To appear in Interseech 2009

Adaptive Individual Background Model for Speaker Verification

Yossi Bar-Yosef, Yuval Bistritz

Department of Electrical Engineering
Tel-Aviv University, Tel-Aviv 69978, Israel
yossibaryosef@gmail.com, bistritz@eng.tau.ac.il

Abstract

Most techniques for speaker verification today use Gaussian Mixture Models (GMMs) and make the decision by comparing the likelihood of the speaker model to the likelihood of a universal background model (UBM). The paper proposes to replace the UBM by an individual background model (IBM) that is generated for each speaker. The IBM is created using the K-nearest cohort models and the UBM by a simple new adaptation algorithm. The new GMM-IBM speaker verification system can also be combined with various score normalization techniques that have been proposed to increase the robustness of the GMM-UBM system. Comparative experiments were held on the NIST-2004-SRE database with a plain system setting (without score normalization) and also with the combination of adaptive test normalization (ATnorm). Results indicated that the proposed GMM-IBM system outperforms a comparable GMM-UBM system.

Index Terms: Model adaptation, Gaussian Mixture Models, Kullback-Leibler divergence, speaker verification, cohort selection, score normalization.

1. Introduction

The Gaussian mixture model (GMM) is the most widely used approach for statistical modeling of a text-independent speaker verification system. In the standard GMM-UBM system, introduced in [1], the verification is based on a likelihood ratio test between a universal background model (UBM) and the speaker model. The UBM is trained with a large amount of speech samples that embodies a large set of speakers in order to represent the alternative to the target speaker. The target speaker GMM is usually obtained by a certain Bayesian adaptation of the UBM parameters to the speaker enrollment data.

The basic configuration was subsequently enhanced using score normalization techniques intended to compensate the models for possible (speaker-dependent and speaker-independent) session variabilities and thus to increase the robustness of the decision. The proposed score normalization techniques include zero-normalization (Znorm), handset-normalization (Hnorm), test-normalization (Tnorm), distance-normalization (Dnorm), and their combinations [2]. The mostly used normalization technique is the Tnorm approach [3] and its two recent variants - the adaptive test-normalization (ATnorm) [4] and the KL-Tnorm [5].

In the Tnorm approach, the normalization parameters are estimated using the log-likelihood scores derived at test time from a set of imposter models. The mean, μ_{Tnorm}, and the standard deviation, σ_{Tnorm}, of the imposter scores are then used to transform the target speaker score as follows

$$S_t(X) = \frac{L_t(X) - \mu_{Tnorm}}{\sigma_{Tnorm}} , \qquad (1)$$

where $L_t(X)$ is the log-likelihood with respect to the target speaker model, for observation set X. Indeed, the Tnorm procedure does add computations to the testing phase, however, it makes it up by improving the robustness of the decision. The amount of computations is usually decreased by using a fast scoring mechanism similar to the scoring used for GMM-UBM in [1]. The ATnorm approach [4] follows earlier observations that better performance is obtained when cohort normalization uses speaker-specific selection of cohorts [6].

The method proposed in this paper may also be regarded as aiming to exploit the advantage of selecting speaker-specific cohorts but it follows an entirely different route that attains, as will be demonstrated, better results. We generate speaker-specific background models that replace the UBM for each speaker with a tighter alternative. The suggested background model provides a mixed representation of the general UBM and a set of K, most similar, imposters related to the target speaker. The IBM is generated by adaptation of the UBM to a set of K-nearest speaker models using a new algorithm that requires only the parameters of these models. In the verification phase, the IBM of the claimed speaker replaces the UBM and thus, with no increase in computation. The new GMM-IBM speaker verification system can be used in conjunction with various score normalization techniques that were proposed for the GMM-UBM verification system. In the reported experiments we studied a plain GMM-IBM verification system and its combination with adaptive test normalization (ATnorm). In both settings (plain and with ATnorm) the GMM-IBM system outperformed a comparable GMM-UBM system. The proposed method of adding score normalization to the GMM-IBM boosts its performance significantly for just a little increase in computational load.

The rest of this paper proceeds as follows. The next section describes the algorithm for adapting a GMM to a set of other models. Section 3 describes the creation of the GMM-IBM speaker verification system and its combination with ATnorm. Section 4 presents and analyzes experimental results on the NIST-2004-SRE database and it is followed by conclusions.

