SVM Kernel Adaptation in Speaker Classification and Verification

Pedro J. Moreno, Purdy P. Ho

HP Cambridge Research Laboratory,
One Cambridge Center, Cambridge MA 02142, U.S.A.
{pedro.moreno,purdy.ho}@hp.com

Abstract

Techniques for speaker classification and verification based on discriminant classifiers such as Support Vector Machines (SVM’s) are becoming more and more popular. However, when compared with state of the art statistical based techniques such as Gaussian Mixture Models, their performance suffer for two main reasons: first their inability to scale up and handle a large number of classes, and second their inability to adapt model parameters. In this paper we address the second issue.

Previously [8] we have introduced a kernel based classifier that combines the best of generative methods and discriminative classifiers. Each utterance is fitted with a generative model such as a Gaussian Mixture Model (GMM) and a kernel distance is defined among GMM’s. In this paper we extend this kernel with the ability to adapt to the speaker utterance by adapting the utterance GMM using Maximum Likelihood Linear Regression (MLLR) techniques. Our experimental results on two different speaker databases show that kernel adaptation is a promising technique highly effective on long utterances when compared with non-adapted kernels.

1. Introduction

Statistical based classifiers such as Gaussian Mixture Models have been successfully used for years and to date remain the technology of choice for speaker classification and verification (see for example [9]). Their ability to model speakers with a small number of statistical parameters (prior probabilities, mean vectors and covariance matrices) and the availability of adaptation techniques [7] to tie parameters make GMM’s extremely attractive and easy to use and implement.

However, during recent years the application of discriminative techniques such as Support Vector Machines (SVM’s) [13] for speaker classification and verification has gain popularity and has been explored in many research papers. For example, among others [2, 11] compare the use of traditional based kernel SVM’s with Gaussian classifiers, examine the use of SVM’s for phonetic classification, and [3] studies the use of SVM’s to classify telephone handsets based on speech signals.

SVM’s are model free methods that do not make any distributional assumptions about the data and at the same time offer a discriminative solution to classification problems with strong bounds on error minimization. The study of kernels has also gained importance in the last years in the machine learning community. Beyond traditional kernels such as linear, Gaussian or polynomial new kernels have been designed to model DNA and protein strings, to model images and audio signals[14]. These new kernels try to take advantage of the nuances of specific signals.

Techniques that attempt to combine both approaches, i.e., generative classifiers such as GMM’s and discriminative methods such as SVM’s have become the focus of intense research in the last year, for example [5, 4, 8] explores the use of kernels defined on exponential family generative models (GMM’s, HMM’s, Naive Bayes, etc) and introduce several distributional distances such as symmet-ric Kullback-Leibler, Bhattacharyya and Expected Likelihood dis-tances to compare distributions.

In this paper we try to bring another advantage of generative classifiers to SVM’s; their ability to adapt model parameters to the individual nuances of the data. Of all possible adaptation schemes we focus on Maximum Likelihood Linear Regression (MLLR), a technique quite popular in the speech and speaker recognition community [7, 6]. Rather than adapting the SVM and its parameters (support vectors, their weights) we focus on adapting kernel parameters such as means and variances.

The outline of this paper is as follows. In section 2 we give a brief introduction to SVM classifiers, the Kullback-Leibler, Bhattacharyya and other generative based kernels. In section 3 we briefly describe the MLLR adaptation scheme and how it is applied to kernel parameters. We follow in section 4 describing the experimental databases and our results on two different speaker corpora. Finally, we conclude the paper and suggestions for future work in section 5.

2. SVM’s and Generative Based Kernels

Support Vector Machines were first introduced by Vapnik and evolved from the theory of Structural Risk Minimization [13]. SVM’s learn the boundary regions between samples belonging to two classes by mapping the input samples into a high dimensional space and seeking a separating hyperplane in this space. The separating hyperplane is chosen in such a way as to maximize its distance from the closest training samples (support vectors). This distance quantity is referred to as the margin.

An SVM classifier has the general form:

\[ f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \]  

(1)

where \( \mathbf{x}_i \in \mathbb{R}^n \), \( i = 1, 2, \ldots, l \) are the training data. Each point of \( \mathbf{x}_i \) belongs to one of the two classes identified by the label \( y_i \in \{-1, 1\} \). The coefficients \( \alpha_i \) and \( b \) are the solutions of a quadratic programming problem [13], \( \alpha_i \) are non-zero for support vectors (SV) and are zero otherwise. \( K \) is the kernel function.

