

SPEAKER VERIFICATION USING PHONEME-ADAPTED GAUSSIAN MIXTURE MODELS

Dan Gutman and Yuval Bistriz

Department of Electrical Engineering
Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT

Despite intuitive expectation and experimental evidence that phonemes contain useful speaker discriminating information, phoneme-based speaker recognition systems reported so far were not found to perform better than phoneme-independent speaker recognition systems based on Gaussian Mixture Model (GMM). The paper proposes a new phoneme-based speaker verification technique that uses models obtained by adaptation of well-trained speaker GMMs. The new proposed system was found to consistently outperform comparable sized phoneme-independent GMM based speaker verification systems in experiments held with clean and telephone speech databases.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking by using speaker specific information included in speech waves. Speaker recognition is classified into two specific tasks: identification and verification e.g. [1]. This work considers speaker verification, that is system that has to accept or reject the identity claim of a speaker.

Gaussian Mixture Models (GMMs) have been successfully applied in speaker verification systems [2]. A GMM consists of a weighted sum of M Gaussians. The model may be collectively represented by:

$$\lambda = \{w_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, \dots, M \quad (1)$$

where $\vec{\mu}_i$, Σ_i represent the mean and covariance of each Gaussian and w_i represents its weight. This work will use only diagonal covariance matrices.

Let \vec{x} denote a feature vector of length D . The mixture density used for the likelihood function is defined as:

$$p(\vec{x} | \lambda) = \sum_{i=1}^M w_i p_i(\vec{x}) \quad (2)$$

where each $p_i(\vec{x})$ is a uni-modal Gaussian density:

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'(\Sigma_i)^{-1}(\vec{x} - \vec{\mu}_i)\right\} \quad (3)$$

The study of GMMs for speaker recognition tasks recognized that it performs best when the Gaussian components are strongly correlated with acoustic events such as phonemes [2]. Phoneme-based methods in speaker recognition have been studied in several previous works, including the following [3]-[6]:

Margin-Chagnoleau et al. [3] have shown that speaker identification performance depends on the phonetic label of the speech

segments used. Utterances containing mostly phonemes, like vowels and nasal consonants, performed better than phonetically balanced utterances.

Newman et al. [4] performed phoneme based speaker recognition by taking general speaker-independent phoneme models and adapting them to each speaker. The results using this method did not outperform phoneme-independent systems based on GMMs.

Olsen [5] introduced a 2-stage approach to phoneme based speaker recognition. In the first stage the speech is segmented according to HMM phone models. In the second stage, the speaker is verified using phoneme dependent radial basis function networks. Performance was significantly improved by applying a phoneme and speaker dependent linear discriminant analysis or Fisher Transform on the feature vectors.

Auckenthaler et al. [6] compared a phoneme-independent approach to speaker verification based on GMMs to a phoneme-based approach using Hidden Markov Models (HMMs) that is based on phonetic classes. The phoneme-independent GMM system has consistently outperformed the phoneme based HMM system. However, it was found that the performance of the phoneme-independent GMM system can be improved by applying phonetic weighting, obtained from the HMM system, to the tested frames of the GMM speaker verification system.

The aforementioned works show that there is a strong connection between the phonetic information and the speaker recognition process. They also show that the performance of the phoneme-independent systems can be significantly improved by using the phonetic information contained in the speech data. However, phoneme-based speaker verification systems have not succeeded so far to outperform good phoneme-independent systems.

This paper follows these studies and proposes a new phoneme-based approach that creates a GMM model for each phoneme of the speaker by applying Bayesian adaptation to the phoneme-independent GMM. In experiments we held using the TIMIT and NTIMIT databases the new speaker verification scheme proposed here consistently outperforms comparable sized phoneme-independent GMM speaker verification system.

2. PHONETIC SEGMENTATION

The first step in the construction of a phoneme-based speaker recognition system is classifying the features extracted from the frames to clusters of phonemes. The classification requires phonetic segmentation of the data. The feature vectors contained in each cluster are then used to develop models for each phoneme.

Let X represents the whole set of feature vectors of a certain speaker, and let K be the number of different phonemes, then at the end of this step we partition X into K groups, say,

$$X \rightarrow X_1, \dots, X_K \quad (4)$$

where some of the X_k , $k = 1, \dots, K$, may be empty if a certain phoneme do not appear in the data.

The phonetic segmentation stage can be carried out independently from the subsequent means used for speaker recognition and use different feature vectors [5].