2. Parametric adaptation of Gaussian mixture models

This section presents a new and efficient algorithm of the adaptation of a prior GMM to a set of K target GMMs. The adaptation does not require any of the original samples or any simulation of distributed observations. It uses directly and only the parameters of the models. The new algorithm is inspired by recent works on simplifying a mixture model using alternating minimization as presented in [7]. It exploits the fact that the Kullback-Leibler (KL) divergence measure between two Gaussian distributions has a closed form expression as follows. For two Gaussian distributions, $N(\mu_1, C_1)$

and $N(\mu_2, C_2)$ with vector dimension d, the KL-divergence $D_{KL}\left[N(\mu_1,\Sigma_1)||N(\mu_2,\Sigma_2)\right]$ is given by

$$\begin{aligned}\tfrac{1}{2}\Big[&\log\tfrac{|C_2|}{|C_1|} + (C_2{}^{-1}C_1) - d \\ &+ (\mu_1-\mu_2)^T C_2{}^{-1}(\mu_1-\mu_2)\Big] \quad .\end{aligned} \tag{2}$$

The parametric adaptation procedure is carried out as follows. Consider a prior GMM $Q(x)$,

$$Q(x) = \sum_{i=1}^{M} \xi_i q_i(x) \quad \text{where} \quad q_i(x) \sim N(\mu_i, C_i^Q), \tag{3}$$

and a set of K target GMMs, $\{P^k(x)\}_{k=1}^{K}$,

$$P^k(x) = \sum_{i=1}^{N_k} w_i^k p_i^k(x) \quad \text{where} \quad p_i^k(x) \sim N(m_i^k, C_i^k). \tag{4}$$

Let us regard the set of K target models as a single large GMM which is a weighted sum of K mixture models,

$$\sum_{k=1}^{K} \gamma_k P^k(x).$$

The mixing coefficients $\gamma_k \in [0,1]$ with $\sum_{k=1}^{K}\gamma_k = 1$ are used to calibrate the contribution of each target GMM to the adapted model. The goal is to adapt the prior model to the mixed set of target models. First we compute the association probabilities of the target Gaussians to the prior model components. For prior component i and target Gaussian component p_j^k the association probability is given by

$$\Pr(i|p_j^k) = \frac{\xi_i e^{-D_{KL}(p_j^k||q_i)}}{\sum_{m=1}^{M}\xi_m e^{-D_{KL}(p_j^k||q_m)}}, \tag{5}$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence (2). Next, $\Pr(i|p_j^k)$ are used to obtain the new estimates of the adapted model in a controlled manner as follows. For each $i = 1,\ldots,M$ we compute the weights by

$$\hat{\xi}_i = (1-\alpha)\xi_i + \alpha\sum_{k=1}^{K}\gamma_k\sum_{j=1}^{N_k}\Pr(i|p_j^k)w_j^k \;; \tag{6}$$

the means by

$$\hat{\mu_i} = \frac{1}{\hat{\xi}_i}\left[(1-\alpha)\xi_i\mu_i + \alpha\sum_{k=1}^{K}\gamma_k\sum_{j=1}^{N_k}\Pr(i|p_j^k)w_j^k m_j^k\right] \;; \tag{7}$$

and the covariance matrices by

$$\begin{aligned}\hat{C_i^Q} &= \frac{1}{\hat{\xi}_i}(1-\alpha)\xi_i\left[C_i^Q + \mu_i\mu_i^T\right] \\ &+ \frac{1}{\hat{\xi}_i}\alpha\sum_{k=1}^{K}\gamma_k\sum_{j=1}^{N_k}\Pr(i|p_j^k)w_j^k\left[C_j^k + m_j^k {m_j^k}^T\right] \\ &- \hat{\mu_i}\hat{\mu_i}^T \;;\end{aligned} \tag{8}$$

where $\alpha \in [0,1]$ is the adaptation coefficient. Here, α is taken for simplicity to be common for all the estimations of the GMM. The adaptation coefficent controls the extent of the adaptation in the final estimation where the use of a value strictly lower than 1 preserves a certain amount of the old parameters.

3. Individual background model for speaker verification

In order to reduce the number of false detect errors we aim to provide a tighter alternative to the hypothesized speaker. We construct an individual background model (IBM), by adapting the parameters of the UBM to the parameters of the K-nearest cohort models related to the specific speaker. By employing this approach, we maintain a coupling relation between the new IBM and the UBM, which proved to be a robust method for making decision using the likelihood ratio test (LRT). Next, the construction of the IBM is described, followed by an efficient method to combine the GMM-IBM system with adaptive test normalization scoring.

