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

*Penny Karanasou, Mark Gales, Philip Woodland*

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{pk407, mjfg, pcw}@eng.cam.ac.uk

## Abstract

I-vectors are a well-known low-dimensional representation of speaker space and are becoming increasingly popular in adaptation of state-of-the-art deep neural network (DNN) acoustic models. One advantage of i-vectors is that they can be used with very little data, for example a single utterance. However, to improve robustness of the i-vector estimates with limited data, a prior is often used. Traditionally, a standard normal prior is applied to i-vectors, which is nevertheless not well suited to the increased variability of short utterances. This paper proposes a more informative prior, derived from the training data. As well as aiming to reduce the non-Gaussian behaviour of the i-vector space, it allows prior information at different levels, for example gender, to be used. Experiments on a US English Broadcast News (BN) transcription task for speaker and utterance i-vector adaptation show that more informative priors reduce the sensitivity to the quantity of data used to estimate the i-vector. The best configuration for this task was utterance-level test i-vectors enhanced with informative priors which gave a 13% relative reduction in word error rate over the baseline (no i-vectors) and a 5% over utterance-level test i-vectors with standard prior.
**Index Terms**: i-vectors, speaker adaptation, prior information, deep neural networks

## 1. Introduction

I-vectors have been used recently with success for the adaptation of hybrid DNN-HMMs acoustic models [1], [2], [3]. They offer a low-dimensional fixed-length representation of speaker-space spanning the dimensions of highest variability, and they are a convenient method for unsupervised adaptation of DNNs. They are appended as auxiliary features to the input of the DNN system, and are estimated independently of the DNN parameters. Since there are only a small number of parameters to estimate, i-vector based adaptation can be suitable even for a limited amount of adaptation data.

The capacity of i-vectors to perform adaptation given small amounts of adaptation data can be useful in many cases. Working at the utterance level enables the use of i-vectors even if speaker labels are not available for the adaptation data. In certain cases extracting i-vectors from smaller segments of speech may also help to avoid overlapping speech and uncertain speaker labelling. In [4] utterance level i-vectors are extracted because of limited coverage in the available corpora in terms of speakers and environmental conditions. Last but not least, working with shorter speech segments offers adaptation with lower latency. The preferable length of adaptation data for the i-vector estimation has been a subject of study for speaker verification with the speaker verification performance declining sharply once utterance lengths fall below 10s [5].

When a limited amount of adaptation data is available, a prior on the i-vector estimation model can make the system more robust. The default is the standard normal prior [6] which is however sensitive to the available amount of data per speaker and to a mismatch between the training and test data length. This is because of a Gaussian assumption over the i-vector space which is not always true; when reducing duration, the variance of the i-vector estimate increases and decisions become error-prone. Heavy-tailed (HT) priors were thus proposed [7] to allow for larger deviations from the mean (e.g., severe channel distortions) and to increase the robustness to outliers in the ML estimation of the model parameters. This approach performed better than the Gaussian prior, hence, providing strong empirical evidence towards non-Gaussian behaviour of speaker and channel effects. In [8], to avoid the complicated HT models, the Gaussian assumptions are kept, but a length-normalisation of the i-vectors is performed. It is shown that the length normalisation approximates the HT to the standard Gaussian distribution. Lastly, in [9] a GMM prior estimated on the training data is incorporated to the basic statistics at test time.

In this paper a count-smoothing framework is adopted for incorporating prior knowledge into i-vector estimation. The smoothing idea, first introduced by [10], is based on the interpolation of observed statistics and prior statistics, both derived from the training data. In [11] it was used successfully in transformation estimation for rapid speaker adaptation in the ASR domain, while in [12] it was applied in text-to-speech (TTS) in a cluster adaptive training (CAT) representation to give robustness over utterance estimations. This is a flexible framework for incorporating priors, not constraining the prior to be static. It may be dynamic and change across utterances, or represent information at different levels. In this work, the prior statistics were first estimated for the entire speaker space offering an average representation of it. Second, gender clustering of the training data was used and the prior statistics of two gender i-vectors were extracted. Our approach integrates prior estimation into EM training of the i-vectors after a normalisation of the prior statistics. A normalised prior estimated using the training data models the actual behaviour of the speaker space and is less sensitive to the quantity used to estimate the i-vector and to the mismatch between training and test data. Thus, it does not degrade the word error rate (WER) without retraining the DNN as will be shown to be the case when the standard normal prior is used (see Section 5). Experiments were conducted using the US

This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Supporting data for this paper is available at the http://www.repository.cam.ac.uk/handle/1810/248387 data repository.

