Gaussian Selection for Speaker Recognition Using Cumulative Vectors

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

Advanced Technologies Application Center
{freyes,jcalvo,gsierra}@cenatav.co.cu
http://www.cenatav.co.cu

Abstract. Speaker recognition systems frequently use GMM - MAP method for modeling speakers. This method represents a speaker using a Gaussian mixture. However in this mixture not all the 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. Speaker verification experiments applying the proposal show a similar performance to baseline; however the speaker models have a reduction of 80 % regarding the speaker model used for baseline. The application of this Gaussian selection 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.

1 Introduction

Speaker recognition state of the art approaches are mainly based on statistical modeling of the acoustical space. Speaker utterances information has been modeled using the Gaussian Mixture Model-Universal Background model (GMM-UBM) Reynolds's paradigm [1]. The usual approach is to train an UBM model, 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 GMM speaker model can be adapted from the UBM using much less data through Maximum a Posteriori (MAP) adaptation of the UBM means [2], while variance and weight are unchanged.

GMM-UBM 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 GMM-UBM speaker model. In real or embedded applications these method is not efficient enough because there are some aspects that increase the computational and memory cost:

(1) The GMM-UBM 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.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 724–731, 2012.

(2) The GMM-UBM 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) A sub-set of Gaussian components represents better the speaker utterance (best discriminative for the target)
   b) Another subset of components represents better the utterances of many speakers (non discriminative between targets)
   c) The rest of the components represents better utterances of other speakers (discriminative for impostors)

So, it is convenient to apply some Gaussian components 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 storage size of the models, too.

Well known and more extended criterion is proposed by Reynolds et al. in [1] which performs at the verification stage, a selection of Gaussian components from the GMM-UBM target model, using the top-C "better classified gaussian components" of the UBM for each feature vector of test signal, creating a sub-model with R components for each feature vector, obtaining as many sub-models as feature vectors contain a test utterance. In this variant it is possible to perform a reuse of Gaussian components in different sub-models causing an increase of the computational load and runtime of the verification stage; this method is used in our work as a baseline for comparison.

A recent article of Saeidi et al. [3] refers a good overview of previous works that deal with another Gaussian selection in GMM models for speaker recognition systems:

- Auckenthaler and Mason [4] applied UBM-like hash model; for each speaker Gaussian component, there is a shortlist of indices of the expected best scoring components of UBM model.
- Xiang and Berger [5] construct 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 an MLP neural network.
- Kinnunen, T. et al. [6] 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.

Previous methods degrade the system performance in exchange for gaining speed-up. Saeidi et al in [3] proposed an optimization of the sorted function exposed by Mohammadi et al. in [7], obtaining better results than GMM-UBM baseline. They use the Particle Swarm Optimization (PSO) method, evaluating the search width in power of 2. Results obtained with search width of 512 are better than those obtained with the sorted function in [7].

More recently, another extension of the method explained in [3] is proposed by Saeidi et al. in [8], 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 ratios of 157:1, 85:1, 37:1, 11:1, 5:1 and 3:1, respectively.

Lately, Liu et al. in [9] proposed a Gaussians selection method using only the components selected by cluster UBM (CUBM) as input for calculating a EM statistic with the objective to improve the speed of estimating the factor analysis model (FA) obtaining a good balance between the efficiency and performance. The efficiency of CUBM-FA is much better than baseline factor analysis (the cost time has been reduced from 9.53 sec to 1.24 sec) while having similar performance (both around 3.8% in EER and 0.02 in minDCF).

These recent methods obtain similar performance than baseline reducing the processing time.

Our work focuses the attention in the redundant information present in speaker models and proposes a method to reduce this, performing a simple selection of Gaussian components of the GMM-UBM and the UBM models, based on cumulative vectors of number of activations of better classified components for each feature vector of the acoustic utterances. As our intention is to evaluate the reduction of redundant information in speaker models only, we will use GMM-UBM speaker verification experimental framework under Reynolds's paradigm [1].

