A Gaussian Selection Method for Speaker Verification with Short Utterances

Flavio J. Reyes Díaz, Gabriel Hernández Sierra, and José Calvo de Lara

Advanced Technologies Application Center (CENATAV), La Havana, Cuba

{freyes, gsierra, jcalvo}@cenatav.co.cu

Abstract. Speaker recognition systems frequently use GMM-MAP method for modeling speakers. This method represents the speaker using a Gaussian mixture. However, in this mixture not all Gaussian components are truly representative of the speaker. In order to remove the model redundancy, this work proposes a Gaussian selection method to achieve a new GMM model only with the more representative Gaussian components. The results of speaker verification experiments applying the proposal show a similar performance to the baseline; however, the speaker models used have a reduction of 80% compared to the speaker model used as the baseline. Our proposal was also applied to speaker recognition system with short test signals of 15, 5 and 3 seconds obtaining an improvement in EER of 0.43%, 2.64% and 1.60%, respectively, compared to the baseline. The application of this method in real or embedded speaker verification systems could be very useful for reducing computational and memory cost.

Keywords. Speaker verification, Gaussian components selection, cumulative vector, short utterance.

Método de selección de gaussianas para la verificación de locutores con señales cortas

Resumen. Los sistemas de reconocimiento de locutores con frecuencia utilizan el método GMM-MAP para modelar locutores. Sin embargo, en estos modelos no todas las componentes gaussianas son representativas del locutor. Con el fin de eliminar dicha redundancia, proponemos un método de selección de gaussianas obteniendo un nuevo modelo con las componentes gaussianas más representativas. Los resultados experimentales muestran un rendimiento similar a la línea de base, no obstante los modelos obtenidos presentan una reducción del 80% respecto al modelo del locutor utilizado en la línea base. Los métodos propuestos son aplicados sobre señales de prueba más cortas, 15, 5 y 3 segundos; mejorando el EER de 0,43%, 2,64% y 1,60% respectivamente en comparación con la línea base. La aplicación del método propuesto en sistemas reales de verificación podría ser muy útil para reducir el costo computacional y la carga en memoria.

Palabras clave. Verificación de locutores, selección de componentes gaussianas, vector acumulativo, señales cortas.

1 Introduction

State of the art approaches in speaker recognition are mainly based on statistical modeling of acoustic space. The usual approach is to train a Universal Background Model (UBM) through the estimation of a large number of Gaussian components, using as much data as possible from many different speakers of impostor's population. Then, each speaker Gaussian mixture model (GMM) can be adapted from the UBM using much less data through Maximum a Posteriori (MAP) adaptation of the UBM means (GMM-MAP) [1], while variance and weight are unchanged. Aiming at producing more effective applications, the idea to use the mean vector of GMM-MAP speaker models as a super vector input data in a Support Vector Machine (GSV-SVM) classifier [2] came up, see Fig. 1.

GMM-MAP method includes a natural hierarchy between the UBM and each speaker model; for each UBM Gaussian component, there is a corresponding adapted component in the speaker model. These methods are not efficient enough because there are some aspects that increase the computational and memory cost:

Computación y Sistemas Vol. 18 No. 2, 2014 pp. 345-358
ISSN 1405-5546
http://dx.doi.org/10.13053/CyS-18-2-2014-036

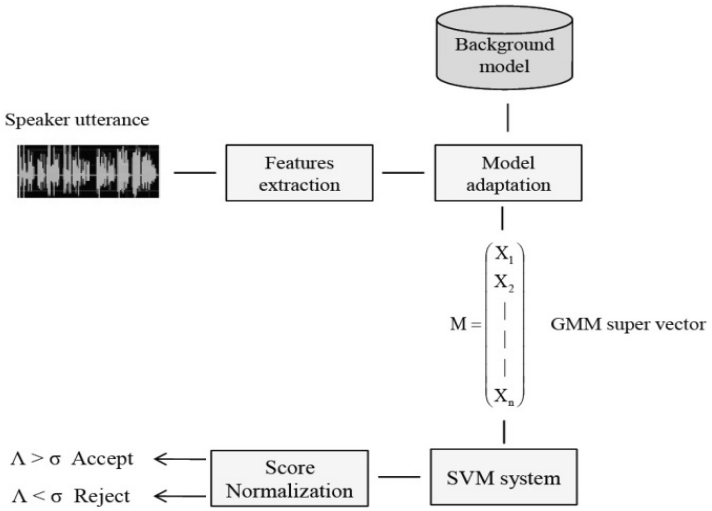

Fig. 1. GSV-SVM method proposed by Campbell *et al*. (2006)

1. The GMM-MAP speaker model has a high number of Gaussian components, commonly M=1024 or 2048, because the MAP adaptation from UBM to speaker data uses all the UBM Gaussian components, then the Super Vector Machine (SVM) take a super vector obtained concatenating all D-dimensional mean vectors of Gaussian components resulting in a high DxM dimension. For example, if D=24 and M=1024, each speaker utterance super vector will have 24576 values.
2. The speaker model has only some Gaussian components that represent better the acoustic space of each speaker utterance and the rest are redundant, in other words:

- a sub-set of Gaussian components represents better the speaker utterance (best discriminative for the target);
- another subset of components represents better the utterances of many speakers (non-discriminative between targets);
- the rest of the components represent better the utterances of other speakers (discriminative for impostors).

So, it is convenient to apply some Gaussian component selection method in order to reduce this redundancy and to bring more effective classification methods, mainly in front of real and embedded applications, and to reduce the store size of the models, too.

