Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

SHORT-DURATION SPEAKER MODELLING WITH PHONE ADAPTIVE TRAINING

Giovanni Soldi[1], *Simon Bozonnet*[2], *Federico Alegre*[1], *Christophe Beaugeant*[3] *and Nicholas Evans*[1]

[1]Multimedia Communications Department, EURECOM, Sophia Antipolis, France
[2]OnMobile Telisma, Paris, France
[3] Intel Mobile Communications, Sophia Antipolis, France,
[1] {soldi,alegre,evans}@eurecom.fr, [2]simon.bozonnet@onmobile.com
[3]christophe.beaugeant@intel.com

ABSTRACT

This paper presents a new approach to feature-level phone normalisation which aims to improve speaker modelling in the case of short-duration training data. The new approach is referred to as phone adaptive training (PAT). Based on constrained maximum likelihood linear regression (cMLLR) and previous work in speaker adaptive training (SAT), PAT learns a set of transforms which project features into a new phone-normalised but speaker-discriminative space. Originally investigated in the context of speaker diarization, this paper presents new work to assess and optimise PAT at the level of speaker modelling and in the context of automatic speaker verification (ASV). Experiments show that PAT improves the performance of a state-of-the-art iVector ASV system by 50% relative to the baseline.

Index Terms— Speaker modelling, short-duration, phone adaptive training, automatic speaker verification

1. INTRODUCTION

Many automatic speech processing applications are required to operate in the face of varying data quantities. When data is plentiful, phonetic or nuisance variation can be implicitly normalised or marginalised and often has limited or no impact on performance. For example, the use of long-duration training and testing data in text-independent speaker verification effectively compensates for the effect of differing phone content. In contrast, when training data is scarce, then performance can degrade drastically if the phonetic variation is dissimilar to that encountered in testing; phonetic variation is no longer marginalised. Short-duration text-independent speaker verification [1–3] and speaker diarization [4, 5] are two such examples in which either speaker models can be trained on low quantities of data or well-trained models can be compared to short test segments. In both cases there is a bias towards the specific phone content [6, 7].

Drawing upon the related work in short-duration automatic speaker verification (ASV) [2, 3, 8] we have thus investigated approaches to normalise phone variation, originally in the scope of speaker diarization. The central aim is to project acoustic features into a new space in which phone discrimination is minimised while speaker discrimination is maximised. The outcome of this work is based on the original idea of speaker adaptive training (SAT) [9], a technique commonly used in automatic speech recognition (ASR) and language recognition [10, 11]. SAT projects speaker-dependent features into a speaker-neutral space in order that recognition may be performed reliably using speaker-independent models. By interchanging the role of phones and speakers, we applied the same ideas to suppress phone variation while emphasising speaker variation. The resulting algorithm, first reported in [12], is referred to as phone adaptive training (PAT). Initial experiments showed that PAT can reduce phone variation by 55% while increasing speaker discrimination by 27%. While subsequent experiments with integrated PAT and speaker diarization showed the potential, results failed to meet early expectations.

While PAT operates at the feature level and targets improved speaker modelling, its use within a speaker diarization framework makes for somewhat troublesome optimisation. Our most recent work has thus sought to assess and optimise PAT in isolation from the convolutive complexities of speaker diarization and under strictly controlled conditions. This paper analyses PAT performance when applied to short-duration, text-independent ASV using a dataset manually labelled at the phone level.

The remainder of this paper is organized as follows. Section 2 outlines previous related work. The principles and implementation of phone adaptive training are described in Section 3. Section 4 describes the experimental setup used to obtain results presented in Section 5. Our conclusions and ideas for future work are presented in Section 6.

2. PRIOR WORK

The influence of phone variation in degrading the performance of short duration speaker recognition and speaker diarization is well acknowledged [3, 6, 7, 13]. The work in [7] illustrates that, as the quantity of data used for model training is reduced, then the phone distribution tends to be more and more dissimilar. In [1], Fauve et al. analysed the impact of short-duration training utterances on two automatic speaker verification (ASV) systems: a Gaussian mixture model system with a universal background model (GMM-UBM) and a GMM supervector system based on a support vector machine (SVM) classifier. They showed that conventional intersession compensation techniques and ASV attain sub-optimal performance when confronted with short-duration training utterances. The same authors following work [8] highlighted the sensitivity of speech activity detection (SAD) and the limitations of maximum a posteriori (MAP) adaptation in the case of short-duration training. Eigenvoice modelling was shown to improve robustness by removing model components which are insufficiently adapted as a result of training data scarcity.

