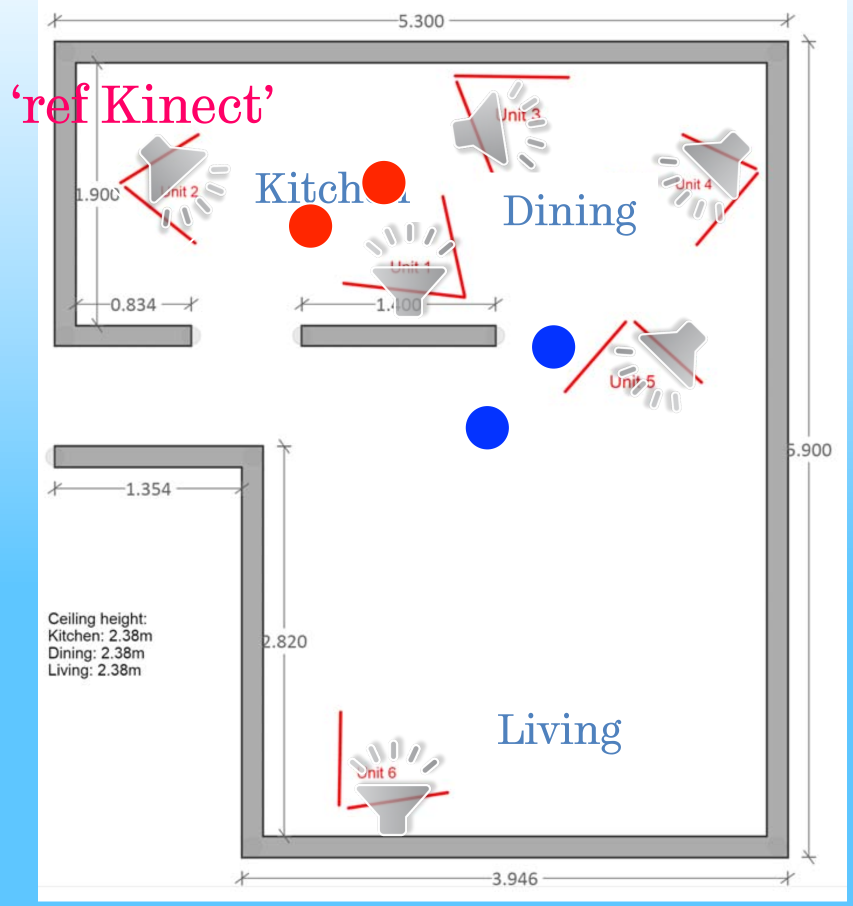
Baseline: LF-MMI TDNN	WER of Dev
Worn mic.	**47.22**
BeamformIt, ref. Kinect	**80.62**
CH1	**80.89**
CH2	**80.63**
CH3	**80.94**
CH4	**80.97**

Motivation

Note, the audio examples in the figure are only available in the original powerpoint slides. The samples demonstrate that when speech overlaps, the perceptually dominant speaker in the mixture will vary depending on the recording device position.