One of the most used learning methods to obtain GMMs is the Expectation Maximization algorithm (EM) [3]. With this method it is very difficult to know how many components are enough to fit the probabilistic distribution of the learning data, usually the number is obtained empirically [4]. Selection of an adequate number of components is an important issue in Gaussian mixture model learning. With too many components, the mixture model would overfit the data; on the other hand, with too few components, it would not be enough to describe the structure of the data.

Various Gaussian selection methods —referred to in [5] as Gaussian-layer methods— for reducing the number of GMM computations in speaker recognition have been proposed.

Better known and more extended criterion is proposed in [1] which performs at the verification stage, a selection of Gaussian components from the GMM-MAP target model using the top-C "better classified components" of the UBM for each feature vector of test signal, where C=10 is recommended; this method is our baseline for

comparison. Concerning the reduction of GMM-MAP models in embedded applications of speaker recognition systems, Reynolds states [6] that one of the most important problems is the size of the models related to memory storage and network traffic. For MAP adapted models, the storage of all parameters is not required because only mean vectors are adapted from UBM. Reynolds proposes to store only the mean difference between the GMM-MAP target model and the UBM model, achieving memory reductions of 56:1 with only an Equal Error Rate (EER) relative increase of 3.2%.

Auckenthaler and Mason [7] applied UBM-like hash model; for each speaker Gaussian component, there is a short list of indices of the expected best scoring components of the UBM model. Using the short list, only the corresponding Gaussian components in the speaker model are then scored, reporting a speed-up factor of about 10:1 with a minor degradation in the verification performance.

Xiang and Berger [8] constructed a tree structure for the UBM and multilevel MAP adaptation is used for generating the speaker model with a tree structure. In the verification phase, target speaker scores and UBM scores are combined using a multi-layer perceptron neural network. They reported a speed-up factor of 17:1 with a 5% relative increase in the EER.

Now we will briefly describe some proposals that "prune" the GMM.

Kinnunen *et al*. in [9] pre-quantize the test sequence prior to matching, reducing the number of test vectors and prune out unlikely speakers during the identification process, generalizing best variants to GMM/UBM based modeling.

Roch in [10] proposed the Gaussian selection to obtain *N*-best hypothesis in a pre-classifier considering that the classification of all tokens increases the computational load exponentially. A criterion, based on pre-classifier without charging the test process, normalizes the number of components and specifies the percentage of the distribution to be selected as queues. This job reduces the pre-classifier cost while keeping the accuracy inside 95% confidence interval.

Previous methods degrade the system performance as they gain speed-up. Other proposals use a different strategy to make a clustering of GMM, now we will consider some of them.

Aronowitz in [11] proposed to obtain an approximation of the GMM score without using test utterance applying Approximate Cross Entropy (ACE). With feature vectors and UBM, it performed a small selection to Gaussian with greatest likelihood and then it was submitted to some cluster in a Vector Quantification Tree (VQ-tree). In addition to this, we propose a GMM compression method thus considerably decreasing the storage space required for the models.

Liu *et al*. in [12] proposed a Gaussian selection method using only the components selected by cluster UBM (CUBM) as input for calculating an EM statistic with the objective of improving the speed of estimating the factor analysis model obtaining a good balance between efficiency and performance. Setting the number of CUBM Gaussian to 16 Gaussian components, the efficiency of CUBM-FA is much better than baseline factor analysis (the time cost has reduced from 9.53 sec to 1.24 sec) while having similar performance (both around 3.8% in EER and 0.02 in minimal of detection cost function (minDCF)).

Recently another proposal that sorts and indexes the GMM is [13]. Saeidi *et al*. proposed an optimization of the sorted function exposed by Mohammedi and Saeidi in [14], obtaining better results than GMM-UBM baseline. They use the Particle Swarm Optimization (PSO) method, evaluating the search width in a power of 2. Results obtained with search width of 512 are better than those obtained with the sorted function in [14] presenting an EER=8.21 and minDCF=0.4024, while the sorted GMM method in [14] presents an EER=8.72 and minDCF=0.4090, and the baseline GMM-MAP method has an EER=8.31 and minDCF=0.3929.

More recently, another extension of the method explained in [13] is proposed by Saeidi *et al*. in [15], using a two-dimensional indexation, allowing simultaneous selection of Gaussian and frames. The evaluation was developed using several values of a control parameter to specify the neighborhood of the optimization (2%, 3%, 5%, 10%, 15% and 20%) obtaining speed-up

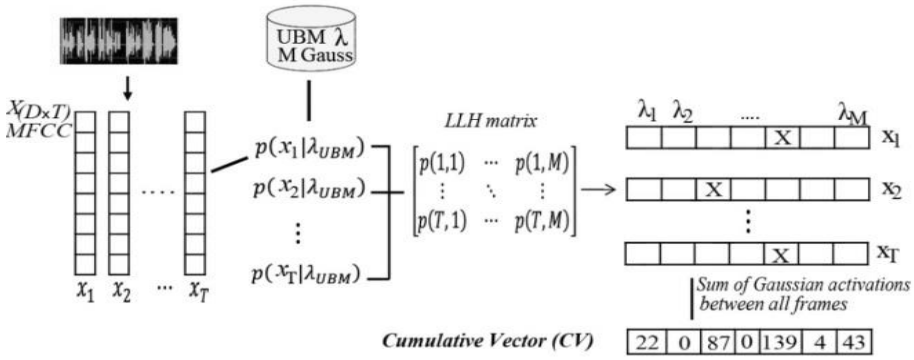

Fig. 2. Cumulative vector method

ratios of 157:1, 85:1, 37:1, 11:1, 5:1 and 3:1, respectively. This method is computationally more efficient than the GMM-MAP baseline because less frames and Gaussian components are evaluated in the test, obtaining similar baseline efficacy only with the control parameter in 20%.

