On The Relationship Between Speaker Informed DNN Training and Linear DNN Input Normalisation

Yulan Liu, Thomas Hain

The University of Sheffield, UK

{acp12yl, t.hain}@sheffield.ac.uk

Abstract

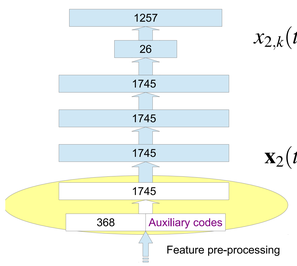
- Recent work showed equivalence between **speaker informed DNN training** and **speaker dependent biases**.
- This work
  - further investigates **speaker based DNN input normalisation** (**additive, multiplicative and in combination**);
  - shows that **additive speaker based input normalisation alone** yields equivalent performance to speaker informed training, while with considerably fewer parameters;
  - also investigates the **combination** of **proposed DNN input normalisation methods** and **informed training methods**, since they all factorise the DNN input layer.

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

Background

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

Speaker informed DNN training

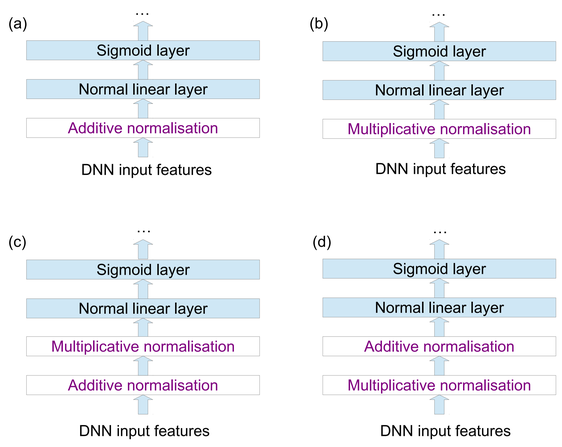


$$x_{2,k}(t) = f\Big(\sum_{m=1}^{M} x_{1,m}(t)w_{1,m,k} + \sum_{l=1}^{L} c_l^s h_{l,k} + b_{1,k}\Big)$$

$$\mathbf{x}_2(t) = f\Big(\mathbf{W}_1\mathbf{x}_1(t) + \beta_{\text{in}}^s\Big)$$

where $\beta_{\text{in}}^s = \mathbf{H}\mathbf{c}^s + \mathbf{b}_1$.

Speaker Normalisation on DNN Input



Additive input normalisation

$$\mathbf{x}_2(t) = f\Big(\mathbf{W}_1\big(\mathbf{x}_1(t) + \gamma^s\big) + \mathbf{b}_1\Big) = f\Big(\mathbf{W}_1\mathbf{x}_1(t) + \beta_{\text{ad}}^s\Big)$$

- Equivalent speaker dependent biases: $\beta_{\text{ad}}^s = \mathbf{W}_1\gamma^s + \mathbf{b}_1$.

Multiplicative input normalisation

$$\mathbf{x}_2(t) = f\Big(\mathbf{W}_1\big(\mathbf{x}_1(t) \otimes \alpha^s\big) + \mathbf{b}_1\Big) = f\Big(\Omega_{\text{mul}}^s\mathbf{x}_1(t) + \mathbf{b}_1\Big)$$

- Equivalent speaker dependent weights: $\Omega_{\text{mul}}^s = \mathbf{W}_1 \otimes \alpha^s$.

Combining additive and multiplicative normalisation (I)

$$\mathbf{x}_2(t) = f\Big(\mathbf{W}_1\big(\mathbf{x}_1(t) + \gamma^s\big) \otimes \alpha^s + \mathbf{b}_1\Big) = f\Big(\Omega_{\text{I}}^s\mathbf{x}_1(t) + \beta_{\text{I}}^s\Big)$$

- Equivalent weights: $\Omega_{\text{I}}^s = \mathbf{W}_1 \otimes \alpha^s$.
- Equivalent biases: $\beta_{\text{I}}^s = \mathbf{W}_1\gamma^s \otimes \alpha^s + \mathbf{b}_1$.

Combining additive and multiplicative normalisation (II)

$$\mathbf{x}_2(t) = f\Big(\mathbf{W}_1\big(\mathbf{x}_1(t) \otimes \alpha^s + \gamma^s\big) + \mathbf{b}_1\Big) = f\Big(\Omega_{\text{II}}^s\mathbf{x}_1(t) + \beta_{\text{II}}^s\Big)$$

- Equivalent weights: $\Omega_{\text{II}}^s = \mathbf{W}_1 \otimes \alpha^s$.
- Equivalent biases: $\beta_{\text{II}}^s = \mathbf{W}_1\gamma^s + \mathbf{b}_1$.

