Combining GMM’s with Support Vector Machines for Text-independent Speaker Verification

Jamal Kharroubi, Dijana Petrovska-Delacrétaz, Gérard Chollet

Ecole Nationale Supérieure des Télécommunications
Signal and Image Department, CNRS- LTCI
46, rue Barrault, 75634 Paris Cedex13, France
{kharroub, petrovsk, chollet}@tsi.enst.fr

Abstract

Current best performing speaker recognition algorithms are based on Gaussian Mixture Models (GMM). Their results are not satisfactory for all experimental conditions, especially for the mismatched (train/test) conditions. Support Vector Machine is a new and very promising technique in statistical learning theory. Recently, this technique produced very interesting results in image processing [2], [3], [4] and for the fusion of experts in biometric authentication [5]. In this paper we address the issue of using the Support Vector Learning technique in combination with the currently well performing GMM models, in order to improve speaker verification results.

1. Introduction

Current best performing speaker recognition algorithms are based on Gaussian Mixture Models (GMM). Although quite high-accuracy results can be reached with this method, it still does not give satisfactory performance for all experimental conditions. There still remains a tremendous performance gap to be bridged between matched and mismatched train/test conditions. Support Vector Learning Theory is a new and promising technique in statistical learning theory. In this paper we address the issue of using the Support Vector Learning technique in combination with the currently well performing GMM models, in order to improve speaker verification results.

There has been some attempts to use Support Vector Machines for speaker verification purposes. In [1] the classification method based on SVM’s uses speech feature vectors based on widely used pre-treatment procedures in speech recognition. These feature vectors are used as input vectors for the Support Vector Machine (SVM). It is well known that these spectral based features convey concomitant speaker and channel information that are bound together in an unknown way. Another way to circumvent this problem is to use the SVM but after the modeling stage. The proposed work is organized as follows: Section 2 gives a short description of the basic principles of the SVM’s. In Section 3 we describe the way we have coupled the SVM’s with the existing GMM speaker modeling methods. The details concerning the database and the experimental protocol are given in Section 4. Experimental results are described in Section 5, and conclusions are given in Section 6.

2. Support Vectors Machines

2.1. The linear separable case

The support vector machines is a new technique of the statistical learning theory proposed by Vapnick in 1995 and developed from the Structural Risk Minimization (SRM) theory [6] [9]. They belong to the family of universal learning machines that implement the strategy of keeping the empirical risk fixed and minimizing the confidence interval. The SVM’s achieve the structural risk minimization inductive principle by mapping the input vector into high-dimensional feature space using a non-linear transformation chosen a priori. In this space an optimal Separating Hyperplane is considered, and the goal is to minimize the bound on the generalization error of a model.

Given a set of training data \( x_1, x_2, \ldots, x_m \), belonging to two different classes labeled +1 and -1, each point \( x_i \) is associated to \( y_i = 1 \) if \( x_i \) belong to the class +1 and \( y_i = -1 \) if \( x_i \) belong to the class -1. These training set is projected to the high dimensional space by the transformation \( \phi \). Then each hyperplane in the feature space \( H: w \cdot \phi(x) + b \) separating the two classes must satisfy the following conditions:

\[
\begin{align*}
w \cdot \phi(x_i) + b & \geq 1 \quad \text{if } y_i = 1; \\
w \cdot \phi(x_i) + b & \leq -1 \quad \text{if } y_i = -1;
\end{align*}
\]

which is equal to:

\[
y_i[w \cdot \phi(x_i) + b] \geq 1.
\]

The optimal separating hyperplane is the one maximizing the margin \( M \) given by the equation:

\[
M = \min_{x_i \in +1} \left[ \frac{w \cdot \phi(x_i) + b}{|w|} \right] - \max_{x_i \in -1} \left[ \frac{w \cdot \phi(x_i) + b}{|w|} \right]
\]

\[
= \frac{2}{|w|}
\]

(4)

to maximize the margin \( M \), one need to minimize

\[
\Phi = \frac{w^2}{2};
\]

(5)

using the Lagrange multipliers and the Kuhn-Tucker theorem. The problem can be translated to the following dual problem [6]:

Minimize:

\[
\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j);
\]

(6)

(7)
subject to the constraints:

\[ \alpha_i \geq 0; \quad \sum_{i=1}^{m} \alpha_i y_i = 0; \]  

where the \( \alpha_i \) are the Lagrangian multipliers with one Lagrange multiplier for each training data point. The support vectors are the point \( x_i \) for which \( \alpha_i \) is strictly positive.

