CARI'12

The Robustness of GMM-SVM in Real World

Applied to Speaker Verification

Nassim ASBAI, Abderrahmane AMROUCHE and Youcef AKLOUF

Speech Com. & Signal Proc. Lab.
Faculty of Electronics and Computer Sciences,
USTHB, Bab Ezzouar, 16 111, Algeria

asbainassim@gmail.com;namrouche@usthb.dz;yaklouf@yahoo.fr

RÉSUMÉ. Les modèles de mélanges gaussiens (GMM) ont montré un très grand succès pour leur utilisation dans la vérification du locuteur. Le principe standard pour les modèles GMM est d'utiliser l'adaptation MAP des moyennes des composantes du mélange basé sur la parole d'un locuteur cible. Dans ce travail, nous étudions les différents modèles (GMM-UBM et GMM-SVM) et leurs applications à la vérification du locuteur. Pour cela, des vecteurs caractéristiques, constitués par les coefficients cepstraux (MFCC), extraits du signal de parole sont utilisés pour entraîner le modèle de mélange gaussien (GMM), dont les moyennes sont ensuite utilisées pour entrainer SVM. Pour les deux systèmes GMM-UBM et GMM-SVM, 2048 composantes sont utilisées pour construire le modèle UBM. La phase de vérification a été testée avec une base de données Aurora avec différents rapport signal/ bruit (SNR) et dans trois milieux bruités.

ABSTRACT. Gaussian mixture models (GMMs) have proven extremely successful for text-independent speaker verification. The standard training method for GMM models is to use MAP adaptation of the means of the mixture components based on speech from a target speaker. In this work we look into the various models (GMM-UBM and GMM-SVM) and their application to speaker verification. In this paper, features vectors, constituted by the Mel Frequency Cepstral Coefficients (MFCC) extracted from the speech signal are used to train the Gaussian mixture model (GMM) and mean vectors issued from GMM-UBM to train SVM. To fit the data around their average the cepstral mean subtraction (CMS) are applied on the MFCC. For both, GMM-UBM and GMM-SVM systems, 2048-mixture UBM is used. The verification phase was tested with Aurora database at different Signal-to-Noise Ratio (SNR) and under three noisy conditions. The experimental results showed the outperformance of GMM-SVM against GMM-UBM in speaker verification especially in noisy environment.

MOTS-CLÉS : Vérification du locuteur, Milieu bruité, MFCC, GMM-UBM, GMM-SVM, Fonctions à noyaux, Aurora.

KEYWORDS: Speaker verification, Noisy environment, MFCC, GMM-UBM, GMM-SVM, Kernel functions, Aurora.

1. Introduction

In the last decade people have come forward to investigate various aspects of speech such as mechanical realization of speech signal, human machine interaction, speech and speaker recognition. Speaker verification has been a wide and attractive area of speaker recognition research. We consider the problem of text-independent speaker verification. That is, given a test utterance, a claim of identity, and the corresponding speaker model, determines if the claim is true or false.

There are many techniques proposed to model the speakers, e.g., vector quantization [1], hidden Markov model [2], neural networks [3] and Gaussian Mixture Model [4].The standard approach to this problem is to model the speaker using an adapted Gaussian mixture model (GMM), which belongs to the stochastic modeling and it is based on the modeling of statistical variations of the features.

In recent years, it is more common to represent speakers with Support Vector Machines (SVM) [5]. SVMs have proven to be a new effective method for speaker recognition [6], [7]. SVMs are a natural solution to the problem, since speaker verification is fundamentally a two-class problem. We want to decide between the hypothesis that the speech is produced from the speaker or the hypothesis that the speech is produced from someone else in the population. SVMs perform a nonlinear mapping from an input space to an SVM feature space. The combination of both methods GMM and SVM has been viewed as an interest direction for speaker verification task. This approach derives a GMM-supervector [8], [9] by stacking the mean vectors of a MAP-adapted GMM [4] that captures the acoustic characteristics of a speaker. The supervector is then presented to a speaker-dependent SVM for scoring.

The focus of this paper is to describe a classification scheme that incorporates both the GMM and the SVM in a way that the robustness advantage of the statistical method GMM favorably combines with the discriminative power of the SVM. This scheme is applied on text-independent speaker verification task, under various mismatched noise conditions. In this way, three types of additive noise (produced by airport, train-station and subway) are added to speech signal issued from the Aurora database.