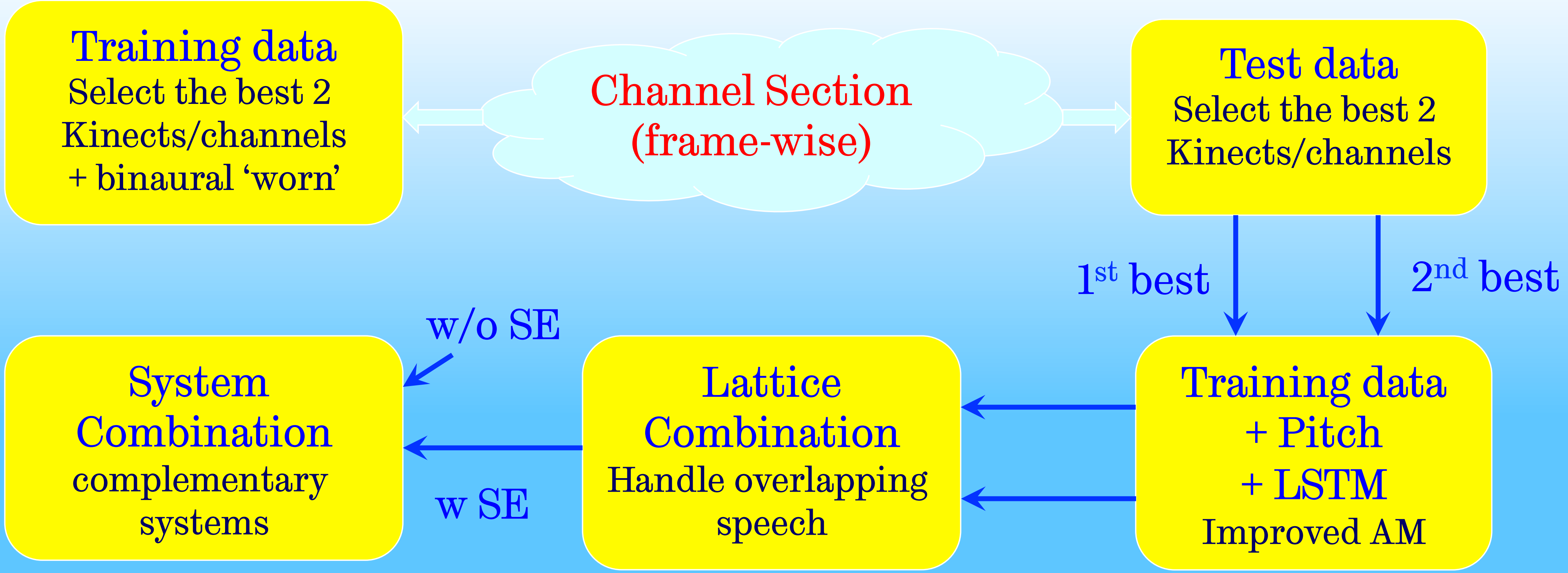
- The nearer the Kinect locates to the target speaker, the more reliable the recognition is ← high SNR and low reverberation

- A large portion of overlapping speech, and the speakers in one Location typically spatially dispersed → using reference Kinect per location seems to be not enough



'dev' 'S02' 'Kitchen' 'CH1'

System

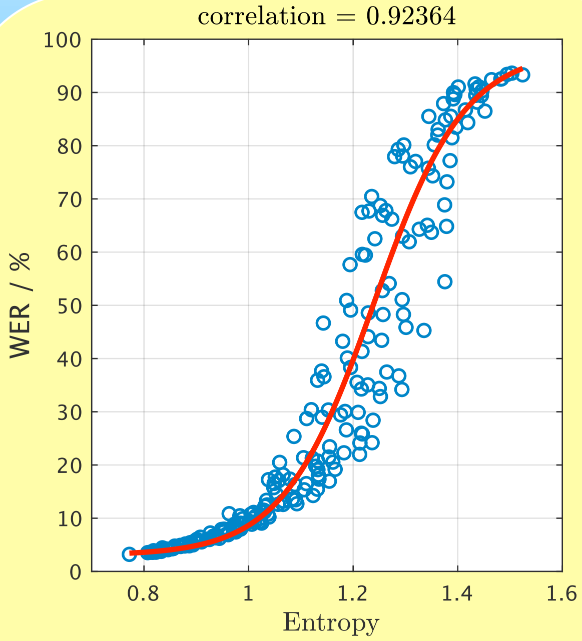


w SE: speech enhancement via baseline BeamformIt
w/o SE: signal from CH2 is used

Channel Selection

- Motivated by the findings in ASR system monitoring
  - Entropy of the DNN posteriors → similarity to the training data → to reflect the final ASR performance without consuming decoding