2 Proposed Methods

2.1 The Cumulative Vector

This algorithm is similar to the recently proposed by Anguera in [10], but using a universal background model (UBM) instead of anchor models. The process consists in obtaining the most likely 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 figure 1.

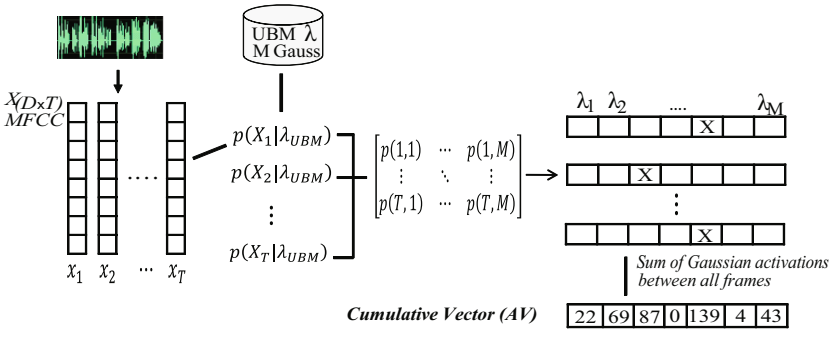

Fig. 1. Cumulative vectors method

The likelihood is calculated for each speech utterance 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 most likely component is done and it is identified as activated, then a sum by columns of activated components is performed (over all frames of utterance), and the result is stored in the cumulative vector (CV). The cumulative vector contains M accumulative values, reflecting the number of activations of each Gaussian component, for the utterance.

2.2 Gaussian Component Selection through Cumulative Vectors

As described above, there are several Gaussian component selection methods based on the feature vector likelihood given the Gaussian component $p(x|\lambda)$; hence the goal of our proposal is to select a set of Gaussian components that better characterize the acoustic classes of a speaker utterance, based on the k greatest accumulative values of cumulative vector. Using the cumulative vector obtained from UBM in 2.1, the Gaussian components with the k greatest accumulative values are selected. See figure 2.

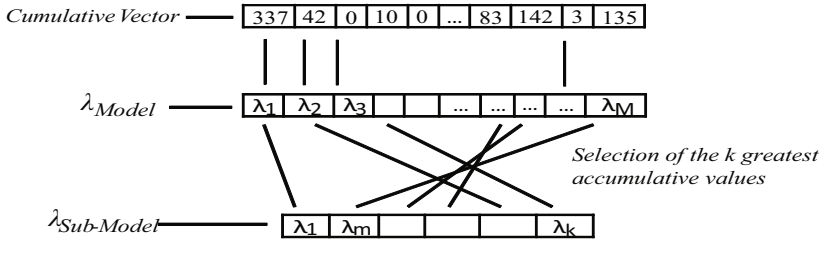

Fig. 2. Gaussian component selection criteria using cumulative vector (GCS-CV)

This method brings an important reduction of model from M to only k components. These k components are the most likely components in all utterances, so the new model would be more discriminative.

2.3 Two New Classification Methods for Speaker Recognition

Two methods of classification using the GMM-UBM framework [2] will be proposed; both methods use the UBM Gaussian component selection based on cumulative vector (GCS-CV) explained above to select the Gaussian components and obtain a reduced model which better represents the speaker utterance. The methods were applied using training and test utterances.

Classification Method Using the Training Utterance. Using the feature vectors of the training utterance and the UBM, a speaker model is obtained with MAP adaptation; simultaneously CV is obtained as explained in section 2.1,

using the same input data. With GMM-UBM model and CV, GCS-CV method is applied as explained in 2.2, obtaining a new k-components model of the training utterance. Lastly, test utterance is classified in GMM-UBM framework, but using the new model of the training utterance. See figure 3A.

