Uncertainty Estimation of DNN Classifiers

Sri Harish Mallidi[1], Tetsuji Ogawa[2], Hynek Hermansky[1]
[1]Johns Hopkins University, [2]Waseda University

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

How reliable are posteriors from a DNN?

- **Proposed two measures to reliability of posteriors from a DNN classifier.**
- **Applied proposed measures to extract noise robust bottleneck features.**

Autoencoder based measure of reliability

Motivation:

- A DNN is best performing and least uncertain about its estimates on the training data.
- Uncertainty of a test data can be measured by computing the deviation in probability estimates of the test and the training data.

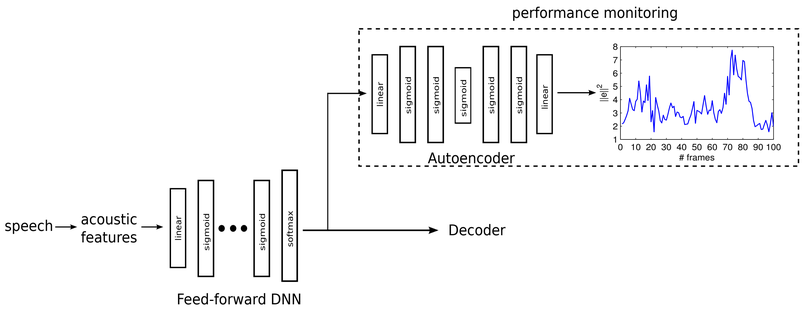

Figure 1: ASR system with Autoencoder based performance monitoring technique.

During training: DNN outputs are modelled using Autoencoders. Parameters of the autoencoder are optimized to minimize squared reconstruction error.

During testing: Uncertainty of each posterior vector of test utterance is proportional to reconstruction error from the autoencoder.

M-Delta measure

Motivation:

- For a "good" posteriogram,

$$Divergence(p_{t1}, p_{t2}) > Divergence(p_{t1,t3}) \quad (1)$$

where $p_{t1}, p_{t2} \in$ same class and $p_{t1}, p_{t2} \in$ different class.

We define

$$\mathcal{M}delta = \mathcal{M}^{\text{ac}} - \mathcal{M}^{\text{wc}} \quad (2)$$

where $\mathcal{M}^{ac}$ is accumulated diveregence of *across-class* posterior vectors and $\mathcal{M}^{\text{wc}}$ is accumulated diveregence of *within-class* posterior vectors.

- In test, we do not know whether $p_{t1}, p_{t2} \in$ same class or different class.
- We assume, accumulated diveregence of posterior vectors, seperated by Δt, can be written as

$$\mathcal{M}(\Delta t) = p^{\text{wc}}(\Delta t) \cdot \mathcal{M}^{\text{wc}} + p^{\text{ac}}(\Delta t) \cdot \mathcal{M}^{\text{ac}} + \epsilon_{\Delta t}, \quad (3)$$

where $\mathcal{M}(\Delta t) = \frac{1}{T-\Delta t} \sum_{t=\Delta t}^{T} \mathcal{D}(p_{t-\Delta t}, p_t)$, and

$p^{\text{wc}}(\Delta t)$ and $p^{\text{ac}}(\Delta t)$ denote the probability of a pair of frames separated by Δt being instances from the same and different phonemes, determined from the training data transcriptions.

Comparison with past measures

	seen noises					unseen noises			
Train \Test	clean	car	babble	bucc1.	bucc2.	destops.	exhall	f16	factory
Multi-condition	8.7	12.5	32.7	43.3	50.0	37.7	32.1	33.3	40.7
Matched condition	6.5	7.0	22.2	24.0	22.4	37.7	32.1	33.3	40.7
Past measures									
Entropy	8.2	12.0	36.3	41.3	48.1	39.7	34.5	33.6	43.0
M measure	6.7	8.9	38.2	30.2	34.2	41.8	35.8	30.2	46.8
Proposed measures									
M-delta measure	6.7	8.3	30.5	26.4	30.4	40.5	31.2	28.4	44.8
Autoencoder	6.7	7.0	22.5	23.9	23.5	37.3	24.8	25.1	39.9

Noise robust experiments

ASR system architecture:

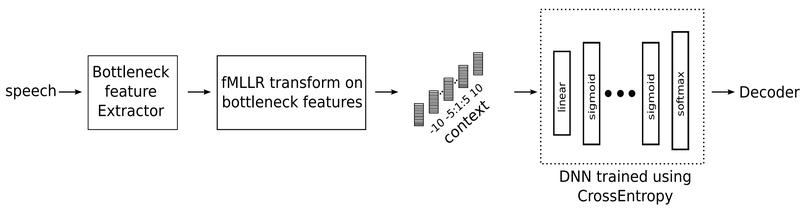

Figure 2: ASR system architecture used for noise robust experiments.

Multistream bottleneck feature extractor

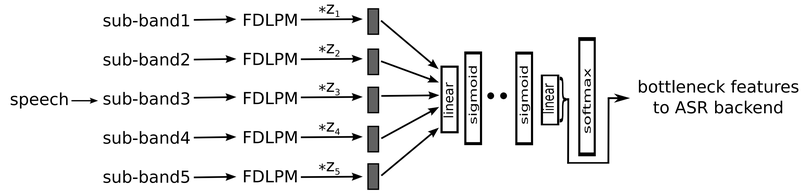

Figure 3: Training stage of Multistream bottleneck feature extractor.

Features of each band are multiplied with mask z_i, which takes $\{0, 1\}$ with equal probability. This allows the network not to break down when we switch-off a stream during test case.

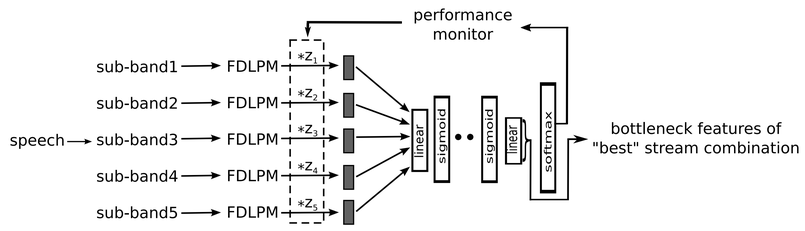

Figure 4: Test stage of Multistream bottleneck feature extractor.

Comparison with baseline bottleneck features (WER (%))

Data: Aurora4 noisy training data, and Aurora4 test set and CHiME3 dev. and eval. sets

Performance at Bottleneck feature extractor stage (WER %)

	Mel Filterbank	FDLPM	Multistream M-delta	Multistream Autoencoder
Aurora4	15.05	13.90	12.06	12.03
CHiME3	34.86	32.73	29.78	30.01

Performance at final classifier stage (WER %)

	Mel Filterbank	FDLPM	Multistream M-delta	Multistream Autoencoder
Aurora4	13.71	11.82	10.53	10.47
CHiME3	32.89	31.31	29.19	28.80