Phone Adaptive Training for Speaker Diarization

Simon Bozonnet, Ravichander Vipperla and Nicholas Evans

EURECOM
BP193, F-06904 Sophia Antipolis Cedex, France
{bozonnet,vipperla,evans}@eurecom.fr

Abstract

The linguistic content of a speech signal is a source of unwanted variation which can degrade speaker diarization performance. This paper presents our latest work to reduce its impact. The new approach, referred to as Phone Adaptive Training (PAT), is analogous to speaker adaptive training used in automatic speech recognition. We report an oracle experiment which shows that PAT has the potential to deliver a 33% relative improvement in the diarization error rate of our baseline system. Practical experiments show significant improvements across two standard, independent evaluation datasets.

Index Terms: Speaker Diarization, Phone Adaptive Training, Speaker Discrimination

1. Introduction

Speaker diarization refers to the 'Who Spoke When?' task and involves the detection of speaker turns within an audio document (segmentation) and the grouping together of all same-speaker segments (clustering). While the state-of-the-art has advanced over recent years, performance can still vary significantly from one audio show to another.

In [1, 2] we show that the linguistic content of a speech signal can be a significant source of unwanted variation. This and other sources of variation can cause a diarization system to converge towards artifacts not related to different speakers and therefore a non-optimal speaker inventory. Errors in the speaker inventory can degrade diarization performance when they relate to speakers with significant floor time.

A small number of approaches have been proposed to take account of linguistic information in speaker diarization. Chen *et al.* [3] propose the modelling of speakers with a phone subspace mixture in order to take account of linguistic variation in the ΔBIC distance measure. Žibert *et al.* [4] present a speech activity detection (SAD) component which uses the linguistic information in the output of an automatic speech recognition (ASR) system to improve performance. These approaches use lexical information only with a single system component (e.g. for cluster fusion, or SAD) whereas typical speaker diarization systems involve several sequential stages.

We have investigated a new approach to linguistic normalisation which reduces the influence of linguistic variation in every diarization processing stage. The new approach, referred to as Phone Adaptive Training (PAT), utilises the output of a speech transcription system to suppress linguistic variation at the feature level while retaining variation related to different speakers. PAT leads to a more speaker-discriminative feature space, and hence better diarization performance.

The remainder of this paper is organised as follows. Section 2 introduces the new PAT approach. Section 3 presents oracle-based experiments which aim to demonstrate the potential of PAT. Section 4 reports real speaker diarization experiments which show improvements in performance when PAT is applied to our baseline speaker diarization system. Our conclusions are presented in Section 5.

2. Phone Adaptive Training

Phone adaptive training (PAT) is based on the analogous idea of speaker adaptive training (SAT) [5], a model estimation framework used in automatic speech recognition (ASR). SAT jointly estimates speaker-dependent transforms and speaker-independent acoustic models. The transforms capture unwanted speaker variability while only the desirable phonetic variation is captured in the acoustic models. PAT is used instead to suppress phonetic variability in order to provide a more speaker-discriminant feature space for the task of speaker diarization.

Consider a training database transcribed at the phone level either manually or through automatic speech recognition. Let the database contain data from a set of R speakers, let the exhaustive phone set be represented by P and let the set of observations for each phone and each speaker be represented by $O^{(r,p)}$, $r \in R$ and $p \in P$. PAT aims to jointly estimate a set of speaker models $\Lambda_{PAT} = (\lambda_{PAT}^{(1)}, \ldots, \lambda_{PAT}^{(R)})$ and a set of phone specific transforms $W = (W^{(1)}, \ldots, W^{(P)})$ in order to maximise the likelihood $\mathcal{L}$ with respect to the training data according to:

$$(\bar{\Lambda}_{PAT}, \bar{W}) = \underset{\Lambda_{PAT}, W}{\operatorname{argmax}} \prod_{r=1}^{R} \prod_{p=1}^{P} \mathcal{L}(O^{(r,p)} | W^{(p)} \lambda_{PAT}^{(r)}) \quad (1)$$

