Phonetic Class-based Speaker Verification

Matthieu Hébert and Larry P. Heck

Nuance Communications
{hebert,heck}@nuance.com

Abstract

Phonetic Class-based Speaker Verification (PCBV) is a natural refinement of the traditional single Gaussian Mixture Model (Single GMM) scheme. The aim is to accurately model the voice characteristics of a user on a per-phonetic class basis. The paper describes briefly the implementation of a representation of the voice characteristics in a hierarchy of phonetic classes. We present a framework to easily study the effect of the modeling on the PCBV. A thorough study of the effect of the modeling complexity, the amount of enrollment data and noise conditions is presented. It is shown that Phoneme-based Verification (PBV), a special case of PCBV, is the optimal modeling scheme and consistently outperforms the state-of-the-art Single GMM modeling even in noisy environments. PBV achieves 9% to 14% relative error rate reduction while cutting the speaker model size by 50% and CPU by $2/3$.

1. Introduction

Improvements in speaker verification performance can come from several different sources: feature extraction, modeling, score normalization, multi-modal cues, etc. Historically, the modeling scheme has involved maximum *a poteriori* (MAP) adaptation [1] of a background [2] (cohort [3] or world [4]) model with user speech to create the speaker model. The underlying model was usually a Single GMM [3] that covers all phonetic events (classes). Let us define, for the purpose of this paper, a phonetic class as representing phonetic events sharing common (linguistic) properties. Examples of phonetic classes are: consonnants, vowels, phonemes, states of the HMMs representing phonemes. The canonical modeling scheme has been refined along two main axis: per-channel and per-phonetic class modeling. These two sets of techniques are in fact the incarnation of the same goal: representing the acoustic space as accurately as possible. The per-channel modeling [5, 6] involves the presence of a set of background models (channel- and gender- dependent) that represent impostor population on a given channel (in this paper we use the term “channel” to describe any combination of acoustic channel and gender). This technique is most powerful in noisy environments and in cross-channel verification attempts. The per-phonetic class modeling [7, 8, 9] uses cues like frame-level phonetic alignment from an automatic speech recognizer to perform verification on a per-phoneme basis before recombining scores in one way or another. This technique is most powerful in clean conditions.

This paper will be centered on a few themes. We will first describe a flexible framework to investigate PCBV. We investigate the modeling granularity to find the optimal units (phonetic classes) for speaker verification performance. We then present a study on the effect of the amount of enrollment data. Next, we’ll report preliminary results on data sharing in the context of PCBV. The last part will report verification results across several test sets and covering several languages. In the conclusion, we analyze the source of gains from PCBV or PBV (noisy environment, gender, language, etc.).

2. Hierarchical representation of speaker models

Our implementation is based on a tree-like structure that represents acoustic events that are to be modeled; the acoustic events are a series of frames that are given a phonetic class label by a speaker-independent speech recognizer [10].

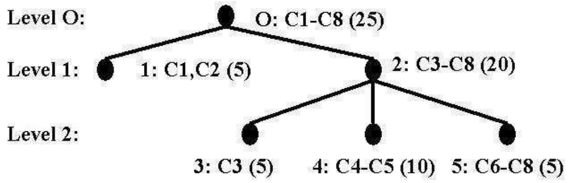

Figure 1: Example of tree-like structure. Each node is defined by *node# : classes (#gaussians),* where *node#* is the internal node number, *classes* are the labels of the collection of phonetic classes represented by that node and *#gaussians* is the number of gaussians for the GMM specific to each phonetic class.

Each node of the tree (Fig. 1) represents a GMM of variable complexity; the per-channel modeling [5, 6] is performed at the node-level enabling transformations of class Ci between channels. The GMMs used in this work are mixtures of gaussians with diagonal covariances. Granularity of the representation of the acoustic space by the tree can be adjusted by selecting the modeling level (see Fig. 1). For example, modeling level 1 uses nodes 1 and 2 only; the GMM associated to node 2 is formed by pooling gaussians from the GMMs at nodes 3, 4 and 5. Terminal nodes (for example node 1) are propagated to subsequent levels; i.e. GMM associated to node 1 is also present at level 2 (it is cloned). This procedure ensures a) that all classes are covered at all levels and b) that the total number of gaussians present on any given level is constant; this is a compact way to represent a phonetic class-based speaker model. This representation of the acoustic space by a tree of phonetic classes is a flexible framework for studying PCBV; by simply varying the modeling level, we can study coarse to refined modeling very easily while keeping the total complexity of the underlying models constant. In the above example, the same exact 25 gaussians are used across all levels; they are assigned differently from level to level. Undesirable acoustic events (such as silence and noise) can be discarded from any processing by omitting them in the tree structure. Gaussian tying has been implemented to allow parameter sharing.