### 2.2. The non-linear separable case

In this case the training samples of two classes are non-linearly separable. To solve this problem, a set of penalty variables \( \Xi = (\xi_1, \ldots, \xi_m) \) are introduced such that the constraints are looser. Then the hyperplane must satisfy the following inequality:

\[ y_i [w \cdot \phi(x_i) + b] \geq 1 - \xi_i. \]  

The objective is to minimize

\[ \Phi = \frac{w^2}{2} + C \sum_{i=1}^{m} \xi_i; \]  

where \( \Xi = (\xi_1, \ldots, \xi_m) \) and \( C \) are constants.

The dual problem corresponding to this case is slightly different from the linear separable case. The goal is to minimize:

\[ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j); \]  

subject to the constraints:

\[ 0 \leq \alpha_i \leq C; \quad \sum_{i=1}^{m} \alpha_i y_i = 0. \]  

In the linear and non-linear cases the optimal separating hyperplane defined by \( u_0 \) and \( b_0 \) is determined as follows:

\[ H_0 : u_0 \cdot \phi(x) + b_0; \]

where

\[ u_0 = \sum_{i=1}^{m} \alpha_i \phi(x_i) y_i; \]

and \( b_0 = 1 - u_0 \cdot x_i \) for \( x_i \) with \( y_i = 1 \).

The classification function is:

\[ \text{class}(x) = \text{Sign}(u_0 \cdot \phi(x) + b_0); \]

\[ = \text{Sign}(b_0 + \sum_{i=1}^{m} \alpha_i y_i \phi(x_i) \cdot \phi(x)); \]

\[ = \text{Sign}(b_0 + \sum_{i=1}^{m} \alpha_i y_i K(x_i, x)). \]  

where \( K \) is the Kernel. Some widely used kernels are:

- Linear: \( K(u, v) = u \cdot v \);
- Polynomial: \( K(u, v) = (u \cdot v + 1)^p \);
- Radial Basis Function (RBF): \( K(u, v) = \exp[-\gamma \|u - v\|^2] \).

### 3. Combining GMM’s and SVM’s for Speaker Verification

In this Section we describe the way we have combined the GMM’s and the Support Vector Machines for the Speaker Verification purpose. Let us remain the goal of speaker verification experiments: given a client speaker, with a limited amount of training data, given a test speech segment, whith a claimed identity, we have to decide if that speech test data belongs to the claimed speaker or not. As already mentioned in Section 1 Gaussian Mixture Models are the state-of-the art for text-independent speaker modeling. They are used to obtain the likelihood ratios of the hypothesized speaker model and the background model. We are going to denote this method as a Likelihood Ratio (LLR) technique. Better performances can be obtained with handset score normalization.

Support Vectors Machines have shown good performance in pattern classification problems, where labelled data for each class are available in the development phase. The motivation behind this work is the good performance of SVMs in image processing and particularly in face recognition [2], [3].

We have combined these two techniques in the following way. We use the GMM technique to construct the input feature representation for the SVM classifier. The speaker GMM are built by adaptation of the background model of the same sex and same handset [8]. We use these speaker and background models to build a new input vector for the SVM classifier. We are going to train this classifier to discriminate between the client access class and the impostor access class in the following way: suppose we have a given speech segment \( X \) and the client model \( \lambda \). First, all the components of the SVM vector are initialized to zero. For each frame \( x_i \) of the speech segment \( X \), the \( \text{Log}[P(x_t / y_t)] \) score is computed for each gaussian \( y_t \) of the client model \( \lambda \) and of the background model. Suppose that the clients and background models have \( n \) gaussians, then the dimension of the vectors to train the SVM model is \( 2 \times n \).

