Towards an Optimal Speaker Modeling in Speaker Verification Systems using Personalized Background Models

Ayoub Bouziane, Jamal Kharroubi, Arsalane Zarghili
Laboratory of Intelligent Systems and Applications, University Sidi Mohamed Ben Abdellah
Fez, Morocco.

Article Info

Article history:

Received Mar 19, 2017
Revised Jun 29, 2017
Accepted Jul 14, 2017

Keyword:

Speaker modeling
Speaker verification
Gaussian mixture models
Universal background model
Maximum a posteriori (MAP)
Personalized background models

ABSTRACT

This paper presents a novel speaker modeling approachfor speaker recognition systems. The basic idea of this approach consists of deriving the target speaker model from a personalized background model, composed only of the UBM Gaussian components which are really present in the speech of the target speaker. The motivation behind the derivation of speakers' models from personalized background models is to exploit the observeddifference insome acoustic-classes between speakers, in order to improve the performance of speaker recognition systems.

The proposed approach was evaluatedfor speaker verification task using various amounts of training and testing speech data. The experimental results showed that the proposed approach is efficientin termsof both verification performance and computational cost during the testing phase of the system, compared to the traditional UBM based speaker recognition systems.

Copyright © 2017Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Ayoub Bouziane,
Laboratory of Intelligent Systems and Applications,
Sidi Mohamed Ben Abdellah University,
P.B: 2202 Immouzer Road, Fez, Morocco.
Email: ayoub.bouziane@usmba.ac.ma

1. INTRODUCTION

The GMM-UBM based speaker modeling approach was firstly introduced to the speaker recognition community by Reynolds, in 2000 [1]–[3]. Since then, it have become the predominant approach for speaker modeling in text-independent speaker recognition systems, and the basis of the most successful approaches that have been emerged in the last decade: the hybrid GMM-SVM approach [4], [5], the joint factor analysis approach [6]–[8], and the recently introduced i-vectors approach [9], [10]. The motivation behind the use of adapted Gaussian mixture models for speaker modeling is generally based on the assumption that Gaussian densities may model a set of hidden acoustical classes that reflect some general speaker dependent vocal tract characteristics.

The main idea of the UBM-based speaker recognition systems consist of deriving the target speaker model from a universal background model that represents the set of all human acoustic classes (e.g., gender dependent acoustic classes, age-dependent acoustic classes, accent dependent acoustic classes…). The target speaker modelwill therefore be composed of an adapted version of the various acoustic classes defined in the universal background model. However, as it sounds to the human ear, speakers don't share the same acoustic classes. For example, the acoustic classes of male speakers are different from those of female speakers and the acoustic classes of adult speakers are different from those of minor speakers, as well as, the acoustic classes of timid speakers are different from those of bold speakers, etc. Thereby, it seems that it is not logical to model a speaker by an adapted version of anacoustic class which may not exist in its voice. Furthermore, the difference of acoustic classes between speakers can be exploited to discriminate between them in speaker recognition systems.

In view of these observations, the present study attempts, firstly,to experimentally investigate the degree of acoustic-classes'difference between speakers,and secondly, to propose a speaker modeling approach that takes into account these observations and exploits the differences in acoustic classes between speakers to improve the performance of speaker recognition systems.

The remainder of this paper is organized as follows. The first section gives a brief overview of the traditional GMM-UBM based Speaker verification systems. Next, the proposed speaker modeling approach, based on personalized background models, is introduced. Afterward, experimental resultsand discussions are presented in the third section. Finally, conclusions and future research directions are drawn in the last section.

2. TRADITIONAL GMM-UBM BASED SPEAKER VERIFICATION SYSTEMS

The GMM-UBM based speaker verification system was originally proposed by Reynolds in 2000 [1], [11]. Since then, it has become the predominant approach for speaker modeling in text-independent speaker recognition systems and the basis of the most successful approaches that have been emerged in the last decade. The main idea of the GMM-UBM approach consists, as shown in Figure 1, of deriving speakers' models from a universal background model using maximum a posteriori (MAP) adaptation [1]. The Universal Background Model (UBM) is typically a Gaussian mixture model (GMM) that represents the distribution of the entire acoustic space of speech [2], [3].