Observe that recent methods obtain similar performance to the baseline reducing the processing speed.

Our work centers the attention on the redundant information present in speaker models and proposes a method to reduce this, performing a selection of Gaussian components of the GMM-MAP and UBM models, based on cumulative vectors of number of activations of better classified components for each feature vector of the acoustic utterances. Since our intention is to evaluate the reduction of redundant information in speaker models, we will perform GMM-MAP speaker verification experiments using two Gaussian component selection approaches to obtain the best model to characterize the speaker with a reduced dimension. We also propose the use of weights for the selected Gaussian components to give a greater emphasis to the most activated Gaussian components of the speaker model.

This paper is organized as follows. Section 2 describes the proposed methods, Section 3 describes the databases and front end used, Section 4 evaluates the two methods experimentally. Then, Section 5 explains the results obtained in the application of one of the methods to speaker verification experiment with short test utterances, and Section 6 concludes the paper and proposes new research lines.

2 Proposed Methods

This section contains the explanation of the proposed methods

- to obtain the cumulative vector,
- to select the Gaussian components,
- to adjust the weights of the Gaussian components,
- to perform classification.

2.1 The Cumulative Vector

Since our aim is to perform a Gaussian selection from UBM that best represents the client, we use recent methods proposed in [16], applying the UBM model instead of the anchor model. The process consists in obtaining the better classified component of the UBM regarding each frame of speech utterance and storing in a vector the sum of reached "activations" in all frames of the utterance, see Fig. 2.

The likelihood is calculated for each frame, regarding all Gaussian components of the UBM, obtaining a likelihood matrix $LLH(X|\lambda_{UBM})_{(T,M)}$, where T is the number of frames and M is the number of Gaussian components of UBM. From the LLH matrix, a row (frame) search of the best classified component (maximal likelihood) is done and it is identified as activated, then a sum by

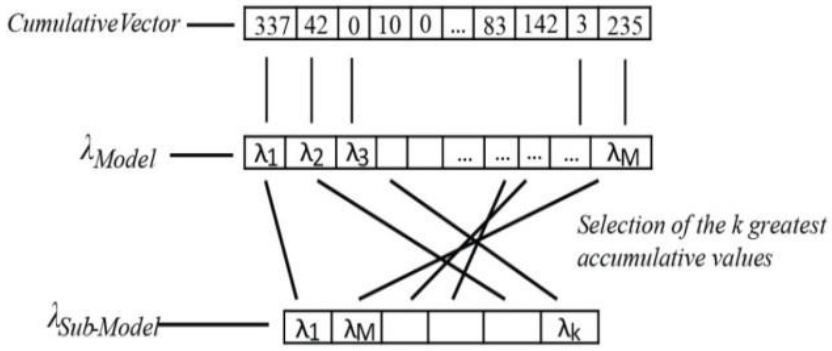

Fig. 3. Gaussian component selection criteria using the cumulative vector (GCS-CV)

columns of activated components is performed (over all frames of utterance), and the result is stored in the cumulative vector. The cumulative vector contains M accumulative values, reflecting the number of activations of each Gaussian component, for all utterances.

2.2 Gaussian Component Selection Criteria Using the Cumulative Vector

As described above, there are several Gaussian component selection criteria [17] based on the feature vector likelihood given the Gaussian component $p(x|\lambda)$. The goal of our proposal is to select a set of Gaussian components that better characterizes the acoustic classes of a speaker utterance, based on the k greatest accumulative values of the cumulative vector. Using the cumulative vector obtained from UBM, for each speech utterance, the Gaussian components with the k greatest accumulative values are ordered and selected, see Fig. 3.

This criterion brings an important reduction of model dimensionality: observe the dimensionality reduction of the speaker model from M to only k components. These k components are the best classified components in all utterances, so the model would be more discriminative.

2.3 Selection Variant Using the Training Utterance

Two variants of classification [17] will be explained using the GMM-MAP framework [1]; both methods use the UBM Gaussian component selection based on the cumulative vector explained above to select the Gaussian components and to obtain a reduced model which better represents the speaker utterance.

a. Selection Variant Using the Training Utterance

The method performs a selection of the Gaussian components using the feature vectors of the training utterance and the UBM. A speaker model is obtained with MAP adaptation; simultaneously the cumulative vector (CV) is obtained, using the same data. With the model and CV, GCS-CV method is applied, obtaining a new *k*-components model of the training utterance. At last, test utterance is classified in GMM-MAP framework, but using the new model of the training utterance, see Fig. 4.

b. Selection Variant Using the Test Utterance

This method performs a selection of the Gaussian components using the feature vectors of the test utterance and GCS-CV with the target model to obtain a new model of the target utterance to make the classification. Using the feature vectors of the training utterance and the UBM, a speaker model is obtained with MAP adaptation; then, using the feature vectors of the test utterance and the UBM, the cumulative vector is obtained. With the model and the cumulative vector, GCS-CV method is applied, obtaining a new *k*-components model of the test utterance. At last, the test

