An Investigation into Speaker Informed DNN Front-end for LVCSR

Yulan Liu[1], *Penny Karanasou*[2], *Thomas Hain*[1]

[1]*The University of Sheffield, UK;* [2]*University of Cambridge, UK.*

[1]{acp12yl, t.hain}@sheffield.ac.uk [2]p.karanasou@eng.cam.ac.uk

Natural Speech Technology SPandH The University Of Sheffield. UNIVERSITY OF CAMBRIDGE

Abstract

- Considerable interest in "**informed training**" of DNNs: DNN input is augmented with **auxiliary codes carrying speaker information**.
- This work
  - shows mathematical equivalence between speaker informed DNN training and "bias adaptation";
  - analyses influential factors such as dimension, discrimination and stability of auxiliary codes;
  - compares different speaker informed DNN training methods in LVCSR task;
  - introduces a system based on speaker classification followed by speaker informed DNN for short utterances.

This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

Background

Signal and feature level	Integration into DNN
VTLN (Grezl2007) fMLLR (Yu2011)	LIN (Neto1995), LON (Li2010), LHN (Gemello2007), fDLR (Yu2011)
Speaker informed DNN training	**Encoding speaker information in DNN topology**
Eigenvectors (Dupont2000), i-vectors (Saon2013), speaker codes (Hamid2013), SSBN or d-vectors (Liu2014)	Output layer (Yao2012), bottleneck layer (Doddipatla2014), LHUC (Swietojanski2014)

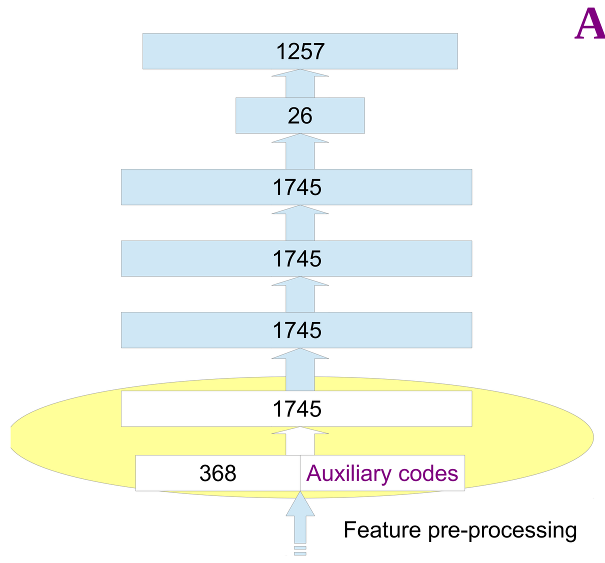
Speaker informed DNN training

DNN input layer:

$$x_{2,k}(t) = f\Big(\sum_{m=1}^{M} x_{1,m}(t) w_{1,m,k} + b_{1,k}\Big)$$

Speaker informed DNN input layer:

$$x'_{2,k}(t) = f\Big(\sum_{m=1}^{M} x_{1,m}(t) w'_{1,m,k} + \sum_{l=1}^{L} c_l(t) h'_{l,k} + b'_{1,k}\Big)$$



Auxiliary codes:

- eigenvectors in speaker space
- speaker i-vectors
- speaker codes
- speaker separation bottleneck features (SSBN)

Equivalent overall bias: $\beta_k(t) = \sum_{l=1}^{L} c_l(t) h'_{l,k} + b'_{1,k}$

- With speaker dependent auxiliary codes →

Speaker dependent equivalent bias: $\beta_k^s = \sum_{l=1}^{L} c_l^s h'_{l,k} + b'_{1,k}$

- Assume optimal bias $\hat{b}_k^s$, ideally $\beta_k^s = \hat{b}_k^s$
- force it to be true for all speakers →

$$\begin{pmatrix} c_1^1 & c_2^1 & \cdots & c_L^1 \\ c_1^2 & c_2^2 & \cdots & c_L^2 \\ \vdots & \vdots & \ddots & \vdots \\ c_1^S & c_2^S & \cdots & c_L^S \end{pmatrix} \begin{pmatrix} h'_{1,k} \\ h'_{2,k} \\ \vdots \\ h'_{L,k} \end{pmatrix} = \begin{pmatrix} \hat{b}_k^1 - b'_{1,k} \\ \hat{b}_k^2 - b'_{1,k} \\ \vdots \\ \hat{b}_k^S - b'_{1,k} \end{pmatrix}$$