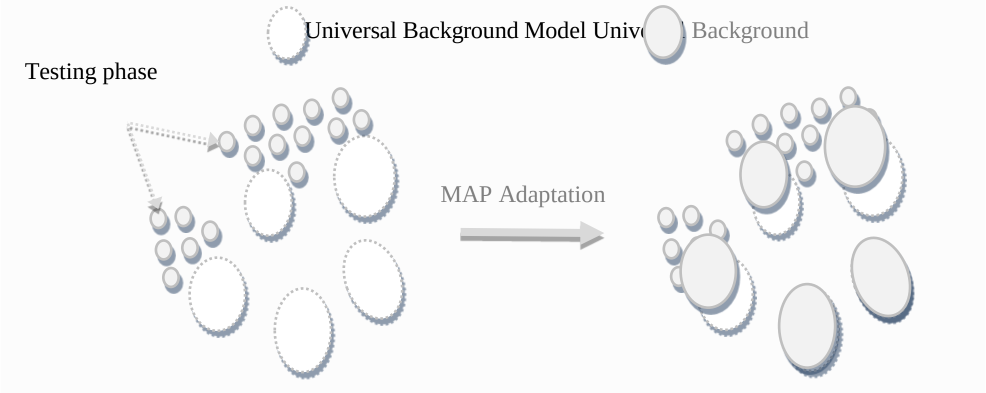


Figure 1. A MAP adaptation of a GMM comprises 5 Gaussian densities. Original Gaussian densities of the UBM are depicted as unfilled ellipses (dotted line), whereas the adapted Gaussian densities are denoted by filled ellipses, and the observed feature vectors are depicted as small circles

The MAP adaptation process is generally composed of two steps. First, the sufficient statistics estimates of the speaker's training data are computed for each mixture in the UBM. Next, the computed sufficient statistics estimates are combined with the old sufficient statistics from the UBM mixture parameters using a data-dependent mixing coefficient. The specifics of the adaptation are as follows. Given that the universal background model is composed of M Gaussian components, each of which is parameterized by a mean vectorμ_i, variance matrix $\sigma^2{}_i$ and its weight$w_i$in the mixture model. Initially, the posteriori probability of each UBM component $\{\mu_i, \sigma_i, w_i\}$ given the feature vector x_t is computed as follows:

$$Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{L} w_i p_j(x_t)} \tag{1}$$

where $p_i(x_t)$ is the density of the probability, given by:

$$p_i(x_t) = \frac{1}{(2\pi)^{D/2}|\sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(x_t-\mu_i)^T\sigma_i^{-1}(x_t-\mu_i)\right\} \tag{2}$$

The sufficient statistics for the weight, mean, and variance parameters are then computed as follows:

$$n_i = \sum_{t=1}^{T} Pr(i|x_t)\ ;\ E_i(x) = \frac{1}{n_i}\sum_{t=1}^{T} Pr(i|x_t)\, x_t\ ;\ E_i(x^2) = \frac{1}{n_i}\sum_{t=1}^{T} Pr(i|x_t)\, {x_t}^2 \tag{3}$$

Thereafter, the computed sufficient statistics are used for estimating the adapted mixture weights $\widehat{w_i}$, means $\widehat{\mu_i}$ and variances $\widehat{\sigma_i}$ of the given speaker:

$$\widehat{\mu_i} = \left[\beta_i^{\mu} E_i(x) + \left(1 - \beta_i^{\mu}\right)\mu_i\right] \tag{4}$$

$$\widehat{w_i} = [\beta_i^{w} n_i)/T + (1 - \beta_i^{w})w_i]\gamma \tag{4}$$

$$\widehat{\sigma^2}_i = \left[\beta_i^{\sigma^2} E_i(x^2) + \left(1 - \beta_i^{\sigma^2}\right)\left({\sigma^2}_i - \beta_i^{\sigma^2}\right) - {\widehat{\mu_i}}^2\right] \tag{5}$$

with, $$\beta_i^{\rho} = \mathrm{n_i}/(\mathrm{n_i} + r^{\rho}), \quad \rho \in \{\mathrm{w}, \mu, \sigma^2\} \tag{6}$$

Here, γ is a scale factor computed over all adapted mixture weights to ensure that they sum to unity, $\beta_i^{\rho}, \rho \in \{w, \mu, \sigma^2\}$ are the adaptation coefficients that control how the adapted GMM parameters will be affected by the observed speaker data, and $r^{\rho}$is a fixed relevance factor for parameter ρ.

