A SESSION-GMM GENERATIVE MODEL USING TEST UTTERANCE GAUSSIAN MIXTURE MODELING FOR SPEAKER VERIFICATION

Hagai Aronowitz[1], David Burshtein[2] and Amihood Amir[1,3]

[1]Department of Computer Science, Bar-Ilan University, Israel
[2]School of Electrical Engineering, Tel-Aviv University, Israel
[3]College of Computing, Georgia Tech, USA
aronowc@cs.biu.ac.il, burstyn@eng.tau.ac.il, amir@cs.biu.ac.il

ABSTRACT

Test-utterance parameterization (TUP) using Gaussian Mixture Models (GMMs) has recently shown to be beneficial for speaker indexing due to its computational efficiency and identical accuracy compared to classic GMM-based recognizers. In this paper we show that TUP can also lead to more accurate speaker recognition. On the NIST-2004 evaluation corpus, recognition error rate was reduced by 8% compared to the classic GMM-based algorithm. Furthermore, we introduce a novel generative statistical model for generation of test utterances by speakers. This model is incorporated naturally into the TUP framework and improves speaker recognition accuracy. On the NIST-2004 evaluation corpus, recognition error rate was reduced by 15% compared to the classic GMM-based algorithm.

1. INTRODUCTION

The GMM algorithm [1-3] has been the state-of-the-art algorithm for speaker recognition for many years. The GMM algorithm calculates the log-likelihood of a test utterance given a target speaker by fitting a parametric model to the target training data and computing the average log-likelihood of the test-utterance feature vectors assuming frame independence. In [4] a new speaker recognition technique was presented. The idea is to train GMMs not only for target speakers but also for the test utterances. The likelihood of a test utterance is approximated using only the GMM of the target speaker and the GMM of the test utterance.

In [4] we addressed the task of speaker indexing in large audio archives. Our motivation for representing a test utterance by a GMM was the distributive nature of the speaker recognition algorithm based on test-utterance parameterization (TUP). The fact that fitting a GMM to a test utterance is independent of the target speaker enables fitting the GMM during the indexing process (before the archive is queried), hence reducing the time complexity of searching for a target speaker. In this paper we explore a different attribute of the TUP algorithm. We claim that TUP is more flexible for implementation of more complex generative models than just the simple frame-independence-based generative model implied by the classic GMM algorithm. In addition, we claim that the TUP algorithm exploits a-priori knowledge about test utterances, namely the smoothness of their distribution. Using universal background model (UBM) MAP-adaptation for fitting the GMM exploits additional a-priori knowledge. Therefore there is a potential for exceeding the classic GMM's accuracy.

In this paper we suggest a novel generative model for generation of test utterances by speakers. We model each speaker by a prior distribution over all GMMs, and assume that at the beginning of a spoken session a GMM is selected from the speaker's prior distribution. The frames are generated independently using the selected GMM. The suggested generative model is a generalization of the simple generative model used by the classic GMM system where the prior distribution assigns a non-zero probability only to a single GMM. The motivation for the new model we present is that there is apparently considerable intra-session dependency that may be attributed to channel, noise, and temporary speaker characteristics (mood, fatigue, etc.). It is reasonable to assume that these factors are constant during a single session but change between sessions.

In this paper we propose a simple prior distribution over the GMM space and present an algorithm to train this distribution and use it to compute the likelihood of a test utterance given a target speaker.

The organization of this paper is as follows: we overview the TUP-based speaker recognition algorithm in section 2. We present the new generative model in section 3. Section 4 describes the experimental corpus, the

experiments, and the results. Finally, section 5 presents conclusions and proposed future work.