$$-\sum_s (P(s,t) \cdot \log_2 P(s,t))$$



Aurora4 noise+reverb 218 sets

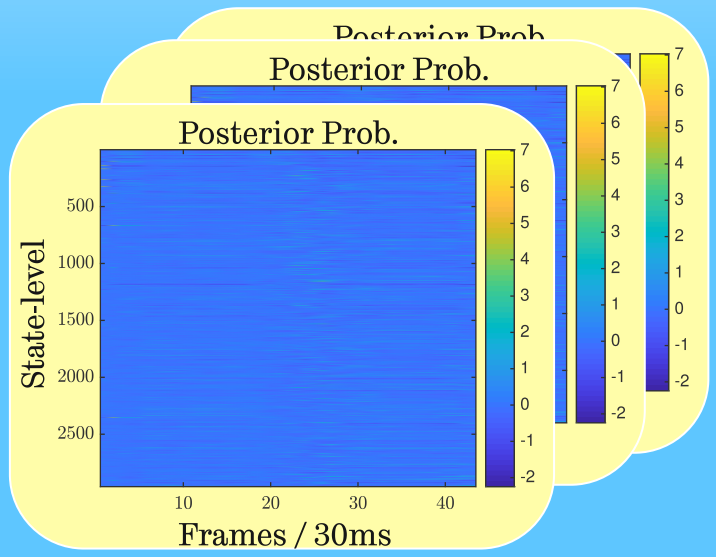
Barker, Williams and Renals. "*Acoustic Confidence Measures for Segmenting Broadcast News*," ICSLP, 1998

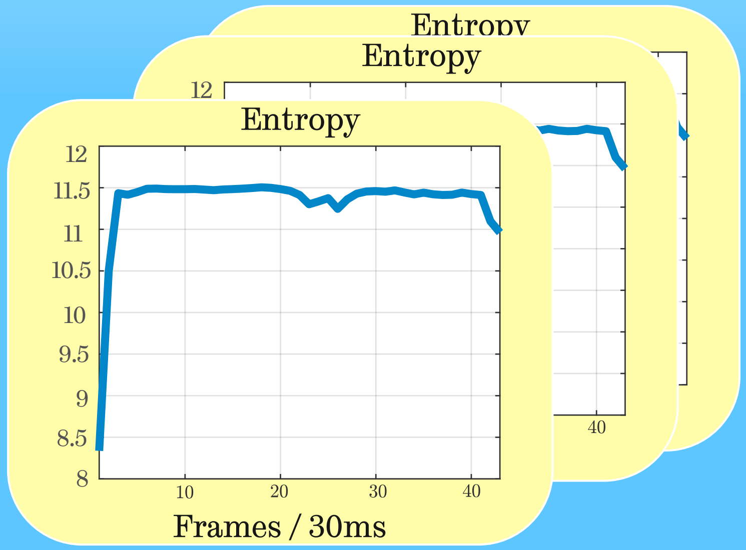
Misra et al. "*New entropy based combination rules in HMM/ANN multi-stream ASR*," ICASSP, 2003