An overall diagram of the GMM-UBM based speaker verification system is shown in Figure 2. The basic operating structure of the system, as shown in Figure 2, is composed of three phases: the training phase, the enrollment phase and the testing phase. During the first phase, i.e. the training phase, a large collection of speech utterances is collected from a background population of speakers, their corresponding feature vectors are extracted and used to train the universal background model. The training process of the UBM is done generally using maximum likelihood estimation via the EM algorithm. In the second phase, i.e. the enrollment phase, speaker models of new client speakers are derived from the universal background model through MAP adaptation using the speakers' training feature vectors [13]. While in the testing phase, the extracted feature vectors of the unknown speaker's utterance $X^u = \{{x^u}_1, {x^u}_2, \ldots, {x^u}_N\}$ are compared against both the claimed target speaker model and the background model. The log likelihood ratio $\mathrm{LLR}(X^u; \lambda_{spk}, \lambda_{UBM})$ between the claimed speaker model and the universal background model is then calculated and used to make a decision about the acceptance/rejection of the claimed identity. The log likelihood ratio (LLR) of the test utterance X^u between the speaker model λ_j and the UBM model λ_{UBM}:

$$LR(X^u; \lambda_{spk}, \lambda_{UBM}) = \frac{1}{N}\left[\sum_{i=1}^{N} log\, p({x^u}_i|\lambda_{spk}) - \sum_{i=1}^{N} log\, P({x^u}_i|\lambda_{UBM})\right] \tag{7}$$

With, $$p(x_t|\lambda) = \sum_{i=1}^{M} w_i b_i(x_t) \tag{8}$$

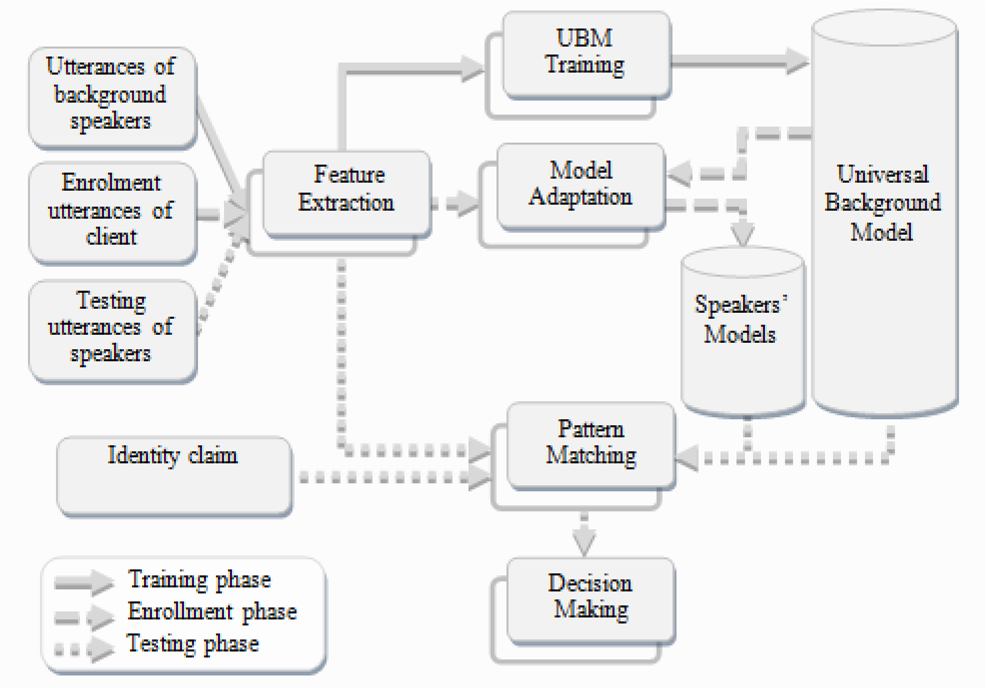


Figure 2. Block diagram of the GMM-UBM based speaker verification system

3. SPEAKER RECOGNITION USING PERSONALIZED BACKGROUND MODELS (PBMS)

The main idea of the proposed PBM-based speaker modeling approach consistsof adapting the target speaker model from a personalized background model (PBM), composed only of the UBM Gaussian components which are actually present in the speaker's speech. The MAP adaptation step of traditional UBM based systems will, therefore, be preceded by a selection step that selectsthe background Gaussian component swhich reflect the general form of the acoustic classes characterizing the speaker, see Figure 3.