2. GMM SCORING USING TEST UTTERANCE PARAMETERIZATION

In this section we overview the TUP based speaker recognition algorithm. Our goal is to simulate the calculation of the log-likelihood of a test utterance X given a GMM Q by using a GMM fitted to the test utterance. The log-likelihood of X given Q is calculated in equation (1):

$$\frac{1}{n} LL(X|Q) = \frac{1}{n} \sum_{i=1}^{n} \log(\Pr(x_i|Q)) \quad (1)$$

2.1. GMM simulation

The vectors $x_1,\ldots,x_n$ of the test utterance are acoustic observation vectors generated by a stochastic process. Let us assume that the true distribution of which the vectors $x_1,\ldots,x_n$ were generated by is P. The average log-likelihood of an utterance X of asymptotically infinite length $|X|$ generated by the distribution P is given in equation (2):

$$\frac{1}{n} LL(X|Q) = \frac{1}{|X|} \sum_{i=1}^{|X|} \log(\Pr(x_i|Q)) \xrightarrow[|X| \to \infty]{} \int_x \Pr(x|P) \log(\Pr(x|Q)) dx \quad (2)$$

2.2. Estimation of distributions *Q and P*

We assume that the test utterance is generated using a true distribution P. In [4] we estimated P by adapting the UBM using MAP. In this paper we also tried to estimate P and Q using several EM-iterations with the UBM used as the prior distribution for each iteration.

2.3. Calculation of $\int_x \Pr(x|P) \log(\Pr(x|Q)) dx$

Definitions:

w_i^P, w_j^Q : The weight of the i[th]/ j[th] Gaussian of distribution P/Q.

$\mu_{i,d}^P, \mu_{j,d}^Q$: The d[th] coordinate of the mean vector of the i[th]/ j[th] Gaussian of distribution P/Q.

$\sigma_{i,d}^P, \sigma_{j,d}^Q$: The d[th] coordinate of the standard deviation vector of the i[th]/j[th] Gaussian of distribution P/Q.

G: The number of Gaussians of distribution P.

dim: The dimension of the acoustic vector space.

C_1-C_3: Constants.

In [4] we show that the average likelihood of an utterance X given a target speaker Q can be approximated using equation (3):

$$\frac{1}{n} LL(X|Q) \cong \int_x \Pr(x|P) \log(\Pr(x|Q)) dx \cong \sum_{i=1}^{G} w_i^P \max_j \left\{ \begin{array}{l} \log w_j^Q - \sum_{d=1}^{\dim} \frac{(\mu_{i,d}^P - \mu_{j,d}^Q)^2}{2\sigma_{j,d}^{Q\,2}} \\ - \sum_{d=1}^{\dim} \log \sigma_{j,d}^Q - \frac{1}{2} \sum_{d=1}^{\dim} \left(\frac{\sigma_{i,d}^P}{\sigma_{j,d}^Q} \right)^2 \end{array} \right\} + C_1 \quad (3)$$

In [4] we have shown an efficient technique for calculating the right hand side of equation (3).

2.4. Global variance models

Global variance GMM models are GMMs with the same diagonal covariance matrix shared among all Gaussians and all speakers. Using global variance GMMs has the advantages of lower time and memory complexity and can also result in better accuracy as can be seen in section 4. Applying the Global variance assumption to equation (3) results in a much simpler equation (4):

$$\frac{1}{n} LL(X|Q) \cong \sum_{i=1}^{G} w_i^P \max_j \left\{ \log w_j^Q - \sum_{d=1}^{\dim} \frac{(\mu_{i,d}^P - \mu_{j,d}^Q)^2}{2\sigma_d^{\,2}} \right\} + C_2 \quad (4)$$

2.5. Analyzing the role of Gaussian weights

For improved time complexity it may be very appealing to use the non-adapted weights of the UBM for every speaker and test utterance or even set all weights to a constant. Several papers such as [5] reported no degradation in accuracy when using the UBM's weights for all target speakers. We tested the sensitivity of both the baseline GMM system and the TUP-based system to using the UBM's weights or replacing the weights by a constant for the target speaker model (Q). We also tried the same techniques for the GMMs fitted to test utterances (P) which may be interpreted as normalizing mismatch between train and test. Assuming all weights to be constant and assuming global variance results in equation (5):

$$\frac{1}{n} LL(X|Q) \cong -\frac{1}{G} \sum_{i=1}^{G} \min_j \left\{ \sum_{d=1}^{\dim} \frac{(\mu_{i,d}^P - \mu_{j,d}^Q)^2}{2\sigma_d^{\,2}} \right\} + C_3 \quad (5)$$

3. SESSION-GMM GENERATIVE MODEL

The classic GMM algorithm assumes that each speaker can be modeled by a single GMM. The generative model implied is that each frame is emitted by that single GMM independently from other frames. Consequently, if 2 utterances are spoken by the same speaker and are long

enough, they should have identical empirical distributions (when length approaches infinity). Unfortunately, that is not the case. In reality there exist session-dependent factors that cause the distribution of different sessions of the same speaker to inherently deviate from each other.

