Clustering Algorithm in Automatic Speaker Verification

Djellali Hayet[1], Laskri Mohamed Tayeb[2]
[1]Network and Security Laboratory
[2]L.R.I Laboratory
Department of Computer Science
Badji Mokhtar University
Annaba, Algeria
hayetdjellali@yahoo.fr, mtlaskri@univ-annaba.dz

ABSTRACT: *We propose a new modeling approach in Automatic Speaker Verification A.S.V based on Gaussians Mixtures Models and Maximum a posteriori adaptation MAP. We propose clustering algorithm for intra and inter speaker's variability in voice module and contribute for Universal Speaker Model design. We compare the traditional approach which uses one specific customer model with the second called Universal speaker model USM (customers families). Voice module is applied for characterizing customers only; Universal Speaker Model is applied when speaker model is weak and designed for computing a reliable score.*

Keywords: Intra Speaker Variability, Gaussian Mixtures Models, MAP Adaptation, Voice Module, Kmeans, Universal Speaker Models

Received: 19 June 2011, Revised 28 July 2011, Accepted 3 August 2011

1. Introduction

Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their larynx sizes, vocal tract shapes, and others parts of their voice production organs are different.

The Automatic Speaker Verification (ASV) can decide for acceptance or rejection if the acoustic signal and the identity proclaimed provided input or not originating from the same person.

The Gaussian mixture model (GMM) is the most popular model for text-independent recognition. In training paradigm, models can also be categorized into generative and discriminative models. The discriminative models such as support vector machines and artificial neural networks(ANNs), in contrast, model the boundary between speakers. The generative models such as GMM and VQ estimate the feature distribution within each speaker.

In Training phase, a preprocessing step and feature extraction is necessary then, modeling speakers client and impostors (called the world models UBM: Universal Background Models) by Gaussian mixture models (GMM).

During the test phase, the parameters of the test signal are extracted and the calculation of the score (derived from the client model and the world) is made which is compared to a threshold decision. The outcome is either acceptance or rejection.

We aim to build a text-independent ASV system based on GMM-UBM and MAP adaptation and characterize each client. Through the voice module that verifies whether the characterization of the customers in terms of pitch and formant, will therefore provide a better accuracy of belonging the test signal to the formant client area.

Two problems remain in ASV, the first is the small quantity of training data, and the second is the transmission channels variations. We want to create a robust model, in small quantity of test data. We constitute customers families which have the closet vocal characteristics. We compare the traditional approach which uses one specific customer model with the second called Universal speaker model USM (customer's families). However, the customer model is kept for comparative study.

The voice module that verifies whether the characterization of the customers in terms of pitch and formant, will, therefore provide a better accuracy of belonging the test signal of client formant domain. The on line customers accesses (for example with Web sites), the required security level is not very constraining. Indeed, certain applications of ASV (others than bank accesses) prefer to authorize an impostor to reject a customer, but if our system is in front of a weak acoustic signal, why not use a group of close customers.

This paper aim to check if the universal customer modeling will give further information better to the level of the score calculation, the universal customer model goes supported the model specific to the customer. We aim to reduce EER in the presence of small training data of each customer.

We organized paper as follows, modeling and characterization speakers are introduced in Section 2, the architecture proposed in Section 3 and finally the experiments in Section 4.

2. Modelling and Speaker Charaterization

2.1 Modeling Speakers

Two models are created, the client model based on these data and the acoustic model of the world UBM whose acoustic vectors are derived from a large population of speakers other than our customers. Learning both GMM models based on the EM algorithm (Estimation-Maximization). The model GMMUBM ML, based on the estimation of Maximum Likelihood ML (Maximum Likelihood).

The Maximum A Posteriori MAP approach is to use the world model and client training data to estimate the client model on the basis of these data learning and MAP Adaptation [1] [2] [3].

2.2 Speaker characterization

Fundamental frequency (F0) is the most important prosodic parameter. Combining F0-related features with spectral features has been shown to be effective, especially in noisy conditions. Hence, an efficient F0 extractor and an accurate F0 estimate calculated can be used in an algorithm for gender identification [4]. "*Pitch*" is the perception of overall frequency in a speech sound. The primary acoustic correlate of pitch is fundamental frequency, which is directly determined by the vocal cord vibrations. Several works have implemented pitch extraction algorithms based on computing the short time autocorrelation function of the speech signal [6].