Other authors have investigated the impact of duration mis-match, namely differences in the data quantities used for modelling and testing. In the context of a joint factor analysis (JFA) system, Vogt et al. [2] showed that ASV performance degrades when speaker and channel sub-spaces are trained on full-length utterances, but short utterances are used for testing. This behaviour is caused by the phone-variation in short utterances which tends to dominate the effects of intersession variability. Improved ASV performance was obtained by training the channel subspace matrix on utterances of duration similar to those encountered during testing. Other work reported in [3,7,13,14] showed similar effects on iVector [15] and probabilistic linear discriminant analysis (PLDA) [16] system variants. Common to all this work is the modelling and testing using similar quantities of data, thereby marginalising to some extent the effects of phone-variation.

The work in [17–19] all investigated approaches to compensate for phone variation in the context of speaker identification (SI). That in [17] investigated the projection of features into a phone-independent subspace in order to improve text-independent SI. Based on the assumption that phone variation dominates speaker variation, the phone-independent subspace is learned using principal component analysis (PCA). Features are then projected onto the eigenvectors which correspond to the lowest eigenvalues. In [18] probabilistic principal component analysis (PPCA) is used instead to learn the phoneme-independent subspace.

Other more generalised techniques such as maximum likelihood linear regression (MLLR) and constrained maximum likelihood linear regression (cMLLR), have been used extensively to improve speaker discriminability and thus to improve ASV performance. Stolcke et al. [20] and Ferras et al. [21] both used speaker dependent MLLR and cMLLR transforms in order to model the difference between speaker independent and dependent models. The estimated transforms capture speaker dependent characteristics and are used themselves as features in order to train SVM-based verification systems. With the aid of ASR transcripts, these approaches can exploit knowledge of the phone content in order to estimate phone-neutral speaker models [20]. Stolcke's later work [22] showed that the same phone content can be used to derive more speaker-discriminative cMLLR transforms using speech-constrained phonetic regions defined by prosodic and phonetic criteria.

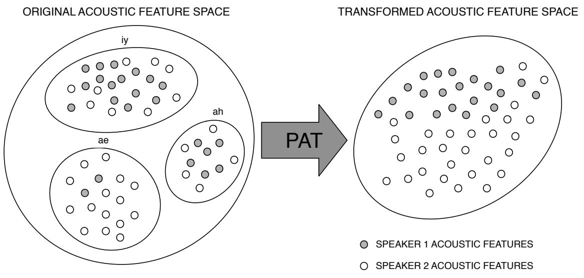

ORIGINAL ACOUSTIC FEATURE SPACE
TRANSFORMED ACOUSTIC FEATURE SPACE
iy
ah
ae
PAT
● SPEAKER 1 ACOUSTIC FEATURES
○ SPEAKER 2 ACOUSTIC FEATURES

Fig. 1: The concept of phone adaptive training (PAT) which projects input features into a new, phone-normalised space.

When training data is scarce, for instance in the case of short-duration ASV or in the case of model initialization in some approaches to speaker diarization, the learning of speaker and phone specific transforms can be impractical. Phone adaptive training (PAT), introduced by Bozonnet et al. [12], differs from the previous work in that phone-dependent cMLLR transforms are learned in a speaker-independent fashion. By using speech data collected offline from a large pool of speakers, the resulting transforms give improved performance in the case of limited speaker-specific data. As illustrated in Figure 1 PAT is used to project acoustic features into a new, phone-normalised space which is more discriminative in terms of speakers. Of particular appeal, the projected features can be used in the place of baseline features with any ASV or diarization system.

PAT was originally investigated within the context of speaker diarization. Intermediate experiments which assessed the impact of PAT in terms of speaker and phone discrimination showed the potential but gave only modest improvements in diarization performance. This paper presents an assessment of PAT for automatic speaker verification at the speaker modelling level, i.e. beyond the basic assessment of discrimination, but independently from diarization.

3. PHONE ADAPTIVE TRAINING

Starting with background MLLR and cMLLR theory, this section describes the principles and specific implementation of PAT used for all experimental work presented in this paper.

3.1. MLLR and cMLLR

Maximum likelihood linear regression (MLLR) is an affine transform approach to model adaptation. The aim is to reduce the mismatch between a model and an adaptation dataset. As detailed in [23, 24], when the model is a GMM with initial model mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then the adapted mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ are estimated according to:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \tag{1}$$
$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{B}\boldsymbol{H}\boldsymbol{B}^T \tag{2}$$

where the transform is characterized by an $n \times n$ regression matrix $\boldsymbol{A}$ (n being the dimension of the feature space), an n-dimensional bias vector $\boldsymbol{b}$ and an $n \times n$ matrix $\boldsymbol{H}$. $\boldsymbol{B}$ is the inverse of the Cholesky factor of $\boldsymbol{\Sigma}^{-1}$. Both $\boldsymbol{A}$ and $\boldsymbol{b}$ are optimized according to a standard expectation maximisation (EM) algorithm [25] to maximize the likelihood of the model with respect to the adaptation data.