English BN transcription task for speaker and utterance level adaptation i-vectors. The best configuration for this task was utterance-level test i-vectors enhanced with informative priors which gave a 13% relative reduction in WER.

The rest of the paper is organised as follows. First, i-vectors are briefly described in Section 2. Next, Section 3 details the proposed informative priors in a count-smoothing framework. Sections 4 and 5 present the experimental setup and the hybrid DNN-HMM decoding results, respectively. The paper concludes in Section 6 with a summary of the presented work and some future work plans.

## 2. I-vector estimation

Following [3], the i-vector approach is presented as a type of model-based CAT estimation [13] where the HMM model is replaced by a GMM model, meaning that no transcriptions of the data are required. The intrinsic phoneme variability is represented by a canonical model $\mathcal{M}$, which here is a GMM universal background model (UBM) with $M$ mixture components [14]. It is defined by a mean supervector of component means $\boldsymbol{\mu}_0^{(m)}$, diagonal component covariance matrices $\boldsymbol{\Sigma}^{(m)}$ and mixture coefficients $\omega^{(m)}$. The input acoustic feature vectors $\boldsymbol{x}_t \in \mathbb{R}^D$ are seen as samples generated by the model $\mathcal{M}$.

We examine the case of having one i-vector per speaker, estimated on all the data of the particular speaker and being constant across all utterances of the speaker. Each speaker is represented by a point in the "speaker eigenspace" spanned by the i-vectors. There is a linear dependence between the speaker-adapted means (i.e. speaker-dependent supervector) and the canonical means, which for a particular Gaussian component $m \in M$ is given by

$$\boldsymbol{\mu}^{(sm)} = \boldsymbol{\mu}_0^{(m)} + \boldsymbol{M}^{(m)}\boldsymbol{\lambda}^{(s)} \tag{1}$$

where $\boldsymbol{\mu}^{(sm)}$ is the $m$-th component of speaker-dependent supervector, $\boldsymbol{M}^{(m)}$ is the factor submatrix for component $m$ of size $D \times P$, representing $P$ bases spanning the subspaces with the highest variability in the mean supervector space, and $\boldsymbol{\lambda}^{(s)}$ is a vector of size $P$ representing the i-vector of speaker $s$.

To extract the initial speaker i-vectors, a speaker-dependent (SD) model using all the data of each speaker is trained and used to extract a mean supervector. Principal component analysis (PCA) is then applied to these supervectors to obtain the speaker i-vectors that span the $P$-space. Next, maximum-likelihood estimation of the model parameters and of the i-vectors is performed. The auxiliary function to be maximised is

$$Q(\mathcal{M}, \boldsymbol{\lambda}^{(s)}; \hat{\mathcal{M}}, \hat{\boldsymbol{\lambda}}^{(s)}) = \\ -\frac{1}{2}\sum_{s,t,m} \gamma_t^{(m)}(s)(\boldsymbol{x}_t - \boldsymbol{\mu}^{(sm)})^T \boldsymbol{\Sigma}^{(m)-1}(\boldsymbol{x}_t - \boldsymbol{\mu}^{(sm)}) \tag{2}$$

where $\mathcal{M}$ is the canonical model to be estimated and $\hat{\mathcal{M}}$ is the "old" model. $\boldsymbol{\lambda}^{(s)}$ are the i-vectors to be estimated and $\hat{\boldsymbol{\lambda}}^{(s)}$ the "old" i-vectors. $\gamma_t^{(m)}(s)$ is the posterior probability of Gaussian component $m$ at time $t$ determined using the canonical model parameters $\hat{\mathcal{M}}$ and the speaker i-vectors $\hat{\boldsymbol{\lambda}}^{(s)}$.

The training procedure uses the Expectation-Maximisation (EM) algorithm to estimate the parameters.