2.1. Training of the background model

For a given channel, the set of GMMs representing the terminal node's classes to be modeled are trained using the speech recognizer to align each frame in an utterance to the corresponding GMM at the highest level (level 2 in our Figure 1 example) of the tree. The collection of these frames are used to train the GMM for that given node using standard vector quantization followed by expectation-maximization techniques [5, 6].

2.2. Enrollment and verification

Both the enrollment and verification procedures use a two-pass approach [5, 6]. The first pass is common to both and is used to identify the current channel. The frame-level alignments from the recognizer are used to score the relevant GMM at the given modeling level for the current phonetic class; this is done for each channel. The selected channel is the one with the highest likelihood.

The second pass of the enrollment procedure uses the frame-level phonetic alignments to MAP adapt the corresponding background model GMM at a given modeling level for the selected channel with the current frame [5, 6]. Sufficient statistics from the accumulated frames for all gaussians of all GMMs for the current modeling level, are saved as the speaker model. Note that the size of the speaker model (complexity) does not depend on the modeling level.

The second pass of the verification procedure also uses the frame-level phonetic alignments to score the relevant GMMs on the selected channel with the current frame. A likelihood ratio scoring scheme is used; and thus the speaker model as well as the background model are scored.

Let $j(t)$ be the frame-level phonetic alignment given by the recognizer; it states that at time t the frame was aligned to class j. Also, let us define $\lambda^{l,c}$ and $\bar{\lambda}^{l,c}$ as the speaker and background models at modeling level l for channel c. Then the likelihood ratio scoring is expressed as

$$\begin{aligned} L(\mathbf{X}|\lambda^{l,c}) &= \log p(\mathbf{X}|\lambda^{l,c}) - \log p(\mathbf{X}|\bar{\lambda}^{l,c}) \\ &= \frac{1}{T}\sum_{t}\left[\log p(\mathbf{x}_t|\lambda^{l,c}_{j(t)}) - \log p(\mathbf{x}_t|\bar{\lambda}^{l,c}_{j(t)})\right] \qquad (1) \end{aligned}$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_1, ..., \mathbf{x}_T, \}$ is the set of feature vectors extracted from the utterance and $\lambda^{l,c}_j$, $\bar{\lambda}^{l,c}_j$ are respectively the GMMs representing the user (speaker model) and background model at level l, for channel c and class j. The modeling level is set *a priori* and the channel is identified on-the-fly as explained above. Finally, the likelihood ratio score is compared to an overall threshold Θ to accept/reject the attempt; the value of Θ sets the operating point of the system.

3. Experiments and testset description

The experiments conducted for this study cover three languages of North America: English, Canadian French and Spanish. A language specific background model is trained for each of them. The data covers a variety of channels conditions. The baseline experiment (state-of-the-art Single GMM for all phonetic classes) is expressed by using a degenerate tree: it has a single level and a single node for all speech phones in the target language.

The trees used in the experiments were similar for all three languages. It was composed of five levels and inspired from linguistic properties [11]. Level 0 has a single node for all speech events. Level 1 segregates frames in vowels, consonants and non-continuants. For level 2, the segragation is finer: front vowels, mid vowels, back vowels, voiced/unvoiced fricatives, affricates, etc. The next level (level 3) has a phone per node. Finally, the last level (level 4) represents each states of each phones.