In contrast to standard MLLR, which requires two different, independently optimised transforms, $(\boldsymbol{A}, \boldsymbol{b})$ and $\boldsymbol{H}$, the constrained MLLR (cMLLR) algorithm requires a single transform $\boldsymbol{W} = (\boldsymbol{A}, \boldsymbol{b})$ to adapt both mean and variance parameters [26]. Equations 1 and 2 then become:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \tag{3}$$
$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T \tag{4}$$

where the transform $\boldsymbol{A}$ and $\boldsymbol{b}$ are the constrained $n \times n$ transform matrix and the n-dimensional bias vector respectively, both still estimated in the maximum likelihood sense from the training data. Since the mean and the variance transforms are tied, in addition to model transformation, cMLLR can also be used to transform an acoustic feature $\boldsymbol{o}$ according to:

$$\hat{\boldsymbol{o}} = \boldsymbol{A}^{-1}\boldsymbol{o} - \boldsymbol{A}^{-1}\boldsymbol{b} \tag{5}$$

The application of cMLLR at the feature level is the starting point for PAT.

3.2. Phone Adaptive Training

The motivation behind PAT stems from the idea behind SAT, a technique commonly used in ASR. SAT aims to decouple speaker and phone variation and to preserve only the latter in order that ASR may be performed reliably using speaker-independent models. In order to decouple speaker and phone variation, SAT jointly estimates a canonical speaker independent acoustic phonetic model $\boldsymbol{\lambda}$ and a set of speaker transforms to capture unwanted speaker variability. In contrast, PAT aims to suppress phone variability in order to provide speaker-discriminative features for speaker modelling.

We suppose a dataset of utterances collected from S different speakers. Each utterance is composed of P different phones such that the global set of acoustic features is represented by $\boldsymbol{O}_{s,p} = (\boldsymbol{o}_{s,p,1}, \ldots, \boldsymbol{o}_{s,p,N_{s,p}})$ where $N_{s,p}$ is the number of acoustic features corresponding to each speaker $s \in S$ and each phone $p \in P$. For each phone p, PAT estimates iteratively a transformation $\tilde{\boldsymbol{W}}_p = (\tilde{\boldsymbol{A}}_p, \tilde{\boldsymbol{b}}_p)$ which captures the phone variation across speakers. Simultaneously, PAT learns a set of phone-normalised speaker models $\tilde{\boldsymbol{\Lambda}} = (\tilde{\boldsymbol{\lambda}}_1, \ldots, \tilde{\boldsymbol{\lambda}}_S)$. The algorithm is thus defined by:

$$\left(\tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{W}}\right) = \underset{\boldsymbol{\Lambda},\boldsymbol{W}}{\arg\max} \prod_{s=1}^{S}\prod_{p=1}^{P} \mathcal{L}\left(\boldsymbol{O}_{s,p}|\boldsymbol{W}_p, \boldsymbol{\lambda}_s\right) \tag{6}$$

where $\tilde{\boldsymbol{W}} = (\tilde{\boldsymbol{W}}_1, \ldots, \tilde{\boldsymbol{W}}_p)$ represents the set of phone transforms. As in Equation 5, phone-normalized features $\tilde{\boldsymbol{O}}_{s,p}$ are then obtained according to:

$$\tilde{\boldsymbol{o}}_{s,p,t} = \tilde{\boldsymbol{A}}_p^{-1}\boldsymbol{o}_{s,p,t} - \tilde{\boldsymbol{A}}_p^{-1}\tilde{\boldsymbol{b}}_p \tag{7}$$

where $t = 1, \ldots, N_{s,p}$ is the feature index. Since there is no closed-form solution, Equation 6 is optimised iteratively.

We denote by $\boldsymbol{O}^{(0)}_{s,p}$ the set of initial acoustic feature vectors for each speaker s and phone p. The initial step consists in training a set of speaker models $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}^{(0)}_1, \ldots, \boldsymbol{\lambda}^{(0)}_S)$ using the initial acoustic features vectors. Then, for each iteration i, the algorithms proceeds as follows:

1. Estimate a cMLLR transform $\boldsymbol{W}^{(i)}_p = (\boldsymbol{A}^{(i)}_p, \boldsymbol{b}^{(i)}_p)$ for each phone p such that:

$$\boldsymbol{W}^{(i)}_p = \underset{\boldsymbol{W}}{\arg\max} \prod_{s=1}^{S} \mathcal{L}\left(\boldsymbol{O}^{(i-1)}_{s,p}\middle|\boldsymbol{W}, \boldsymbol{\lambda}^{(i-1)}_s\right) \tag{8}$$