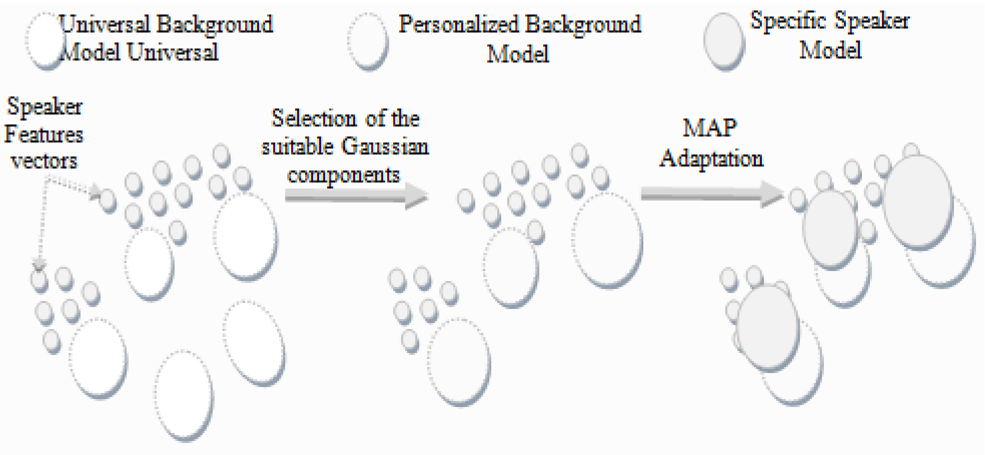

Universal Background Model Universal
Personalized Background Model
Specific Speaker Model
Speaker Features vectors
Selection of the suitable Gaussian components
MAP Adaptation

Figure 3. A MAP adaptation of a GMM comprises 5 Gaussian densities. Original Gaussian densities of the UBM are depicted as unfilled ellipses (dotted line), whereas the adapted Gaussian densities are denoted by filled ellipses, and the observed feature vectors are depicted as small circles.

Given a UBM of M Gaussian component $\lambda_{UBM} = \{\mu_{UBM_i}, \sigma_{UBM_i}, w_{UBM_i}\}/\ i \in \{1,2, \dots, N\}$, and M training feature vectors $X = \{x_1,\ x_2, \dots ,\ x_N\}$, extracted from the target speaker's speech.The PBMGaussian components are generally chosen from the UBM using a winner-take-all based strategy. A UBM Gaussian component $\theta_{\mathrm{UBM}_i} = \left(\mu_{\mathrm{UBM}_i}, \sigma_{\mathrm{UBM}_i}, \mathrm{w}_{\mathrm{UBM}_i}\right)$ isselected to belong to the personalized background model λ_{PBM} of the target speaker, if there is at least one feature vector $\mathrm{x}_\mathrm{n} \in \mathrm{X}$, where the UBM Gaussian component $\theta_{\mathrm{UBM}_\mathrm{i}}$achieves the maximum posterior probability of x_n belongingness:

$$Pr\left(\theta_{UBM_i}|x_\mathrm{n}\right) \geq Pr\left(\theta_{UBM_l}|x_\mathrm{n}\right), \forall l \in \{1,2,\dots,\mathrm{N}\} \qquad (9)$$

Once the PBM Gaussian components are selected, the weights $\mathrm{w}_{UBM_\mathrm{j}}$of the selected components are divided by their sum so that the total weight is equal to unity.

A block diagram of the speaker modeling process using personalized background models is shown in Figure 4. Firstly, the training feature vectors of the target speaker are extracted from itsenrollment utterances. Next, the extracted feature vectors are used to select the UBM Gaussian components which will compose the speaker's personalized background model.Afterwards, the composed personalized background model is utilized do derive the speaker model using the MAP adaptation procedure. Finally, the adapted model is stored together with the corresponding indices of the PBM Gaussian components in the UBM.

An example of a two-dimensional projectionof two speakers' features and the means of their corresponding UBM and PBM adapted modelsis shown in Figure 5. As it can be seen from this figure, the means of the PBM adapted models fit the speakers' features better than the means of the traditional UBM adapted models. Moreover, it seems that the adaptation of the UBM Gaussian components which haven't any relationship with the target speakerinfluence on the feature vectors belongingness to the appropriate Gaussian components, which therefore affect negatively the adapted model.