The speaker models ($\lambda^{(r)}$) are standard continuous density Gausssian mixture models. The phone specific information is modeled by constrained maximum likelihood linear regression (CMLLR) matrices [6] $W^{(p)} = [A^{(p)}\ b^{(p)}]$, where $A^{(p)}$ and $b^{(p)}$ represent the transformation matrix and the bias for the phone p respectively. One of the advantages of the CMLLR formulation is that it can be viewed as normalisation in feature space according to:

$$\hat{o}^{(r,p)} = {A^{(p)}}^{-1} o^{(r,p)} + {A^{(p)}}^{-1} b_c \quad (2)$$

The optimization problem in Equation 1 is non-convex and can only be solved using an iterative update approach. Feature space normalisation using CMLLR provides a simplified framework to marginalise undesired phonetic variation in the feature space. The iterative estimation procedure is outlined below:

1. Train a phone independent acoustic model ($\lambda^{(r)}$) for each speaker in the training set.
2. Using speaker models obtained in 1, estimate a set of phone specific CMLLR adaptation transforms (W) to maximise the likelihood of the training set according to:

$$W = \underset{W}{\operatorname{argmax}} \prod_{r=1}^{R} \prod_{p=1}^{P} \mathcal{L}(O^{(r,p)} | W^{(p)} \lambda_{PAT}^{(r)})$$

3. Normalise all the feature vectors for a phone in the training set with the corresponding phone transform estimated in 2.
4. Retrain the acoustic model for each speaker from the phone normalised feature vectors from 3.
5. Repeat steps 1 to 4 until the likelihood scores converge.

The resulting phone transforms $\bar{W}$ capture the common linguistic component for each phone across all speakers whereas $\bar{\Lambda}_{PAT}$ are the phone normalised speaker models. While it is desirable to estimate a separate transform for each phone, the amount of training data is often insufficient. In such scenarios, the above framework allows clustering of phones to more generic classes and class specific transforms can then be estimated in the place of phone specific transforms. Binary regression trees based on linguistic analysis can be used in such clustering to balance the need for sufficient data per class for training accurate CMLLR transforms and the need for enough acoustic classes so that phonetic variations are well modeled.

3. Oracle Experiments

PAT requires a speech and speaker transcription. In this section we report oracle experiments which use ground-truth transcriptions[1] to assess the potential of PAT under ideal conditions. The experimental setup is described in Section 3.1. Results in terms of a speaker discrimination metric and diarization performance are presented in Sections 3.2 and 3.3 respectively.

3.1. Experimental setup

Oracle experiments to investigate speaker and phone discrimination were performed on a development dataset containing 9 shows from the NIST RT'05 and RT'06 evaluation datasets. Evaluation work with a full diarization system was performed on the full RT'07 and RT'09 datasets. In all cases the signal is first characterized by 20 un-normalised linear frequency cepstral coefficients (LFCCs) plus energy coefficients computed every 10ms using a 20ms window. PAT is then applied directly to each show as described in Section 2 where the speaker models of step (1) are 16-component GMMs which are MAP adapted from a universal background model (UBM) according to the ground-truth segmentation for each speaker. The global process (steps 1 to 5) is repeated 20 times. Due to the limited quantity of data in each show for each speaker and each phone, a regression tree is applied to control the number of acoustic classes.

3.2. Speaker and phone discrimination

In order to decouple the performance of a speaker diarization system and that of PAT, speaker and phone discrimination assessments were conducted independently from full speaker diarization experiments. Discrimination is measured using the ratio of inter and intra class variance, where classes are either speakers or phones. This is the Fisher score defined as follows:

$$score_{Fisher} = \frac{\sum_{i=1}^{R} \sum_{j=1}^{R} (\mu_i - \mu_j)(\mu_i - \mu_j)^T}{\sum_{i=1}^{R} \sum_{\forall o_k \in O^{(r=i)}} (o_k - \mu_i)^2} \quad (3)$$

where $O^{(r=i)}$ represents the ensemble of observations attributed to speaker i, o is a single sample feature, μ is its mean value for speaker i, or j.

Initial experiments showed that in the order of 25 acoustic classes are required for optimal results. Speaker

[1] In practice the speech ground-truth transcription is obtained by a forced alignment of the phone transcription for each utterance in the ground-truth references.