Optimisation

- Normalisation parameters are optimised with back-propagation in the same way like other DNN parameters.

$$\gamma^s(\tau) = \gamma^s(\tau-1) - a_{\text{add}} \cdot \frac{\partial Q}{\partial \gamma^s(\tau)}$$

$$\alpha^s(\tau) = \alpha^s(\tau-1) - a_{\text{mul}} \cdot \frac{\partial Q}{\partial \alpha^s(\tau)}$$

$$\frac{\partial Q}{\partial \gamma^s(\tau)} = \frac{\partial Q}{\partial \mathbf{x}_2(t)} \cdot \frac{\partial \mathbf{x}_2(t)}{\partial \gamma^s(\tau)}, \qquad \frac{\partial Q}{\partial \alpha^s(\tau)} = \frac{\partial Q}{\partial \mathbf{x}_2(t)} \cdot \frac{\partial \mathbf{x}_2(t)}{\partial \alpha^s(\tau)}$$

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.
- Speaker labels in test set are classified by SSDNN per utternace before selecting speaker normalisation parameters.

Speaker based mean and variance normalisation

ID	method	seen	unseen	overall
I0	*SI baseline*	*21.5*	*25.0*	*23.8*
O1	speaker mean norm	**21.9**	**24.4**	**23.5**
O2	speaker var norm	22.7	25.3	24.4
O3	utterance mean+var norm	22.6	25.9	24.7

Speaker based DNN input normalisation

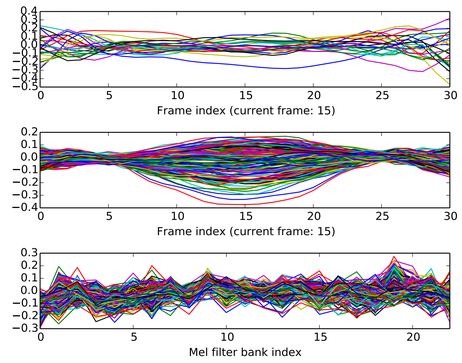
ID	method	seen	unseen	overall
I0	*SI baseline*	*21.5*	*25.0*	*23.8*
I1	informed training, UBIC [1]	**19.3**	26.8	24.1
I2	informed training, IV13 [1]	**19.3**	26.3	**23.8**
I3	AN ($a_{\text{add},0}$=0.008)	19.6	30.4	26.5
I4	AN ($a_{\text{add},0}$=0.001)	**19.3**	29.0	25.6
I5	MUN-L ($a_{\text{mul},0}$=0.008)	20.1	29.2	25.9
I6	MUN-S ($a_{\text{mul},0}$=0.008)	20.0	28.7	25.6

[1] Y. Liu, P. Karanasou, and T. Hain, "An investigation into speaker informed DNN front-end for LVCSR," in ICASSP 2015.

Combination of different DNN input normalistion

ID	methods	seen	unseen	overall
I0	*SI baseline*	*21.5*	*25.0*	*23.8*
C1	$[\mathbf{x}_1(t) + \gamma^s] \otimes \sigma(\alpha^s)$	**19.1**	30.9	26.6
C2	$\mathbf{x}_1(t) \otimes \sigma(\alpha^s) + \gamma^s$	19.6	33.0	28.3
C3	AN ($a_{\text{add},0}$=0.008), UBIC	19.2	31.0	26.7
C4	AN ($a_{\text{add},0}$=0.001), UBIC	19.2	**30.1**	**26.2**
C5	$[\mathbf{x}_1(t) + \gamma^s] \otimes \sigma(\alpha^s)$, UBIC	19.4	30.8	26.7

Visualising trained additive normalisation parameters



Summary

- **Speaker informed DNN training**: equivalent to **speaker based additive normalisation** over DNN input.
  - Both **factorises linearly** the overall biases in input layer;
  - Both reduced WER by 10.2% relative on seen speakers;
  - Combining two methods does not improve further.
- **Speaker based DNN input normalisation**:
  - **Both additive and multiplicative** normalisation are effective on seen speakers, and outperform mean and variance normalisation;
  - **Mutual optimisation** of speaker normalisation parameters with DNN is crucial;
  - **Additive normalisation** generally outperforms multiplicative normalisation;
  - Combining additive and multiplicative normalisations brings slight improvement further on seen speakers;
  - Combining with informed training does not improve further.