22nd IEEE Convention of Electrical and Electronic Engineers in Israel, Dec. 2002, Tel Aviv, Israel.

PHONEME BASED SPEAKER VERIFICATION BY GAUSSIAN MIXTURE MODELS WITH ADAPTATION OF SUBSETS OF DOMINANT PARAMETERS AND PHONEMES

Dan Gutman and Yuval Bistritz

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

ABSTRACT

The paper considers improvement of Phoneme Adapetd GMM (PA-GMM) speaker verification systems by applying adaptation to only a selected subsets of the most discriminative parameters and phonemes. PA-GMM's are basically GMM's developed for phonemes by Bayesian adaptation of a general phoneme-independent GMM. Speaker verification systems using PA-GMM's have shown to perform better than comparable sized phoneme-independent GMM systems in experiments held on both 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 the speech waves. Several good texts with background and surveys on speaker recognition are available, see e.g. the recent account in [1]. This work considers speaker verification, that is the task where a system has to either accept or reject the claimed identity of a speaker.

In the past, there has been only moderate interest in phoneme based speaker recognition systems mainly due to the fact that these systems didn't perform as well as phoneme-independent systems [2]. Yet, it has been found that the linguistic content of the speech signal has a profound impact on the performance of speaker recognition systems [3] with vowels or nasal consonants phonemes performing better than fricatives phonemes.

Considering the different abilities of phonemes in discriminating between speakers, phoneme-based speaker recognition systems can be improved by taking a subset of the most discriminative phonemes or by applying weighting to the different phonemes [4] [5].

This paper examines speaker verification system where the speaker's phonemes are represented by Phoneme Adapted Gaussian Mixture Models (PA-GMM) that are basically models for phonemes of speakers obtained by Bayesian adaptation of a general phoneme-independent speaker models. Speaker verification done using this new phoeneme based scheme has consistently outperformed comparable sized phoneme-independent GMM system on experiments held on the TIMIT and NTIMIT databases [6]. In this paper, further improvement in performance of the PA-GMM system was reached by applying the adaptation process only to a subset of the most discriminative phonemes and parameters.

2. PHONEME ADAPTED GMM

PA-GMM is a GMM created for a specific phoneme by adaptation of an original phoneme-independent GMM [6]. The outline of the training procedure for a PA-GMM is as follows:

1. A phoneme-independent GMM, denoted by $\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M$, is created for a specific speaker using the whole training data of the speaker (including all phonemes).
2. The training feature vectors of the speaker are clustered into K phoneme groups. Each group $X_k = \{x_1, \ldots, x_{T_k}\} \quad k = 1, \ldots, K$ contains the feature vectors of phoneme k.
3. **"Expectation" step**: For each phoneme k, a new set of parameters $\{n_{i,k}, E_{i,k}(x), E_{i,k}(diag(xx'))\}$ is estimated via Bayesian adaptation to the phoneme-independent speaker's GMM (λ) as follows:

$$Pr(i \mid x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{1}$$

$$n_{i,k} = \sum_{t=1}^{T_k} Pr(i \mid x_t) \tag{2}$$

$$E_{i,k}(x) = \frac{1}{n_{i,k}} \sum_{t=1}^{T_k} Pr(i \mid x_t) x_t \tag{3}$$

$$E_{i,k}(diag(xx')) = \frac{1}{n_{i,k}} \sum_{t=1}^{T_k} Pr(i \mid x_t) diag(x_t x_t{}') \tag{4}$$

where $p_i(x_t)$ denotes the Gaussian density of mixture i of the phoneme-independent GMM.

4. **"Combination" step**: For each phoneme k, the new estimated parameters are combined with the original phoneme-independent GMM parameters to form the final set of parameters $\{\widehat{w}_{i,k}, \widehat{\mu}_{i,k}, \widehat{\sigma}^2_{i,k}\}$ as follows:

$$\alpha_{i,k} = \frac{n_{i,k}}{n_{i,k} + r_k} \tag{5}$$

where r_k is a fixed relevance factor for phoneme k.

$$\widehat{w}_{i,k} = [\alpha_{i,k} n_{i,k}/T_k + (1 - \alpha_{i,k}) w_i]\gamma_k \tag{6}$$

$$\widehat{\mu}_{i,k} = \alpha_{i,k} E_{i,k}(x) + (1 - \alpha_{i,k})\mu_i \tag{7}$$

$$\widehat{\sigma}^2_{i,k} = \alpha_{i,k} E_{i,k}(diag(xx')) + (1 - \alpha_{i,k})(\sigma_i^2 + diag(\mu_i \mu_i')) - diag(\widehat{\mu}_{i,k}\widehat{\mu}_{i,k}') \tag{8}$$

where γ_k is a scale factor used to ensure that the weights of the Gaussian components sum to unity. The adaptation factor $\alpha_{i,k}$ determines the balance between the new adapted parameters and the phoneme-independent GMM parameters.