3. PHONEME-BASED SPEAKER VERIFICATION

We initially tried a direct approach to phoneme-based speaker recognition as follows: We partitioned the data for each speaker into clusters of phonemes (in number that may vary from speaker to speaker). Then we built a GMM for each phoneme (using 2 Gaussians for vowels and single Gaussian for consonants). The performance of this system was found to be low compared to a phoneme-independent speaker verification system. We next tried to improve the performance of the above system by increasing weights in the scoring process of phonemes that were found to be more discriminating than others. This weighting improved the performance slightly, but it still remained below the performance of a phoneme-independent speaker verification system.

Several reasons may be raised to explain the degradation in performance observed in the above phoneme-based speaker verification systems, as well as in the method examined in [4]. First comes to mind, the effect of segmentation error. We discarded this reason after comparing the segmentation software that we used with the phonetic segmentation provided with the used databases (TIMIT and NTIMIT). A second possible cause is that other general acoustical events that contribute to speaker recognition and are related to correlation and transition between phonemes, are lost when handling each phoneme separately. A third cause may be that there is not enough data to train some phonemes and over-training the data of some models. The new technique described in the next section was devised with anticipation that it might combat the two latter possible causes for degraded performance.

4. PHONEME ADAPTED GMM

The new approach to phoneme-based speaker verification presented starts with a statistically well-trained phoneme-independent GMM for each speaker. Next, a GMM model for each phoneme found in the training data is created by certain adaptation of the speaker's model. This approach differs from the technique described in our experiments above and from the method reported in [4] (a speaker-independent models for phonemes adapted to each speaker) in that modeling of the phonemes occurs *after* modeling the speakers and via adaptation of each speaker's model to speaker dependent phoneme models for phoneme available in the training data. The potential of such approach to improve verification performance stems from the fact that it starts with models that already discriminate speakers and then adapts itself to phonemes abundant in the training data while handles gracefully adaptation for phonemes scarcely or non available in the training data. A more detailed description of the phoneme-adapted scheme proposed is described in the following:

First, a phoneme-independent GMM is created for each speaker using the whole training data of the speaker. Let the phoneme-independent GMM of speaker s be denoted by $\lambda_s = \{w_i, \mu_i, \Sigma_i\}$ $i = 1, \dots, M$.

Next, the training feature vectors of speaker s are clustered into K phoneme groups. Each group $X_k = \{\vec{x}_1, \dots, \vec{x}_{T_k}\}$ $k = 1, \dots, K$ contains the feature vectors of phoneme k . X_k is an empty group if phoneme k does not appear in the training data.

For each phoneme k , a new GMM is developed by adaptation (as shown in a moment) of the phoneme-independent GMM. The resulting new GMM model for phoneme k , denoted by $\lambda_{s,k} = \{\hat{w}_{ik}, \hat{\mu}_{ik}, \hat{\Sigma}_{ik}\}$ $i = 1, \dots, M$, $k = 1, \dots, K$, has the same size as λ_s , the phoneme-independent GMM of speaker s .

The adaptation technique that was used is similar to the technique described in [7]. The difference is that in [7] it is used to adapt an universal background model (UBM), representing a large group of background speakers, to create a GMM for each speaker and here the technique is used to adapt a phoneme-independent GMM of a certain speaker to create a model for each phoneme. It includes 2 steps: "Expectation" step in which a new set of GMM parameters is estimated and "Combination" step in which the new estimated set is combined with the original phoneme-independent GMM parameters as follows.

"Expectation" step: Compute $n_{ik}, E_{ik}(\vec{x}), E_{ik}(diag(\vec{x}\vec{x}'))$, the new estimated weight, mean and variance parameters for phoneme k and mixture component i of the GMM:

$$Pr(i | \vec{x}_t) = \frac{w_i p_i(\vec{x}_t)}{\sum_{j=1}^M w_j p_j(\vec{x}_t)} \quad (5)$$

$$n_{ik} = \sum_{t=1}^{T_k} Pr(i | \vec{x}_t) \quad (6)$$

$$E_{ik}(\vec{x}) = \frac{1}{n_{ik}} \sum_{t=1}^{T_k} Pr(i | \vec{x}_t) \vec{x}_t \quad (7)$$

$$E_{ik}(diag(\vec{x}\vec{x}')) = \frac{1}{n_{ik}} \sum_{t=1}^{T_k} Pr(i | \vec{x}_t) diag(\vec{x}_t \vec{x}_t') \quad (8)$$

where $p_i(\vec{x}_t)$ denotes the Gaussian density of mixture i of the phoneme-independent GMM (3).