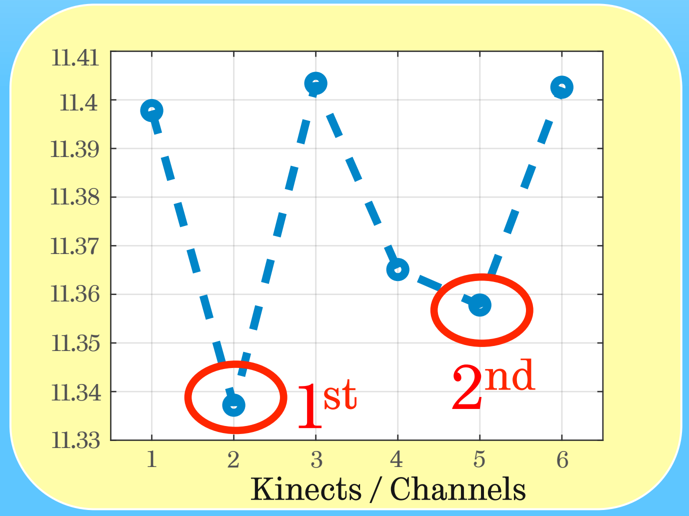
Wang, Li, Hermansky, Interseech 2018

Channel Selection

- Train a DNN model only with binaural 'worn' speech signals

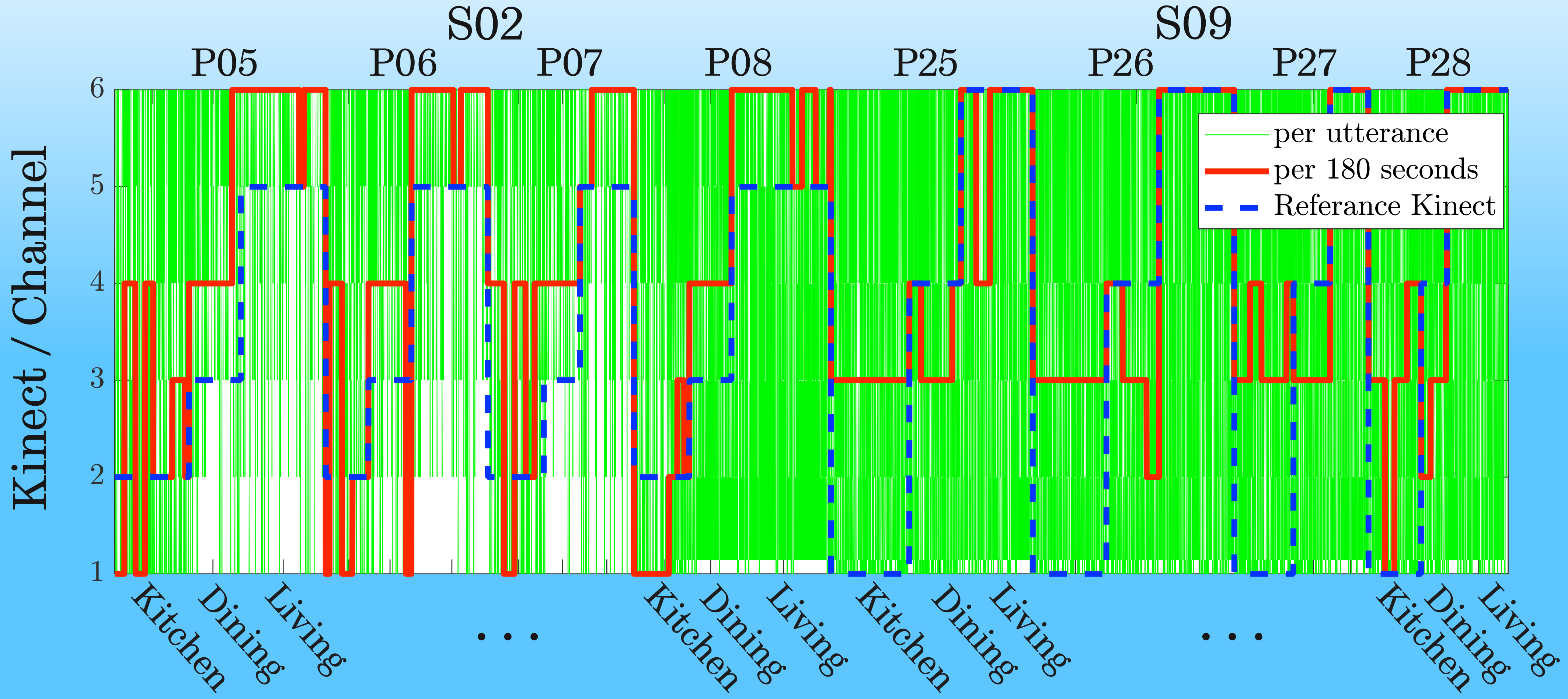






Channel Selection

- Dev set (7437 utterances)



Channel Selection

- Dev set (7437 utterances)

Utterance-based: more promising

Risk: potential different best Kinect for different speaker

Baseline	Avg.	S02 Kitchen	S02 Dinning	S02 Living	S09 Kitchen	S09 Dinning	S09 Living
Ref. Kinect	**80.62**	86.50	78.89	78.64	81.39	79.60	76.65
Per utterance	**78.85**	83.16	79.21	75.21	79.55	78.65	77.96
Per 180 seconds	**79.51**	84.80	79.26	76.42	80.27	79.29	76.60
STOI on Dev	**76.18**	76.50	78.31	72.82	78.58	77.19	76.60

Risk: potential non-accurate selection (short average window)
Session: only one potential best Kinect for all 4 speakers

Channel Selection

- Dev set (7437 utterances)
- Per utterance

Risk: potential different best Kinect for different speaker

Speaker	ref	S02 Dining 1st best	2nd best	Lattice Combination
P05 (f)	88.34	88.52	**88.15**	86.00
P06 (m)	71.00	**70.16**	71.07	66.09
P07 (m)	75.20	**75.20**	75.80	73.57
P08 (f)	90.03	93.72	**90.40**	90.04