3. EXPERIMENTAL CONFIGURATION

The experiments reported in this paper were conducted on the NTIMIT database. The NTIMIT database [7] contains speech recorded from 438 male speakers over the telephone network. 350 speakers were used for training and 88 speakers were used to train the UBM. Training was done using 8 utterances with a total duration of about 20 seconds for each speaker and duration of about 30 minutes for the UBM. Testing was done on the remaining 2 utterances (each of duration of about 2-3 seconds). In each experiment 350 tests were conducted with true speakers and 350 with impostors. Speech was parameterized by 12 mel-cepstrum coefficients concatenated with 12 delta mel-cepstrum coefficients. Mean removal was applied on the parameters to reduce channel noise. Features were extracted using 32ms hamming window and 16ms frame period.

The NTIMIT databases comes with phonetic transcription and segmentation. We used the phonetic transcription but not the segmentation data. Thus the experiments may present a realistic system where known text admits reliable and relatively simple segmentation procedure, e.g., a text prompted speaker verification schemes. Segmentation was carried out using the segment program of [8]. It uses the phoneme sequence files provided with the database and applies Viterbi algorithm to carry out the segmentation.

4. EXPERIMENTS AND RESULTS

In this section we present experiments and results using the Phoneme-Adapted GMM system described in section 2.

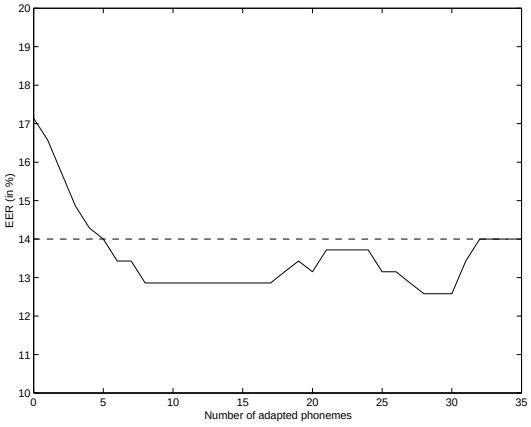

Figure 1: Speaker verification results using subsets of all phonemes.

4.1. Adapting subsets of phonemes

The first set of experiments were designed to explore whether adaptation of an appropriate subset of the phonemes improves performance scores. One granted advantage of using subsets of phonemes is reducing the size of the speaker models. In all the experiments, phonemes not in the subset that passes adaptation assume the speaker's phoneme-independent GMM model. Figure 1 shows the performance of a phoneme-adapted GMM system as a function of the number of adapted phonemes. The experiments were carried out using a 32 size GMM and relevance factor of 12. It is seen that the performance achieved by using all the phonemes (EER of 14%) is equally achieved by adapting only 5 phonemes. It is also observed that a better results of 12.86% EER is obtained for subsets of 8-17 phonemes.

4.2. Adapting partial sets of parameters

So far, the adaption was always applied to all the parameters of the model: weights (w), mean values (m) and variance values (v). In the next set of experiments we examined the effect of applying adaptation to only subsets of the complete set of parameters. The result of experiments carried out using a 32 sized GMM and relevance factor of 12 (with the number of adapted phonemes is not constrained) is depicted in figure 2. The figure reveals that some partial combinations of parameters perform better than the complete set of parameters. For example, the EER for adapting only variances (v) is 12.29%, for adapting only means and weights (m,w) is 13.15% compared to EER of 14% for adapting the complete set of parameters. Adaptation of partial set of parameters reduces the model size stored for each speaker.

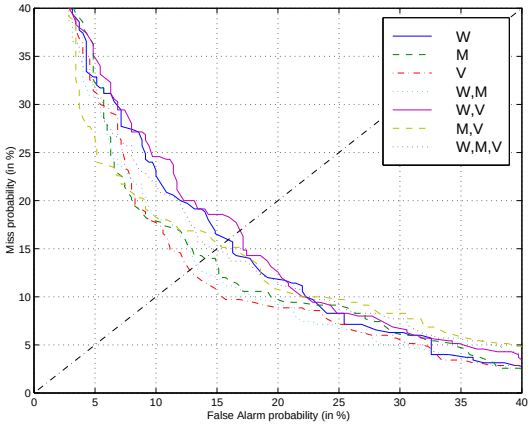

Figure 2: Speaker verification results using different sets of parameters. 'w' - denotes weights, 'm' - means and 'v' - variances.