The remainder of the paper is structured as follows. In sections 2 and 3, we discuss the GMM and SVM classification methods and briefly describe the principles of GMM-UBM at section 4. In section 5, the experimental protocols used in this work are described. In section 6, experimental results of the speaker verification in noisy environment using GMM-UBM and GMM-SVM systems based using Aurora database are presented. Finally, a conclusion is given in Section 7.

2. Gaussian Mixture Model (GMM)

In GMM model [10], there exist k underlying components $\{ \omega_1, \omega_2, \ldots, \omega_k \}$ in a d-dimensional data set. Each component follows some Gaussian distribution in the space. The parameters of the component ω_j include $\lambda_j = \{\mu_j, \Sigma_j, \pi_j\}$, in which μ_j = (μ_j [1], . . . , μ_j [d]) is the center of the Gaussian distribution, Σ_j is the covariance matrix of the distribution and π_j is the probability of the component ω_j. Based on the parameters, the probability of a point coming from component ω_j appearing at $x = (x[1],...,x[d])$ can be represented by

$$p(x/\lambda_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma^{-1}(x-\mu_j)\right\} \tag{1}$$

Thus, given the component parameter set $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_k\}$ but without any component information on an observation point x, the probability of observing x is estimated by

$$p(x/\lambda) = \sum_{j=1}^{k} p(x/\lambda_j)\pi_j \tag{2}$$

Under the assumption of independent feature vectors, the log-likelihood of a model λ for a sequence of feature vectors X = { x_1, x_2 , . . . , x_n }, is computed as follows:

$$\log p(X/\lambda) = \frac{1}{n}\sum_t \log p(x_t/\lambda) \tag{3}$$

where $p(x_t/\lambda)$ is computed as in equation (2). Note that the average log-likelihood value is used so as to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by *n* can be considered a rough compensation factor.

3. Support Vector Machines (SVMs)

Support vector machine (SVM) [9] is one of the most robust classifiers in speaker identification. It has been applied both with spectral, prosodic, and high-level features. SVM has been successfully combined with GMM to increase accuracy.

One reason for the popularity of SVM is its good generalization performance to classify unseen data. The SVM, is a binary classifier which models the decision boundary between two classes as a separating hyperplane. In speaker verification, one class consists of the target speaker training vectors (labeled as +1), and the other class consists of the training vectors from an ‘‘impostor” (background) population (labeled as -1). Using the labeled training vectors, SVM optimizer finds a separating hyperplane that maximizes the margin of separation between these two classes. Formally, the discriminate function of SVM is given by equation below,

$$f(x) = class(x) = sign\left[\sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d\right]. \qquad (4)$$

where $t_i \in \{+1, -1\}$ are the ideal output values, $\sum_{i=1}^{N} \alpha_i t_i = 0$ and $\alpha_i > 0$.

The support vectors x_i, their corresponding weights α_i and the bias term d, are determined from a training set using an optimization process. The kernel function $K(.,.)$ is designed so that it can be expressed as $K(x, y) = \Phi(x)^T \Phi(y)$ where $\Phi(x)$ is a mapping from the input space to kernel feature space of high dimensionality. The kernel function allows computing inner products of two vectors in the kernel feature space. In a high-dimensional space, the two classes are easier to separate with a hyperplane. To calculate the classification function class (x) we use the dot product in feature space that can also be expressed in the input space by the kernel [7]. SVMs were originally designed primarily for binary classification [8].

4. Gaussian Supervector SVM

The supervectors of a GMM-UBM [9], are formed by concatenating the mean of each Gaussian component [10]. For each enrolment utterance, a GMM is trained with the extracted spectral features, and the corresponding supervector is obtained. Instead of training the GMM via EM algorithm, we adapt the GMM from a universal background model (UBM), which is widely used in speaker recognition. The UBM is a GMM trained via EM algorithm using speech from a large number of speakers. The adaptation of each utterance’s GMM is performed with maximum *a posteriori* (MAP) algorithm , and only the means are adapted. Each of the corresponding Gaussian components has the same weight and covariance matrix, therefore the derived GMMs’ supervectors are comparable in the supervector space. The GMM supervector can be considered as a mapping from the spectral features of an utterance to a high-dimensional feature vector. This mapping allows the production of features with a fixed dimension for all the utterances. Therefore, we can use the GMM supervectors as input for SVM learning.