- Special case: auxiliary code matrix is a unit matrix: **Unique Binary Index Codes** (**UBIC**)

Auxiliary codes

1. **Dimension**
   - As the number of speakers increases, the dimension of auxiliary codes should also increase.
2. **Discrimination**
   - **Linear separability and orthogonality**.
   - Higher discriminability → lower condition number of auxiliary code matrix.
   - Related to speaker separation using the auxiliary codes.
3. **Stability**
   - Using only local information enables fast estimation.
   - Temporal noise in auxiliary codes estimation degrades numerical stability in training, and the approximation to optimal speaker dependent biases in test.

Experiments

- AMI corpus, IHM, DNN-HMM-GMM (6 layered DNN);
- Training: 77.5h from 170 speakers;
- Test: 6.9h from 27 speakers (2.5h: 10 seen speakers; 4.4h: 17 unseen speakers);
- Average utterance length: 4.2s in training, 5s in test.

Auxiliary codes investigated

- SSBN, SSDNN posteriors [1]
- Speaker i-vectors [2]
- Hand-crafted codes

	8 dim binary index	170 dim UBIC	188 dim UBIC
Speaker 1 (seen)	00000000	000...001	000...001
Speaker 2 (seen)	00000001	000...010	000...010
Speaker 3 (seen)	00000010	000...100	000...100
...			
Speaker 174 (unseen)	10101101	000...000	000...010...000
Speaker 175 (unseen)	10101110	000...000	000...100...000
...			

- 170dim UBIC + SSDNN

[1] Y. Liu, P. Zhang and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in ICASSP2014.

[2] P. Karanasou, Y. Wang, M. Gales, and P. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in Interspeech 2014.

Frame-wise auxiliary codes

	SSBN	Seen Spkr	Unseen Spkr	Overall
baseline	–	*21.5*	*25.0*	*23.8*
SSBN	13	**20.3** (5.6%↓)	25.5	23.6
	40	20.4	**25.3** (1.2%↑)	**23.5** (1.2%↓)
	60	20.4	26.9	24.5
	80	20.5	25.9	23.9
	100	21.0	25.9	24.1
SSDNN posteriors	13	20.0	25.8	23.7
	40	20.5	**25.5** (2.0%↑)	23.7
	60	**19.8** (7.9%↓)	26.0	23.8
	80	20.1	25.9	23.8
	100	19.9	25.6	**23.5** (1.2%↓)

i-vectors

	Dim	Seen Spkr	Unseen Spkr	Overall
baseline	–	*21.5*	*25.0*	*23.8*
i-vectors	13	**19.3** (10.2%↓)	26.3	23.8
	40	19.6	**25.4** (1.6%↑)	**23.3** (2.1%↓)
	60	20.6	26.6	24.4
	80	19.6	**25.4** (1.6%↑)	**23.3** (2.1%↓)
	100	19.5	26.5	24.0

Hand-crafted codes

	SSBN	Seen Spkr	Unseen Spkr	Overall
baseline	–	*21.5*	*25.0*	*23.8*
8 dim codes	–	19.6	**25.6** (2.4%↑)	**23.4** (1.7%↓)
170dim UBIC	–	**19.3** (10.2%↓)	28.8	25.4
188dim UBIC	–	19.4	28.3	25.1
170dim UBIC (estimated)	13	19.4	26.4	23.8
	40	**19.3** (10.2%↓)	26.8	24.1
	60	**19.3** (10.2%↓)	26.8	24.1
	80	**19.3** (10.2%↓)	26.8	24.1
	100	19.4	26.7	24.1

Summary

- **Speaker informed DNN training**: common mathematical framework for auxiliary codes in DNN input.
  - Equivalent to using speaker dependent biases;
  - The dimension, discriminability and stability of auxiliary codes all influence the performance in practice.
- **Performance**
  - Seen speakers
    - i-vectors and UBIC based methods achieved equivalent and the best performance;
    - SSDNN-UBIC structure enables fast adaptation on short utterances without performance degradation.
  - Unseen speakers: no improvement potentially because:
    - i-vectors: insufficient speaker diversity in training data;
    - rest methods: lacking information about unseen speakers, system overfits to training.
  - Overall: i-vectors achieved best performance, followed by 8 dim hand-crafted binary index codes.