PHONE ADAPTIVE TRAINING FOR SHORT-DURATION SPEAKER VERIFICATION

Giovanni Soldi[1], *Simon Bozonnet*[2], *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,evans}@eurecom.fr, [2]simon.bozonnet@onmobile.com
[3]christophe.beaugeant@intel.com

ABSTRACT

Phone adaptive training (PAT) aims to derive a new acoustic feature space in which the influence of phone variation is minimised while that of speaker variation is maximised. Originally proposed in the context of speaker diarization, our most recent work showed the utility of PAT in short-duration, automatic speaker verification where phone variation typically degrades performance. New to this contribution is the assessment of PAT utilising automatically generated acoustic class transcriptions whose number is controlled by regression tree analysis. Experimental results using a standard database show that PAT delivers significant improvements in the performance of a state-of-the-art iVector speaker verification system.

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

1. INTRODUCTION

Many automatic speech processing applications involve the learning or training of models using variable quantities of speech data. When data is plentiful, unwanted or nuisance variation can be normalised or marginalised and thus it does not necessarily impact on performance. Examples include text-independent speaker verification where the use of long-duration training and testing data effectively neutralises the effect of differing phone content.

In contrast, when training data is scarce, then performance can degrade significantly in the face of nuisance variation which is not otherwise marginalised. Speaker diarization [1, 2] and short-duration text-independent speaker verification [3, 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].

A number of approaches to attenuate phone bias in speaker modelling have been proposed. Stolcke et al. [8, 9] and Ferras et al. [10] both investigated approaches to increase speaker discrimination in SVM-based speaker verification systems through the learning of phone-neutral speaker models. Both approaches derive speaker-dependent transforms using constrained maximum likelihood linear regression (cMLLR) [11, 12] and phone transcripts derived through automatic speech recognition (ASR) [8]. The two approaches concentrate on concatenating the parameters of estimated cMLLR transforms into high dimensional feature vectors which are then used to train speaker discriminant SVM-based ASV systems.

Our own approach to attenuate phone bias operates entirely at the feature level. Based on speaker adaptive training (SAT) [13], phone adaptive training (PAT) [14] estimates a set of phone-specific transforms which are used to project acoustic features into a new feature space in which phone discrimination is minimised while speaker discrimination is maximised. Our recent work [15] shows that PAT is successful in marginalising phone variation in an automatic speaker verification (ASV) framework, always under strictly controlled conditions, including the use of manually derived phone transcripts. This paper reports our continued work to develop PAT into a fully unsupervised system. New contributions include an approach to automatic acoustic class transcription using regression tree analysis. We assess the performance of PAT as a function of model complexity and for varying quantities of training data. Additional new experiments show that PAT performs well even when the number of acoustic classes is reduced well below the number of phones.

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

2. PHONE ADAPTIVE TRAINING

This section describes the principles and specific implementation of PAT used for all experimental work presented in this paper. The motivation behind PAT stems from the idea behind speaker adaptive training (SAT) [13], a technique com-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

monly used in speaker-independent automatic speech recognition (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 contrast, PAT aims to suppress phone variability and to maximise speaker-discrimination for more reliable 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 $\boldsymbol{O}_{s,p} = (\boldsymbol{o}_{s,p,1}, \ldots, \boldsymbol{o}_{s,p,N_{s,p}})$ for all speakers $s \in S$ and phones $p \in P$. For each phone p, PAT estimates iteratively a constrained maximum likelihood linear regression (cMLLR) transformation $\tilde{\boldsymbol{W}}_p = (\tilde{\boldsymbol{A}}_p, \tilde{\boldsymbol{b}}_p)$ which captures the phone variation across speakers with $\tilde{\boldsymbol{A}}_p$ being a $n \times n$ regression matrix (n being the dimension of the feature space), and $\tilde{\boldsymbol{b}}_p$ an n-dimensional bias vector. 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:

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

where $\tilde{\boldsymbol{W}} = (\tilde{\boldsymbol{W}}_1, \ldots, \tilde{\boldsymbol{W}}_p)$ represents the set of phone transforms. The main advantage of using CMLLR transforms is that phone-normalized features $\tilde{\boldsymbol{O}}_{s,p}$ can then be obtained according to:

$$\tilde{\boldsymbol{o}}_{s,p,t} = \tilde{\boldsymbol{A}}_p^{-1} \boldsymbol{o}_{s,p,t} + \tilde{\boldsymbol{A}}_p^{-1} \tilde{\boldsymbol{b}}_p \qquad (2)$$

where $t = 1, \ldots, N_{s,p}$ is the feature index. Since there is no closed-form solution, Equation 1 is optimised iteratively [14].

