I-Vector Estimation Using Informative Priors for Adaptation of Deep Neural Networks

P. Karanasou, M. Gales, P. Woodland
University of Cambridge

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

UNIVERSITY OF CAMBRIDGE

Introduction

- I-vectors: low-dimensional representation of speaker space
- Popular in unsupervised adaptation of DNN acoustic models
- A small number of parameters to estimate => can be used with very little data, i.e. a single utterance
- Prior needed to improve robustness of i-vector estimates with limited data

Use of priors in i-vector estimation

- Default standard normal prior [Kenny, 2006] - Gaussian assumption distorts speaker space
- Heavy-Tailed (HT) prior [Kenny, 2010] allows larger deviations from the mean - works better but complicated model
- Keep the Gaussian assumption and perform length-normalisation of i-vectors [Garcia-Romero and Espy-Wilson, 2011]
- GMM prior estimated on training data - not integrated to i-vector training [Travadi et al., 2014]

Our approach

- Use an **informative prior** that models the actual behaviour of speaker space
  - Derived from training data and normalised
  - Incorporated into i-vector estimation using a count-smoothing framework [Gales, 1997], [Breslin et al., 2010]
- Less sensitive to the quantity of data used to estimate the i-vector and to the mismatch between training and test data
- Allows prior information at different levels, for ex. average over speaker-space, gender-dependent, etc.

I-vector estimation

- One i-vector per speaker, estimated on all data of the speaker

$$\boldsymbol{\mu}^{(sm)} = \boldsymbol{\mu}_0^{(m)} + \boldsymbol{M}^{(m)}\boldsymbol{\lambda}^{(s)}$$

,where $\boldsymbol{\mu}^{(sm)}$ the speaker-dependent supervector, $\boldsymbol{M}^{(m)}$ a low-rank matrix $D \times P$, $\boldsymbol{\lambda}^{(s)}$ the i-vector of size P of speaker s.

- ML estimation of i-vectors and model parameters. The i-vector for speaker s is given by

$$\boldsymbol{\lambda}^{(s)} = \boldsymbol{G}_\lambda^{(s)-1}\boldsymbol{k}_\lambda^{(s)}$$

where

$$\boldsymbol{G}_\lambda^{(s)} = \sum_{m,t} \gamma_t^{(m)}(s)\boldsymbol{M}^{(m)T}\boldsymbol{\Sigma}^{(m)-1}\boldsymbol{M}^{(m)}$$

$$\boldsymbol{k}_\lambda^{(s)} = \sum_{m} \boldsymbol{M}^{(m)T}\boldsymbol{\Sigma}^{(m)-1}\sum_{t} \gamma_t^{(m)}(s)(\boldsymbol{x}_t - \boldsymbol{\mu}_0^{(m)})$$

Informative priors for i-vector estimation

- Each prior estimates an i-vector that should well represent the speaker space
- ID prior : i-vector $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- SI prior : $\boldsymbol{\lambda}_{SI}^{(s)} = \boldsymbol{G}_{\lambda(\mathtt{SI})}^{-1}\boldsymbol{k}_{\lambda(\mathtt{SI})}$
- Gender prior : $\boldsymbol{\lambda}_{Gender}^{(s)} = \boldsymbol{G}_{\lambda(\mathtt{Gender})}^{-1}\boldsymbol{k}_{\lambda(\mathtt{Gender})}$ for Gender $\in \{M, F\}$
- Should also cover some "variance"

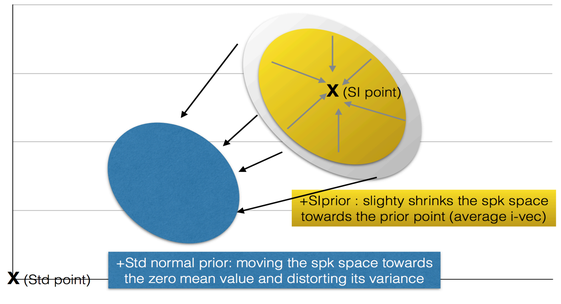
- Add prior information to accumulates used for i-vector estimation
- ID prior

$$\boldsymbol{G}_\lambda^{'(s)} = \boldsymbol{G}_\lambda^{(s)} + \tau\boldsymbol{I},\ \boldsymbol{k}_\lambda^{'(s)} = \boldsymbol{k}_\lambda^{(s)} + \tau * 0$$

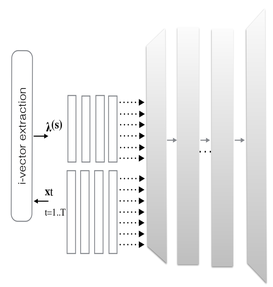
- Count-smoothing prior: Interpolate basic accumulates with priors estimated on the training data

$$\boldsymbol{G}_\lambda^{'(s)} = \boldsymbol{G}_\lambda^{(s)} + \tau\frac{\boldsymbol{G}_{\lambda(\mathtt{pr})}}{\sum_{m,t}\gamma_t^{(m)}},\ \boldsymbol{k}_\lambda^{'(s)} = \boldsymbol{k}_\lambda^{(s)} + \tau\frac{\boldsymbol{k}_{\lambda(\mathtt{pr})}}{\sum_{m,t}\gamma_t^{(m)}}$$

where $\boldsymbol{G}_{\lambda(\mathtt{pr})}$ and $\boldsymbol{k}_{\lambda(\mathtt{pr})}$ can be speaker-independent (SI), gender-dependent(Gender)...



Appending i-vectors to DNN input



- Training on US English BN Corpus: $144h$, $\sim 8k$ speakers
- Evaluation on dev03 set, manually and automatically segmented: $\sim 4h$
- Average utter: training $11.6s$, dev03-manual $16.1s$, dev03-auto $8.7s$

Results of hybrid decoding

Table : Hybrid decoding results for DNNs with SI input features (WER %)

System	dev03-manual	dev03-auto
Baseline	12.7	12.9
+iv-spk	**11.9**	15.6
+iv-utter	**11.5**	**11.8**
+iv-utter-Stdprior	14.2	14.2
+iv-utter-SIprior	11.5	11.8
+iv-utter-Genderprior	11.6	11.9
+iv-utter-Stdprior-retrain	11.6	12.5
+iv-utter-SIprior-trn-retrain	**11.1**	**11.6**
+iv-utter-Genderprior-retrain	**11.1**	**11.4**

- "+iv-utter": append utterance-level test i-vectors to DNN input
- Compare "+iv-utter-Stdprior" and "+iv-utter-Stdprior-retrain": Std normal prior sensitive to mismatch of trn and test i-vector spaces
- Best performance with utter-level test i-vectors with informative priors ("+iv-utter-{SIprior,Genderprior}-retrain")

Conclusions and Future work plans

- Use of informative priors for i-vector estimation; derived from training data and better model the behaviour of speaker space
- Best performance on US BN data with utterance level test i-vectors enhanced with the informative priors
- Informative priors to factorised i-vectors [Karanasou et al., 2014]
- Use of other prior sources within the count-smoothing framework