3.1. IBM construction

Our individual background model (IBM) is a model

$$\hat{Q}(x) = \sum_{i=1}^{M}\hat{\xi}_i q_i(x) \quad \text{where} \quad q_i(x) \sim N(\hat{\mu_i}, \hat{C_i^Q}), \tag{9}$$

generated for each speaker by adapting the parameters of a UBM, $Q(x)$ as in Eq. (3), to the parameters of the K-nearest speaker models, $\{P^k(x)\}_{k=1}^{K}$ of the form given in Eq. (4), in a manner described in section 2. In order to select the nearest models, we used the matched-based KL approximation described in [8] to compute the distance between the speaker model and the imposter model.

The IBM works with a smaller number of cohorts than might be expected from the number of cohorts involved in score normalization. In preliminary experiments, our setting performed well with 8 to 16 nearest cohorts, figures that are by far lower than 50 to 75 cohort models that were used in the ATnorm approach in [4] and [5]. We used $K = 15$ models in the subsequently reported experiments. The smaller number of nearest cohorts means a tighter alternative to the target speaker compared to the ATnorm approach. A value of $\alpha < 1$ is of course crucial in order to preserve the information of the universal background model, which may be regarded as a far cohort representation. In our experimental setting, good results were obtained with values between 0.7-0.9. In the results reported below $\alpha = 0.8$ was used. An appropriate value for α depends on other choices of the parameters, like the choice of K, the number of nearest cohorts. This issue needs further investigation. We might add that we found α to also depend on whether the models were created by update of means only or by adaptation of the variances and weights in addition. The trend seems to be that increasing the number of adapted parameters requires lower values of the adapting factor. In the experiments reported in this paper we used an equal weighting, $\gamma_k = \frac{1}{K}$, for all cohort models. In the generation of both, the speaker model and the IBM, only the means were updated.

3.2. IBM and ATnorm

The advantage of the Tnorm over the cohort normalization technique is in that it also considers the variance of the log-likelihoods of the tested sample among the large set of imposter models [3]. Therefore, the performance of our GMM-IBM system should also benefit from applying to it the ATnorm, in the sense of obtaining a more robust decision threshold. In the IBM case, every target speaker model is provided with a different background model. Consequently, the normalization has to be applied over the log of likelihood ratio scores. We start by generating an IBM for each speaker in the cohort set. Note that in the case of an operational system, the cohort will be most likely selected from the pool of target speakers. For each speaker (target or cohort), denoted by index s, the log-likelihood difference is computed by

$$\Lambda_s(X) = L_s(X) - L_s^{ibm}(X) \tag{10}$$

where $L_s^{ibm}(X)$ denotes the log-likelihood with respect to the speaker's IBM. When denoting a target speaker by index t, and its cohort models by indices c_t, the normalization parameters, μ_t and σ_t (mean and standard deviation, respectively), are computed over the scores $\Lambda_{c_t}(X)$ of the selected cohort set for the specific target speaker, as in ATnorm. The normalized score of the target speaker is then computed by

$$S_t(X) = \frac{\Lambda_t(X) - \mu_t(X)}{\sigma_t(X)}. \tag{11}$$

As already mentioned, a typical size of cohort set for ATnorm, C, may run between 50 to 75. In GMM-IBM combined with ATnorm, we apply a fast scoring mechanism by using the original UBM as a reference as follows. First, the top-N components are determined by scoring the observation against the UBM. Next, all log-likelihoods of the speaker models and of their related IBMs are scored using only the corresponding N components. For a UBM of order M and ATnorm-cohort of size C, the target score requires $M + 2(1 + C)N$ Gaussian computations. Thus, an extra of $(1+C)N$ computations are required above a standard GMM-UBM system with ATnorm. In a typical setting where $M = 2048$, $C = 50$, and $N = 10$: the resulting overhead for incorporating ATnorm with GMM-IBM is only 20%.