Other prosodic features for speaker recognition have included duration (phone duration), speaking rate, and energy distribution modulations among others (Adami et al., 2003[1]; Reynolds et al. 2003[7]). In Shriberg et al 2005[8] it was found out, among a number of other observations, that F0 related features yielded the best accuracy, followed by energy and duration features.

3. Proposed A.S.V Architecture

We propose a system based on GMM-MAP helped by a voice module. We describe the different modules (Figure 1) of our ASV architecture which includes:

3.1 Training Phase

We first build UBM model, both speaker model(SM) and Universal speakers models USM by MAP Adaptation.

3.1.1 Preprocessing and Features Extraction P.F.E

Silence Detection SD: We remove the frames of silence and noise that decrease the ASV system performance. The energy and

ZCR (zero crossing rate criterion are used to select the frames of words (high energy) and remove frames of silence (low energy).

• Features Extraction FE: Cepstral analysis is used due to its robust estimation of noisy signal[2]. We extracted 13 cepstral coefficients and their derivatives and second derivative every 10ms calculated on an analysis window of 25ms hamming error. The cepstral mean is applied (Cepstral Mean Subtraction), removing the average distribution of each cepstral parameters.

3.1.2 Modeling

• GMM-UBM-ML Modeling: It makes learning UBM male model with Gaussian mixture model and the other female UBM (from female speech). The model parameters (mean, covariance and weight of the Gaussian) are trained with the EM algorithm (Expectation-Maximization).

• GMM-MAP Speaker Adaptation: The client model is derived from the world model by adapting the GMM parameters (mean, covariance, weights) are estimated. However, experimentally, only the averages of GMM are adapted [3]. The maximization of average parameter is expressed as follows: for a Gaussian i of GMM, expressed as:

$$\mu t = \alpha t \ \mu_i^o + (1 - \alpha t) \mu_i^w \quad (1)$$

μ_i^o is client mean, μ_i^w USM mean α : is a weight that allows you to assign more or less weight, the parameters a priori by the parameters estimated on the training data. It is defined by:

$$\alpha t = ni / (ni + \tau) \quad (2)$$

ni number of frames assigned to a gaussian *i*.

τ : The relevance factor, it controls the degree of adaptation of each Gaussian in terms of frames allocated.

We build two USM DG dependent gender model by Maximum a posteriori Adaptation, Vocal module compute gender for each client. All speech of target speakers are trained. We compare the target test signal from USM GD(Gender dependent) and SM models.

3.1.3 Voice Module

The voice module use a matlab program based on calculating the average pitch AvgF0. The pitch is extracted with autocorrelation method [5].

For each client, we extract acoustic features from customer's signal, F0 and formant parameters: F1, F2, F3, F4, then, comparing these parameters extracted from the test phase, we eliminate those whose gender is different from the test signal gender. We use Kmeans algorithm to classify target speaker and his close speakers, the pseudo code is:

3.1.4 Speakers Clustering Algorithm

Assume the data lives in a Euclidean space and we want k classes for each speaker. We use a form of Biclustering (subset of 2 speakers).

Algorithm 1: Speaker Intra variability Algorithm

We define k = 2 clusters;
For every speakers:
- Assume we start with randomly located speaker cluster centers.

The algorithm alternates between two steps:
- Assignment step: Assign each f0 & formant(f0, F1, F2, F3, F4) to the closest cluster.
- Refitting step: Move each cluster center to the center of gravity of the data assigned to it.

Endfor;

We obtain speaker centroids dimension= cd = k*5 = 10 values. The kmeans algorithm give us cd*n vectors (n is the number of speakers).

Algorithm 2: Speaker Inter Variability Algorithm

We then apply kmeans again between two different speakers:
Spk1(f0, f1, f2, f3, F4; f0', f1', f2', f3', F4');
Train Phase:
For T = 1 to n-1 do *Begin*
For j:= T + 1 to n do
Begin
Apply kmeans between speaker T and j
with k = 2 clusters; Store centroid;
End;
We select speaker centroid constrant to: minimal distance between speaker T and all others speakers j
End
During Test we Estimate minimal distance for test vector among speakers centroids; we select this subset (2 speakers) and create their models with GMM MAP.

3.2 Test Phase

• Parameterization

The same treatment as when learning is done in test phase for the acoustic data of the speaker.

• Decision

To make a decision, the score will be calculated as follows:

$\text{Log}\,(p\,(X \mid \lambda_{client})$: Client Model Score proclaimed

$\text{Log}\,(p\,(X \mid \lambda_{USM})$: client model Score calculated by the voice module(close speakers).