By differentiating Equation 2 with respect to the i-vector of a particular speaker and equating to zero, the i-vector for speaker $s$ may be shown to be:

$$\boldsymbol{\lambda}^{(s)} = \boldsymbol{G}_\lambda^{(s)-1}\boldsymbol{k}_\lambda^{(s)} \tag{3}$$

where $\boldsymbol{G}_\lambda^{(s)}$ and $\boldsymbol{k}_\lambda^{(s)}$ are given by

$$\boldsymbol{G}_\lambda^{(s)} = \sum_{m,t} \gamma_t^{(m)}(s)\boldsymbol{M}^{(m)T}\boldsymbol{\Sigma}^{(m)-1}\boldsymbol{M}^{(m)} \tag{4}$$

$$\boldsymbol{k}_\lambda^{(s)} = \sum_m \boldsymbol{M}^{(m)T}\boldsymbol{\Sigma}^{(m)-1}\sum_t \gamma_t^{(m)}(s)(\boldsymbol{x}_t - \boldsymbol{\mu}_0^{(m)}) \tag{5}$$

To estimate the factor matrix $\boldsymbol{M}^{(m)}$, it suffices to differentiate Equation 2 with respect to $\boldsymbol{M}^{(m)}$ and equate to zero. Doing so, the sufficient statistics are collected:

$$\boldsymbol{G}_{\mathtt{M}}^{(m)} = \sum_{s,t} \gamma_t^{(m)}(s)\boldsymbol{\lambda}^{(s)}\boldsymbol{\lambda}^{(s)T} \tag{6}$$

$$\boldsymbol{K}_{\mathtt{M}}^{(m)} = \sum_{s,t} \gamma_t^{(m)}(s)(\boldsymbol{x}_t - \boldsymbol{\mu}_0^{(m)})\boldsymbol{\lambda}^{(s)T} \tag{7}$$

The factor matrix $\boldsymbol{M}^{(m)}$ is estimated as:

$$\boldsymbol{M}^{(m)} = \boldsymbol{K}_{\mathtt{M}}^{(m)}\boldsymbol{G}_{\mathtt{M}}^{(m)-1} \tag{8}$$

## 3. Informative priors for i-vector estimation

No prior is used in the above presented model (Section 2). The most commonly used prior assumes a standard normal Gaussian distribution $P(\boldsymbol{\lambda}^{(s)}) = \mathcal{N}(0, \boldsymbol{I})$ over the i-vectors ("Stdprior"). This prior can be incorporated in the accumulates statistics used for i-vector estimation as

$$\boldsymbol{G}_\lambda^{'(s)} = \boldsymbol{G}_\lambda^{(s)} + \tau\boldsymbol{I} \tag{9}$$

$\boldsymbol{k}_\lambda^{(s)}$ is not altered because the assumed prior has a zero mean. $\tau$ sets the weight of the contribution of the prior to the final statistics. The higher the value of $\tau$, the bigger the contribution of the prior to the i-vector estimation.

As already mentioned in Section 1, this Gaussian assumption distorts the speaker space. It assumes that the i-vectors extracted from utterances of the same speaker form separable clusters. This is however not always the case, especially when working with short utterances with high within-class variability in estimated sufficient statistics. In this paper a more informative alternative is proposed to incorporate prior knowledge estimated from the data. A count-smoothing framework interpolates observed and prior statistics:

$$\boldsymbol{G}_\lambda^{'(s)} = \boldsymbol{G}_\lambda^{(s)} + \tau\frac{\boldsymbol{G}_{\lambda(\mathtt{pr})}}{\sum_{m,t}\gamma_t^{(m)}} \tag{10}$$

$$\boldsymbol{k}_\lambda^{'(s)} = \boldsymbol{k}_\lambda^{(s)} + \tau\frac{\boldsymbol{k}_{\lambda(\mathtt{pr})}}{\sum_{m,t}\gamma_t^{(m)}} \tag{11}$$

The prior statistics $\boldsymbol{G}_{\lambda(\mathtt{pr})}$ and $\boldsymbol{k}_{\lambda(\mathtt{pr})}$ are weighted by a factor $\tau$ so that they effectively contribute $\tau$ frames to the final statistics. They are also normalised by the total occupancy counts of the data $\sum_{m,t}\gamma_t^{(m)}$. The prior statistics are integrated into the EM estimation of the i-vectors (Section 2). The count-smoothing can be seen as a form of Maximum a Posteriori (MAP) framework, where a Gaussian prior is used at the "M-step" of the EM algorithm. In this work, the prior statistics were first estimated across all training speakers ("Siprior"). This prior represents a speaker-independent i-vector $\boldsymbol{\lambda}^{(\mathtt{SIpr})} = \boldsymbol{G}_{\lambda(\mathtt{SIpr})}^{-1}\boldsymbol{k}_{\lambda(\mathtt{SIpr})}$, estimated on all training utterances without the use of any speaker labels. A more specific prior was also applied, estimated on the gender clusters of the training data ("Genderprior"). In this case, two prior i-vectors $\boldsymbol{\lambda}^{(\mathtt{Genderpr})}$ are estimated on the training data, one for each gender.