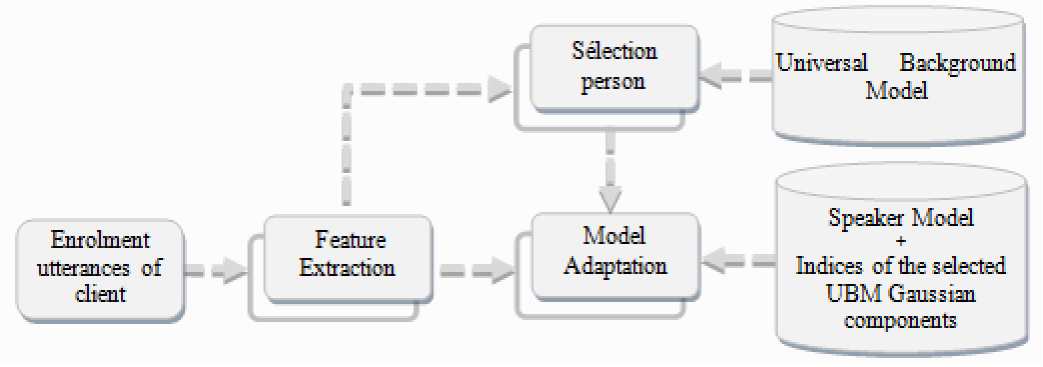
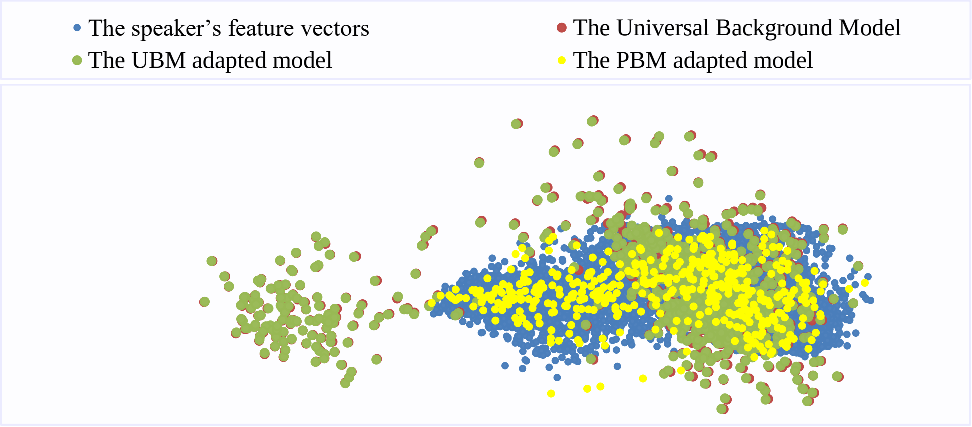


Figure 4. Block diagram of speaker modeling process using personalized background models



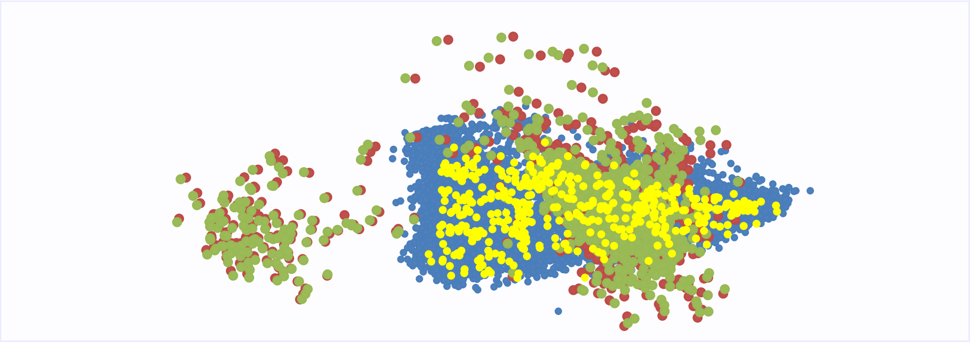


Figure 5. A two-dimensional projection of two speakers' features (blue points), the means of their corresponding UBM and PBM adapted models (in green and yellow points, respectively) and the means of the UBM model (red points)

During the test phase, the log-likelihood ratio $\mathrm{LLR}\left(\mathrm{X^u};\lambda_{\mathrm{spk}},\lambda_{\mathrm{PBM}}\right)$ between the claimed speaker model and the personalized background model is used to make a decision about the acceptance or the rejection of the claimed identity:

$$LR\left(X^u;\lambda_{spk},\lambda_{PBM}\right)=\frac{1}{N}\left[\sum_{i=1}^{N} log\, p({x^u}_i|\lambda_{spk})-\sum_{i=1}^{N} log\, P({x^u}_i|\lambda_{PBM})\right] \quad (10)$$

With, $$p(x_t|\lambda)=\sum_{i=1}^{M} w_i b_i(x_t) \quad (11)$$

The motivation behind the use of the personalized background modelinstead of the universal background model is to penalize the decision score of impostor speakers who don't share the same acoustic classes with the target speaker.