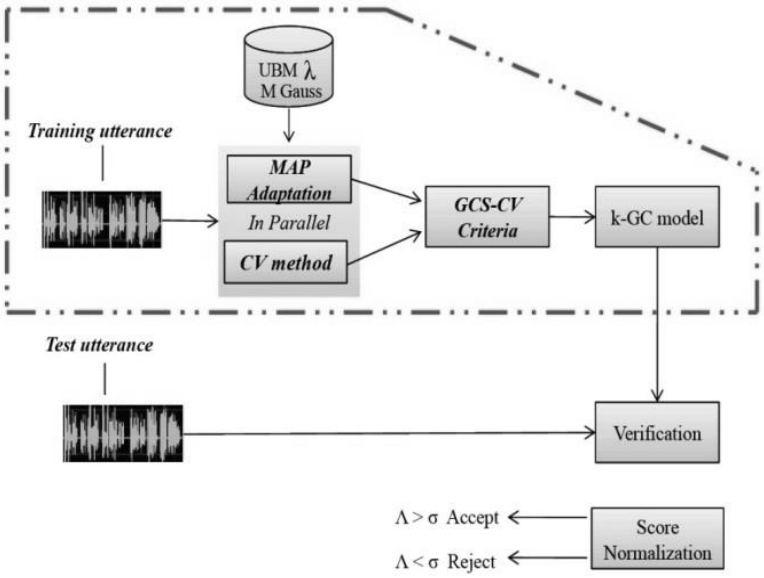

Fig. 4. Classification method using the training utterance

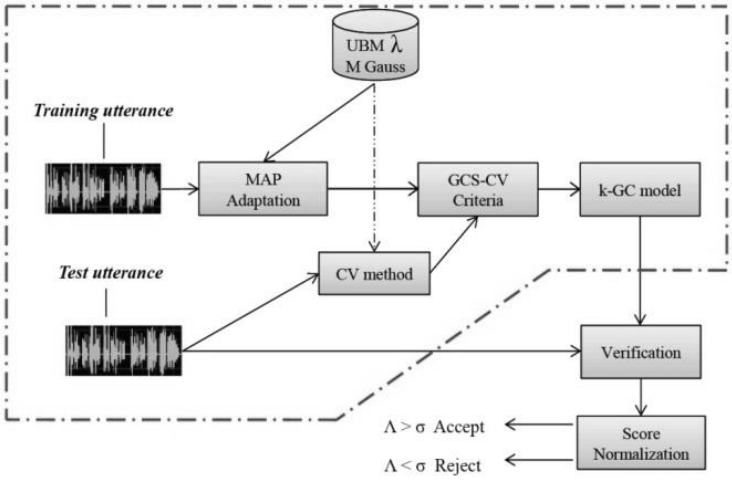

Fig. 5. Classification method using the test utterance

utterance is classified in GMM-MAP framework, but using the new representation, see Fig. 5.

c. Replacement of the Weights of the Selected Component Based on the Number of Activations

The complete adaptation to the features of the training utterance causes in the model the existence of sub-sets of Gaussian components with different acoustic information in the speaker's space: target information, impostor's information and information common to many speakers as we hypothesized in the introduction. In the model, those Gaussian components that carry common information for many speakers are less discriminative and more redundant. These Gaussian components could reflect high probabilities for many speakers, being less

discriminative for a specific speaker, even being adapted to its utterance.

Taking into account that these Gaussian components could present a similar number of activations, the better reflection of the difference between speakers is present if we substitute the original weight of the Gaussian component by the number of activations, which are not common Gaussian components. For this reason we propose to use the accumulative values av_m of the CV of the utterance as a weight for the Gaussian components that model the utterance. In practice, after the selection of k Gaussian components that present higher values of activation, each weight of the Gaussian component is replaced by the corresponding normalized activation value:

$$av_m = \frac{av_m}{\sum_{m=1}^{k} av_m}. \quad (1)$$

3 Experimental Results

3.1 Database

The UBM model is obtained from "SALA: SpeechDat across Latin America" telephone speech database (Venezuelan version) [18]. It contains 1000 speakers (503 male, 497 female) uttering in each telephone call a total of 45 read and spontaneous items.

For training and test utterances, NIST2001 Ahumada database was used [19]. Ahumada is a speech database of 104 male Spanish speakers, designed and acquired under controlled conditions for speaker characterization and identification, which incorporates several speech variability factors. Each speaker in the database expresses five types of utterances (digits sequences, balanced phrases, balanced and random text and spontaneous speech) in seven microphone sessions and three telephone sessions, with a time interval between them. Conventional telephone land line was used. In session $T1$ every speaker was calling from the same telephone in an internal-routing call. In session $T2$, all speakers were requested to make a call from their own home telephone, trying to search a quiet environment, so the channel and handset characteristics are unknown. In session $T3$, a local call was made from a quiet room using 9 randomly selected standard handsets. Each speaker utters a spontaneous expression of about 60 sec. in each telephone session; eliminating the pauses, speech is about 48 sec. as average, in each utterance.

3.2 Front End

Well known Mel-Frequency Cepstrals Coefficient (MFCC) features [20] have been used to represent the short time speech spectra. As shown in Fig. 6, all telephone speech signals are quantized at 16 $bits$ at 8000 Hz sample rate, pre-emphasized with a factor of 0.95, and an energy based silence removal scheme is used. A Hamming window with 20ms window length with 50% overlap is applied to each speech frame and a short time spectrum is obtained applying a FFT. The magnitude spectrum is processed using a 24 Mel-spaced filter bank, the log-energy filter outputs are then cosine transformed to obtain a standard set of 12 MFCC, the zero cepstral coefficients is not used [21].