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. The root node of the tree is initialised 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 separate transforms $\tilde{\boldsymbol{W}}_p$ are determined. The split is made according to that which maximises the data likelihood in Equation 1. 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 [16].

Table 1: The 39 phones used for generating acoustic class transcriptions and for PAT.

39 ENGLISH-LANGUAGE PHONES
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, sil

3. EXPERIMENTAL SETUP

In line with our previous work [15], that reported in this paper was performed on the TIMIT database and in the context of automatic speaker verification (ASV).

3.1. Database

The TIMIT database [16] 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. In the same way as reported in [17], we reduced the 61 phonetic labels in the TIMIT transcriptions to the 39 phones illustrated in Table 1. All data from a subset of 462 speakers (136 female, 326 male) is set aside for the learning of a UBM and acoustic class models used for acoustic class transcription (4620 speech recordings in total). 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.

3.2. Feature extraction and acoustic class transcription

Non-speech segments are removed from all TIMIT sentences according to the ground-truth transcriptions. Remaining speech segments are then parametrised with 12 mel-scaled frequency cepstral coefficients (MFCCs) augmented by normalised energy, delta and acceleration coefficients, thereby obtaining a feature vector with a total of 39 coefficients.

Using the pool of acoustic features extracted from the UBM and acoustic class training dataset, and by varying the likelihood threshold, the 38 phones in Table 1 (without silence) are reduced to between 5 and 38 acoustic classes through automatic regression tree analysis. For each number of acoustic classes, the phone labels in the phonetic transcriptions are replaced by their corresponding acoustic class labels.

The acoustic class models are 3-state hidden Markov models (HMMs) where each state is characterised by a Gaussian mixture model (GMM). Each acoustic class model is first initialized with a single Gaussian component whose

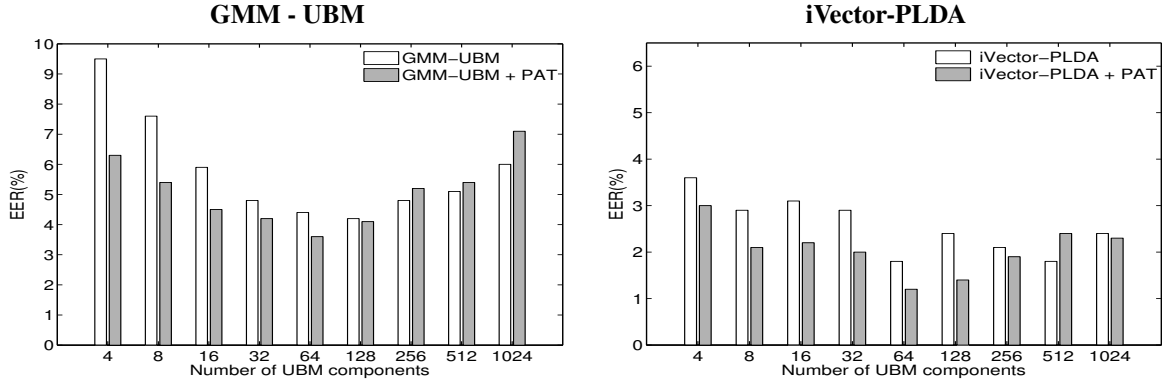

Fig. 1: An illustration of ASV performance for GMM-UBM and iVector-PLDA ASV systems with 21 and 25 acoustic classes respectively and for training data of 1 TIMIT sentence. Plots show the best obtained EER for baseline systems (clear bars) and the same systems with 5 iterations of PAT (shaded bars).

mean and variance are set to that of the global class data. Subsequently, six iterations of embedded training are performed. The number of Gaussian components is doubled and embedded training is performed again on the new, larger model. This procedure is repeated until the number of Gaussian components reaches 128. The dataset used for ASV experiments is transcribed automatically using the given set of acoustic classes and corresponding models. Both training and decoding phases were implemented with the Hidden Markov Model Toolkit (HTK) [18].