$\text{Log}\,(p(X \mid \lambda_{UBMF}))$: Score from the female world model. Log

$(p(X \mid \lambda_{UBMM}))$: Score from the model of the world Male:

First case :

If $(\text{LLR}\,(p\,(X \mid \lambda_{client}) > \text{LLR}\,(p\,(X \mid \lambda_{USM}))$ then $\Lambda\,(X)$ is computed like this :

$\Lambda\,(X) = \Lambda 1(X)$ if Voice Module determins a men

$\Lambda\,(X) = \Lambda 2(X)$ Else : women

Knowing that $\Lambda 1(X)$ et $\Lambda 2(X)$ are calculated as follows :

$$\Lambda 1(X) = \log p\,(X \mid \lambda_{client}) - \log p\,(X \mid \lambda_{UBMM}) \quad (3)$$

$$\Lambda 2\,(X) = \log (p\,(X \mid \lambda_{client}) - \log (p\,(X \mid \lambda_{UBMF})) \quad (4)$$

We compared to a threshold θ:

If $\Lambda\,(X) > \theta$ client acces Else impostor.

Second Case : with univesal speaker models

$$\Lambda 3\,(X) = \log (p\,(X \mid \lambda_{USMM}) - \log (p\,(X \mid \lambda_{UBMM})) \quad (5)$$

$$\Lambda 4\,(X) = \log (p\,(X \mid \lambda_{USF}) - \log (p\,(X \mid \lambda_{UBMF})) \quad (6)$$

Any speaker is near to the target speaker, if $(\text{LLR}(p\,(X \mid \lambda_{client}) < \text{LLR}\,(p\,(X \mid \lambda_{USM}))$ we consider that either give the test data are corrupt or the deviation between the training data and test. In this case, we compute the score with formula (5) and (6).

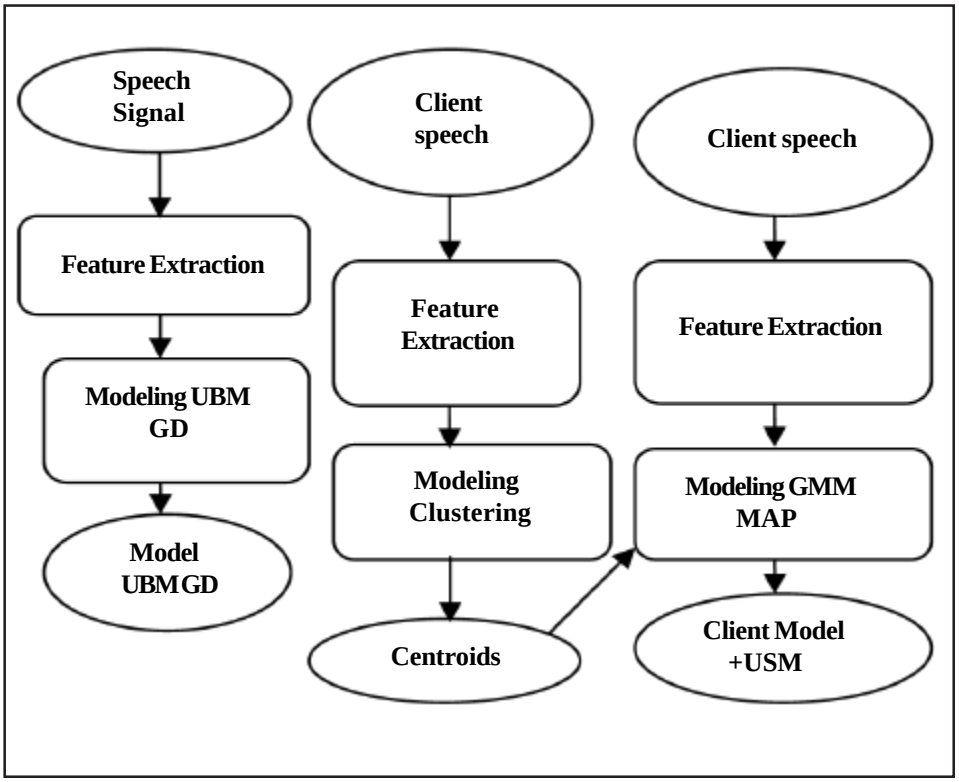

Figure 1. Proposed Automatic Speaker Verification Architecture

4. Protocol Experiment

4.1 Database and Baseline System

The database is recorded in Goldwave frequency 16KHz for a period of 60s for each speaker when learning and 30s in the testing phase. The UBM population is 15 men's and 15 women. Three sessions are recorded for each speaker at an interval of 1 month. Ten clients are registered in the database (5 men and 5 women). The ASV reference system GMMUBM- IG is independent gender obtained by merging the two models male and female speakers [7].