"Combination" step: Combine the new estimated parameters with the original phoneme-independent GMM parameters to form the final set of parameters for phoneme k and mixture i : $\hat{w}_{ik}, \hat{\mu}_{ik}, \hat{\sigma}_{ik}^2$, where $\hat{\sigma}_{ik}^2$ denotes the diagonal covariances of mixture i and phoneme k obtained from $\hat{\Sigma}_{ik}$, as follows: For each mixture i and phoneme k compute: The adaptation factors

$$\alpha_{ik} = \frac{n_{ik}}{n_{ik} + r_k}, \quad (9)$$

where r_k is a fixed relevance factor for phoneme k ; Weights

$$\hat{w}_{ik} = [\alpha_{ik} n_{ik} / T_k + (1 - \alpha_{ik}) w_{ik}] \gamma_k, \quad (10)$$

where γ_k is a scale factor that ensures that the weights of the Gaussian components sum to unity; Means and covariances

$$\hat{\mu}_{ik} = \alpha_{ik} E_{ik}(\vec{x}) + (1 - \alpha_{ik}) \mu_{ik} \quad (11)$$

$$\hat{\sigma}_{ik}^2 = \alpha_{ik} E_{ik}(diag(\vec{x}\vec{x}')) + (1 - \alpha_{ik})(\sigma_{ik}^2 + diag(\mu_{ik} \mu_{ik}')) - diag(\hat{\mu}_{ik} \hat{\mu}_{ik}') \quad (12)$$

The adaptation factor α_{ik} determines the balance between the new adapted parameters and the phoneme-independent GMM parameters. Large n_{ik} value (well-trained Gaussian component) brings α_{ik} closer to 1 and thus more weight is given to the new adapted parameters. Low n_{ik} value (under-trained Gaussian component) sets α_{ik} closer to 0 and thus more weight is given to the original phoneme-independent GMM parameters. Phonemes not encountered in the training will set α_{ik} to zero and thus will be given the original phoneme-independent model without any adaptation. The adaptation factor ensures that phonemes with small amount of training data will remain close to the phoneme-independent model and are granted not to be missing in the model; Phonemes with large amount of training data will be more adapted to the specific way it is uttered by the speaker but remain faithful to the phoneme-independent GMM speaker model that contains the global phonetic events that are essential for speaker recognition.

The background model used for testing consisted universal phoneme models created by adaptation to phonemes of a phoneme-independent UBM. First, a phoneme-independent UBM was created using data speech from a large group of speakers, this model will be denoted by λ_{UBM} . Then Bayesian adaptation was applied (as in equations (5) - (12)) to create the universal background models for each phoneme, denoted by $\lambda_{UBM,k}$ where $k = 1, \dots, K$.

At testing, comparison is held for each frame between the likelihoods of its modeling by the speaker's adapted phoneme model and its modeling by the phoneme-adapted UBM of the same phoneme. The resulting scoring for a speaker is obtained by summing the log-likelihood ratios along the tested sequence of feature vectors $X = \{\vec{x}_1, \dots, \vec{x}_T\}$ as follows:

$$\Lambda(X) = \sum_{t=1}^T \log p(\vec{x}_t | \lambda_{s,k_t}) - \log p(\vec{x}_t | \lambda_{UBM,k_t}) \quad (13)$$

assuming feature vector \vec{x}_t has been associated with phoneme k_t of speaker s .

5. EXPERIMENTAL CONFIGURATION

The experiments reported in this paper were conducted on the TIMIT and NTIMIT databases. The TIMIT [8] database contains clean speech recorded from 438 male speakers. The NTIMIT [9] database contains the same utterances recorded over the telephone network. For each speaker there are 10 different utterances of 2-3 seconds duration each. 350 speakers were used for training and the remaining 88 speakers were used to train the Universal Background Model (UBM). Training was done using 2, 4 and 8 utterances with a total duration of 5, 10 and 20 seconds for each speaker and duration of 8, 15 and 30 minutes for the UBM. Testing was done on the remaining 2 utterances. In each experiment 350 tests were conducted with true speakers and 350 with impostors.

Speech was parameterized using 12 mel-cepstrum coefficients concatenated with 12 delta mel-cepstrum coefficients (total of 24 coefficients). Mean removal was applied on the parameters to reduce channel noise. Features were extracted using 32ms hamming window and 16ms frame period.