4. EXPERIMENTS, RESULTS AND DISCUSSION

4.1. The Experimental Protocol

The performed experiments in this study were conducted on the THUYG-20 SRE database [14]. This database isgenerally composed of 353 speakers, collected in a controlled environment (silent office by the samecarbon Microphone). The entire speech corpuswas divided into three data sets: the first dataset consists of 200 genderbalanced speakers (100 Male and 100 Female) and devoted totrain the Universal Background Model (UBM), the second and the third data sets are composed of the same set of 153 client speakers.The first dataset comprises the training speech data, whereas the second comprises the testing speech data of the 153 speakers.During the testing phase of the system, the client speakers were tested against each other,resulting in total of 896,886 trials of 4 seconds and 441,558 trials of 8 seconds.

The feature vectors of the overall speech utterances were extracted using the MFCC approach: the digitized speech is firstly emphasized using a simple first order digital filter with transfer function H (z) = 1 – 0.95z. Next, the emphasized speech signal is blocked into Hamming-windowed frames of 25 ms (400 samples) in length with 10 ms (160 samples) overlap between any two adjacent frames.Finally, 19 Mel-Frequency Cepstral Coefficients were extracted from each frame [15]. During the training phase, a universal background model (UBM) of 1024 Gaussian components was trained on the overall training data (7.5 hours of speech) using the EM algorithm.

4.2. Investigation on the Degree of Difference in Acoustic Classes between Speakers

Theaim of the performed experiments in this sectionisto investigate the degree of difference in acoustic classesbetween speakers. For this purpose, we have carried out several speaker verification experiments in which we have basedonly on the difference between the acoustic classes present in the target speaker's speech and those present in the claimed speaker's speech. To proceed, we have represented the training speech utterance(s) of the target speaker and the test speech utterance of the claimed speaker by histograms. Each histogramis composed of 1024 bins, whereeach bin represents aUBM Gaussian component.

The value of each bin is defined as the number of times that the corresponding UBM Gaussian component has the maximum posterior probability over the feature vectors of the speaker. The specifics of the histogram construction process are shown in Algorithm 1.

Algorithm 1 The histogram construction process

```
Input: The feature vectors $X = \{x_1, x_2, \ldots, x_N\}$ of the speech utterance,
       The universal background model $\lambda_{UBM} = \{\mu_{UBM_i}, \sigma_{UBM_i}, w_{UBM_i}\} / i \in \{1,2,\ldots,N\}$.
Output: The corresponding histogram $\mathcal{H}$ of the speech utterance.
    $\mathcal{H} = zeros(1,1024)$
    FOR EACH $x_i$ IN X
        $j = \arg\max_i Pr(i|x_t)$
        $\mathcal{H}(j) = \mathcal{H}(j) + 1$
END
```

Once the histogram of the target and the claimed speakers are constructed, $\mathcal{H}_T$ and $\mathcal{H}_C$ respectively, the comparison between themis done using the Bhattacharyya distance:

$$D(\mathcal{H}_T, \mathcal{H}_C) = -\log \sum_{i=1}^{1024} \sqrt{\mathcal{H}_T(i).\mathcal{H}_C(i)} \tag{12}$$

The computed distance $D(\mathcal{H}_T, \mathcal{H}_C)$ is then used to make a decision about the acceptance or the rejection of the claimed speaker.The obtained results of the performed experiments, while varying the amount of training and testing speech data, are shown in Table1.

Table 1. The obtained EERs using several amount of training and testing speech data.

		Amount of Enrollment Speech Data		
		20 seconds	40 seconds	60 seconds
Amount of Testing Speech Data	4 seconds	6.38	5.90	4.64
	8 seconds	5.86	5.31	3.98

First and foremost, as it can be seen from Table 1, the obtained results are highly encouraging. The lower obtainedequal error rates, based only on the difference in hidden acoustic classes between speakers, reflect a great difference in those hidden acoustic classes between speakers. Additionally, it appears that each increase in the amount of training or testing speech data is translated into better verification performance.This proportional relationshipbetween the amounts of speech data and the performance of the system reflects the fact that the overall acoustic classes of a speaker cannot be assembled in its pronunciation of one or two utterances.

4.3. Assessment of the Proposed GMM-PBM Approach Compared to the Traditional GMM-UBM Approach

The performed experiments in this section attemptto assess the performance of the proposed GMM-PBM based speaker modeling approach, compared to the traditional GMM-UBM based approach. Hence, various experiments were carried out using the two approaches while varying the amount of training and testing speech data. The obtained results are illustrated in Figure 6.