2. Apply the transform $\boldsymbol{W}^{(i)}_p$ obtained in step 1 to the set of acoustic features resulting from iteration $i-1$ to obtain a new set of phone-normalised acoustic features for each speaker s and phone p:

$$\boldsymbol{o}^{(i)}_{s,p,t} = \boldsymbol{A}^{(i)^{-1}}_p \boldsymbol{o}^{(i-1)}_{s,p,t} - \boldsymbol{A}^{(i)^{-1}}_p \boldsymbol{b}^{(i)}_p \tag{9}$$

3. Through MAP adaptation of speaker models $\boldsymbol{\Lambda}^{(i-1)}$ obtained at step $i-1$, estimate a new set of normalised speaker models $\boldsymbol{\Lambda}^{(i)} = (\boldsymbol{\lambda}^{(i)}_1, \ldots, \boldsymbol{\lambda}^{(i)}_S)$ for each speaker s, using the phone-normalised acoustic features $\boldsymbol{O}^{(i)}_{s,p}$ obtained in step 2.

4. Increase i to $i+1$ and iterate from step 1 until a maximum number of iterations is reached.

For each speaker s and phone p, the final iteration produces phone-normalised acoustic features $\tilde{\boldsymbol{O}}_{s,p}$, cMLLR phone transforms $\tilde{\boldsymbol{W}}_p$ and phone-normalised speaker models $\tilde{\boldsymbol{\Lambda}}$.

In practice, due to data limitations, it can be preferable to learn transforms $\tilde{\boldsymbol{W}}_p$ for groups of phones, often referred to as phone classes or acoustic classes, instead of individual phones. Based on linguistic analysis, suitable classes can be learned with a binary regression tree. As illustrated in Figure 2, the root node is initialized with a single acoustic class containing the full set of phones illustrated in Table 1. Each node is progressively split into smaller sub-classes for which

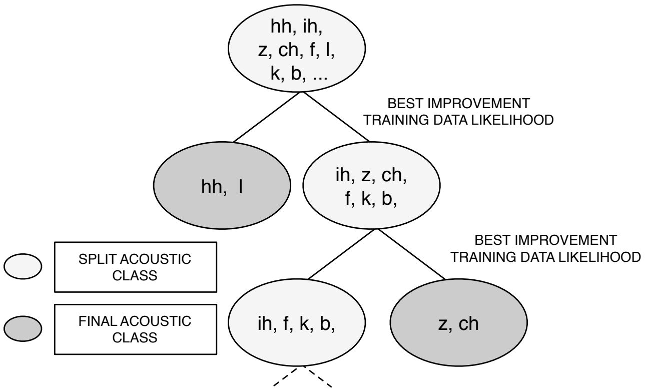


Fig. 2: An illustration of regression tree analysis used to identify suitable acoustic classes or groups of phones for PAT.

separate transforms $\tilde{\boldsymbol{W}}_p$ are determined. The split is made according to that which maximises the data likelihood in Equation 6. The pooling of data according to acoustic classes, instead of phones, allows the reliable estimation of a smaller set of transforms with less data.

PAT thus results in phone-normalised acoustic features from which more discriminant, phone-normalised speaker models can be learned. In the following we seek to assess PAT performance through a series of experiments performed on the TIMIT database [27] which is manually labelled at the phone level.

4. EXPERIMENTAL SETUP

In contrast to our previous work [12] which was performed on the NIST Rich Transcription datasets in the context of speaker diarization, the work reported in this paper was performed on the TIMIT database and in the context of ASV. The full experimental setup is described here.

4.1. Database

The TIMIT database [27] is composed of high-quality, read speech collected from a total of 630 speakers (192 female, 438 male). Each speaker contributes 10 short, phonetically-rich English language sentences whose average duration is 3 seconds. All data from a subset of 462 speakers (136 female, 326 male) is set aside for the learning of a UBM (4620 speech recordings in total) whereas that from the remaining 168 speakers (56 female, 112 male) is used for ASV experiments. One sentence per speaker is randomly selected and set aside for testing. In order to assess PAT performance in the case of varying quantities of training data, between 1 and 7 of the remaining sentences are randomly selected and used to learn speaker models.

Table 1: The setup of 38 phones used for PAT.

ENGLISH-LANGUAGE PHONES IN TIMIT ANNOTATIONS
hh, ih, z, eh, f, l, aa, b, ae, k, dh, dx, er, iy, m, n, g, r, ey, w, v, ah, y, uw, d, s, t, ng, p, sh, uh, ch, ay, ow, aw, th, jh, oy

4.2. Feature extraction and PAT

According to the ground-truth TIMIT transcriptions, all intervals of non-speech are first removed. Remaining speech is then parametrised by 12 mel-scaled frequency cepstral coefficients (MFCCs) augmented by normalized energy, delta and acceleration coefficients thereby obtaining a feature vector with a total of 39 coefficients.

