NOVEL NEURAL NETWORK BASED FUSION FOR MULTISTREAM ASR

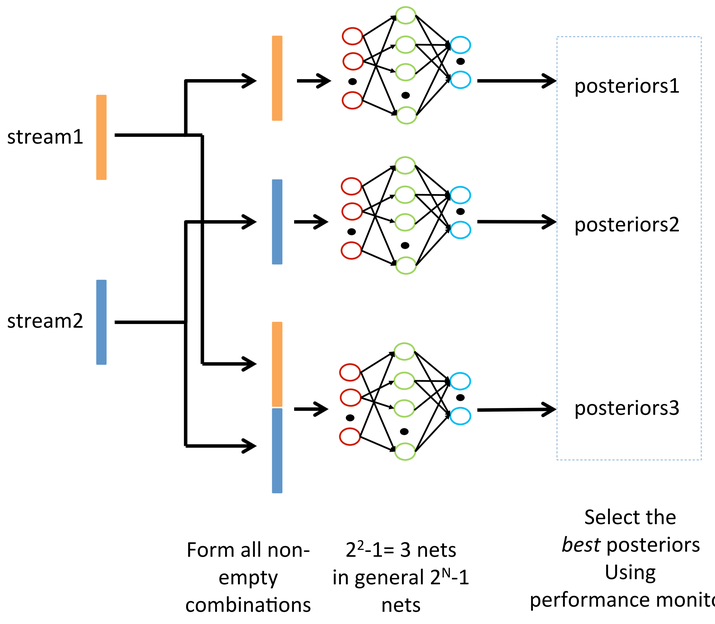
Sri Harish Mallidi, Hynek Hermansky
Johns Hopkins University, Baltimore, MD, USA

The Center For Language and Speech Processing at the Johns Hopkins University

Overview

- Improving robustness of ASR systems using Multistream framework
- Past multistream approaches involve training large number of fusion DNNs
- We propose a new multistream architecture with single fusion DNN
- Robust bottleneck features from the proposed multistream approach.
  -- Improvements in IARPA BABEL and Aurora4 ASR tasks

Full Combination Multistream System[1-2]:



Drawbacks:
Too many neural networks
N streams result in 2^N-1 neural networks
2 streams → 3 networks
5 streams → 31 networks
7 streams → 127 networks
Not suitable for bottleneck feature (BNF) extraction
Different weights results in different transforms
BNFs are crucial for many speech applications, e.g. Speaker ID/Language ID

References

[1] S. Tiberwala and H. Hermansky,"Sub-band based recognition of noisy speech," ICASSP, 1997.

[2] A. Morris, A. Hagen, H. Glotin, and H. Bourlard,"Multi-stream adaptive evidence combination for noise robust-ASR,"Speech Commun. 2001.

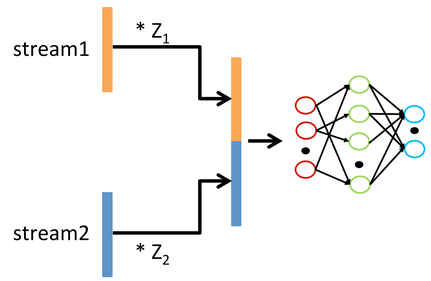
[3] S. Mallidi, T. Ogawa and H. Hermansky,"Uncertainty estimation of DNN classifiers," ASRU 2015.

[4] S. Mallidi et. al. "Autoencoder based multi-stream combination for noise robust speech recognition," Interspeech, 2015.

Proposed architecture:

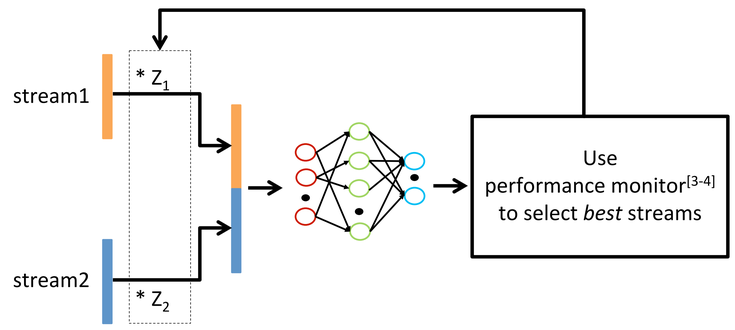
Replace multiple DNNs in Full combination architecture with single DNN

Training stage:



- Z_i is a binary switch
- Block Dropout features of a stream
- DNN randomly sees one of the following input patterns:
  - [stream1; 0]
  - [0; stream2]
  - [stream1; stream2]

Testing stage:



		Low Freq+High Freq	Low Freq	High Freq
1	Full Combination	4.95	8.39	12.07
2	Proposed DNN	5.06	8.56	13.19

- 3 DNNs in Full Combination arch. can be replaced with 1 DNN
- Row 1: WERs in each column are obtained from 3 different DNNs
- Row 2: WERS in each column are obtained from 1 DNN, but using masks
  -- [1 , 1], [1, 0] and [0, 1]

Streams:

- Sub-band streams. Each stream covers 2 Bark bands. Total 9 streams
- TRAP features in each sub-band
  - 11 frame context log-Mel filterbank, projected to 6 DCT basis

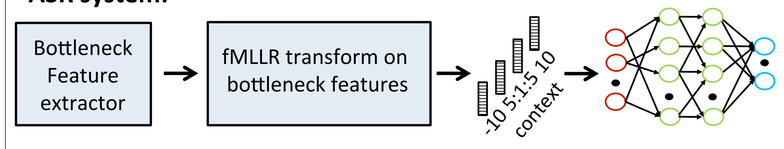
Results in Synthetic band-limited noises:

	20 dB	10 dB	0 dB
DNN	42.65	48.71	55.45
Mstrm DNN	37.49	41.12	46.26
Mstrm DNN + Perf. Monitor	35.39	36.39	38.57

Results in Additive noises:

	subway	volvo	factory	babble
DNN	71.3	79.6	78.0	78.2
Mstrm DNN	68.8	77.8	75.8	76.6
Mstrm DNN + Perf. Monitor	64.1	73.9	74.8	75.6

ASR system:



Results:

IARPA BABEL Year4 Languages:

	Igbo	Javanese	Guarani	Amharic	Mongolian
Baseline	60.5	58.0	46.7	43.6	52.2
Proposed	59.7	57.3	46.1	43.5	51.5

Aurora4

	A	B	C	D	Average
Baseline	3.16	5.09	5.70	16.32	9.83
Proposed	2.43	4.54	3.72	14.03	8.42

This work was supported in parts by the National Science Foundation via award number IIA-0530118, .Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013, by Google via Google faculty award to Hynek Hermansky. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Google, NSF, IARPA, DoD/ ARL, or the U.S. Government.