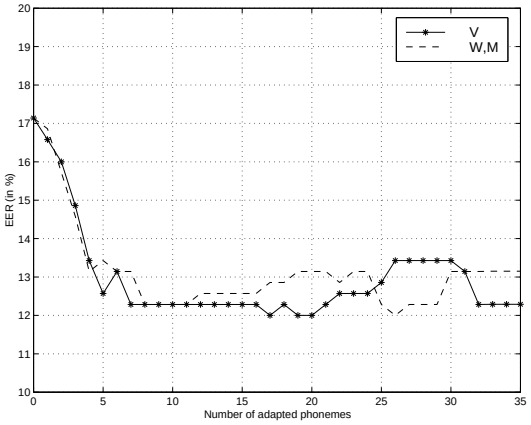

Figure 3: Speaker verification results using different sets of parameters and subset of phonemes.

4.3. Partial sets of parameters and subsets of phonemes

In the last set of experiments we explored the combination of adaptation of a partial set of parameters with adaptation of subsets of phonemes. Using insight gathered from the better performing subsets in the experiments reported above, we examined adaptation of variance values only (v) and adaptation of mean and weight values only (w,m) while changing the size of the subset of phonemes. The results that were obtained are reported in figure 3. It can be seen that adaptation of a subset of the most discriminative phonemes in a system using a partial set of parameters may yield results that are equal or better than adapting all phonemes. For example, for the case of adaptation of only variance values (v), an EER (12.29%) equal to when using all phonemes has been equally obtained by adapting only 7-16 phonemes. Also, for the case of adapting only mean and weight values (w,m), an EER of 12.29% for 8-11 phonemes that is better than EER of 14% obtained when using all phonemes was observed.

5. CONCLUSIONS

This paper explored phoneme-adapted GMM (PA-GMM) systems for speaker verification. In this new phoneme based approach a GMM for each phoneme is obtained by Bayesian adaptation of a usual (phoneme-independent) GMM of the speaker. The extent of the adaptation is controlled by an adaptation factor α which is a function of the amount of data available on the phoneme. The PA-GMM system was found to perform better than a regular phoneme-independent GMM systems in all experiments held.

The paper has demonstrated that adaptation of only subsets of phonemes and parameters may reduce models size without degradation and with even improvement of verification scores. The latter outcome is possible in principle because different phonemes and parameters carry different speaker discriminating abilities. The proposed PA-GMM succeeds to capture this potential and exploit it to create equal or better performing speaker verification systems with lower training and storage requirements because the untrained parts default to the speaker's phoneme-independent models and the trained parts converge gracefully (when the available training data is low) to the speaker's phoneme-independent models.

6. REFERENCES

[1] S. Furui, "Recent advances in speaker recognition", *Pattern Recognition Letters 18*, 1997, pp 859-872.

[2] M. Newman , L. Gillick , Y. Ito , D. McAllaster and B. Peskin "Speaker verification through large vocabulary countinous speech recognition", *ICSLP'1996*, pp. 2419-2422.

[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", *Eurospeech'95*, pp. 337-340.

[4] R. Auckenthaler , E.S. Parris and M.J. Carey , "Improving a GMM speaker verification system by phonetic weighting", *IEEE Int. Conference on Acoustics, Speech, Signal Processing*, 1999.

[5] J.Ø. Olsen, "A two-stage phone based speaker verification", *Pattern Recognition Letters* Vol. 18, pp. 889-897, 1997.

[6] D. Gutman and Y. Bistritz, "Speaker verification using phoneme-adapted gaussian mixture models", *EUSIPCO'2002*, Toulouse, France.

[7] C. Jankowski , A. Kalyanswamy , S. Basson and J. Spitz, "NTIMIT: A phonetically balanced continuous speech, telephone bandwidth speech database", *IEEE Int. Conference on Acoustics, Speech, Signal Processing*, pp 109-112, 1990.

[8] C. Becchetti and L.P. Ricotti, *Speech Recognition - Theory and C++ Implementation*, Wiley, 1999.