3.3. Phone adaptive training and speaker verification

We investigated PAT performance using two different ASV systems: a traditional GMM-UBM system and a state-of the art iVector-PLDA system. Speaker models with between 4 and 1024 Gaussian components are derived from the UBM using conventional maximum a posteriori (MAP) adaptation. The features used to train the UBM are treated with PAT which is applied using the TIMIT ground-truth transcriptions. All remaining data used for ASV experiments (model training and testing) is instead treated with PAT applied using automatically generated acoustic class transcriptions. In both cases Equation 1 is applied with 5 iterations. As for acoustic class segmentation, PAT was also implemented with the Hidden Markov Model Toolkit (HTK) [18].

Baseline ASV experiments were performed using the initial set of features $\boldsymbol{O}_{s,p}$ (or derived iVectors) while PAT performance was assessed using different numbers of acoustic classes and corresponding normalised features. 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 only to observe the difference in ASV performance with PAT, the PLDA model is learned with the same development iVectors.

4. EXPERIMENTAL RESULTS

Figure 1 illustrates the performance of GMM-UBM (left) and iVector-PLDA (right) systems, with and without PAT, for model sizes between 4 and 1024 components and using 21 and 25 acoustic classes respectively. Results indicate the equal error rate (EER) when speaker models are trained with only a single TIMIT sentence. In all cases, baseline performance is illustrated with clear bars whereas that with 5 iterations of PAT is illustrated with shaded bars. Noting the difference in scale between plots for each system, we see that the iVector-PLDA system largely outperforms the GMM-UBM system (EERs in the order of 1 to 3.5% cf. 3.5 to 9.5%). The performance envelope for the GMM-UBM systems are convex with minima at 128 components for the baseline and 64 components with PAT. The iVector-PLDA profiles are somewhat noisy, mostly likely due to the lack of sufficient data to train the total variability matrix. Optimal performance with PAT is again achieved for a model with 64 components.

Figure 2 illustrates PAT performance for the GMM-UBM system (clear bars) and the iVector-PLDA systems (shaded bars) with different numbers of acoustic classes. The complexity of both systems is fixed to 64 components. While the profile envelopes are somewhat non-convex, most likely again due to lack of training data, the application of PAT results in better performance than the respective baselines (solid and dashed horizontal lines). These observations indicate that PAT is beneficial even without reliable phone transcriptions. With 15 and 25 acoustic classes respectively the relative improvement in performance is 18% for the GMM-UBM system and 33% for the iVector-PLDA system.

Detection error trade-off (DET) profiles for the iVector-PLDA system using 25 acoustic classes is illustrated in Figure 3. The profiles illustrate performance for speaker models of size 64 trained using only a single TIMIT sentence, with

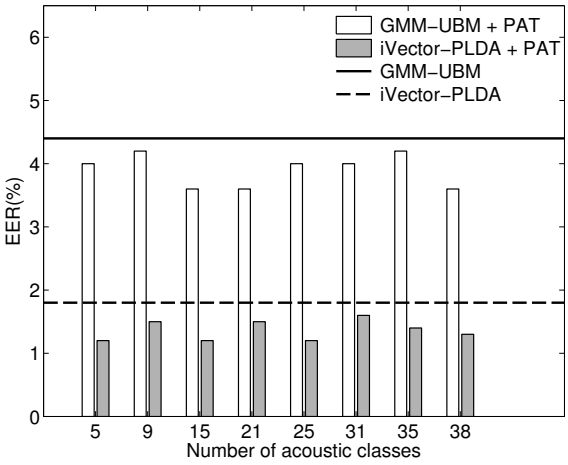

Fig. 2: An illustration of ASV performance for GMM-UBM and iVector-PLDA systems with 5 iterations of PAT for different numbers of acoustic classes, all for training data of 1 TIMIT sentence and for 64 UBM components. The baseline performance for GMM-UBM and iVector-PLDA systems are represented respectively by the solid and dashed horizontal lines.