We investigated PAT performance using speaker models of between 4 and 1024 GMM components. Models are derived from the UBM using conventional maximum a posteriori (MAP) adaptation. PAT transforms $\tilde{\boldsymbol{W}}_p$ for each phone $p \in P$ are then learned from a set of acoustic classes derived from the initial set of 38 phones illustrated in Table 1. A number of acoustic classes is controlled in the conventional manner with a regression tree. Independent transforms are learned for male and female speakers and for the set of utterances used to train the UBM and for ASV experiments.

The global PAT process (steps from 1 to 4) described in Section 3 was implemented with the Hidden Markov Model Toolkit (HTK) [28], in particular for creating the binary regression tree and for estimating the cMLLR transforms by solving Equation 8.

4.3. Speaker verification systems

We assessed PAT performance on two different ASV systems: a traditional GMM-UBM system and a state-of the art iVector-PLDA system. Baseline experiments were performed using the initial set of features $\boldsymbol{O}_{s,p}$ (or derived iVectors) used in PAT initialisation while ASV experiments with PAT are performed using the phone-normalised speaker features $\tilde{\boldsymbol{O}}_{s,p}$ (or derived iVectors) previously defined in section 3. For the iVector-PLDA system we estimated the total variability matrix using the same data used to estimate the UBM. Due to data limitations and since we do not aim to optimise ASV, but to observe the difference in ASV performance with PAT, the PLDA model is learned with the same development iVectors.

5. EXPERIMENTAL RESULTS

PAT performance is analyzed first, in terms of speaker and phone discrimination statistics, and second, in terms of its impact on ASV performance.

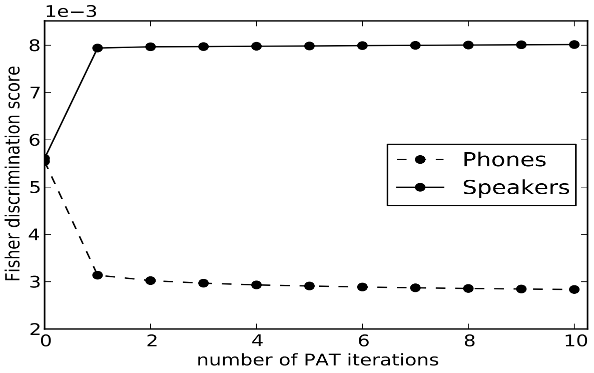


Fig. 3: Average phone and speaker discrimination for up to 10 iterations of PAT. Results shown for the 112 male speakers in the test dataset.

5.1. Speaker and phone discrimination

As reported previously in [12,29], speaker and phone discrimination can be assessed at the feature level in terms of Fisher scores. They reflect the ratio of inter and intra class variance, where classes infer the subset of features corresponding to distinct speakers or distinct phones.

Given $C_i,\ i = 1, \ldots, S$ classes (phones or speakers) and a set of N labelled features $\boldsymbol{o}_t,\ t = 1, \ldots, N$ with $T_i = \{t|\boldsymbol{o}_t \in C_i\}$, the Fisher score is defined as follows:

$$S_{Fisher} = \frac{\sum\limits_{i=1}^{S}\sum\limits_{j=1}^{S}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\sum\limits_{l=1}^{S}\sum\limits_{t\in T_l}(\boldsymbol{o}_t - \boldsymbol{\mu}_l)^T(\boldsymbol{o}_t - \boldsymbol{\mu}_l)} \tag{10}$$

where $\boldsymbol{\mu}_i$ is the mean for class C_i and $\boldsymbol{o}_t$ is the t-th feature in the subset corresponding to class C_l.

Figure 3 illustrates average phone and speaker discrimination for the 112 male speakers in the test dataset. Discrimination is plotted as a function of PAT iterations. As expected, PAT reduces the phone discrimination (dashed profile) significantly. A rapid drop occurs after a single iteration. The algorithm converges with 10 iterations, after which the phone discrimination is approximately 50% lower than without PAT. Importantly, PAT also enhances speaker discrimination (solid profile). Figure 3 shows that after 10 iterations, speaker discrimination increases by approximately 43%.

Features exhibiting lower phone discrimination but higher speaker discrimination should result in more discriminative speaker models. While improvements in ASV performance might be modest when training data is plentiful (models will be inherently phone-normalised without PAT), performance should improve in the case of limited training data. We now seek to verify this hypothesis with ASV experiments.