5. Experimental protocol

The speech database used in this work is issued from the AURORA database. It consists of a set of 10 digits of the English language (zero to nine + letter O) spoken in sequences, by 104 speakers of both genders (52 male + 52 female) with eight sequences for each speaker (i.e. five sequences (104x5=520 utterances) for training set and three sequences (i.e. 104x3=312 utterances) for test set). This database was recorded in **“**.08**”** format, with a sampling frequency equal to 8 kHz. To simulate the impostors, UBMs with 2048 mixture Gauss number were trained using EM to model 100 unknown speakers (50 female and 50 male), with five phrases spoken in English by each unknown speaker. To simulate the real environment we used noises specific to the database AURORA (train-station, airport and subway). In parameterization phase, we specified the feature space used. Indeed, as the speech signal is dynamic and variable, we presented the observation sequences of various sizes by vectors of fixed size. Each vector is given by the concatenation of the coefficients mel cepstrum MFCC [13] (12 coefficients), these first and second derivatives (24 coefficients), extracted from the middle window every 10 ms. A cepstral mean subtraction (CMS) [10] is applied to these features in goal to fit the data around their average. We used a Detection Cost Function (DCF) and Equal Error Rate (EER) as the evaluations metric. Score normalization (T-norm) is applied to the scores issued from the GMM-UBM model, in goal to improve the verification rate. To calculate the classification function class (x) in SVM model, we used the RBF kernel.

6. Experiment results

6.1. Speaker verification in quiet environment using GMM-UBM and GMM-SVM approaches

The results in terms of Equal-Error Rate (EER) shown by the DET curve in Figure 1:

1- Used GMM-UBM is 8.24**%**.

2- Used GMM supper vector means adapted by MAP estimation, as input to the support vectors machines SVMs (GMM/SVM) is 2.08**%**.

From the same figure, it can be observed that, the performance of the GMM/SVM is superior to the GMM-UBM in terms of EER. It is also noticed, the integration of GMMs in SVMs brings improvement in accuracy rate in quiet environment.

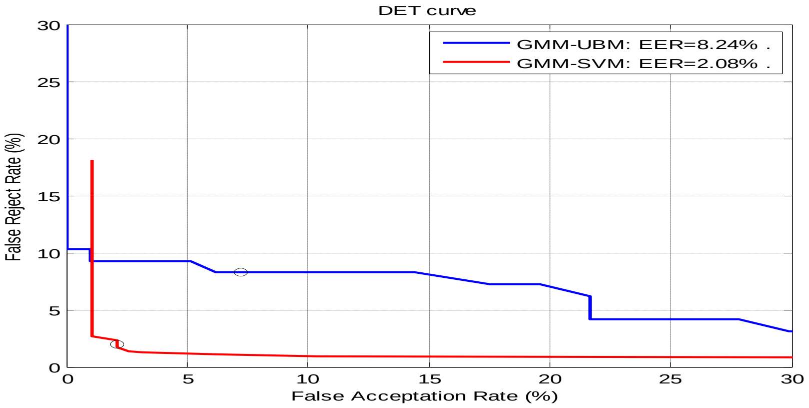

Figure 1. *Speaker verification detection error tradeoff (DET) curves for the Aurora corpus, tested on all 104 speakers*

6.2. Speaker verification in noisy environment using GMM-UBM and GMM-SVM approaches

The goal of the experiments doing in this section is to evaluate the verification performance of GMM-UBM, GMM–SVM when the quality of the speech data test is contaminated with different levels of different noises specified to AURORA database. This provides a range of speech SNRs (0, 5, 10 and 15 dB). As expected, it is seen that there is a drop in accuracy for these approaches with decreasing SNR. From The table below, it is seen that the performance of GMM-SVM appears better than GMM-UBM. Because SVM scoring approach is superior to the conventional likelihood-ratio scoring (GMM-UBM). Indeed, the contribution of individual background speakers and the target speaker to the verification scores can be optimally weighted by the Lagrange multipliers of the target-speaker's SVM. Otherwise to say, maximum likelihood convergence does not translate to optimal classification if a priori assumptions about the data are not correct. So, the problem of finding the optimal decision boundary still remains (zone of confusion) in GMM-UBM. However, when we weight the supervectors issued from GMM-UBM by Lagrange multipliers α_i (see section 3, eq. (4)), the decision boundary is directly learned from the data (GSV). Otherwise, SVM eliminates zone of confusion between classes by finding a good hyperplan which separates classes.