3.1. Suggested generative model

We define the term session-GMM as the GMM distribution used to generate the frames of a single session. We model each speaker as a prior distribution over session-GMMs. Therefore, the likelihood of an utterance *X* given speaker *S* is:

$$\Pr(X|S) = \int_{GMM} \Pr(X|GMM)\Pr(GMM|S)\,dGMM \quad (6)$$

Correspondingly, the likelihood of an utterance *X* given the UBM is:

$$\Pr(X|UBM) = \int_{GMM} \Pr(X|GMM)\Pr(GMM|UBM)\,dGMM \quad (7)$$

In order to develop simple and tractable training and scoring algorithms we approximate equations (6, 7). We assume that the distribution Pr(*X*|*GMM*) is much sharper than distributions Pr(*GMM*|*S*) and Pr(*GMM*|*UBM*). Therefore, defining:

$$P = \arg\max_{GMM}\{\Pr(X|GMM)\} \quad (8)$$

Equations (6, 7) can be approximated:

$$\Pr(X|S) \cong \Pr(P|S) \quad (9)$$

$$\Pr(X|UBM) \cong \Pr(P|UBM) \quad (10)$$

Note that assuming that distribution Pr(*GMM*|*S*) is much sharper than distribution Pr(*X*|*GMM*) results in the classic GMM algorithm.

3.2. Generative model for a session-GMM

According to our experimental results (section 4) each session-GMM is represented by a global variance model with equal weights. Therefore we only have to model the means on the GMM. We assume that the dth coefficient of the mean vector of the ith Gaussian of speaker *S* distributes normally with a mean $\mu_{i,d}^{S}$ and a variance $(\sigma_{i,d})^2$. Note that we chose the variance to be speaker-independent because currently we want to avoid using several training sessions per speaker.

3.2. Training the prior distribution models

In order to train the means of the distribution $\{\mu_{i,d}^{S}\}$ we train a GMM Q_S for target speaker *S* and the means of Q_S are the Maximum Likelihood estimate for $\{\mu_{i,d}^{S}\}$.

In order to train the speaker-independent standard deviations $\{\sigma_{i,d}\}$ we take pairs of same speaker sessions from a development corpus. For each pair we train GMMs and calculate the difference of the corresponding means of the GMMs: $\delta_{i,d} = \mu_{i,d}^{1} - \mu_{i,d}^{2}$. $\delta_{i,d}$ is a random variable with zero mean and variance= $2(\sigma_{i,d})^2$. Therefore we can estimate $\sigma_{i,d}$ from $\{\delta_{i,d}\}$ calculated over pairs of different speakers.

3.3. Calculation of $P = \arg\max_{GMM}\{\Pr(X|GMM)\}$

P is estimated using MAP adaptation of the UBM model.

3.4. Calculation of $\Pr(X|S)$, $\Pr(X|UBM)$

According to equations (9, 10), the log-likelihood of utterance *X* given speaker *S* is presented in equation (11):

$$\log\Pr(X|S) = -\frac{\dim}{2}\log 2\pi - \sum_{i=1}^{G}\sum_{d=1}^{\dim}\left(\log\sigma_{i,d} - \frac{\left(\mu_{i,d}^{P} - \mu_{i,d}^{S}\right)^2}{2(\sigma_{i,d})^2}\right) \quad (11)$$

The likelihood of utterance *X* given the UBM is similar to equation (11) with trivial modifications.