4.2 Voice Module

Pitch Extraction for gender detection: The speech signal is divided into segments of 60 ms, each segment is extracted every 50ms interval and requires a function autocorrelation "*pitch*" to estimate the fundamental Frequency of this segment. This algorithm was tested on speech samples from people of different gender from the basis with 16khz sampling frequency. Errors are 2%.

We used two main matlab programs, one for extracting the fundamental frequency (average pitch)1 and the second calculate the parameters F1, F2, F3, F4, gender. For each speaker, we tried 5 enrollments, therefore, there are intervals for which formant belongs and used to identify the speaker. Table 1 and 2 show the result of 3 males (M) speakers and 3 females speaker(F), Average F0 by Ellis[1] program and praat[2] , three formant F1, F2, F3, F4.

4.3 Speakers Models

We carry out the training from the nearest customer by voice module (F0 and formant F1,F2,F3,F4). We must try several

[1] Design of speaker recognition
http://www.oppapers.com/Design-Speaaker-Recognition- System-Matlab/69120.

[2] Praat http://www.fon.hum.uva.nl/praat/

customers numbers (two or more but not exceed the maximum number of client by gender). Each client is trained by MAP adaptation.

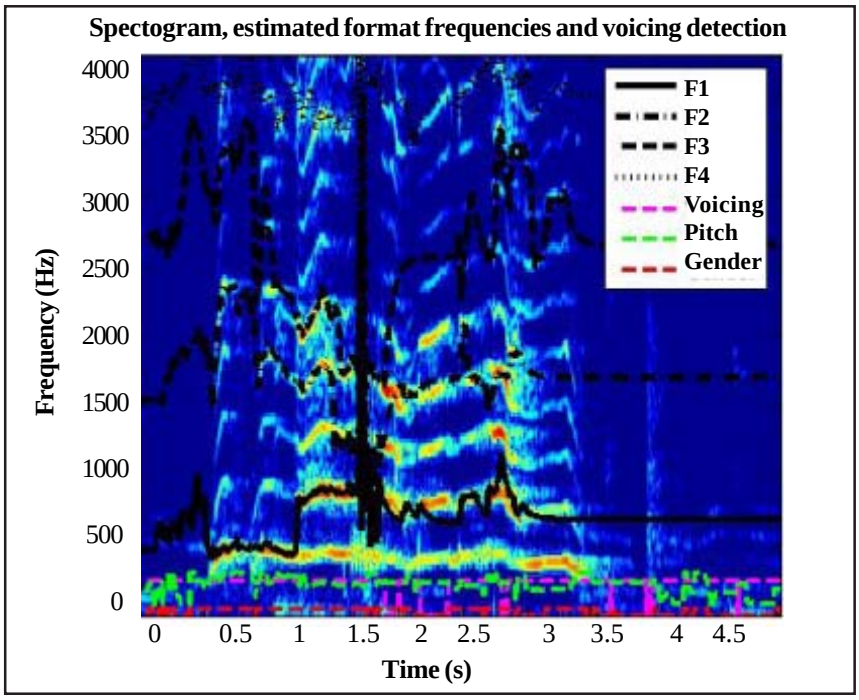

Figure 2. Female, language: Arabic, F1, F2, F3, F4

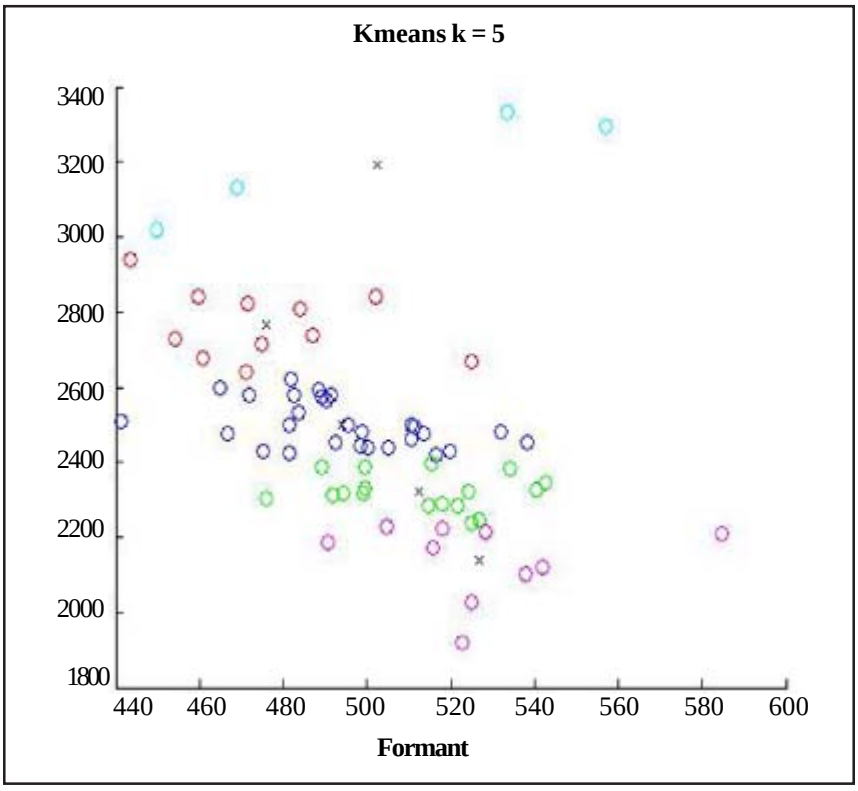

Figure 3. kmeans k = 5, F1 = function (F2)

Number gender	AverageF0 Ellis	F0 Praat
1M	146,217	174,763
2M	213,188	229,254
3M	237,225	278,615
1F	180,775	172,379
2F	194,959	202,008
3F	193,960	199,219

Table 1. Pitch Values by Program and Speakers

We built UBM models from 30 arabic speakers; UBM male with 15 male speakers and UBM female from 15 female speakers.

We project testing 8, 16, 32, 64, 128 Gaussians and classify them by gender (male, female) with vocal module. The global threshold is computed from other database: 8 male and 8 female speakers. We get for GMM MAP models the result in table 3 with only 8 mixtures.

# mixtures	8 mixtures
EER%	24%

Table 2. Equal Error Rate

4.4 Discussion

Equal error rate is high, first 8 mixtures is not enough for well speakers modeling second we didn't apply any normalization like Tnorm and finally we must test 16, 32, 128 gaussians.