# 4. Experiments

## 4.1. English Broadcast News Corpus

The US English BN transcription task was chosen to evaluate the effectiveness of i-vector adaptation on a large vocabulary continuous speech recognition task. The training set includes the 144h 1996 [15] & 1997 [16] Hub-4 English BN Speech dataset (LDC97S44, LDC98S71), containing 288 shows with $\sim$8k speakers. The speakers are distributed in a very unbalanced way with a few dominant speakers and many speakers with limited data. Another difficulty of the data is that they include seven so called "focus conditions", corresponding to a mix of noise conditions and speech style. Working with a corpus with different noise conditions and speech styles and with an unbalanced speaker distribution enables us to investigate the effectiveness of i-vectors in a real world scenario.

Two versions of the DARPA RT03 dev03 set were used for evaluation (approx. 4h of speech). The first version is based on a manual segmentation and true speaker labels, while the second version is automatically segmented and clustered using the RT04 Cambridge segmentation system [17]. The average utterance duration is 11.6s for the training set, 16.1s for the manually segmented dev03 set and 8.7s for the automatically segmented dev03 set. All presented decoding results were produced after the lattices were rescored with the trigram language model used in the Cambridge RT04 transcription system.

## 4.2. Hybrid DNN baseline

Hybrid DNN-HMM systems [18] with speaker-independet (SI) and speaker-adapted (SA) features were built using the BN corpus and serve as baselines for this work. The SI input acoustic features were 52-dimensional, consisting of 12 PLP coefficients, the zeroth cepstrum, the delta, the delta-delta and the delta-delta-delta coefficients, processed by global cepstral mean normalisation (CMN) and cepstral variance normalisation (CVN). To extract the SA features, a speaker-level Constrained Maximum Likelihood Linear Regression (CMLLR) transform [19] was applied to the input features using a GMM-HMM system. For all DNNs (baselines and later with appended i-vectors), the input features used a temporal context of 9 frames, the number of hidden layers was set to 5 with 1000 units each, while the output layer had 6000 units. The output units correspond to context-dependent sub-phone units derived by decision-tree state tying [20]. The DNNs were trained with an extension of ICSI's QuickNet software [21]. Sigmoid and softmax functions were used for the nonlinearities in the hidden and output layers, respectively. The parameters of the network were initialised using a discriminative layer-by-layer pre-training algorithm [22], followed by "fine tuning" of the full network. The standard back propagation procedure was used to optimise the DNN weights with cross entropy as the objective function.

## 4.3. Estimation of training and test i-vectors

To model the feature space from which the i-vectors were estimated, an SI UBM GMM model with 2048 mixture components was trained on the BN training corpus. Each component corresponded to a 39-dimensional front-end feature vector, consisting of 12 PLP coefficients appended with the energy, the delta and the delta-delta coefficients. SD models were trained on all utterances of each speaker, from which the speaker i-vectors were extracted. The training of the speaker i-vectors was then done following the procedure described in Section 2. In parallel, the informative priors were estimated on a randomly selected subset of the training data (around $1/10$ of the total training data).The $\tau$ value (see Section 3) was set to $40$, which is the minimum number of frames of speech for a speaker to be taken into account into the estimation of the i-vector space.

Two sets of experiments were conducted for the test i-vectors. First, one i-vector was extracted for each speaker. Next, each test utterance was treated as a separate entity (i.e speaker) and utterance level i-vectors were extracted. This second set of experiments is very useful in the case of the automatically segmented dev03 where the speaker clustering is not error-free, as will be further shown in Section 5. In addition, working with shorter data more clearly shows the influence of the priors on the adaptation. To estimate the test i-vectors in all cases, we used the model parameters and the priors learned during the i-vector training and we updated the i-vector weights. The choice of training the i-vectors at the speaker level, while the test i-vectors are extracted either at speaker or utterance level, is made in order to achieve a better estimate of the speaker space making use of the speaker information available for the training data. In [2] both train and test i-vectors were extracted at the utterance level, which results in an independent of within speaker variability but more noisy estimation (especially for short utterances).

The i-vectors were concatenated with the acoustic features to form the input for hybrid DNN training and decoding. The dimension of the i-vectors was set to 30. Before the concatenation, the i-vectors were normalised to zero mean and unit variance on the training data, as normally done for DNN training.