4. EXPERIMENT AND RESULTS

4.1. The SPIDRE and the NIST-2004 corpuses

The GMM baseline and the TUP system were first tuned on the SPIDRE corpus [7]. Experiments were done on the NIST-2004 speaker evaluation data set [8]. The primary data set was used for selecting both target speakers and test data. The data set consists of 616 1-sided single conversations for training 616 target models, and 1174 1-sided test conversations. All conversations are about 5 minutes long and originate from various channels and handset types. In order to increase the number of trials, each target model was tested against every test session. The SPIDRE corpus was used for training the UBM, for training the speaker-independent variances of the prior distribution models and for development data.

4.2. The baseline GMM system

The baseline GMM system in this paper was inspired by the GMM-UBM system described in [1-3]. A detailed description of the baseline system can be found in [4]. The baseline system is based on an ETSI-MFCC [6] front-end

+ derivatives and an energy based voice activity detector. In the verification stage, the log likelihood of each conversation side given a target speaker is divided by the length of the conversation and normalized by the UBM score. The resulting score is then normalized using z-norm [2].

4.2. Accuracy of the TUP system compared to the GMM baseline system

In table (1) we summarize selected results of our experiments on the TUP system compared to the baseline system.

	Miss probability (in %)		
	fa=1%	fa=5%	fa=10%
Baseline GMM	43.1	27.6	19.3
GMM non-GVAR	46.2	27.9	20.1
TUP	44.5	25.9	19.5
TUP + EM training	41.9	25.2	18.1
TUP + no weights for parameterization of test utterances	40.1	25.3	17.8
(Error reduction compared to baseline)	(7%)	(8%)	(8%)

Table 1: Results of selected TUP experiments compared to the baseline GMM system.

Note that the classic non-global variance GMM system performed worst than the global variance system. The results of the TUP system are very similar to the results of the baseline GMM.

4.2.1. EM training

EM training does improve performance when training GMMs for test utterances (3-7% reduction in misdetection). Trying to train models for target speakers using the EM algorithm degrades accuracy for both the baseline system and the TUP system.

4.2.1. Constant weights

Using constant weight for parameterization of test utterances improves performance (2-10% reduction in misdetection). Trying to use the same technique for target speaker models or just using the non-adapted weights of the UBM degraded accuracy for both the baseline system and the TUP system.

4.3. Accuracy of the session-GMM generative model

In table (2) we present results for the session-GMM generative model compared to the baseline GMM.

	Miss probability (in %)		
	fa=1%	fa=5%	fa=10%
Baseline GMM	43.1	27.6	19.3
Session-GMM generative model	37.2	23.3	16.7
Error reduction	14%	16%	13%

Table 2: Results of the session-GMM generative model compared to the baseline GMM system.

5. CONCLUSIONS

We have proposed two speaker recognition algorithms: a modified TUP based algorithm and a session-GMM-generative-model-based algorithm. The first one reduces recognition error by about 8% in various false acceptance rates and the second one reduces recognition error by about 15%. The Session-GMM generative model we implemented was a simple one and we intend to explore more complex models, for example by modeling correlation between the Gaussian means.

6. REFERENCES

[1] Reynolds D. A., T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing* , Vol. 10, No.1-3, pp. 19-41, 2000.

[2] Reynolds, D. A., "Comparison of background normalization methods for text-independent speaker verification", in Proc. Eurospeech, pp.963-966, 1997.

[3] McLaughlin J., D. A. Reynolds, and T. Gleason, "A study of computation speed-ups of the GMM-UBM speaker recognition system", in Proc. Eurospeech, pp.1215-1218, 1999.

[4] Aronowitz H., D. Burshtein D., A. Amir, "Speaker indexing in audio archives using test utterance Gaussian mixture modeling", to appear in Proc. ICSLP, 2004.

[5] Ben M., F. Bimbot, "D-MAP: A distance-normalized MAP estimation of speaker models for automatic speaker verification", in Proc. ICASSP, pp. 69-72, 2003.

[6] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," ETSI Standard: ETSI-ES-201-108-v1.1.2, 2000, http://www.etsi.org/stq.

[7] Linguistic Data Consortium, SPIDRE documentation file, http://www.ldc.upenn.edu/Catalog/readme_files/.

[8] "The NIST Year 2004 Speaker Recognition Evaluation Plan", http://www.nist.gov/speech/tests/spk/2004/.