Classification of a test data point \( \mathbf{x} \) is performed by computing the right-hand side of Eq. (1).

Much of the flexibility and classification power of SVM’s resides in the choice of kernel. Some examples are linear, polynomial degree \( p \), and Gaussian. These kernel functions have two main disadvantages for speech signals. First they only model individual data points as opposed to an ensemble of vectors which speech classification decisions must be based on. Secondly these kernels are quite generic and do not take advantage of the statistics of the individual speech signals we are targeting.
2.1. Probabilistic Distance Kernels

We start with a statistical model $p(x|\theta_i)$ of the data, i.e., for each utterance $X_i = \{x_1, x_2, \ldots, x_l\}$ we estimate the parameters $\theta_i$ of a generic probability density function (PDF). We pick PDF's that have been shown over the years to be quite effective at modeling speech patterns. In particular we use diagonal Gaussian mixture models. The parameters $\theta_i$ are priors, mean vectors, and diagonal covariance matrices.

$$p(x|\theta_i) = \sum_{k=1}^{K} \alpha_k N(x; \mu_k, \Sigma_k)$$

Once the PDF $p(x|\theta_i)$ has been estimated for each training and testing utterance we replace the kernel computation in the original utterance space by a kernel computation in the PDF space:

$$K(X_i, X_j) \Rightarrow K(p(x|\theta_i), p(x|\theta_j))$$

To compute the $\theta_i$ parameters for a given utterance $X_i$ we use a maximum likelihood approach. For diagonal mixture models there is no analytical solution for $\theta_i$ and we use the Expectation Maximization algorithm. Effectively we are proposing to map the input space $X_i$ to a new feature space $\theta_i$.

The next step is to define the kernel distance in this new feature space. Because of the statistical nature of the feature space a natural choice for a distance metric is one that compares PDF's. From the standard statistical literature there are several possible choices; however, in this paper we only report our results on the symmetric Kullback-Leibler (KL) divergence also known as the Jeffrey's divergence.

$$D(p(x|\theta_i), p(x|\theta_j)) = \int_{-\infty}^{\infty} p(x|\theta_i) \log \frac{p(x|\theta_i)}{p(x|\theta_j)} dx + \int_{-\infty}^{\infty} p(x|\theta_j) \log \frac{p(x|\theta_j)}{p(x|\theta_i)} dx$$

Because a matrix of kernel distances directly based on symmetric KL divergence does not satisfy the Mercer conditions, i.e., it is not a positive semidefinite matrix, we need a further step to generate a valid kernel. Among many possibilities we simply exponentiate the symmetric KL divergence and scale it (A multiplicative factor below)

$$K(X_i, X_j) \Rightarrow e^{-A D(p(x|\theta_i), p(x|\theta_j))}$$

In the case of Gaussian mixture models the computation of the KL divergence is non-trivial. In fact there is no analytical solution to equation 4 and we have to resort to Monte Carlo methods or numerical approximations.

3. Kernel Adaptation

As we can see for every utterance we learn a Gaussian mixture model. In the case of long utterances we might have enough feature vectors to effectively learn a GMM. However, in many other situations we might not have enough data to accurately learn model parameters. In these cases adaptation of the GMM parameters (means, variances) is a good alternative.

Of all possible adaptation approaches, in this paper we focus on the Maximum Likelihood Linear Regression scheme proposed by Leggetter [6], MLLR estimates a set of linear transformations for the mean and variance parameters of each GMM. It assumes that all mean vectors have been subjected to a global linear transformation. This transformation is encoded by a shift vector $b$ and a rotation matrix $A$. The covariance matrix of each Gaussian mixture are also assumed to be linearly transformed and this transformation is encoded in a rotation matrix $H$. The new mean and covariance matrices for each individual Gaussian are hence defined as

$$\mu_i = A \mu_i + b$$

$$\Sigma_i = B_i H B_i$$

where $B_i$ is the inverse of the Choleski factor of $\Sigma_i$, therefore $\Sigma_i^{-1} = C_i C_i^T$ and $B_i = C_i^{-1}$. Notice, however, that if the original covariance matrices $\Sigma_i$ are diagonal the above equations are further simplified. Also, the resulting adapted covariances are non-diagonal, however, the non-diagonal elements can be set to zero with no loss in performance.

