USING NEURAL NETWORK FRONT-ENDS ON FAR FIELD MULTIPLE MICROPHONES BASED SPEECH RECOGNITION

Yulan Liu[1], ***Pengyuan Zhang***[2], ***Thomas Hain***[1]

[1]*University of Sheffield, UK;* [2]*Key Laboratory of Speech Acoustics and Content Understanding, IACAS, China.*

[1]{acp12yl, t.hain}@sheffield.ac.uk [2]pzhang@hccl.ioa.ac.cn

Abstract

- Meeting recognition with far-field recordings has been a challenging topic of wide research interest.
- We showed on average 25% relative WER reduction by using bottleneck features in tandem structure compared to using PLP features.
- Direct channel concatenation can outperform standard beamforming in utilizing multiple channel data to train DNN front-end.
- Adding meta-information (e.g. speaker information) in DNN front-end can further improve performance.

The research is supported by the EPSRC Programme Grant EP/1031022/1 (Natural Speech Technology project) and the Chinese Academy of Sciences Fellowship for Visiting Scholars.

Meeting Recognition on AMI Corpus

AMI Corpus

- Meeting corpus with multi-channel recordings: headset (IHM), distant microphones (SDM, MDM).
- Multi-level annotation on meta-information like head and body movement of speakers.
- Baseline using PLP features and HMM-GMMs

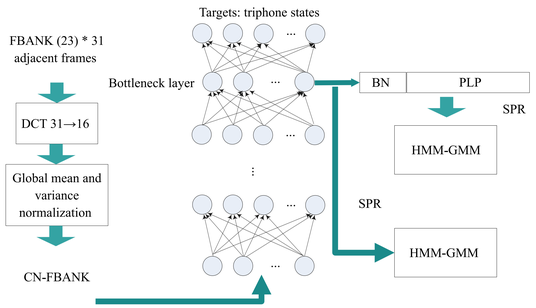
Train	Test	IHM	SDM	2bmit	4bmit	8bmit
o0	o0	35.6	66.3	61.8	60.5	58.2
o4	o0	32.3	61.3	57.1	56.0	53.8
	o4	35.4	65.1	60.4	59.8	58.2

o0: non-overlapping speech; o4: overlapping speech from maximally 4 speakers simultaneously. Beamforming is performed with toolkit BeamformIt.

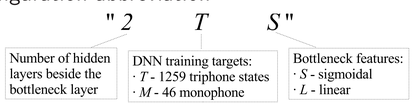
- Training and test sets used:

Dataset	Time	#Utt.	#Words	Description
acftrain	87.8h	73173	863726	Full training set (o4).
acntrain	15.8h	12876	152876	o0 from *acftrain*.
acftest	6.1h	4633	54820	Full test set (o4).
acntest	1.9h	1188	17536	o0 from *acftest*.

DNN Front-end



- Input features: log filter bank.
- Hidden layers: 1745 neurons in all but the bottleneck layer of 26 neurons.
- TNET toolkit, GTX690 based GPUs.
- Configuration abbreviation



- Performance using bottleneck features in HMM-GMMs

Feature	Train	Test	IHM	SDM	2bmit	4bmit	8bmit
BN-2TL	o0	o0	26.6	49.5	46.8	46.3	45.6
PLP+BN-2TS	o0	o0	26.7	49.9	46.9	46.8	45.3
	o4	o0	22.1	43.5	41.8	41.2	39.5
		o4	23.9	48.5	46.8	46.9	45.1

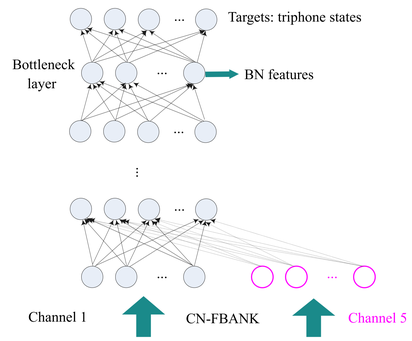
Progress in Channel Concatenation

- Channel concatenation can achieve equivalent or better performance compared with beamforming in both HMM-GMM system [1] and hybrid system [2].

[1] D. Marino and T. Hain, "An analysis of automatic speech recognition with multiple microphones", in INTERSPEECH 2011.

[2] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition", in ASRU 2013.

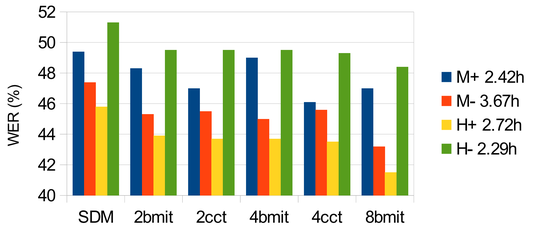
Channel Concatenation in DNN



- Augment the DNN input with features from non-neighbouring multiple channels.
- Performance (%WER)

Feature	Train	Test	2cct	4cct	8cct
BN-2TL	o0	o0	46.0	45.5	-
PLP+BN-2TS	o0	o0	46.8	46.5	47.4
	o4	o0	41.1	40.3	41.7
		o4	46.4	46.2	47.8

- Direct concatenation of 2 or 4 non-neighbouring distant channels outperforms beamforming, while 8 channel degrades because of large input dimension.
- Analysis
  - Distant channels: channel concatenation improved moving speech recognition.

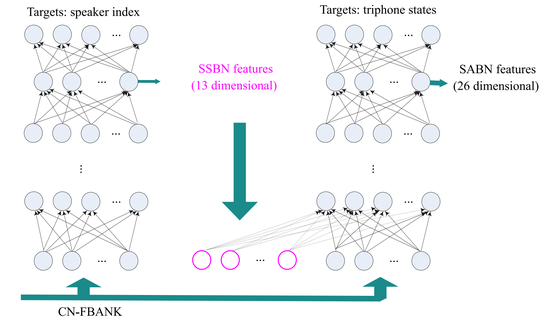


M+: body moving; *M-*: body not moving; *H+*: head moving; *H-*: head not moving.

  - IHM channel: cognitive load when head moves.

	M+	M-	H+	H-
IHM (%WER)	22.4	24.6	23.3	25.7

Speaker Awareness Deep Neural Network (SADNN)



- Augment DNN input with 13 dimensional *Speaker Separation Bottleneck* (SSBN) features, to generate the *Speaker Awareness Bottleneck* (SABN) features.
- Performance (%WER, trained and tested on o0)

Feature	IHM	SDM	2bmit	4bmit	8bmit
SABN-2TL	26.5	48.9	47.4	46.0	44.8
PLP+SABN-2TS	26.1	49.8	47.4	46.0	44.7

- With channel concatenation (%WER)

Feature	Train	Test	2cct	4cct
SABN-2TL	o0	o0	45.7	44.8
PLP+SABN-2TS	o0	o0	46.8	45.5

Adding Other Meta-information

- Location: Adding TDOA over channel concatenation degraded the performance, to 46.8% WER on 4 channels.
- Global information: Adding PLP based global speaker-GMM means degraded performance substantially.

Summary

- 2 and 4 channel concatenation in DNN front-end outperforms beamforming in tandem system.
- Adding speaker information with SSBN leads to further WER reduction.
- In total 2.5% relative WER reduction is observed across different channels by using both channel concatenation and SADNN structure.