The experimental results show, across the various amounts of training and testing speech data that our proposed approach has achieved a better verification performance compared to the traditional UBM based approach. Even in little amounts of training speech data, where the speaker's acoustic classes may not be all present, our proposed approach demonstrates its enhanced verification performance as compared with the UBM based approach. Furthermore, we can see that the obtained performance under the UBM based system using test utterances of 8 seconds was obtained under our proposed PBM based system using test utterances of only 4 seconds. Moreover, it can be seen that the relative error reduction was doubled when we have doubled the amount of testing utterances.

In addition to its performance advantage, the proposed approach can significantly reduce the CPU time required for speaker verification during the testing phase of the system compared to the traditional system, see Figure 7. In fact, the derivation of speakers' models from personalized background models reduces the order of their adapted models, which consequently, reduces theirstorage and computational costs.

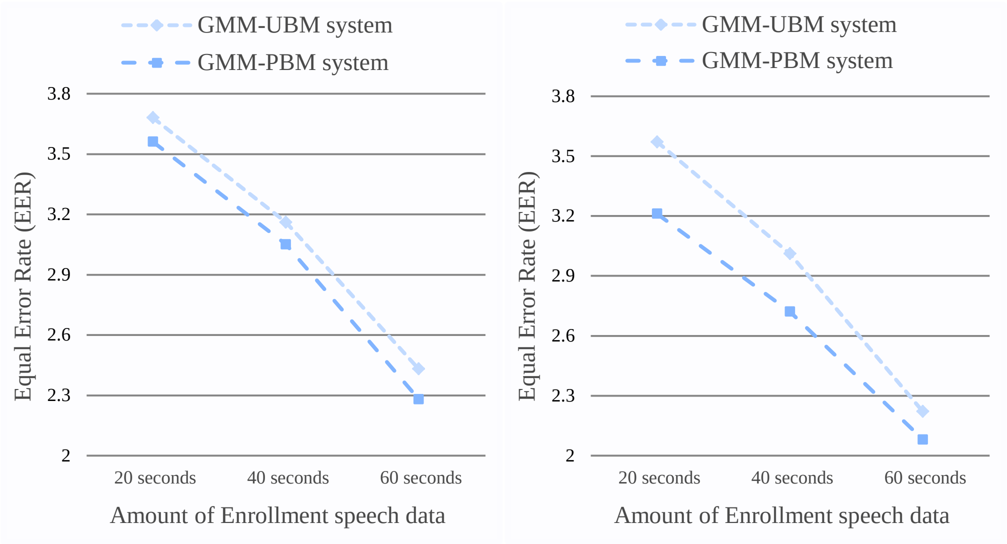


Figure 6. The obtained Equal Error Rates using test utterances of 4 seconds (left figure) and 8 seconds (right figure)

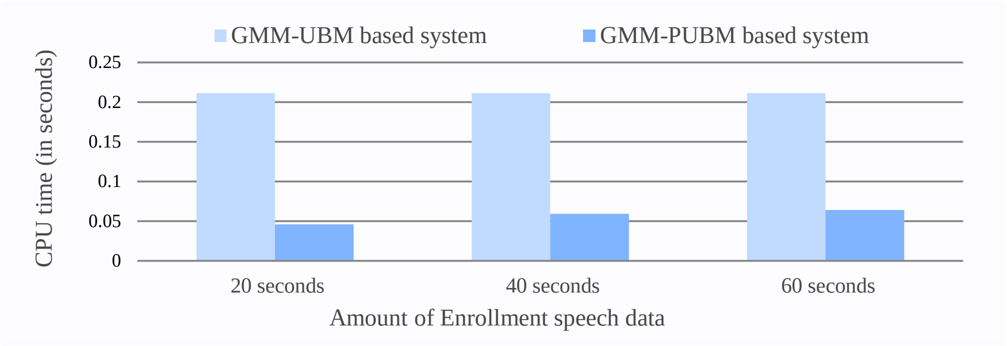


Figure 7. CPU time required for speaker verification using the UBM based approach and the PBM based approach

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The aim of the present study was two-fold. The first aim was to investigate the degree of difference in hidden acoustic-classes between speakers. The second aim was to propose a novel speaker modeling approach that exploits this difference to improve the performance of speaker recognition systems. The findings of the study revealed that there is a great difference in hidden acoustic classes between speakers. Additionally, the evaluation of the proposed approach demonstrates its efficiency in terms of both verification performance and computational cost during the verification phase of the system, compared to the traditional approach.Future researchwill concentrate on applying the proposed approach within the hybrid GMM-SVM and the i-vectors based systems.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digit. Signal Process*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[2] D. Reynolds, “Universal Background Models”, in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds. Springer US, 2015, pp. 1547–1550.