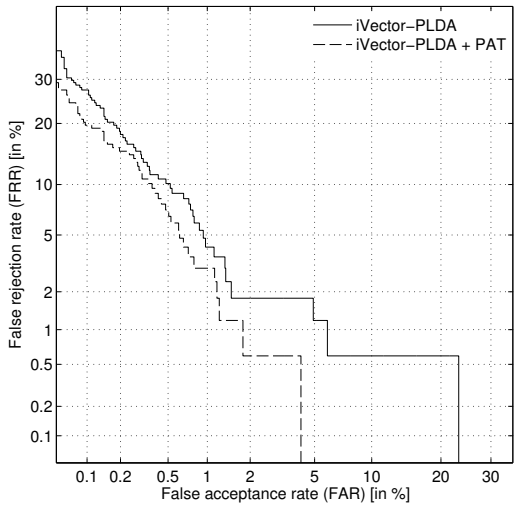

Fig. 3: Detection error trade-off (DET) plots for iVector-PLDA systems with and without 5 iterations of PAT using 25 acoustic classes and for speaker models trained with a single TIMIT sentence.

and without PAT. The baseline EER of 1.8% is shown to fall to 1.2% with the application of PAT, i.e. the same relative improvement of 33%.

Table 2 illustrates a summary of performance for the iVector-PLDA systems for optimal model sizes and for different quantities of training data, namely 1 to 7 TIMIT sentences. Baseline results are illustrated in the second column whereas those for PAT with 25 acoustic classes are illustrated in the third column. We see that, as the quantity of training data increases, then the difference between baseline and PAT performance decreases. This is to be expected since larger quantities of training data will inherently reduce the phone bias and have the same normalising effect as PAT. PAT thus delivers the most significant improvements in ASV performance in the case of short-duration training where the phone bias is otherwise the most pronounced.

Number of sentences for speaker model training	Baseline (EER %)	Baseline + PAT (EER %)
1	1.8	1.2
3	1.2	0.7
5	0.6	0.5
7	0.7	0.5

Table 2: EERs for the iVector-PLDA system with and without 5 iterations of PAT with 25 acoustic classes. Performance is illustrated for varying quantities of training data and for optimal model sizes.

5. CONCLUSIONS AND FUTURE WORK

Based on the application of constrained maximum likelihood linear regression (cMLLR), phone adaptive training (PAT) aims to reduce the influence of phone variation at the feature level, while simultaneously emphasising speaker discrimination. This paper reports our most recent work to assess an automatic approach to acoustic class transcription using regression tree analysis. Results using two different approaches to automatic speaker verification, one at the state of the art, show that PAT improves on baseline performance for all experiments with different numbers of acoustic classes and model complexities. Of particular note, the number of acoustic classes can be reduced significantly meaning that PAT is effective even without reliable phone transcriptions. This may ease the application of PAT to more realistic, noisy data where phone transcription can be troublesome. Finally, and as could have been expected, the improvement in performance tends to reduce as the amount of training data increases, meaning that PAT is most beneficial when training data is scarce.

Our future work will take two directions. First, given the effective performance of PAT with relatively few acoustic classes and the ability to obtain the same performance with less complex models, we will investigate the utility of PAT in low-resource, embedded mobile applications. Second, we intend to return to the original focus of this work in marginalising phone bias in order to improve performance in speaker diarization.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] G. Soldi, S. Bozonnet, Alegre F., C. Beaugeant, and N. Evans, "Short-duration speaker modelling with phone adaptive training," *Odyssey*, 2014.

[16] 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.

[17] S. Fernndez, A. Graves, and J. Schmidhuber, "Phoneme recognition in timit with blstm-ctc," 2008.

[18] 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.