Replacing $\mu_i$ and $\Sigma_i$ by their adapted counterparts in the maximum likelihood formulation, and using the EM algorithm we can estimate the new transform parameters $A, H, b$ that maximized the likelihood on the adapted data

$$p(x|\theta_i) = \sum_{k=1}^{K} \alpha_k N(x; \mu_k + b, B_i H B_i)$$

We can define the log-likelihood of the utterance data being generated by the PDF with parameters $\theta'$ as

$$\mathcal{L}(X_i = \{x_1, x_2, \ldots, x_l\} | \theta'_i) = \sum_{i=1}^{l} \log(p(x_i|\theta'_i))$$

and using the EM algorithm we can find the adaptation parameters $A, b, H$ that maximize the likelihood. EM determines the estimated parameters of a model such that the newly estimated parameters are guaranteed to increase the value of the likelihood function. This is done by maximizing an auxiliary function that is more computationally tractable than the original one

$$Q(\theta, \theta') = \sum_{i=1}^{l} p(x_i|\theta) \log(p(x_i|\theta))$$

For further details of the whole derivation we refer the interested reader to [6].

The means and variances are adapted in two separate stages. First for each utterance the new means are found and then, given these new means, the variances are updated. This results in one set of MLLR adaptation for each training and testing utterance. Notice that for our purposes we can start from a single generic GMM and adapt it for each new utterance, or alternatively start from several different GMMs. In the latter case we can pick this starting GMM based on gender identity of perhaps with some guess of the speaker identity or even using speaker clusters.

It is also possible to apply different transformations to each individual gaussian component. However, in our case we focus on learning a single global transformation per utterance. Also the adaptation scheme can be iterated several times, i.e., given new adapted Gaussian parameters we can learn new transformation parameters and iterate.

4. Experiments

We chose the HUB4-96 [12] News Broadcasting corpus and the Narrowband version of the KINT corpus [1] to train and test
our algorithms such that we could compare the performance on broadcasting-quality (16kHz) speech and telephone-quality (8kHz)
speech. HUB4 is not a common corpus for speaker identification and verification. However, it contains a large number of broadcast-
quality utterances from speakers and it was readily available to us.

The HUB4 corpus has over 2000 speakers. However, we only used the 50 speakers that appeared most frequently in this corpus. The training set contains about 25 utterances (each 3-7 seconds long) from each of the 50 speakers resulting in 1198 utterances. The test set contains the rest of the utterances from these 50 speakers resulting in 15325 utterances.

The KING corpus is a standard speaker identification and verification corpus and was specifically designed for this purpose. We use the narrowband version of the corpus. In order to match with the HUB4 experiments, we also picked 50 speakers in KING for training and testing. The training set contains 4 utterances from each speaker, randomly chosen from S1-S10, and the test set contains 6 utterances (excluding the training set) from each speaker. Utterances had an average duration of 40 seconds. This produced a total of 200 training utterances and 300 testing utterances. We use standard Mel-Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives to compose a 39 dimensional feature vector in all our experiments.

We explore one type of MLLR adaptation scheme. A single generic Gaussian Mixture at high resolution (large number of Gaussians) is trained for all training data, regardless of class identity. Then for each training and testing utterance we adapt its mean vectors and covariance matrices. We explore the effect of the number of rounds of MLLR adaptation. For all experiments we use the Kullback-Leibler kernel.

Our experiments trained and tested using four different types of classifiers: Baseline GMM, GMM with MLLR adaptation, SVM using GMM/KL divergence based kernels, and SVM using GMM MLLR adapted/KL divergence kernels. We compared the performance of all these classifiers in both speaker identification and verification.

In order to identify the 50 speakers from HUB4, 50 SVMs were trained by the 1-vs-rest approach, i.e., one speaker vs. the rest of the 49 speakers. We used a modified version of SVMF1 [10] to train and test our new kernels. We tested these SVM's and each returned a score for each of the 15325 test utterances. The KING speaker SVMs were trained in the same way. For speaker verification using GMM's the speaker score had to be compared with a background score. This score is computed as the arithmetic mean of the 49 speaker scores that did not belong to the actual labeled speaker. This background score is subtracted from the actual speaker score and compared to a threshold $\Theta$.

$$\text{Score}_t = \frac{1}{49} \sum_{t \neq i} \text{Score}_t > \Theta \quad (10)$$

The Detection Error Tradeoff (DET) curves as shown in Figs. 1 and 2 are computed by varying $\Theta$. They were computed by using all the 50 speakers in the HUB4 and KING corpora respectively. Each utterance was tested against all the 50 classifiers.