Algorithm 1: The values of formants from one session to another for the same speaker are close to each others. As example the figure 3 show us the intra speaker clustering is not significant for k greater than 3 , for this reason, we choose k = 2 in other to get the distance as wide as possible for keeping the maximum variance of data(formant). The clustering built a subset of customers and contributes in well speakers modeling. We cannot prove yet if Universal Speaker model is a solution for us because not finished experiments.

5. Conclusion

The experimental section is in progress, we aim to improve the score and thus the final decision with Universal speaker models helped by voice module. The idea is to construct a family of speakers near the customer can replace our target speaker model if client is inadequate (no sufficient data or bad records).

References

[1] Adami, A., Mihaescu, R., Reynolds, D. A., Godfrey, J. J. (2003). Modeling Prosodic Dynamics for Speaker Recognition. *In*: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong.

[2] Bôstjan Vesnicer, France Mihelic. (2008). The likelihood ratio decision criterion for nuisance attribute projection in GMM speaker Verification, *Eurasip Journal On advances in Signal Processing* volume.

[3] Alexandre Preti. (2008). Thesis Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur. Académie d'Aix Marseille, Laboratoire d'informatique d'Avignon.

[4] Ville Hautomaki, Ismo Karkkainen,Tomi Kinnunen. (2008). Maximuma posteriori adaptation of the centroid model for speaker verification, *IEEE Signal Processing letters*, 15.

[5] Kamran Mustapha, Bruce, I. C. (2006). Robust Formant Tracking for continous speech with speaker variability, IEEE Transactions on Speech and audio processing. Recognition, *Speech Communication,* 46 (3-4) 455-472.

[6] Kavita Kasi B.Eng. (2002). Yet Another Algorithm For Pitch Tracking, Andhra University, India. *Master of Science Electrical Engineering*.

[7] Douglas Reynolds, Quateri Thomas, F., Dunn Robert, B. (2000). Speaker verification using Gaussian mixture models, *Signal Processing* 10.19.41.

[8] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke. Modeling Prosodic Feature Sequences for Speaker.