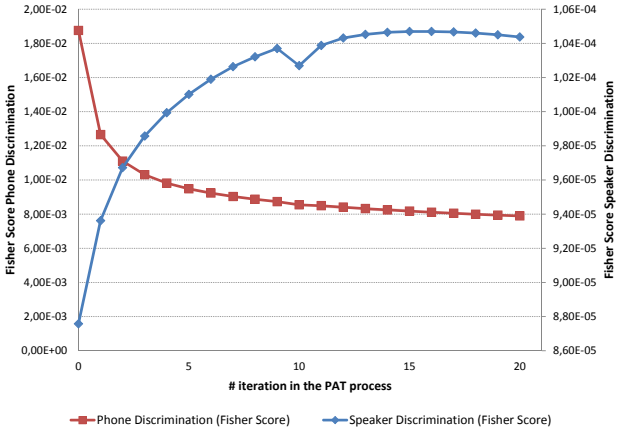

Figure 1: An illustration of speaker and phone discrimination as a function of the number of iterations of phone adaptive training (PAT). Phone discrimination illustrated in red/squares, speaker discrimination illustrated in blue/diamonds.

and phone discrimination results for an average of 25 acoustic classes[2] are illustrated in Figure 1 as a function of the number of iterations. The red/square profile in Figure 1 illustrates phone discrimination. There is an immediate, rapid drop in discrimination stemming from the use of acoustic classes which group some phones together. Phone discrimination then drops further as a direct consequence of PAT and converges after approximately 10 iterations. The blue/diamond curve in Figure 1 illustrates speaker discrimination which increases significantly over the first 10 iterations. Together these results show that PAT successfully marginalises phone variation while emphasising speaker discrimination.

3.3. Effect on Diarization Performance

We now investigate the effect of phone normalization on diarization performance. This work was conducted with our baseline top-down speaker diarization system described in [7]. Two experiments were conducted, first with conventional LFCC features and second with features normalised with PAT.

The diarization system involves the use of a UBM in the MAP adaptation of speaker models. In our previous work [7], the UBM used in the baseline experiments was trained on telephony data from the NIST speaker recognition evaluations (SREs). For PAT experiments, however, the UBM needs to be trained on phone-normalised features. This requires the transcription of all UBM data and thus, for consistency, all these experiments (including baseline system and oracle experiments) we used a new UBM trained on the NIST RT'04 dataset of 14 shows for which transcriptions are available. This dataset is fully independent of our diarization development set and both

[2]The exact number is controlled by the regression tree and varies according to speaker and phone specific statistics in each show.

	Baseline	Oracle	Practical
Dev. Set	23.9	16.1	19.0
RT07	17.1	12.9	15.9
RT09	22.6	20.2	21.5

Table 1: Speaker diarization results (DER) for baseline, oracle and practical setups for the development set and the NIST RT'07 and RT'09 evaluation datasets. All results are for SDM conditions, without the scoring of overlapping speech.

evaluation sets. Feature space is then the only difference between the baseline and PAT setups.

Table 1 presents diarization performance in terms of DER for the development dataset and the two separate evaluation datasets. The second column presents the respective baseline performance obtained with conventional LFCC features. The third column of Table 1 shows performance in terms of DER where all features used in PAT come from transforms learned using the oracle setup. On the development set, a drop in the baseline DER of 23.9% to 16.1% corresponds to a relative improvement of 33%. While a similar improvement is observed on the RT'07 dataset (25% relative improvement), a relative improvement of only 10% is achieved with the RT'09 dataset. The lower performance on the RT'09 dataset can be explained by the high degree of overlapping speech which brings some artifacts in the captured phonetic components, and the increased number of speakers which leaves less training data for each speaker. These results nonetheless show that improvements in speaker discrimination translate to improved speaker diarization performance. In the next section we present the evaluation in DER with a more practical setup, i.e. where ground-truth speaker segmentations are replaced with real diarization outputs.