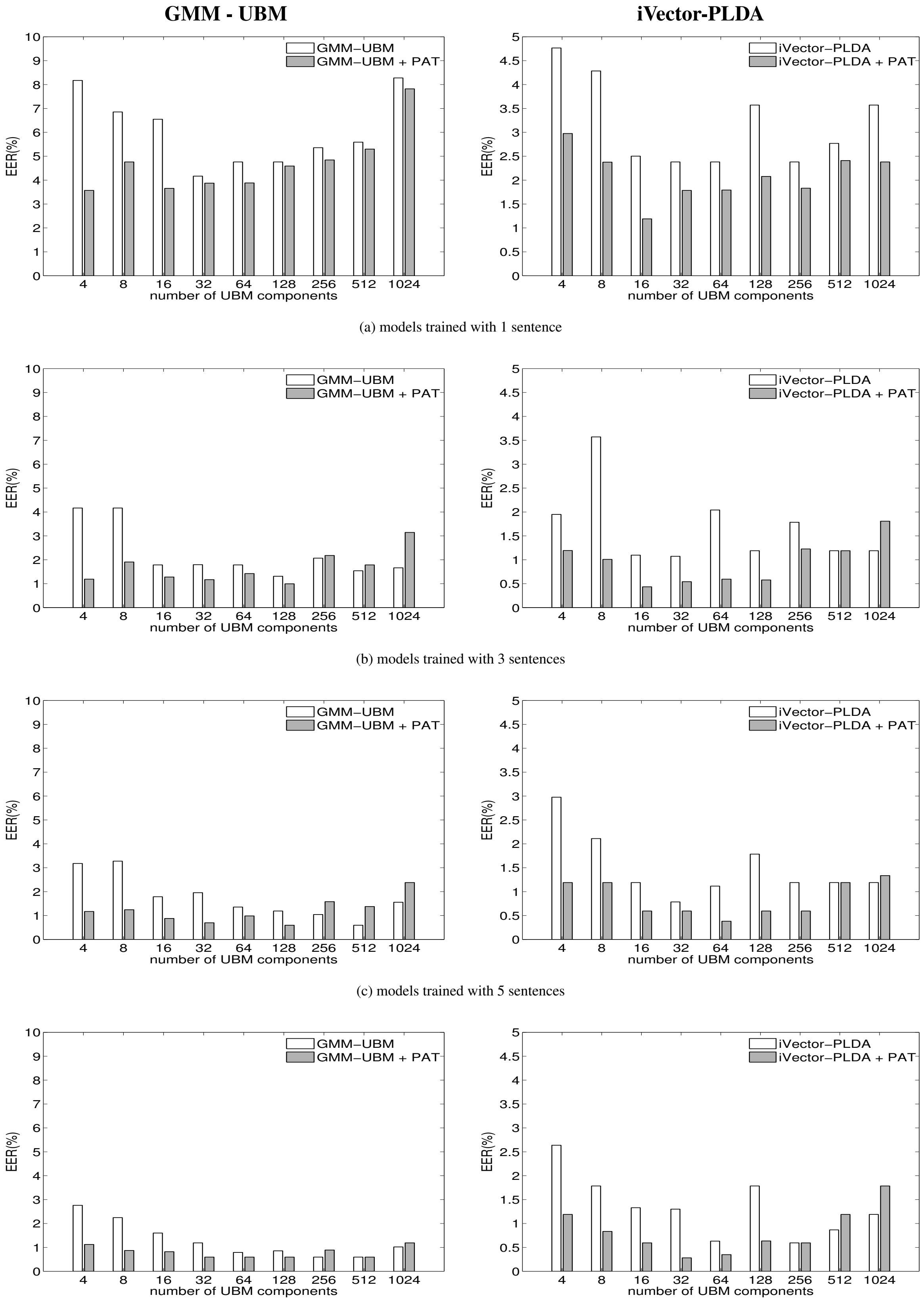
5.2. Automatic Speaker verification

Figure 4 illustrates the performance of GMM-UBM (left) and iVector-PLDA (right) systems, with and without PAT, for model sizes between 4 and 1024 components. Results indicate the equal error rate (EER) and are shown for models trained with 1, 3, 5 or 7 TIMIT sentences per speaker as described in section 4.1. In all cases, baseline performance is illustrated with clear bars. In general, as the amount of training data increases, then better performance is obtained with increasingly complex models. Noting the difference in scale between plots for each system, we also see that the iVector-PLDA system outperforms the GMM-UBM system when models are trained with relatively little data (top plots), whereas similar level of performance are achieved when larger quantities are used (bottom plots).

We now turn to the assessment of PAT performance illustrated in Figure 4 by shaded bars. In general, for smaller model sizes and for both GMM-UBM and iVector-PLDA systems, performance with PAT is better than without – shaded bars are lower than clear bars. While improvements are mostly greater in the case of low quantities of training data (top plots), modest improvements are also observed for the greatest quantities of training data (bottom plots). In some cases, for higher model sizes, PAT degrades performance. While it is difficult to explain these observations precisely, we expect this behaviour to be the result of over-fitting; with PAT, features are phone-normalised and accordingly require models of less complexity. Indeed optimal baseline performance is generally obtained with models of greater complexity than obtained by the same system with PAT.