Language	baseline #gaussians	PCBV #phones in phone set	PCBV #gauss. per phone	#train. utts
English	200	40	5	37k
French	200	38	5	98k
Spanish	200	23	9	79k

Table 1: Description of complexity of modeling for different languages and amount of training data used for the background model (all channel pooled). Each language has between 6 and 8 channels including landline, cellular, etc. for each gender. Note that the complexities of the models are roughly equivalent.

A total of 8 test sets were used in the evaluation of the performance of our implementation of PCBV (see Table 2). They are all set-up to exercise verification in the text-dependent scheme with a maximum of three repetitions of the password in enrollment and one in verification. The average duration of the passwords is $\sim$ 3 seconds.

Name	true sp. trials	impostor trials	Note
E_d1	117688	13671	10-digits telephone #
E_d2	2610	9669	10-digits account #
E_t1	1787	15334	Company names
F_d1	100341	14321	10-digits telephone #
F_d2	47759	47474	8-digits account #
F_t1	47499	47499	City names, dates, names
S_d1	49759	49851	8-digits account #
S_t1	45036	44810	City names, dates, names

Table 2: Description of the test sets used in this study. The name of the test sets are coded as follow: language_task where the language is E for English, F for Canadian French, etc. and task is d for digits and t for text.

3.1. Modeling granularity

First, let's explore the different modeling levels at which the background model can be trained (Section 2.1) to find the optimal units (phonetic classes) for speaker verification performance.

We have used the Canadian French test sets for this study. We have trained background models at each level from 0 to 4 of the tree. For each of these (training levels), we ran the verification varying the level at which the enrollment and verification was performed. Recall that, by construction, a background model trained at a given level can be tested at that level, as well as all levels above in the regression tree. Table 3 shows the performance as well as the corresponding training and testing levels. Note that the overall best performance is at the phone level (level 3): just statistically different from the performance at level 4. The other interesting fact is that the optimal performance for the background model trained at level 4 (state level) is at testing level 3 (phone level). These results seem to indicate that the phone level is optimal. The general features of Table 3 are the result of two opposing tendencies. On one hand, the

	0	1	2	3	4
0	6.3				
1	6.0	5.8			
2	6.2	6.0	5.5		
3	6.4	6.0	5.4	4.8	
4	6.8	6.3	5.6	5.0	5.2

Table 3: Performance (in % EER) as a function of modeling level. Level for training of the background model is the first column, while the level for testing is the first line.

finer the modeling gets (high modeling level), the more accurate and restricted our description of the acoustic space is. On the other hand, the more restricted our modeling gets, the less data sharing we have (a frame being used to train several gaussians). Another avenue to explain the behavior between level 3 and 4 is the reliability of the boundaries between classes: boundaries between phones are a) less frequent than between states and b) likely more reliable; leading to better performance.

3.2. Effect of the amount of enrollment data

Speaker verification is a task that operates in what can be called a "data-starved" paradigm for the enrollment of the speaker models. Studying the effect of the amount of enrollment data on the performance of the verification system at different modeling levels for the PCBV is relevant.

We have also used the Canadian French test sets for this study. The experiments were run with a background model trained at the level 4 of the tree (the state level) to strenghten the claims in Section 3.1.

The results shown in Figure 2 are average Equal Error Rates (EERs) across the 3 test sets for an average of 7.94, 5.30, 2.67, 1.33 and 0.89 seconds for enrollment.

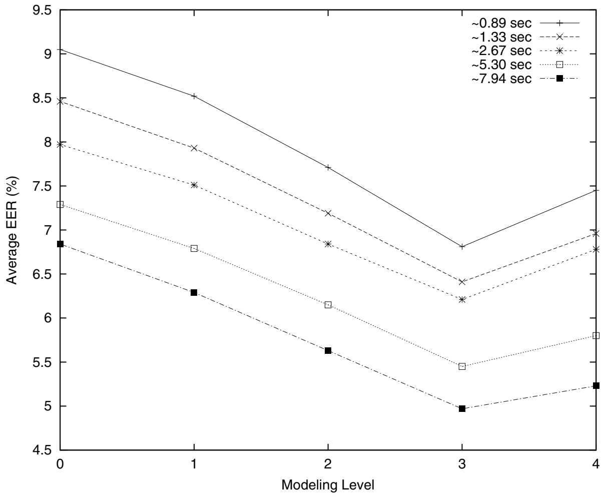

Figure 2: EERs as a function of the modeling level for different amount of enrollment data.