In order to reduce the influence of mismatch between training and testing acoustic conditions in a telephonic environment, a robust feature normalization method for reducing noise and/or channel effects is applied to MFCC features, the Cepstral Mean and Variance Normalization (CMVN) proposed by Viikki and Laurila in [22]. This method normalizes the gross spectral distribution of the utterance and reduces long term intra-speaker spectral variability. Assuming Gaussian distributions of features, CMVN

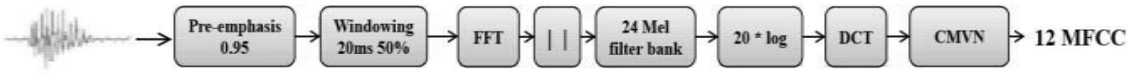

Fig. 6. Normalized MFCC front end

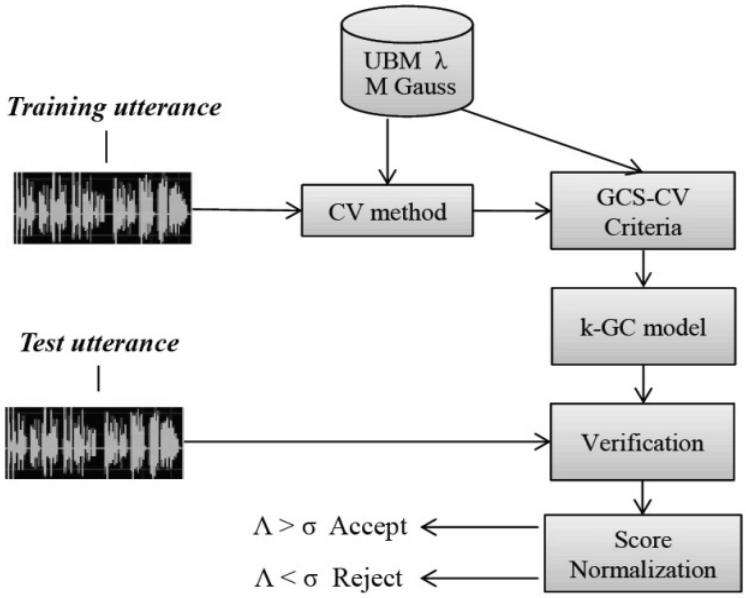

Fig. 7. Selection of Gaussian components directly from the UBM

normalizes each component of the feature vector according to the expression

$$\hat{c}_i[t] = \frac{c_i[t] - \mu_i}{\sigma_i}, \quad (2)$$

where $c_i[t]$ and $\hat{c}_i[t]$ are the ith coefficient of the feature vector at each frame t before and after normalization, respectively, and μ_i and σ_i are the mean and variance estimates of the time sequence of each coefficient c_i. Fig. 6 represents a complete scheme of normalized MFCC front end.

At last, the Δ cepstral features from MFCC normalized cepstral feature are obtained and appended to MFCC features [23] conforming a *24*-dimensional $MFCC + \Delta$ feature vector.

3.3 Score Normalization

Score normalization method for small evaluation databases is proposed in [24]. For each score LLH between a target X_A and a test X_B, the normalized score is

$$LLH_N(X_A, X_B) = LLH(X_A, X_B) - mean\big(LLH(X_A, I_S)\big), \quad (3)$$

where I_S is a subset of impostors. Since the evaluation database is small, we divided the experiment into two subsets, a and b, each of them composed of half of the speakers.

When the subset a is used to perform the speaker recognition test, the speakers from the subset b are used as impostors for normalization and vice versa. The test from the two subsets is polled together in order to obtain the global performance of a given system.

4 Speaker Verification Experiments

4.1 GMM-MAP Speaker Verification Baseline

First, a GMM-MAP speaker verification baseline using the data and methods explained was established. A UBM model with *M=2048* Gaussian components was trained using expectation-maximization (EM) algorithm [3] with 1989 male speech utterances from SALA database. GMM-MAP models of 100 speakers were MAP adapted [1] using the spontaneous utterances of session T1 of Ahumada database as training utterances. For the verification step,

Table 1. EER and DCF results of experiments based on the UBM model1

	Selection		Selection and weight replacement	
k	EER	minDCF	EER	minDCF
150	15.0	7.23	8.28	5.68
200	17.0	7.08	8	5.67
250	18.0	7.67	9	6.11
300	18.18	7.89	8	6.03
350	22.0	8.37	7.81	6.16
400	21.43	8.3	7.87	6.17
500	22.50	8.67	7.80	6.15

testing spontaneous utterances was obtained from the same speakers but in session *T2* of Ahumada database, and the comparison was based on the criteria to reduce the verification load proposed in [1]; it performs a search of the Gaussian components of the UBM, which are the most likely for each feature vector of the test utterance. These Gaussian components from the GMM-MAP target model are selected, creating a sub-model with C=10 Gaussian components for each feature vector of the test utterance, obtaining as many sub-models as feature vectors. With each of the obtained sub-models and their corresponding feature vectors, vector-based likelihood is calculated; the likelihood of the test utterance with respect to all sub-models will be the mean values of the vector-based likelihoods. Score normalization is applied and the results are evaluated on DET curve [25], obtaining an EER=4%; the NIST evaluation criteria, minimal of "detection cost function" was evaluated too, minDCF=2.29%.

4.2 Speaker Verification with Component Selection from the UBM Model

In order to evaluate our initial hypotheses and compare with the rest of experimental results, a simple selection of Gaussian components directly from the UBM was done, based on cumulative vector and component selection methods, see Fig. 7.