Solution: to combine complementary channels

Channel Selection (Training)

- Training data selection
  - Baseline: randomly choosing 100K utterances from Kinects' signal + binaural 'worn' (L + R)
  - To rank the 6 Kinects/Channels (after BeamformIt)
  - To select 74728 utterances (1st best) + 74728 utterances (2nd best) + binaural 'worn' (L + R)

LF-MMI TDNN	Dev (ref)
Baseline (worn L R + 100K)	**80.62 (+SE)**
Channel selection on training (+SE + 1st best)	**79.92 (+SE)**
Channel selection on training (CH2 + 1st best)	**80.35 (CH2)**
Channel selection on training (+SE + 1st best + 2nd best)	**79.42 (+SE)**
Channel selection on training (CH2 + 1st best + 2nd best)	**79.40 (CH2)**
Channel selection on training (+SE + all 6 channels)	**80.83 (+SE)**

Channel Selection (Test)

- Test data selection *(Multiple-Array Track)*
  - Ref. Kinect *(Single-Array Track)*
  - To rank the 6 Kinects/Channels (after BeamformIt) ← per utterance

LF-MMI TDNN	Dev	**Lattice Combination**	**STOI-based selection *(1st best as oracle)***
Baseline (worn L R + 100K)	**80.62 (ref)**		
+ Channel selection on test (1st best)	**78.85**	**77.44 (weigh 0.5:0.5)**	**76.18**
+ Channel selection on test (2nd best)	**81.17**		
Channel selection on training	**79.42 (ref)**		**on training**
+ Channel selection on test (1st best)	**77.40**	**76.17 (weigh 0.5:0.5)**	**74.82**
+ Channel selection on test (2nd best)	**80.12**		

Channel Selection (Test)

- Test data selection *(Multiple-Array Track)*
  - Ref. Kinect *(Single-Array Track)*
  - To rank the 6 Kinects/Channels (after BeamformIt) ← per utterance

LF-MMI TDNN	Dev	Lattice Combination	STOI-based selection *(1st best as oracle)*
[illegible]	[illegible]	[illegible] (weigh 0.5:0.5)	76.18
[illegible]	[illegible]	[illegible]	on training
+ Channel selection on test (1st best)	[illegible]	76.17 (weigh 0.5:0.5)	74.82
+ Channel selection on test (2nd best)	**80.12**		

Barker et al. "*The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes*," Computer Speech and Language, 2017

Taal et al. "*An algorithm for intelligibility prediction of time-frequency weighted noisy speech*," IEEE TASLP, 2011

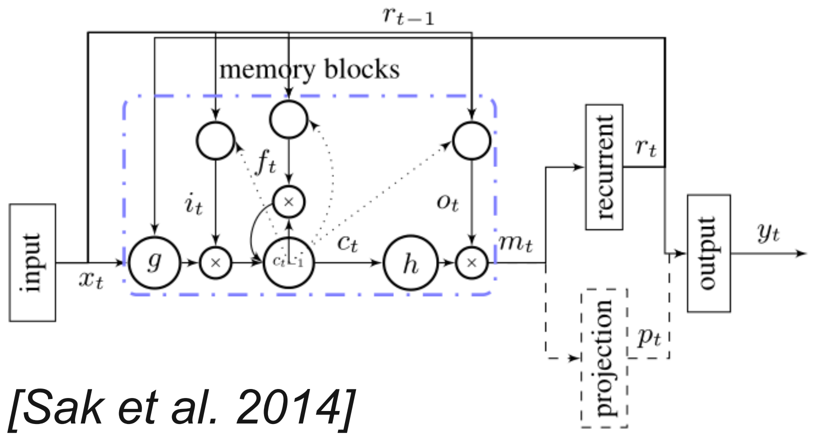
Pitch Features

- Implemented in Kaldi
  - 3-dimentional features: probabilities of voices, log-pitch (1.5s window), and delta-pitch
- To improve the speaker characteristics
  - I-vectors: efficient to capture speaker information
  - Pitch features: slight further improvements

LF-MMI TDNN	Dev (ref)
Baseline	**80.62**
- w/o I-vectors(#100)	**84.09**
+ I-vectors(#100) + Pitch(#3)	**80.36**