[3] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors”, *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[4] N. Dehak and G. Chollet, “Support vector GMMs for speaker verification”, in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, 2006, pp. 1–4.

[5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification”, *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, mai 2006.

[6] N. Dehak *et al.*, “Support vector machines and Joint Factor Analysis for speaker verification”, in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4237–4240.

[7] N. Dehak, R. Dehak, P. Kenny, and P. Dumouchel, “Comparison between factor analysis and GMM support vector machines for speaker verification”, in *Odyssey*, 2008, p. 9.

[8] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms”, *CRIM MontrealReport CRIM-0608-13*, 2005.

[9] P. Kenny, “A small footprint i-vector extractor”, in *Odyssey*, 2012, pp. 1–6.

[10] D. Wu, J. Cao, and H. Wang, “Speaker Recognition Based on i-vector and Improved Local Preserving Projection”, *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 6, pp. 4299–4305, Jun. 2014.

[11] H. Beigi, *Fundamentals of Speaker Recognition*. Springer, 2011.

[12] D. A. Reynolds, “Comparison of background normalization methods for text-independent speaker verification”, in *Eurospeech*, 1997.

[13] D. Reynolds, “Gaussian Mixture Models”, in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds. Boston, MA: Springer US, 2015, pp. 827–832.

[14] A. Rozi, D. Wang, Z. Zhang, and T. F. Zheng, “An open/free database and Benchmark for Uyghur speaker recognition”, in *Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015 International Conference*, 2015, pp. 81–85.

[15] B. Ayoub, K. Jamal, and Z. Arsalane, “An analysis and comparative evaluation of MFCC variants for speaker identification over VoIP networks”, in *2015 World Congress on Information Technology and Computer Applications Congress (WCITCA)*, 2015, pp. 1–6.

BIOGRAPHIES OF AUTHORS

Ayoub Bouziane received the M.Sc. degree in intelligent systems and networks from the Faculty of Sciences and Technologies, Fez, Morocco, in 2012. He is currently pursuing the Ph.D. degree in the intelligent systems and applications laboratory, Sidi Mohammed ben Abdullah University, Morocco. His research interests cover signal processing and machine learning, mostly with applications in automatic speaker recognition. In parallel with his research activities, he teaches undergraduate level courses in computer science, at the Faculty of Sciences and Technologies, Fez. Additionally, He is a member of IEEE Signal Processing Society, IEEE Computational Intelligence Society, IEEE Computer Society and the International Speech and Communication Association (ISCA).

Jamal Kharroubi has his B.Sc. in Computer Sciencefrom Sidi Mohamed Ben Abdellah University (Fez-Morocco) in 1996. Two years after, he got his postgraduate degree in the domain of Artificial Intelligence from Galilee’s Institute - Paris XIII University. In 2002, Hereceived his Ph.D. degree in automatic speaker recognition systems from Telecom Paris Tech (Ecole Nationale Supérieure des Télécommunications de Paris-France)”. Since January 2003, he isan associate professor in the Department of Computer Scienceat the Faculty of Science and Technology. In 2008, he received his HDR diploma (Habilitation to conduct research). Additionally, He iscurrently the coordinator of the Master of Intelligent Systems and Networks. Moreover, he is the author of more than thirty publications in peer-reviewed scientific journals & conference proceedings. His research interests are focused onsignal andimage processing, pattern recognition, etc.

Arsalane Zarghili is a Doctor of Sciencefrom Sidi Mohamed Ben Abdellah University (Fez-Morocco). He received his Ph.D. in 2001 and joined the same University in 2002 as Professor at thecomputer science department of the Faculty of Science and Technology (FST). In 2007 he was head of the computer sciences department andchair of the Software Quality Master in the FST-Fez. Helectures Programming, Distributed, compilation and Information processing, for both under graduate and masterlevels. In 2008 he obtained his HDR in information processing. In 2011, he is the co-founder and the head of the Laboratory of Intelligent Systems and Applications in the FST of Fez. He is amember of the steering committee of the department of computer sciences and was a member of the faculty board. He is also IEEE member since 2011. His main research is about pattern recognition, image indexing and retrieval systems incultural heritage, biometric, etc.