Two experiments were done in order to evaluate the influence of the replacement of weight by activation numbers; Table 1 presents the results of both experiments for different *k*.

Table 2. EER and minDCF results of experiments

Method1	Selection		Selection and weight replacement	
k	EER	minDCF	EER	minDCF
150	5.0	2.69	4.53	2.6
200	4.88	2.53	4.77	2.43
250	5.0	2.47	4.48	2.33
300	4.52	2.39	5.0	2.41
350	4.0	2.22	5.0	2.44
400	5.0	2.29	5.0	2.34
500	5.0	2.23	5.0	2.35
Method2	**Selection**		**Selection and weight replacement**	
k	EER	minDCF	EER	minDCF
150	4.89	2.58	5	2.94
200	4.04	2.32	4.18	2.68
250	4.0	2.28	4.50	2.63
300	4.36	2.41	4.64	2.63
350	4.05	2.28	4.74	2.62
400	4.20	2.22	4.75	2.63
500	4.18	2.21	4.89	2.64

Results shown in Table 1 demonstrate that a simple selection of Gaussian components of the UBM, without any kind of model adaptation of training utterances, introduces a great variability in the performance of the classifier depending on the dimension *k* of the selected components. As k increases, % EER and minDCF increases too, indicating that more non-discriminative or common components were selected and included in the speaker *k-GC* model. Besides, if weight replacement by activation numbers of selected Gaussian components is applied, an increase in the discrimination between speakers is appreciated with a reduction of the EER and the minDCF, for the same *k* selected components.

This experiment supports the hypotheses expressed before, confirming our supposition that the GCS-CV method applied to MAP adapted models using training or test utterances could

increase the performance of the classifier with an important reduction of Gaussian components.

4.3 Speaker Verification Experiments Using Training and Test Data to Select Gaussian Components from the UBM Model

Two experiments were performed with both selection variants explained in Section 2, selecting k=150, 200, 250, 300, 350, 400 and 500 Gaussian components; the combination with the proposed replacement of weights by the activations number was also evaluated.

Our experiments using the proposed methods showed the following results.

Redundancy reduction in the selected Gaussian components. As shown, the experiment using Method 1 with k=350 Gaussian components and experiment using Method 2 with k=250 Gaussian components get the same % EER and less % minDCF related to the GMM-MAP baseline, with a respective reduction of 82.9% and 87.7% of the Gaussian components of the original speaker model (2048). The non-selected Gaussian components are less discriminative of the speaker or not discriminative at all. This reduction of information lowers the verification phase computational burden, due to the use of less number of Gaussian components.

Classification method using the test utterance is better. Method 2 obtains similar results to Method 1 with less Gaussian components (250 vs. 350); this method is more adjusted to the speaker because it selects the components of the model from the test utterance, very similar to Reynolds method [1] but less expensive.

Weight replacement by activation numbers of selected components is not always good. The weight replacement reflects its efficacy only for k<300 components in Method 1, for greater k, the weight replacement does not increase the performance, because the appended new components are less discriminative than the first 300. As Method 2 is more adjusted to speaker, the weight replacement doesn't take the desired effect, because the selected components are the most discriminative ones, their applications produce a reduction in the classifier performance as a consequence.

4.4 Score Fusion and Normalization

Finally, a fusion of scores was done; observe in Table 3 a light increase of performance, EER and minDCF in both approaches using only the selection or the selection with the weight replacement by the activation numbers. The score normalization is done only once, after fusion of scores. The score fusion expression is

$$scorefusion = \frac{1}{2}(scoreM1) + \frac{1}{2}(scoreM2). \qquad (4)$$

5 Application of the GSC-CV Criteria to Speaker Verification with Short Signals

A current problem in real speaker recognition systems is the duration of the utterance to verify [18]. Considering that the proposed methods perform a selection of Gaussian components more representative of the speaker and reduce the common information between all speakers, we carried out speaker verification experiments similar to the ones described in Section 4, with the main difference that duration of test utterance is reduced to 3, 5 and 15 seconds.

Test utterance is very short and does not submit the necessary information to perform an

Table 3 EER and minDCF results of score fusion of experiments

	Method 1-2		Method 1-2 with weight replacement	
k	**EER**	**minDCF**	**EER**	**minDCF**
150	5.0	2.42	4.81	2.19
200	4.0	2.32	4.24	2.19
250	4.0	2.34	4.18	2.24
300	4.0	2.30	4.06	2.23
350	4.0	2.22	4.0	2.22
400	4.8	2.26	4.0	2.20
500	4.58	2.24	4.0	2.20