## 4.4. Prior representation of the speaker space

In order to investigate the informative power of the different priors used in this work, decoding was conducted on the hybrid DNN-HMM system with SI features using only the prior statistics for the test i-vectors, instead of combining the prior with the observed statistics. This means that in the case of "SIprior" for example, the SI (estimated across all training speakers) i-vector was assigned to all test speakers. The results of these illustrative experiments on manual dev03 set are presented in Table 1. The last row of the Table ("Randomiv") presents the case of assigning an i-vector randomly drawn from the training i-vectors set. This is used as a control for the experiments, since any prior is expected to be more representative of the speaker space than a random i-vector, which is the case as can be seen in Table 1.

| System | dev03-manual |
|---|---|
| Baseline (no ivec) | 12.7 |
| Stdprior | 18.9 |
| SIprior | 18.7 |
| Genderprior | 16.4 |
| Randomiv | 22.1 |

Table 1: Hybrid decoding with SI features using only prior statistics for test i-vectors (WER %)

Decoding with prior-only i-vectors over a hybrid system trained with i-vectors without any prior induces a mismatch between the training and the test speaker space, which normally degrades the WER. However, the more informative the prior the better it should model the i-vector space limiting its distortion. This is indeed the case with "Genderprior" which gives the best results thanks to the fact that is uses information of a first clustering of the speaker space. It is also interesting that there is not a big difference in WER between the "Stdprior" and the "SIprior". This is because of the global normalisation of the

standard prior (originally an all-zero i-vector) over the training data, which moves its speaker space closer to the training space. Without this normalisation, it degrades severely the WER.

## 5. Results

Table 2 present the decoding results for hybrid systems using SI input features for training and decoding, with and without appended i-vectors, for the manual dev03 set ("dev03-manual") and the automatically segmented and clustered dev03 set ("dev03-auto"). It can first be seen that the use of speaker i-vectors improves the baseline for "dev03-manual" (row "+iv-spk"), but increases the WER for "dev03-auto". This may be because of errors in the automatic segmentation and clustering procedure, which show a sensitivity of the i-vectors to the speaker clustering. Adding priors to the speaker i-vectors for "dev03-auto" is not further presented since the results do not seem indicative because of the clustering issues.

| System | dev03-manual | dev03-auto |
|---|---|---|
| Baseline | 12.7 | 12.9 |
| +iv-spk | **11.9** | 15.6 |
| +iv-spk-Stdprior | 16.0 | - |
| +iv-spk-SIprior | **11.8** | - |
| +iv-spk-Genderprior | 11.9 | - |
| +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** |

Table 2: Hybrid decoding results for DNNs with SI input features (WER %)

In the second block of the table, the utterance level test i-vectors improve further the WER ("+iv-utter") for both versions of dev03 set, maybe because of less confusions related to noise variations. The remaining rows of the first two blocks present the results when priors are incorporated in the i-vector estimation but without training the hybrid system with the prior-enhanced i-vectors (i.e. the hybrid system is trained with the speaker i-vectors without any prior). This set of experiments is done to investigate the behaviour of different priors. An ideal prior would not result in any distortion of the speaker space and should be able to provide improvements when used in decoding, even without retraining. However, the standard prior degrades in performance. This is explained by the distortion of the i-vector space made by the standard Gaussian prior assumption which results in a mismatch of the training and test i-vector space. The distortion created by informative priors is limited since they are estimated on the training data and normalised and, thus, do not degrade the baseline. This behaviour is consistent for both versions of the dev03 set.

In the third block of Table 2, the hybrid system is trained with prior-enhanced i-vectors. Thus, there is no longer a mismatch between the training and test i-vector space. As already mentioned, the standard prior is particularly sensitive to such mismatch. Now that this mismatch is removed ("+iv-utter-Stdprior-retrain"), the WER does not degrade. As shown, the informative priors do not create a large mismatch, they slightly shrinks the i-vector space towards the prior points (SI or Gender respectively) though. When this slight distortion is corrected with the retraining of the hybrid system ("+iv-utter-SIprior-trn-retrain", "+iv-utter-Genderprior-retrain"), a 13% relative improvement over the "dev03-manual" baseline is achieved (4% relative improvement is due to adding the prior over the utterance i-vector). For "dev03-auto", the gender-based prior gave further improvement over the SI informative prior. This may be again because of the error-prone automatic segmentation that may benefit more by extra information on the speaker space.