The TIMIT and NTIMIT databases come with phonetic transcription and segmentation. We used the phonetic transcription but not the segmentation data. Thus the experiments simulate a real world system where known text admits reliable and relatively

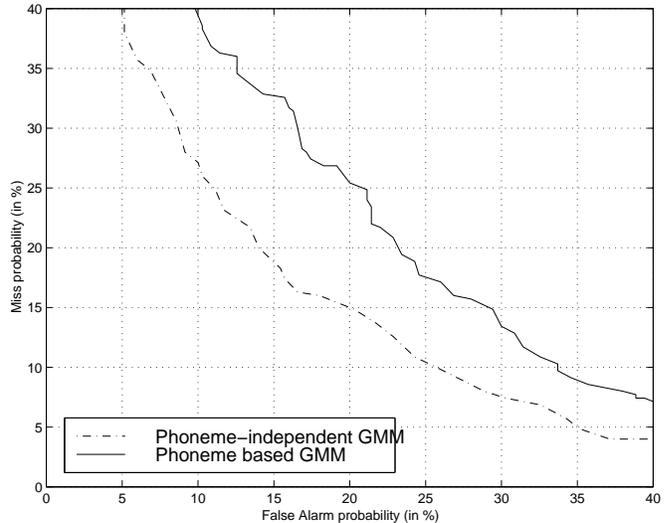


Figure 1: Speaker verification results for a 30-35 phoneme-based verification system compared to a 32 phoneme-independent GMM system.

simple segmentation procedure, e.g., a text prompted speaker verification schemes. Segmentation was carried out using the segment program of [10]. This program uses speaker-independent phoneme models consisting of HMM with 3 states. It uses the phoneme sequence files provided with the TIMIT and NTIMIT databases and applies Viterbi algorithm with forward search to carry out the segmentation.

6. EXPERIMENTS AND RESULTS

In the first of our experiments we examined the performance of the phoneme-based system described in section 3 and compared it to a phoneme-independent GMM system. Performance was tested on the NTIMIT database using 8 utterances for training for each speaker. In the phoneme-based system, GMMs of size 2 were used to model vowels and uni-modal Gaussians were used for consonants. Assuming 13 different vowels and 25 different consonants included in the database, a total number of 51 Gaussian components are required. Since the training data of each speaker doesn't include all the possible phonemes, the actual numbers of Gaussians for each speaker were in the range 30-35. For comparison a GMM with a comparable size of 32 was chosen for the phoneme-independent system. Figure 1 shows the trade-off between miss and false alarm rates for the phoneme-based system compared to the phoneme-independent GMM system. A large drop in performance is noted in the phoneme-based system. The Equal Error Rates (EERs) are 17.1% for the phoneme-independent system and 22.0% for the phoneme-based system.

In the next set of experiments, we applied the phoneme-adapted method described in section 4. The experiments were conducted on both the TIMIT and NTIMIT database using 2, 4 and 8 utterances for training. We used GMMs of size 16 for the 2 utterances experiment and GMMs of size 32 for the 4 and 8 utterances experiments. For the phoneme-adapted systems we used a fixed relevance factor r_k that has been chosen in the range 8-12 for achieving best results. Figures 2 and 3 show the trade-off curves between miss and false alarm error rates for the phoneme-

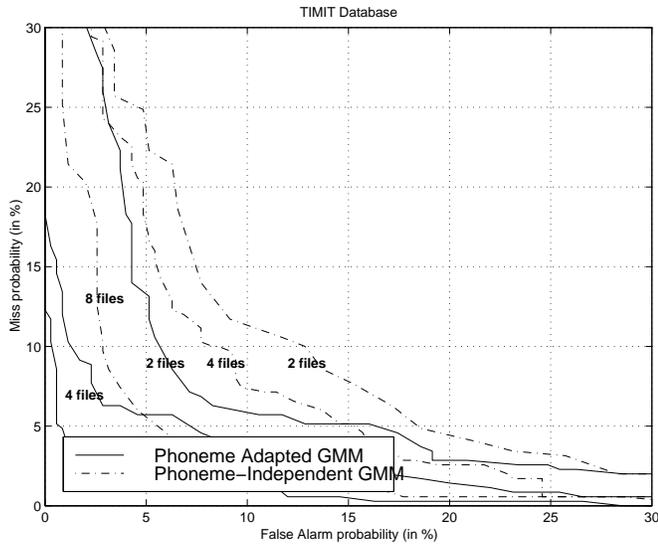


Figure 2: Speaker verification results for phoneme-adapted GMM systems compared to phoneme-independent GMM systems tested on the TIMIT database.