The features of the curves in Fig. 2 are similar to those described in the preceding Section. Also, note that these features are consistent across enrollment conditions: the phone level performs better regardless of the amount of enrollment data.

In Fig. 2, we see that the performance degrades as the amount of enrollment data diminishes; that is not a surprise. The surprise comes from the evidence that the optimal modeling level is, for all of these conditions, the phone level (level 3). This also shows that all levels of modeling, from Single GMM (level 0) to state level (level 4), are affected in roughly the same way by the reduction of the amount of enrollment data.

3.3. Effect of tying

The tying of gaussians enables data sharing during the training of the background models and during the enrollment of the speaker models. In this initial study, we wanted to test two extremes of data sharing during the training of the background models, as well as two conditions of data starving during enrollment of the speaker models. We'll use two background models in this Section to represent two extremes: Constrained is PBV trained at the phone level and Unconstrained is a Single GMM (trained at level 0). The tying is determined *a posteriori* by calculating the top N most likely gaussians (out of all available gaussians) to represent each phone. Figure 3 shows that the optimal performance (across number of tied gaussians) is independent of the way the background model was trained. The optimal number of tied gaussians is around 15 for the Constrained case, whereas it is 200 (all gaussians) for the Unconstrained case. The optimal performance, however, never beats the best PBV (without tying) performance which are 6.04% and 7.36% for three and one repetition of the password during enrollment. The general shape of the curves are not dependent on the amount of enrollment data (they are only shifted up in the case of low amount of enrollment data). For the Constrained case, the optimal performance being around 15 tied gaussians seems to indicate that the number of gaussians to support each phone is not constant. We have investigated that without success; however our tests were limited to verification without online-adaptation [12, 13] which might change the conclusion in that case. We defer further discussion on all of these topics to another report.

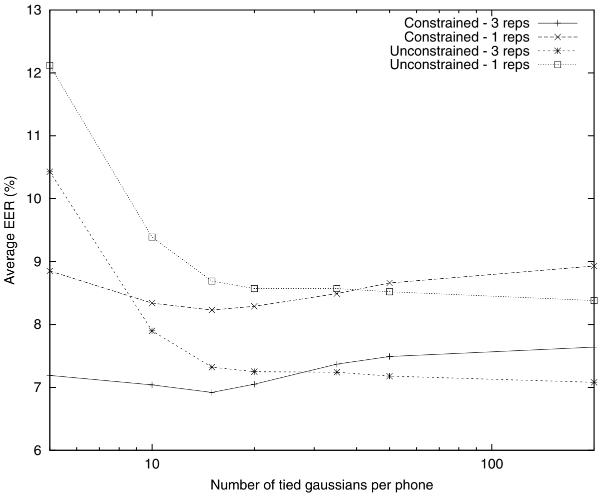

Figure 3: Performance as a function of the number of tied gaussians in the phone-level GMMs on the F_d1 test set. Performance is shown for two different amount of enrollment data (one and three repetitions of the password, respectively 2.67 and 7.94 seconds)

3.4. Results across multiple databases and languages

For the rest of this paper, we are going to concentrate on the PCBV at level 3 (PBV) since it gives the best performance across levels. Experiments were conducted on the test sets presented (Table 2) above for the baseline system and the PBV system. We also added experiments to the baseline system with background models trained with a sub-set of the original training set ("specialized" Single GMM); that sub-set was selected to match as closely as possible the target lexicon (digits only training set for a target digits test set, etc.). This is an easy way of introducing lexical information (or constraints) within the Single GMM scheme.

Name	Single GMM	Specialized single GMM	PBV (level 3)
E_d1	3.17	3.30	3.13
E_d2	3.92	3.53	3.42
E_t1	2.53	2.33	2.81
F_d1	2.67	2.52	2.17
F_d2	7.61	6.37	6.15
F_t1	8.51	8.15	6.21
S_d1	3.76	3.62	2.88
S_t1	5.10	5.48	4.51

Table 4: Results %EER.