| System | dev03-manual | dev03-auto |
|---|---|---|
| Baseline | 11.9 | 12.1 |
| +iv-spk | **11.6** | 14.6 |
| +iv-spk-Stdprior | 13.2 | - |
| +iv-spk-SIprior | **11.5** | - |
| +iv-spk-Genderprior | 11.6 | - |
| +iv-utter | **11.3** | **11.5** |
| +iv-utter-Stdprior | 12.6 | 13.0 |
| +iv-utter-SIprior | 11.3 | 11.5 |
| +iv-utter-Genderprior | 11.4 | 11.6 |

Table 3: Hybrid decoding results for DNNs with SA input features (WER %)

Table 3 presents the decoding results for hybrid systems using SA input features in training and decoding. Combining the SA input features with speaker-level i-vectors reduces the WER for "dev03-manual but not for "dev03-auto" ("dev03-auto" degrades for the reasons explained for Table 2). Utterance level i-vectors again perform better than speaker i-vectors and improve the baseline for both versions of the dev03 set. Thus, speaker level as well as utterance level i-vectors combine well also with the speaker-transformed feature space and reduce the WER.

It is also interesting to compare rows "+iv-utter-Stdprior" in Tables 2 and 3; the degradation caused by this prior is smaller in the case of the speaker-adapted feature space. This is because this space is less noisy and combines better with the normal Gaussian distribution of the standard prior. The informative priors without retraining the hybrid system with prior-enhanced i-vectors do not degrade nor improve the WER as in Table 2. No retraining of the hybrid system with prior-enhanced priors was performed since the configuration of speaker-adapted acoustic features combined with i-vectors performing utterance level adaptation is somehow inconsistent, making the experiments of the second block only illustrative. Finally, comparing Tables 2 and 3, it can be seen that i-vector adaptation can be a good low-latency alternative to speaker adaptive training of acoustic features for a hybrid DNN system.

## 6. Conclusions

In this paper the use of informative priors for i-vector estimation was proposed. These priors are derived from the training data and better model the behaviour of the speaker space. They are more robust than the standard normal prior for noisy short utterances and are less sensitive to the mismatch between training and test data length distributions. On the experiments on US BN data, best performance was achieved when using utterance level test i-vectors enhanced with the informative priors. These attributes may also be useful in applications such as speaker recognition where robust speaker verification on short utterances remains a key consideration. In the future we also plan to integrate this prior information to the factorised i-vectors approach presented in [3]. Another research direction to investigate is the use of other prior sources which can be easily integrated in the adopted count-smoothing framework.

# 7. References


[1] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, 2013, pp. 55–59.

[2] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014, pp. 225–229.

[3] P. Karanasou, Y. Wang, M. J. F. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc. Interspeech*, 2014, pp. 2180–2184.

[4] O. Siohan and M. Bacchiani, "iVector-based acoustic data selection," in *Proc. Interspeech*, 2013, pp. 657–661.

[5] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-Vector based speaker recognition on short utterances," in *Proc. Interspeech*, 2011, pp. 2341–2344.

[6] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," in *Technical Report CRIM-06/08-14*, 2006.

[7] ——, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey-10*, 2010.

[8] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[9] R. Travadi, M. V. Segbroeck, and S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *Proc. Interspeech*, 2014, pp. 3037–3041.

[10] M. J. F. Gales, "Transformation smoothing for speaker and environmental adaptation," in *Proc. Eurospeech*, 1997.

[11] C. Breslin, K. Chin, M. Gales, K. Knill, and H. Xu, "Prior information for rapid speaker adaptation," in *Proc. Interspeech*, 2010, pp. 1644–1647.

[12] L. Chen, M. J. F. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proc. Interspeech*, 2012, pp. 959–962.

[13] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 1999.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[15] D. Graff, "The 1996 broadcast news speech and language-model corpus," in *Proc. 1997 DARPA Speech Recognition Workshop*, 1997, pp. 11–14.

[16] D. S. Pallett, J. G. Fiscus, A. Martin, and M. A. Przybocki, "1997 broadcast news benchmark test results: English and non-english," in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 5–11.

[17] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland, "The development of the Cambridge University RT-04 diarisation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[19] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 75–98, 1998.

[20] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[21] D. Johnson, "Quicknet," *www1.icsi.berkeley.edu/Speech/qn.html*.

[22] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, 2011, pp. 24–29.