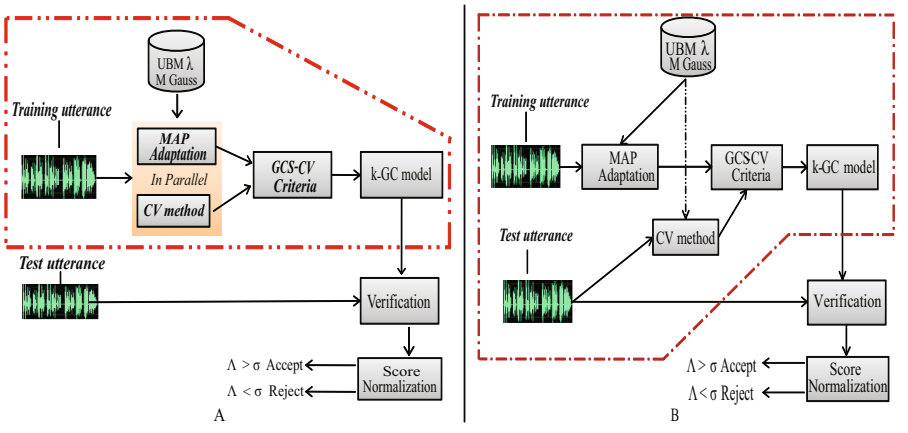

Fig. 3. Both classification method

Classification Method Using the Test Utterance. This method makes the selection of the Gaussian components using the feature vectors of the test utterance, based on the method proposed by Reynolds in [1].

Hence, we propose the use of GCS-CV method on the test utterance together with the target model to obtain a new model of the target and make the classification. In contrast to the one proposed by Reynolds, here we obtain only one model for test utterance.

Using the feature vectors of the training utterance and the UBM, a speaker model is obtained with MAP adaptation; in the test session, using the feature vectors of the test utterance and the UBM, CV is obtained as explained in 2.1. With GMM-UBM model and CV, GCS-CV method is applied as explained in 2.2, obtaining a new k-components model of the training utterance. Finally, the test utterance is classified in GMM-UBM framework, but using the new model of the training utterance. See figure 3B.

3 Experimental Results

Databases, front-end and score normalization used for all experiments are explained in this section.

The UBM model is obtained from "SALA: SpeechDat across Latin America" telephone speech database (Venezuelan version). It contains 1000 speakers uttering in each telephone call a total of 45 read and spontaneous items. For training and test utterances, NIST2001 Ahumada database was used. Ahumada

is a speech database of 104 male Spanish speakers. 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.

Well known MFCC features have been used to represent the short time speech spectra. All telephone speech signals are quantized at 16 $bits$ at 8000 Hz sample rate, pre-emphasized with a factor of 0.97, and an energy based silence removal scheme is used. At last, the Δ cepstral features from MFCC normalized cepstral feature are obtained and appended to MFCC features conforming a 24-dimensional $MFCC + \Delta$ feature vector.

Score normalization method for small evaluation databases is proposed. For each score L between a target X_A and a test X_B, the normalized score is

$$Lnorm(X_A, X_B) = L(X_A, X_B) - mean(L(X_A, I_s)) \tag{1}$$

Where I_s is a subset of impostors. As 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 the normalization and viceversa. The test from the two subset is polled together in order to obtain the global performance of a given system.

4 Speaker Verification Experiments

4.1 GMM-UBM Speaker Verification Baseline

First, a GMM-UBM speaker verification baseline using the data and methods explained in Section 3 was established. A UBM model with $M = 2048$ Gaussian components was trained with 1989 speech utterances from SALA database. GMM-UBM models of 100 speakers were MAP adapted [2] using as training utterances their spontaneous utterances of session T1 of Ahumada database. For verification, 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] and explained in section 2.3, using R=10 components. Score normalization is applied as described in section 3. Results were evaluated on DET curve, obtaining and EER = 4%; the NIST evaluation criteria, minimal of "detection cost function" was evaluated too, minDCF= 2.29%.