Table 1. EERs in speaker verification for GMM–UBM and GMM–SVM under mismatched data conditions using real world noise.

GMM-UBM								
Noises	Test data							
	SNR : 0 dB		SNR : 5 dB		SNR : 10 dB		SNR : 15 dB	
	EER %	minDCF	EER %	minDCF	EER %	minDCF	EER %	minDCF
Airport	48.45	0.153	36.66	0.120	18.33	0.098	12 .56	0.074
Train-station	49.33	0.193	37.33	0.153	25.67	0.101	14.97	0.099
Subway	46.2	0.123	25	0.123	16.33	0.088	11.08	0.061
GMM-SVM								
Noises	Test data							
	SNR : 0 dB		SNR : 5 dB		SNR : 10 dB		SNR : 15 dB	
	EER %	minDCF	EER %	minDCF	EER %	minDCF	EER %	minDCF
Airport	7	0.094	5 .49	0.092	3.44	0.027	3.02	0.02
Train-station	8.12	0.098	5.79	0.097	3.63	0.069	3.32	0.044
Subway	7.31	0.094	5.58	0.047	4.06	0.047	3.68	0.047

7. Conclusion

The aim of our study in this paper was to evaluate the contribution of kernel methods in improving system performance of automatic speaker verification in the real environment, often represented by an acoustic environment highly degraded. Indeed, the determination of physical characteristics discriminating one speaker from another is a very difficult task, especially in adverse environment. For this, we developed a system of automatic speaker verification on text independent mode, part of which verification is based on classifier using GMM-UBM, especially the system hybrid GMM-SVM, which the vector means extracted from GMM-UBM with 2048 mixtures for UBM in step of modeling are inputs for SVMs in phase of decision. The results we have achieved conform all that GMM-SVM technique is very interesting and promising especially for tasks such as verification in noisy environments.

8. Bibliographie

[1] K. Yu, J. S. Mason, J. Oglesby, “Speaker recognition using hidden Markov models, dynamic time warping and vector quantization“, IEE vision, image and signal processing, Berlin, 1995.

[2] N. Minh, M. Do. “Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models”, IEEE Signal Processing Letters,vol. 10, no. 4, pp. 115–118, 2003.

[3] A. Amrouche .“Reconnaissance automatique de la parole par les modèles connexionnistes“. Thèse de doctorat, faculté d'électronique et d'informatique, USTHB. 2007.

[4] D. A. Reynolds, T. F. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Dig. Signal Process.*, vol. 10, no.1-3, pp. 19–41, 2000.

[5] N. Cristianni, J. Shawe-taylor, “An introduction to support vector machines and other kernel-based learning methods“, Cambridge University Press, 2000

[6] W. M. Campbell, “Generalized linear discriminant sequence kernels for speaker recognition,” in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, pp. 161–164, 2002

[7] V. Wan and S. Renals, “Speaker verification using sequence discriminant support vector machines,” *IEEE Trans. Speech Audio Processing*, vol.13, no. 2, pp. 203–210, Mar. 2005.

[8] W.M. Campbell, D.E. Sturim, D.A. Reynolds, “Support vector machines using GMM supervectors for speaker verification”. IEEE Signal Process. Lett. 13, 308–311, 2006.

[9] W.M. Campbell, D.E. Sturim, D.A. Reynolds, A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation”. In: Proc. ICASSP'06, Vol. 1, pp. 97–100, 2006.

[10] G. McLachlan, D. Peel, “Finite mixture models“, Wiley-Interscience. 2000.

[11] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, ”Score normalization for text-independent speaker verification systems”. Digital Signal Process. 10 (1–3), 42–54, 2000.

[12] M. Ben, F. Bimbot, “D-MAP: a Distance-Normalized MAP Estimation of Speaker Models for Automatic Speaker Verification,” in *ICASSP*, vol. 2, pp. 69–72, 2003.

[13] T. Kinnunen , H. Li , “An overview of text independent speaker recognition: From features to supervectors”, Speech Communication 52 (2010) 12–40, ScienceDirect, August 2009.