4. Experimental setup and results

Experiments were conducted, using the NIST-2004-SRE data set [9], to examine the proposed GMM-IBM configuration. A standard GMM-UBM was used as a baseline for performance comparisons. A 25-dimensional feature vector based on Mel-frequency cepstrum processing (MFCC) was used, including 12 cepstral coefficients (energy excluded), delta-log energy, and 12 cepstral derivatives, according to the ETSI standard [10]. Cepstral mean substraction and feature warping with 3 seconds window [11] were applied. An energy based voice activity detector was used to remove non-speech frames. A gender-independent universal background model (UBM) of 2048 Gaussians was trained, and used to generate the target speakers' GMMs, by MAP adaptation of the means as in [1]. The UBM was also used to generate the IBM as presented in section 3. The SPIDRE corpus [12] and a randomly selected subset of the NIST-2004 data set were used for generating the UBM and for the generation of the cohort models. For reference, we used ATnorm [4] for score normalization, where the C-nearest cohort models were selected using match-based KL-approximation [8] also referred to as KL-Tnorm in [5]. Gender-dependent cohort model pools were used (including 145 male and 148 female models) for model selection. A cohort set of size $C = 50$ was used for the ATnorm configuration.

The setting was comprised of 400 target models (148 male, 252 female) that were trained using 400 single-sided conversations, and 698 single-sided test conversations (279 male, 419 female) drawn from the NIST-2004 database [9]. The duration of each conversation unit was approximately 5 minutes taken from various channel and handset type sources. In the NIST-2004 data set, multi-lingual speakers were included (Arabic, Mandarin, Russian, and Spanish along with English). The reported tests were performed over all conditions (which include cross-language trials). In order to increase the number of trials, every target model was tested against every available test session of the same gender. A total of 41292 male trials (of which 513 same speaker) and 105588 female trials (of which 824 same speaker) were performed.

We begin by illustrating the impact of choosing a different adaptation coefficient, α, in the generation of the IBM. Figure 1 brings detection error tradeoff (DET) plots of a basic GMM-IBM system obtained for some different values of α. The best performance was found to be in the range $[0.7, 0.9]$ presented in this figure by the value $\alpha = 0.8$.

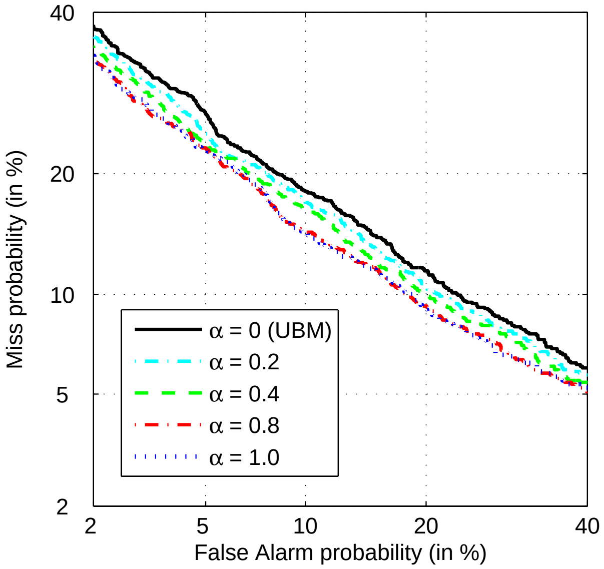

Figure 1: Results of GMM-IBM with variable adaptation coefficient, α, on NIST-2004. The cohort set size for the IBM is $K = 15$.

The performance of comparable GMM-IBM and GMM-UBM systems in our experiments are presented in Table 1 and in Figure 2. The generation of the IBM in these experiments used the value $\alpha = 0.8$ for the adaptation coefficient and a close-cohort set of size $K = 15$. Table 1 presents the equal error rate (EER) (in %) and the minimum decision cost function (DCF) as defined in NIST-2004 evaluation [9]. Figure 2 brings the corresponding DET plots.

The GMM-IBM configuration is seen to perform better than a standard GMM-UBM system. The plain GMM-IBM even compares well with a GMM-UBM system that uses ATnorm attaining a similar EER with only a slightly lower DCF figure. However, the DET curve tends in favor of the GMM-IBM toward the low miss detection rates. When the IBM is combined with ATnorm, further improvement is achieved and all figures outperform the figures of a GMM-UBM with ATnorm. Again, it is observed that the ATnorm affects the performance mostly in the area of lower false alarm rates. In both cases, the new GMM-IBM configuration offers a substantial improvement over a corresponding GMM-UBM configuration.