Detection error trade-off (DET) profiles for both (a) GMM-UBM and (b) iVector-PLDA systems are illustrated in Figure 5. The two plots illustrate performance when speaker models of optimal size in each case are learned with only a single sentence (and thus corresponds to Figure 4a), with or without PAT. Baseline EERs of 4.2% and 2.4% are shown to fall to 3.6% and 1.2% with the application of PAT. PAT thus delivers significant improvements in ASV performance in the case of short-duration training.

Table 2 illustrates a summary of performance for both GMM-UBM and iVector-PLDA systems for different quantities of training data. Results correspond to optimal model sizes in each case. When speaker models are trained on a single sentence, the baseline iVector-PLDA system outperforms the baseline GMM-UBM system by 43% relative (EERs of 4.2% c.f. 2.4%). When 7 sentences are used, both systems attain the same baseline EER of 0.6%. PAT leads to better or equivalent performance in all cases. When speaker models are learned with only a single sentence, baseline EERs decrease to 3.6% and 1.2% for the GMM-UBM and iVector-PLDA systems respectively. Of particular note, the greatest improvements in ASV performance are obtained for the iVector-PLDA system where performance is improved by 50% relative, irrespective of the quantity of training data.



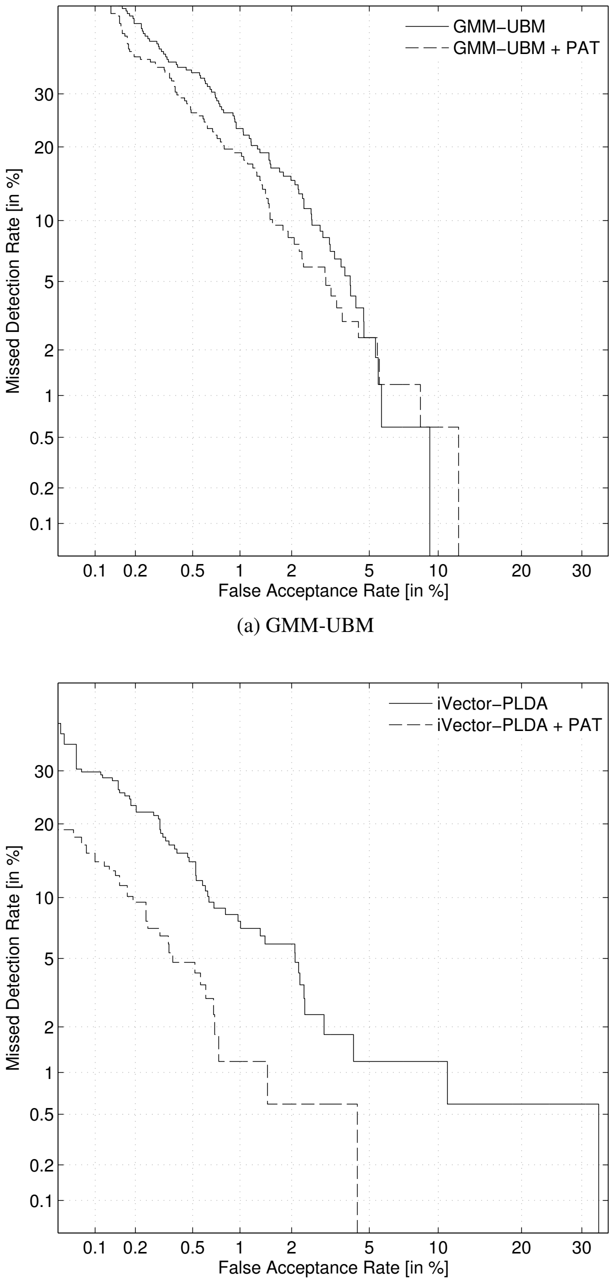
(a) models trained with 1 sentence

(b) models trained with 3 sentences

(c) models trained with 5 sentences

(d) models trained with 7 sentences

Fig. 4: An illustration of ASV performance for different model complexities (4-1024) and for varying quantities of training data (1-7 TIMIT sentences). Plots show the EER for GMM-UBM (left) and iVector-PLDA (right) systems with (shaded bars) and without 5 iterations of PAT (clear bars).



(a) GMM-UBM

(b) iVector-PLDA

Fig. 5: Detection error trade-off (DET) plots for GMM-UBM and iVector-PLDA systems with and without 5 iterations of PAT and for models trained with a single TIMIT sentence.