Tables 1 and 2 show the equal error rates (EER's) of speaker verification and the accuracies of speaker identification for both corpora when trained and tested with all 50 speakers in the HUB4 and KING corpora.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Type of Classifier & HUB4 Acc & HUB4 EER \\
\hline
GMM NG=256 & 85.8 & 7.7 \\
GMM NG=1024 MLLR=1 & 85.2 & 8.6 \\
SVM NG=16KL & 83.8 & 6.9 \\
SVM NG=256 MLLR/KL & 86.0 & 6.3 \\
\hline
\end{tabular}
\caption{Comparison of all the classifiers used on the HUB4 corpus. Both classification accuracy (Acc) and equal error rates (EER) are reported in percentage points.}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure.png}
\caption{Speaker verification detection error tradeoff (DET) curves for the HUB4 corpus, tested on all 50 speakers.}
\end{figure}

Our approach using the adapted GMM/KL SVM kernels is quite promising. In the case of the HUB corpus as shown on table 1 and Fig. 1 our new classifier improves over all other classifiers. The DET shows better performance at all operating points. For the KING corpus as shown on table 2 and Fig. 2 our adapted GMM SVM methods outperform the generative classifiers significantly. We believe this is because in the KING corpus the utterances are longer providing more data to perform effective MLLR adaptation. However, as in the HUB corpus even when the utterances are short and not much data is provided for adaptation, our approach still improves performance.

In our experiments we also explored the effect of the number of MLLR adaptation iterations on classifier performance. We observed that a single round of adaptation works extremely well. We also explored adapting the means only and in general obtain slightly worse results.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Type of Classifier & KING Acc & KING EER \\
\hline
GMM NG=256 & 58.7 & 22.4 \\
GMM NG=512 MLLR & 63.3 & 22.6 \\
SVM NG=12KL & 78.0 & 11.3 \\
SVM NG=256 MLLR/KL & 84.7 & 8.4 \\
\hline
\end{tabular}
\caption{Comparison of all the classifiers used on the KING corpus. Both classification accuracy (Acc) and equal error rates (EER) are reported in percentage points.}
\end{table}

1Our experiments with other generative based kernels (Blattner, Expected Likelihood, etc) have in general resulted in lower classification results.

2Experiments were done where we varied the number of Gaussians. We only report results on the best GMM configuration.
5. Conclusions and Future Work

In this paper we have introduced the concept of kernel adaptation in SVM/KL based classifiers. This idea brings the best of both generative and discriminative classifiers into the realm of speaker classification.

Our approach is quite simple. We first learn a single generic high resolution GMM (with 256 or more components). Then for each training and testing utterance we learn a new GMM, rather than learning the GMM parameters \( \theta \) (priors, mean vectors, covariance matrices) directly from the utterance feature vectors we propose to adapt the parameters of the high resolution generic GMM using techniques such as MLLR. Once the collection of adapted GMMs for each training/testing utterance is computed we define a kernel distance among GMMs based on the symmetric Kullback-Leibler (KL) divergence. The main role of the MLLR adaptation is to get a high resolution GMM fitted to the data with a limited amount of data available; just one utterance.

In our experiments we have validated this new approach to speaker identification and verification comparing its performance with traditional generative classifiers such as GMM's with and without MLLR adaptation. We have also compare the performance of adapted kernels to non-adapted kernels and shown significant improvements in performance. Improvements are more noticeable in the KING database where utterances tend to be longer in the HUB4 datasets. In general MLLR adapted kernels offer the best classification and verification performance when compared to all other classification approaches.

When representing each utterance by a GMM we can use MLLR adaptation techniques, which have been successfully validated over the years in speech and speaker recognition technology. This adaptation step allows us to fit a high resolution Gaussian mixture to each utterance despite limited amounts of data. In turn this high resolution GMM allows us to compute a more accurate kernel distance among GMMs improving the performance of our SVM classifier.

In the future we plan to study the application of these adaptation techniques to image classification and retrieval tasks. We also plan to explore the use of multiple MLLR clusters to start the adaptation from multiple generic GMM's rather than a single one.

6. References