From Table 4, we can see that the average error rate reduction across all test sets is around 14% relative when the training sets of the background models are identical (Single GMM compared to PBV). When we compare PBV with the "specialized" Single GMM results, the error rate reduction is more modest, but still roughly 9% relative. The important point to note here is that since we are operating in the text-dependent mode of verification, our lexicon is constant between enrollment and verification and thus not all of the phones are enrolled. This leads to a lower average size of the speaker models: a 50% reduction in speaker model size. Also, since a preliminary alignment is done by the recognizer, this leads to 2/3 improvements in CPU for the verification part alone (ignoring the CPU to generate the alignments).

4. Discussion and conclusion

We have broken down the results across all test sets and averaged the error rate reduction (on EER between the Single GMM and PBV) as a way to gain insight on the source of the improvement.

	Matched	Mismatched
All data	16%	11%
Male	21%	19%
Female	14%	13%
Landline	18%	18%
Cellular	17%	14%

Table 5: Relative improvements between Single GMM and PBV (level 3) on different breakdowns of the data.

From Table 5, we see that PBV improves performance in matched conditions (matched conditions are when enrollment and verification are done on the same channel, as determined by the system). Surprisingly, the improvement is more important for males than for females. The improvements also seem to be consistent across noise (channel) conditions.

We have presented a framework to easily study phonetic class-based verification. It has been shown that the phoneme level of modeling (PBV) is optimal in a wide range of enrollment conditions. We report gains with PBV ranging between 9% to 14% over the state-of-the-art Single GMM with a substantial reduction of speaker model size and CPU load. We have also noted that the improvements from PBV are most spectacular in matched conditions, for males and/or new languages.

5. References

[1] Gauvain J.-L. and Lee C.-H. (1994) "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains" *IEEE Trans. Speech Audio Process.* **2**, pp. 291–298.

[2] Rosenberg A. E. and Parthasarathy P. (1996) "Speaker background models for connected digit password speaker verification" *Proc. ICASSP '96*, pp. 81–84.

[3] Reynolds D. A. (1995) "Speaker identification and verification using Gaussian mixture speaker models" *Speech Comm.* **17**, pp. 91–108.

[4] Carey M. J., Parris E. S. and Bridle J. S. (1991) "A speaker verification system using alpha-nets" *Proc. ICASSP '91*, pp. 397–400.

[5] Heck L. P. and Weintraub M. (1997) "Handset-Dependent Background Models for Robust Text-Independent Speaker Recognition" *Proc. ICASSP '97*, pp. 1071–1074.

[6] Teunen R., Shahshahani B., and Heck L. P. (2000) "A Model-Based Transformational Approach to Robust Speaker Recognition" *Proc. ICSLP '00*, pp. II-495-498.

[7] Matsui T. and Furui S. (1993) "Concatenated Phoneme Models for Text-Variable Speaker Verification" *Proc. ICASSP '93*, pp. II-391-394.

[8] Chaudhari U.V., Navrátil J. and Maes S. (2003) "Multi-grained Modeling With Pattern Specific Maximum Likelihood Transformations for Text-Independent Speaker Recognition" *IEEE Trans. Speech and Audio Proc.* vol. 11, pp 61-69.

[9] Auckenthaler R., Parris E. S. and Carey M. J. (1999) "Improving a GMM speaker verification system by phonetic weighting" *Proc. ICASSP '99*, pp. .

[10] Digalakis V., Monaco P. and Murveit H. (1996) "Genones: Generalized Mixture Models Tying in Continuous Hidden Markov Model-Based Speech Recognizers" *IEEE Trans. Speech and Audio Processing*, pp. 281-289.

[11] Deller J.R., Proakis J.G. and Hansen J.H. (1987) "Discrete-Time Processing of Speech Signals" *Prenttice-Hall.*

[12] Moreau N., Charlet D. and Jouvet D. (2000) "Confidence measure and incremental adaptation for the rejection of incorrect data" *Proc. ICASSP '00*, pp. 1807-1810.

[13] Heck L. P. and Mirghafori N. (2000) "Unsupervised On-Line Adaptation in Speaker Verification" *Proc. ICSLP '00*, pp. II-454-457.