LSTM Projected RNN

- Integrate 3 LSTM projected layers into TDNN in Kaldi
  - 512 neurons, 128 projection, 128 recurrent
  - To further improve the capability of capturing temporal dynamics of features



[Sak et al. 2014]

	Dev (ref)
Baseline	**80.62**
+ Channel selection on training	**79.42**
+ LSTM	**77.09**
+ Pitch(#3) + LSTM	**76.24**

Sak et al. "*Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*," arXiv, 2014

... Together

Ranking A	Dev (SE)	Dev (CH2)
Baseline (worn L R + 100K)	**80.62**	**80.63**
+ Channel selection on training	79.42	79.40
+ Pitch	79.17	79.18
+ LSTM	76.24	76.30
System combination (weigh 0.5:0.5)	**73.53**	

→ Single-Array Track
7.1% WER reduction

+ Channel selection on test (1st best)	74.49	75.01
+ Channel selection on test (2nd best)	76.95	77.77
+ Lattice combination	72.44	73.75
System combination (weigh 0.6:0.4)	**71.39**	

→ Multiple-Array Track
9.2% WER reduction

Results (Baseline)

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	86.50	78.89	78.64	80.62
		S09	81.39	79.60	76.65	
	Eval	S01	82.80	67.13	81.75	73.29
		S21	78.10	65.56	69.97	

Dev → Eval:
Non-consistence exists

Results (Ranking A)

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	80.89	72.61	70.37	73.53
		S09	73.02	73.02	69.48	
	Eval	S01	74.40	58.86	75.69	65.25
		S21	68.89	57.64	62.02	
Multiple	Dev	S02	77.41	71.30	67.57	71.39
		S09	71.58	69.61	70.38	
	Eval	S01	75.64	58.18	75.64	66.27
		S21	68.38	61.14	66.24	

Results (Ranking A)

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	**80.89**	72.61	70.37	73.53
		S09	73.02	**73.02**	69.48	
	Eval	S01	74.40	58.86	75.69	65.25
		S21	68.89	57.64	62.02	
Multiple	Dev	S02	**77.41**	71.30	67.57	71.39
		S09	71.58	**69.61**	70.38	
	Eval					.27

Single-Array → Multiple-Array:
~3.5% improvements due to potential distributed diversity gain

Results (Ranking A)

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	80.89	72.61	70.37	73.53
		S09	73.02	73.02	69.48	
	Eval	S01	74.40	58.86	75.69	65.25
		S21	68.89	57.64	62.02	
Multiple	Dev	S02	77.41	71.30	67.57	71.39
		S09	71.58	69.61	70.38	
	Eval	S01	75.64	58.18	75.64	66.27
		S21	68.38	61.14	66.24	

Results (Ranking A)

Risk: the given reference Kinect is accurate enough to provide the best performance among distributed arrays → no enough room for improvement due to Kinects distribution in 2 sessions in Eval

Trac…						
Single						
	Eval	S01	74.40	58.8…		65.25
		S21	68.89	57.64		
Multiple	Dev	S02	77.41	71.30	67.5…	71.39
		S09	71.58	69.61	70.38	
	Eval	S01	75.64	58.18	75.64	66.27
		S21	68.38	61.14	66.24	

Conclusions

- A simple yet effective channel selection scheme
  - Entropy of DNN posterior probabilities
  - Meaningful for scenarios with distributed microphones/arrays available (Not for: only one consistent best microphone/array during dinner party)
- Be important to extract speaker characteristics
  - I-vectors, Pitch (future: speaker diarisation?)
- Temporal dynamics in feature extraction
  - LSTM is more efficient than TDNN (future: CNN for spectral dynamics?)
- Complementary knowledge for combination

Acknowledgements

- The authors acknowledge the support of Toshiba Research Europe Limited for valuable discussions and financial support of J. Zhang.
- The authors would like to thank Dr. Lin Wang for valuable discussions about speech separation using distributed microphones.
- B. T. Meyer was supported by the Cluster of Excellence 1077/1 Hearing4all.
- F. Xiong is supported by a Google Faculty Research Award.

Thanks for your listening !

Questions ?