The first \( n \) components of the SVM input vectors correspond to the client gaussian model and the last \( n \) components to the background gaussians model. Then the argument of the gaussian maximizing those scores is incremented by the maximum score. At the end, we normalize all components of the SVM vector obtained by the number of the segment frames. If the segment \( X \) was pronounced by the client \( \lambda \) the vector will be labeled by 1. Otherwise the vector will be labeled by -1. The SVM classifier is trained using all the vectors obtained from all the labelled client and impostor access from the development set.

In the evaluation phase, we first proceed to the adaptation of the background GMM model, with the client training data. Then for each test segment \( Y \) and proclaimed identity \( \beta \), we construct one vector as for the training phase. If the score obtained by testing this vector using the SVM model obtained on development phase is positive, the system decides that the segment test \( Y \) is produced by the client. Otherwise the system decides that the segment is produced by an impostor.

For baseline comparison we are going to use the Likelihood Ratio (LLR) technique, using GMM speaker modeling.

### 4. Database and experimental protocol

The complete NIST’99 evaluation data is divided into three disjoint sets: a development set (50 female and 50 male speakers), an evaluation set (50 female and 50 male speakers), and a set of 150 speakers for the background models. Four gender and
handset dependent background models with 128 gaussians are trained with speech data coming from the background speakers. We apply standard front-end processing to the speech data leading to Linear Filterbank Cepstral Coefficients (LFCC). The dimension of the input feature vectors for the GMM is 33 (16 cepstral coefficients, 16 delta coefficients, and the delta of the energy). The speaker GMM are built by adaptation of the background model, according to [8], using approximately 2min of speech data per speaker. The supervised training of the SVM classifier, as described in the previous section, is realized with about 5190 impostors access and 519 clients access. The speech duration of the test segments is 3-45s. For the evaluation phase, about 4990 impostor tests and 499 true-speaker tests were used.

Because we want to study in more details the mismatched train/test conditions, the test data are pooled in two cases: DNDT=Different-Number-Different-Type and DNST=Different-Number-Same-Type.

The SVM classifier is the one from the SvmFu package that is available as shareware on http://svm.first.gmd.de/ with a linear SVM kernel.

5. Experimental Results

In this paper we are going to compare the speaker verification results based on our new hybrid GMM-SVM method, with the current state-of-the art GMM method using LLR technique. We are also going experimente the normalization using our new hybrid GMM-SVM method.

Figure 1: Det curves on a sub-set of NIST 1999 data using a hybrid GMM-SVM method and the classical LLR technique without normalization

Figure 1 presents four Det curves [10] corresponding to the performances obtained by our new hybrid GMM-SVM method and the current state-of-the art GMM method using LLR technique on two different conditions DNDT and DNST. The curves labeled dndt-SVM and dnst-SVM has been obtained without normalisation. The curves labeled dndt-LLR and dnst-LLR correspond to the performances of the reference system obtained using classical Log-likelihood ratio technique. Figure 1 shows that the performances of the system using the SVM technique are better for the two different conditions. The ERR concerning the dndt condition is about 21% compared to 27% obtained by the LLR system. The ERR is about 15% obtained by the SVM system compared to 16% obtained by the LLR system for the dnst condition.

Figure 2: Det curves on a sub-set of NIST 1999 data using a hybrid GMM-SVM method without and with hnorm normalisation

Figure 2 presents four Det curves [10] corresponding to the performances obtained by our new hybrid GMM-SVM method without and with hnorm normalisation [11] on two different conditions DNDT and DNST. The curves labeled dndt-SVM and dnst-SVM has been obtained without normalisation. The curves labeled dndt-SVM-norm and dnst-SVM-norm has been obtained using the hnorm normalisation. Figure 2 shows that the performances obtained by hnorm are better than DNST condition. The ERR is about 13.5% compared to 15% obtained by the GMM-SVM method without normalisation. Concerning the DNDT condition the curves obtained shows that there is no improvement using hnorm. The ERR is about 21%.

6. Conclusion

Taking account that it is a one of the first approach using the support vector machines technique in text-independent speaker verification, the results obtained seem to be very promising. There is a lot of potential for improvement of the SVM system. Different kernel types and parameters need to be experimented. In the near future, SVMs will be adjusted for each client and some other normalisation techniques like znorm, tnorm will be applied.

7. References