4.2 Speaker Verification Experiments Using Training and Test Data to Select Gaussian Components from de UBM Model

Two experiments were performed with both classification methods explained in section 2.3, selecting k= 250, 300, 350, 400 and 500 Gaussian components. Table 1 reflects the results of both experiments for different k.

Table 1. EER and minDCF results of experiments

	Selection with Method 1		Selection with Method 2	
k	% EER	% minDCF	% EER	% minDCF
250	5.00	2.47	4.00	2.28
300	4.52	2.39	4.36	2.41
350	4.00	2.22	4.05	2.28
400	5.00	2.29	4.20	2.22
500	5.00	2.23	4.18	2.21

Experimental results of the proposed methods show:

a. Redundancy reduction in the selected Gaussian components:
   As shown, experiment using method 1 with $k = 350$ Gaussian component and experiment using method 2 with $k = 250$ Gaussian components, get the same % EER and less % minDCF related to the GMM-UBM baseline, with a respective reduction of 82.9% and 87.7% of the Gaussian components of the original GMM-UBM speaker model (2048). The non-selected Gaussian components are less discriminative of the speaker or not discriminative at all. This reduction of information lowers verification phase computational burden, due to the use of fewer number of Gaussian components.
b. Classification method using the test utterance is better:
   Method 2 obtains similar results as method 1 with less Gaussian components (250 vs. 350); this method is more adjusted to the test speaker because it selects the components of the GMM-UBM model from the test utterance, very similar to Reynolds method [1] but less expensive.

5 Conclusion and Future Work

In the presence of real or embedded applications of speaker verification, classical GMM-UBM [1] method is not sufficient enough because of the high dimensionality of the GMM-UBM model and the existence of non-discriminative and redundant information in them.

Experimental results show that GMM-MAP adaptation of UBM model represent speech utterances not efficiently , containing many non-discrimative and useless Gaussian components. So we can argue that only about 20% of the Gaussian components of the speaker model is as effective as all the model. In conclusion, this results using GSC-CV criteria show that an important reduction of the models, more than 80%, is reached, with similar performance in speaker verification experiments. Of course, the volume reduction will depend on the databases used, but it is present. The use of the GSC-CV method of Gaussian components selection would reduce the computational and memory cost of classifying stage in real applications of speaker verification.

As 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.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17(1), 91–108 (1995)
2. Reynolds, D.A., Quatieri, T., Dunn, R.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10(1), 19–41 (2000)
3. Saeidi, R., Sadegh Mohammadi, H.R., Ganchev, T., Rodman, R.D.: Particle Swarm Optimization for Sorted Adapted Gaussian Mixture Models. IEEE Trans. on Audio, Speech, and Language Processing 17(2), 344–353 (2009)
4. Auckenthaler, R., Mason, J.: Gaussian selection applied to text independent speaker verification. In: Proceedings of Speaker Odyssey: the Speaker Recognition Workshop, Crete, Greece, pp. 83–88 (2001)
5. Xiang, B., Berger, T.: Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. IEEE Transactions on Speech and Audio Processing, 447–456 (2003)
6. Kinnunen, T., Karpov, E., Franti, P.: Real-time speaker identification and verification. IEEE Transaction on Audio, Speech and Language Processing 14(1), 277–288 (2006)
7. Mohammadi, H.R.S., Saeidi, R.: Efficient implementation of GMM based speaker verification using sorted Gaussian mixture model. In: Proc. EUSIPCO 2006, Florence, Italy (2006)
8. Saeidi, R., Kinnunen, T., Mohammadi, H.R.S., Rodman, R., Fränti, P.: Joint frame and gaussian selection for text independent speaker verification. In: IEEE Trans. ICASSP 2010, pp. 4530–4533 (2010)
9. Liu, Q., Huang, W., Xu, D., Cai, H., Dai, B.: A fast implementation of factor analysis for speaker verification. In: Interspeech 2010, pp. 1077–1080 (2010)
10. Anguera, X., Bonastre, J.F.: A Novel Speaker Binary Key Derived from Anchor Models. In: Proceedings of Interspeech (2010)