The explanation for the improved performance of the

System	EER (%)	improv. over UBM	minDCF	improv. over UBM
UBM (baseline)	14.5	-	0.0550	-
IBM	13.0	10%	0.0515	6.4%
UBM+ATnorm	13.2	9%	0.0490	11%
IBM+ATnorm	12.1	16.5%	0.0460	16.4%

Table 1: EER and minDCF for GMM-IBM system compared to the baseline GMM-UBM without score normalization and GMM-IBM compared to GMM-UBM with ATnorm.

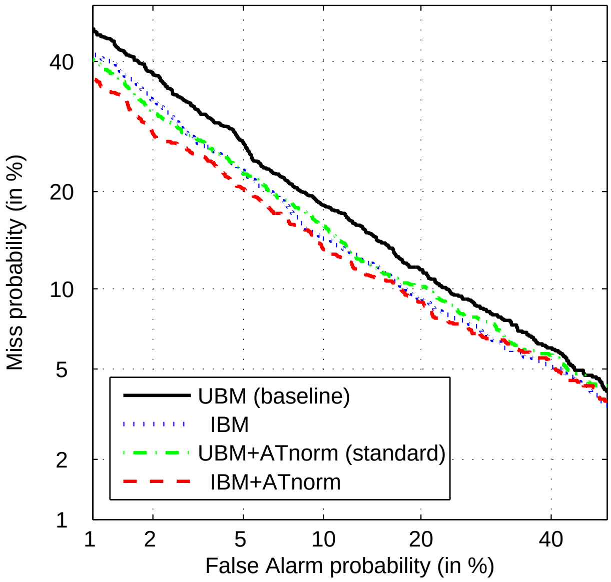

Figure 2: DET curves comparing GMM-IBM to GMM-UBM without and with adaptive test-normalization (ATnorm).

GMM-IBM (in the framework of likelihood ratio test with a single decision threshold) is that the IBM is strongly coupled to the UBM as it is generated by adaptation of the UBM to cohort models that were also adapted from the same UBM. Hence, the resulting score variability is not increased dramatically when using different backgrounds in testing. However, the GMM-IBM system provides a better discrimination between a true speaker and close imposters.

5. Conclusions

The paper proposed a GMM-based speaker verification system with individual background model constructed for each target speaker. The individual background model (IBM) is generated by adaptation of the universal background model (UBM) to the parameters of K-nearest cohort models using a new parametric adaptation method. The paper studied the performance of the proposed GMM-IBM system in a plain setting (without score normalization) and also in combination with adaptive test normalization (ATnorm). In experiments that were conducted on a customary speaker verification task in both settings (plain and with ATnorm), the GMM-IBM system outperformed the corresponding GMM-UBM system.

6. References

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing* 10, pp. 19-41, Jan. 2000.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A tutorial on text-independent speaker verification”, EURASIP J. Appl. Signal Process., vol. 4, pp. 430-451, 2004.

[3] C. Auckenthaler and L. Thomas, “Score normalization for text-independent speaker verification systems”, *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42-54, 2000.

[4] D. Sturim and D. A. Reynolds, “Speaker adaptive cohort selection for Tnorm in text-independent speaker verification”, *Proc. of ICASSP*, pp. 741-744, 2005.

[5] D. R. Castro, J. F. Aguilar, J. G. Rodriguez, and J. O. Garcia, “Speaker verification using speaker- and test-dependent fast score normalization”, *Pattern Recogn. Lett.*, vol. 28, no. 1, pp. 90-98, 2007.

[6] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, *Speech Communication* vol. 17, pp. 91-108, 1995.

[7] J. Goldberger, H.K. Greenspan, and J. Dreyfuss, “Simplifying mixture models using the unscented transform”, *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1496-1502, Aug. 2008.

[8] J. Goldberger and H. Aronowitz, “A distance measure betweem GMMs based on the unscented transform and its application to speaker recognition”, *Proc. of INTERSPEECH*, pp. 1985-1988, 2005.

[9] “The NIST Year 2004 Speaker Recognition Evaluation Plan”, NIST, Gaithersburg, MD. [Online]. Available: http://www.nist.gov/speech/ tests/spk/2004

[10] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms, [Online]. Available: http://www.etsi.org/stq

[11] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification”, in Proc. ISCA Odyssey Workshop, 2001, pp. 213-218.

[12] Linguistic Data Consortium, SPIDRE documentation file, http://www.ldc.upenn.edu/Catalog/readme-files/spidre.readme.html