adapted systems compared to the phoneme-independent GMM system. It can be seen that the phoneme-adapted systems consistently outperform the phoneme independent system. The EERs measured for the 8 utterances experiment on TIMIT database was 5.42% for the phoneme-independent system compared to 2.28% for the phoneme-adapted system. For the NTIMIT database EER of 17.14% was measured for the phoneme-independent system compared to 15.72% for the phoneme-adapted system.

7. CONCLUSIONS

The paper presented a new phoneme-adapted GMM-based speaker verification method. It begins with a well-trained phoneme-independent GMM for each speaker and from it adapts GMMs for the phonemes. An adaptation factor adjusts the extent that the phoneme model is allowed to deviate from the phoneme-independent model. It ensures that phonemes with small amount of training data remains close to the phoneme-independent GMM speaker model and that no phonemes are missing in the model. At the other end, phonemes with large amount of training data may go through more modification but still remain in the vicinity of the regular phoneme-independent speaker model well trained for discrimination among speakers.

Experimental results reported showed that the proposed phoneme-adapted speaker verification method outperforms comparable GMM-based phoneme-independent speaker verification systems.

8. REFERENCES

[1] S. Furui, "Recent advances in speaker recognition", *Pattern Recognition Letters* 18, 1997, pp 859-872.
 [2] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication* Vol. 17, pp 91-108, 1995.

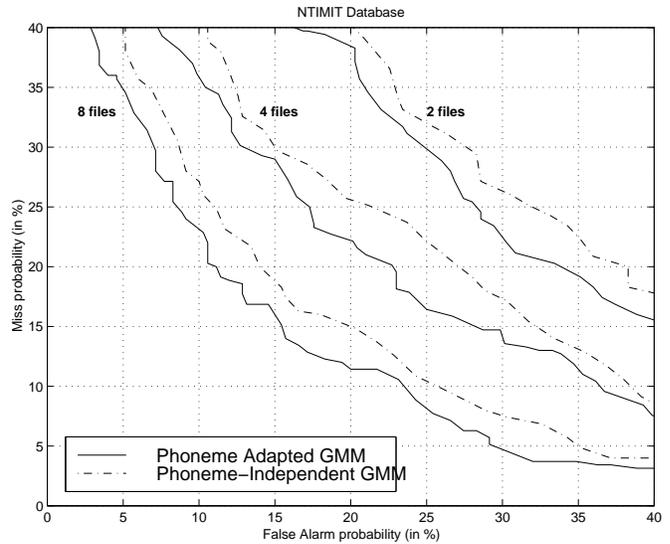


Figure 3: Speaker verification results for phoneme-adapted GMM systems compared to phoneme-independent GMM systems tested on the NTIMIT database.

[3] I. Magrin-Chagnoleau, J. F. Bonastre and F. Bimbot, "Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods", *Proc of the European Conference on Speech Communication and Technology, Eurospeech'95*, Vol. 1, pp. 337-340, Madrid, Spain, September 1995.
 [4] M. Newman, L. Gillick, Y. Ito, D. McAllaster and B. Piskin, "Speaker verification through large vocabulary continuous speech recognition", *Proc. of the International Conference on Spoken Language Processing*, pp. 2419-2422, 1996.
 [5] J. Ø. Olsen, "A two-stage phone based speaker verification", *Pattern Recognition Letters* Vol. 18, pp. 889-897, 1997.
 [6] R. Auckenthaler, E. S. Parris and M. J. Carey, "Improving a GMM speaker verification system by phonetic weighting", *Proc. of the 1999 IEEE Int. Conference on Acoustics, Speech, and Signal Processing*.
 [7] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing Review Journal*, Jan. 2000.
 [8] W. M. Fisher, G. R. Doddington and K.M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status", *Proc. DARPA Workshop on Speech Recognition*, Feb. 1986, pp 93-99.
 [9] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A phonetically balanced continuous speech, telephone bandwidth speech database", *Proc. of the 1990 IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, pp 109-112.
 [10] C. Becchetti and L. P. Ricotti, *Speech Recognition - Theory and C++ Implementation*, Wiley, 1999.