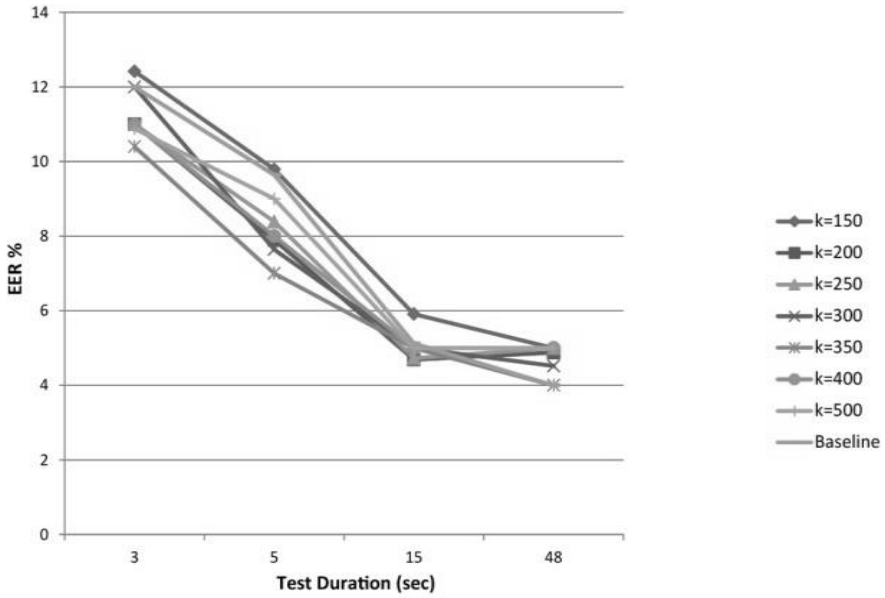

Fig. 8. Comparison of EER using Method 1 with different duration of test utterances

Table 4. EER and minDCF results of experiments with short test utterances

	15 sec		5 sec		3 sec	
K	**EER**	**DCF**	**EER**	**DCF**	**EER**	**DCF**
150	5.9	3.25	9.79	4.15	12.42	5.53
200	4.6	3.16	7.90	4.37	11.0	5.43
250	4.7	2.96	8.39	4.47	11.0	5.31
300	5.0	2.98	7.64	4.2	12.0	5.36
350	5.0	2.86	7.0	4.28	10.40	5.26
400	5.0	2.84	8.0	4.27	11.0	5.15
500	5.0	2.72	9.0	4.23	10.88	5.09
BL	5.1	2.67	9.64	4.11	12.0	5.01

adequate Gaussian components selection, and the resulting model is not representative enough of the speaker. So, we chose Method 1 without weight substitution for the experiments, because it performs the selection from the training utterance.

Experiments were performed with the method explained in Section 2, selecting k=150, 200, 250, 300, 350, 400 and 500 Gaussian components. Table 4 presents the results of the experiments for different k and durations of the test.

Experimental results show that speaker verification with GSC-CV criteria performs better than baseline for short test utterances with k>150; then, the reduction of redundancy in the model implies better efficacy in the verification with short test utterances. Observe that, if the duration of test utterances is reduced, more Gaussian components are necessary to obtain better results, k is between 200 and 250 for 15 sec, k is between 300 and 350 for 5 sec and k is between 350 and 450 for 3 sec. It means that the speaker model requires more Gaussian components as its test utterance is shorter, to obtain better performance; so it is necessary to have a little increase in redundancy in the speaker model to deal with the short test utterances.

Fig. 8 compares the EER results obtained from these experiments with test utterances of 3, 5 and 15 sec. of speech and obtained in experiment and shown in Table 2, for test utterances of about 48 seconds of speech. Both experiments use Method 1 without weight substitution.

Observe that method with k=350 is the best for any test duration; also observe that the differences in efficacy between 15 sec. and 48 sec. for all the methods are very few, so the GSC-CV criteria applied to reduce redundancy in speaker models provoke very similar behavior in speaker verification for test utterances with duration greater than 15 sec., then the use of this method would reduce the processing cost of test utterances in speaker verification.

6 Conclusions and Future Work

In the presence of real or embedded applications of speaker verification, the classical GMM-MAP [1] and GSV-SVM [2] methods are not sufficient enough. It is so, because the classical GMM-MAP obtains a large number of components from the set of Gaussian components selected for each feature vector, it is possible to perform a reuse of Gaussian components in different sub-models, increasing the computational load and runtime of the verification stage, at the same time GSV presents high dimensionality, too. Both methods use non-discriminative and redundant information.

Experimental results using GSC-CV criteria show that an important reduction of the models, more than 80% regarding the number of Gaussian components used in the baseline models (2048) explained in Section 4.1 is reached, with similar performance in speaker verification experiments. Of course, the volume reduction will depend on the databases. In all experiments, the use of the GSC-CV method of Gaussian component selection would reduce the computational and memory cost of classifying stage in real applications of speaker verification in relation to the baseline, because for each frame this method selects the 10 most likely Gaussian components. Also, the use of the proposed method, by selecting 200 or more Gaussian components as shown in Table 4, increases the efficacy of speaker verification with short test utterances compared to the baseline, an aspect very common in forensic situations.

As a future work, we propose to obtain another method to select the Gaussian components of the model, using an Adaboosting classifier, considering the Gaussian component as weak classifiers and utterances of target and impostors speakers as positive and negative samples. The proposal would be to obtain an optimal value of k Gaussian components as a strong classifier of each target speaker to be used as speaker model for speaker verification experiments.

References

1. **Reynolds, D.A., Quatieri, T.F., & Dunn, R.B. (2000).** Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3), 19–41.
2. **Campbell, W.M., Sturim, D.E., & Reynolds, D.A. (2006).** Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5), 308–311.
3. **Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977).** Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B,* 39(1), 1–38.
4. **Cheng, S.S., Wang, H.M., & Fu, H.C. (2004).** A model-selection-based self-splitting Gaussian mixture learning with application to speaker identification. *EURASIP Journal on Applied Signal Processing*, 2004, 2626–2639.
5. **Chan, A., Ravishankar, M., Rudnicky, A., & Sherwani, J. (2004).** Four-layer categorization scheme of fast GMM computation techniques in large vocabulary continuous speech recognition systems. *INTERSPEECH 2004-ICSLP*, Lisbon, Portugal.
6. **Reynolds, D.A. (2003).** Model Compression for GMM based Speaker Recognition Systems. *INTERSPEECH 2003,* Geneva, Switzerland.
7. **Auckenthaler, R. & Mason, J. (2001).** Gaussian selection applied to text-independent speaker verification. *Speaker Odyssey: the Speaker Recognition Workshop*, Crete, Greece, 83–88.
8. **Xiang, B. & Berger, T. (2003).** Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. *IEEE Transactions on Speech and Audio Processing*, 11(5), 447–456.