Number of sentences for speaker model training	Baseline (EER %)	Baseline + PAT (EER %)
1	4.2	3.6
3	1.8	1.0
5	0.6	0.6
7	0.6	0.6

(a) GMM-UBM

Number of sentences for speaker model training	Baseline (EER %)	Baseline + PAT (EER %)
1	2.4	1.2
3	1.1	0.4
5	1.1	0.4
7	0.6	0.3

(b) iVector-PLDA

Table 2: An illustration of EERs for the GMM-UBM and the iVector-PLDA systems with varying quantities of training data. Results shown for optimal model sizes in each case.

6. CONCLUSIONS AND FUTURE WORK

This paper addresses the problem of speaker modelling in the case where training data is scarce. It reports new experiments to assess and optimise a new approach to phone normalisation referred to as Phone Adaptive Training (PAT). PAT is based on the application of constrained maximum likelihood linear regression (cMLLR) to reduce phone influence at the feature level, while simultaneously emphasising speaker discrimination. As such, PAT has utility in any application involving speaker modelling, for example speaker diarization or speaker recognition.

In contrast to previous work which used PAT for speaker diarization, this paper presents our first work to optimise and evaluate PAT at the speaker modelling level, with small-scale automatic speaker verification (ASV) experiments performed on the TIMIT database. We show that PAT is successful in reducing phone bias and that it improves significantly the performance of both traditional GMM-UBM and iVector-PLDA ASV systems in the case of short-duration training. Also of appeal, PAT can typically achieve better performance with less complex models, requiring only the application of a straightforward feature-level linear transform prior to verification.

Our future work will continue the exploration of automatic approaches to acoustic-class transcription; PAT does not necessarily require phone-level transcriptions. Also, while PAT has particular appeal for short-duration ASV, our original goal involves speaker diarization. Other future work will therefore explore the use of speaker level transcriptions derived automatically through diarization and alternative, speaker-independent approaches to phone normalisation. Finally, with an interest in embedded, mobile device applications, we are also investigating the potential for implementing PAT in real-time.

7. REFERENCES

[1] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre, and J. Mason, "Influence of task duration in text-independent speaker verification.," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*. 2007, pp. 794–797, ISCA.

[2] R. J. Vogt, B. J. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2008, pp. 853–856.

[3] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "i-vector based speaker recognition on short utterances," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2011, pp. 2341–2344.

[4] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[5] N. Evans, S. Bozonnet, Dong Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 382–392, 2012.

[6] S. Bozonnet, Dong Wang, N. Evans, and R. Troncy, "Linguistic influences on bottom-up and top-down clustering for speaker diarization," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2011, pp. 4424–4427.

[7] T. Hasan, R. Saeidi, J. H. L. Hansen, and D.A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2013, pp. 7663–7667.

[8] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM- and GMM-based speaker verification," in *Proc. Odyssey*, 2008.

[9] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1996, vol. 2, pp. 1137–1140.

[10] W. Shen and D. A. Reynolds, "Improving phonotactic language recognition with acoustic adaptation," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2007, pp. 358–361.

[11] L. Cheung-Chi, M. Bin, and L. Haizhou, "Parallel acoustic model adaptation for improving phonotactic language recognition," in *Proc. Odyssey*, 2010, p. 41.

[12] S. Bozonnet, R. Vipperla, and N. Evans, "Phone adaptive training for speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2012.

[13] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2013, pp. 7649–7653.

[14] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification.," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2012.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[16] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE Int. Conf. Computer Vision (ICCV)*, 2007, pp. 1–8.

[17] Haoze Lu, H. Okamoto, M. Nishida, Y. Horiuchi, and S. Kuroiwa, "Text-independent speaker identification based on feature transformation to phoneme-independent subspace," in *Communication Technology, 2008. ICCT 2008. 11th IEEE International Conference on*, 2008, pp. 692–695.

[18] X.-C. Lu, J.-X. Yin, and W.-P. Hu, "A text-independent speaker recognition system based on probabilistic principle component analysis," in *Int. Conf. on System Science, Engineering Design and Manufacturing Informatization (ICSEM)*, 2012, vol. 1, pp. 255–260.

[19] J. Wang, A. Ji, and M. T. Johnson, "Features for phoneme independent speaker identification," in *Int. Conf. Audio, Language, Image Processing (ICALIP)*, 2012, pp. 1141–1145.

[20] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 2425–2428.

[21] M. Ferras, C.-C. Leung, C. Barras, and J. Gauvain, "Constrained MLLR for speaker recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2007, vol. 4.

[22] A. Stolcke, A. Mandal, and E. Shriberg, "Speaker recognition with region-constrained MLLR transforms," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2012, pp. 4397–4400.

[23] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[24] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[26] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[28] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, 2006.

[29] S. Bozonnet, *New insights into hierarchical clustering and linguistic normalization for speaker diarization*, Ph.D. thesis, EURECOM, 2012.