4. Practical Experiments

Oracle experiments presented above confirm the potential of PAT to improve speaker discrimination and speaker diarization performance. Those experiments, however, consider the ground-truth speaker transcription to be known, while this is the final objective of the diarization task. In this section we evaluate PAT performance when the ground-truth speaker transcription is replaced with that from a practical speaker diarization system. We stress that these experiments still use reference phonetic transcriptions.

4.1. Experimental setup

For all experiments reported here PAT was performed with speaker segmentations produced automatically using a segmental EM algorithm [8]. It is initialized with 30 clusters which are aligned to the data through a set of

training/realignment iterations.

While PAT is relatively insensitive to under-clustering, it is adversely affected by over-clustering since speaker variability is then treated in the same way as phone variability and is suppressed. Indeed, in the case where several clusters represent the same speaker, PAT training is not directly affected except, eventually, by a smaller quantity of data being available for each cluster[3].

It is therefore necessary to protect speaker discrimination by preventing over-clustering and in the case of the diarization system in [8]. This is achieved simply by deactivating the clustering component. We note that this modification does not limit the application of PAT to specific speaker diarization systems; it can be applied readily to any system so long as the initial speaker segmentation used for PAT is adapted to avoid over-clustering.

The final speaker diarization stage uses PAT features in exactly the way described in Section 3 with the system in [7]. The only difference between the oracle and more practical systems is thus the use of either ground-truth or automatically derived speaker transcription used in PAT.

4.2. Speaker diarization performance

Results are illustrated in the last column of Table 1. For the development set, the baseline DER of 23.9% falls to 19.0% after the application of PAT. This corresponds to a relative improvement of 21% and is not too far from that of the optimal oracle system (33%). Improvements in DER for the RT'07 and RT'09 datasets are less significant (7% and 5% over baseline performances respectively) but do show consistent behavior. Comparisons with the performance of the optimal oracle system in each case show that there is still some potential to further improve performance in all cases.

We note that the only difference between each of the three experiments reported in Table 1 involves the use of different features; the final speaker diarization system used to generate all experimental results (not that used for PAT) is exactly identical. Further system optimisation is likely to better exploit the more speaker-discriminant features produced through PAT.

5. Conclusions

This paper introduces a new phone adaptive training (PAT) approach which aims to suppress phonetic variation in speaker diarization feature space. Experiments show that PAT leads to a new, phone-normalized feature space which is more speaker-discriminative. Oracle speaker diarization experiments show potential for significant improvements in diarization performance.

[3]In the case of an empirically derived speaker segmentation, rather than one obtained directly from ground-truth references, we refer to 'clusters' rather than 'speakers' since clusters do not necessarily correspond to genuine speakers on account of likely under/over-clustering.

Practical experiments are also reported where the speaker ground-truth of the oracle setup is replaced with an automatically derived segmentation. Without any other modifications to our baseline speaker diarization system, results show significant improvements across two standard, independent evaluation datasets.

We acknowledge that results reported in this paper involve the use of ground-truth transcriptions and we are currently investigating the influence on the performance when using real ASR as speech transcripts instead of the ground-truth.

6. References

[1] S. Bozonnet, D. Wang, N. W. D. Evans, and R. Troncy, "Linguistic influences on bottom-up and top-down clustering for speaker diarization," in *ICASSP 2011, 36th International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

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

[3] I. Chen, S. Cheng, and H. Wang, "Phonetic subspace mixture model for speaker diarization," in *INTERSPEECH*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010, pp. 2298–2301.

[4] J. Žibert, N. Pavesić, and F. Mihelič, "Speech/non-speech segmentation based on phoneme recognition features," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 47–47, Jan. 2006.

[5] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[6] V. Digalakis, D. Rtischev, L. Neumeyer, and E. Sa, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions on Speech Audio and Language Processing*, vol. 3, pp. 357–366, 1995.

[7] C. Fredouille, S. Bozonnet, and N. Evans, "The LIA-EURECOM RT09 Speaker Diarization System," in *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*, 2009.

[8] T. H. Nguyen, H. Sun, S. K. Zhao, S. Z. K. Khine, H. D. Tran, T. L. N. Ma, B. Ma, E. S. Chng, and H. Li, "The IIR-NTU Speaker Diarization Systems for RT 2009," in *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*, 2009.