9. **Kinnunen, T., Karpov, E., & Franti, P. (2006).** Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech and Language Processing,* 14(1), 277–288.
10. **Roch, M. (2006).** Gaussian-selection-based non-optimal search for speaker identification. *Speech Communication*, 48(1), 85–95.
11. **Aronowitz, H. & Burshtein, D. (2007).** Efficient speaker recognition using approximated cross entropy (ACE). *IEEE Transactions on Audio, Speech and Language Processing*, 15(7), 2033–2043.
12. **Liu Q., Huang W., Xu, D., Cai, H., & Dai, B. (2010).** A Fast Implementation of Factor Analysis for Speaker Verification. *INTERSPEECH 2010 ISCA*, Makuhari, Japan, 1077–1080.
13. **Saeidi, R., Mohammadi, H.R.S., Ganchev, T., & Rodman, R.D. (2009).** Particle Swarm Optimization for Sorted Adapted Gaussian Mixture Models. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2), 344–353.
14. **Mohammadi, H.R.S. & Saeidi, R. (2006).** Efficient implementation of GMM based speaker verification using sorted Gaussian mixture model. *14th European Signal Processing Conference (EUSIPCO'06),* Florence, Italy.
15. **Saeidi, R., Kinnunen, T., Mohammadi, H.R.S., Rodman, R., & Franti, P. (2010).** Joint frame and gaussian selection for text independent speaker verification. *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, 4530–4533.
16. **Anguera, X. & Bonastre, J.F. (2010).** A Novel Speaker Binary Key Derived from Anchor Models. *INTERSPEECH 2010*, Makuhari, Japan, 2118–2121.
17. **Reyes, F.J., Calvo, J.R., & Hernández-Sierra, G. (2012).** Gaussian selection for speaker recognition using cumulative vectors. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science,* 7441, 724–731.
18. **Moreno, A., Comeyne, R., Haslam, K., van den Heuvel, H., Höge, H., Horbach, S., & Micca, G. (2000).** SALA: Speechdat across Latin America. Results of the First Phase. *Seventh International Conference on Language Resources and Evaluation (LREC 2010),* Valletta, Malta.
19. **Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V., Diaz-Gomez, J.J., Garcia-Jimenez, R., Lucena-Molina, J., & Sanchez-Molero, J.A.G. (1998).** AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification. *1998 IEEE International Conference on Acoustics, Speech and SignalProcessing*, Seattle, WA, USA, 2, 773–776.
20. **Davis, S.B. & Mermelstein, P. (1980).** Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28 (4), 357–366.
21. **Campbell, J.P., Jr. (1997).** Speaker Recognition: A tutorial. *Proceedings of the IEEE*, 85(9), 1437–1462.
22. **Viikki, O. & Laurila, K. (1998).** Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3), 133–147.
23. **Furui, S. (1981).** Cepstral analysis technique for automatic speaker verification. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 29(2), 254–272.
24. **Reynolds, D.A. (1997).** Comparison of background normalization methods for text-independent speaker verification. *EUROSPEECH 1997,* Rhodes, Greece.
25. **Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997).** The DET curve in assessment of detection task performance. *EUROSPEECH 1997*, Rhodes, Greece.

Flavio J. Reyes Díaz is a researcher at the Department of Images and Signals of the Advanced Technologies Application Center (CENA-TAV), Cuba. He graduated from Informatics Sciences University (UCI, at La Havana, Cuba) as Engineer in Informatics Sciences in 2009. He has been a member of the Cuban Society of Mathematics and Computation and of the Cuban Association for Pattern Recognition since 2009. Flavio J. Reyes Díaz focuses his research on speaker recognition system, and his issues of interest are related to processing and analysis of digital signals, speaker modelling and methods of variability compensation of speech utterances.

Gabriel Hernández-Sierra is a researcher at the Department of Images and Signals of the Advanced Technologies Application Center (CENATAV), Cuba. He received his B.Sc. in Computer Sciences from Havana University, Cuba, in 2005, and he is currently a Ph.D. student in Automatics and Computing at the Polytechnic University of Havana (CUJAE), Cuba, and University of Avignon, LIA, France. He has been a member of the Cuban Society of Mathematics and Computation and of the Cuban Association for Pattern Recognition since 2005. Gabriel Hernández-Sierra focuses his researches on speaker recognition systems, but his interests also cover other areas of pattern recognition such as digital signal processing and analyzing, speech recognition and language recognition.

José Ramón Calvo de Lara graduated in Telecommunications Engineering from Polytechnic Institute José A. Echeverría in 1978. He received his Ph.D. degree in Technical Sciences in 2003 and the Scientific Researcher Category in 2004. Now he is a researcher at the Department of Signals and Images of the Advanced Technologies Application Center (CENATAV), leading the speech and language processing research. He is a member of the Cuban Society of Mathematics and Computation and the Cuban Association for Pattern Recognition. His research interests are extraction, selection and modeling of speech features for speaker, language and speech recognition, as well as noise compensation methods and speech quality measures.

Article received on 25/04/2013, accepted on 14/03/2014.