A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels

Pedro J. Moreno, Purdy P. Ho
Cambridge Research Laboratory
HP Laboratories Cambridge
HPL-2004-7
January 9th , 2004*

support vector machine, SVM, speaker identification, speaker verification, KL divergence, Kullback-Leibler divergence, probabilistic distance kernels, multimedia

One major SVM weakness has been the use of generic kernel functions to compute distances among data points. Polynomial, linear, and Gaussian are typical examples. They do not take full advantage of the inherent probability distributions of the data. Focusing on audio speaker identification and verification, we propose to explore the use of novel kernel functions that take full advantage of good probabilistic and descriptive models of audio data. We explore the use of *generative* speaker identification models such as Gaussian Mixture Models and derive a kernel distance based on the Kullback-Leibler (KL) divergence between generative models. In effect our approach combines the best of both generative and discriminative methods. Our results show that these new kernels perform as well as baseline GMM classifiers and outperform generic kernel based SVM's in both speaker identification and verification on two different audio databases.

* Internal Accession Date Only
Approved for External Publication
Published in Eurospeech 2003, 1-4 September 2003, Geneva, Switzerland

1 Introduction

During the last years Support Vector Machines (SVM's) [1] have become extremely successful discriminative approaches to pattern classification and regression problems. Excellent results have been reported in applying SVM's in multiple classification and regression benchmarks. In the general area of speech and speaker recognition SVM's have also been studied over the last years. For example, among others [2] compares the use of traditional based kernel SVM's with Gaussian classifiers, [3] examines the use of SVM's for phonetic classification, and [4] 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. Most research activities however have been focused on the underlying learning algorithms but not on the kernels themselves. Standard kernels such as linear, Gaussian, or polynomial don't take full advantage of the nuances of speech signals. An example of previous attempts in speaker identification and verification using these kernels is described in [5].

On the other hand statistical models such as Gaussian Mixture Models (GMM) or Hidden Markov Models make strong assumptions about the data, are simple to learn and estimate, and are well understood by the research community. It is therefore attractive to explore methods that combine generative models and discriminative models. We propose an approach that combines both discriminative and generative methods to classification. Instead of using these traditional kernels, we customized them for better speaker characteristics representation. We take advantage of diagonal covariance GMM's and full covariance Gaussian models to better represent speech utterances. We use a distance derived from the symmetric Kullback-Leibler (KL) divergence to effectively compare models.

The outline of this paper is as follows. In section 2 we give a brief introduction to SVM classifiers and the Fisher kernel. In section 3 we describe in detail the new kernels we introduce for audio data. 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 Kernels for SVM's

Support Vector Machines were first introduced by Vapnik and evolved from the theory of Structural Risk Minimization [1]. 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:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{1}$$

where $\mathbf{x}_i \in R^n$, $i = 1, 2, \ldots, l$ 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 α_i and *b* are the solutions of a quadratic programming problem [1]. α_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.

The Fisher kernel approach [6] is a first attempt at solving these two issues. It assumes the existence of generative model that explains well all possible data. For example, in the case of speech signals the generative model $p(\mathbf{x}|\boldsymbol{\theta})$ is often a Gaussian mixture. Where the $\boldsymbol{\theta}$ model parameters are priors, means, and diagonal covariance matrices.

For any given sequence of vectors defining an utterance $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ and assuming that each vector in the sequence is independent and identically distributed, we can easily define the likelihood of the ensemble being generated by $p(\mathbf{x}|\boldsymbol{\theta})$ as $P(X|\boldsymbol{\theta}) = \prod_{i=1}^{l} p(\mathbf{x}_i|\boldsymbol{\theta})$. The Fisher kernel maps utterances $\{X_1, \ldots, X_n\}$, each composed of a different number of feature vectors, into a linear space of fixed dimension.

We define a new feature vector, the Fisher score, as

$$U_X = \nabla_{\boldsymbol{\theta}} log(P(X|\boldsymbol{\theta})) \tag{2}$$

Each component of U_X is a derivative of the log-likelihood of the audio sequence X with respect to a particular parameter of the generative model. In our case the parameters $\boldsymbol{\theta}$ of the generative model are chosen from either the prior probabilities, the mean vector or the diagonal covariance matrix of each individual Gaussian in the mixture model. For example, if we use the mean vectors as our model parameters $\boldsymbol{\theta}$, *i.e.*, for $\boldsymbol{\theta} = \boldsymbol{\mu}_k$ out of K possible mixtures, then the Fisher score is

$$\nabla_{\boldsymbol{\mu}_k} log(P(X|\boldsymbol{\mu}_k)) = \sum_{t=1}^{l} P(k|\mathbf{x}_t) \Sigma_k^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_k) \tag{3}$$

where $P(k|\mathbf{x}_t)$ represents the *a posteriori* probability of mixture k given the observed feature vector $\mathbf{x}_t$. Effectively we transform each utterance X of variable length into a single vector U_X. For more details the reader is refered to [6]. The Fisher kernel approach has been successfully applied to speech signals before, see [7, 3].

3 Probabilistic Distance Kernels

Our new algorithm starts with a statistical model $p(\mathbf{x}|\boldsymbol{\theta}_i)$ of the data, *i.e.*, for each utterance $X_i = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\}$ we estimate the parameters $\boldsymbol{\theta}_i$ 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 and single full covariance Gaussian model. In the first case the parameters $\boldsymbol{\theta}_i$ are priors, mean vectors, and diagonal covariance matrices while in the second case the parameters $\boldsymbol{\theta}_i$ are the mean vector and full covariance matrix.
Once the PDF $p(\mathbf{x}|\boldsymbol{\theta}_i)$ 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:

$$K(X_i, X_j) \Longrightarrow K(p(\mathbf{x}|\boldsymbol{\theta}_i), p(\mathbf{x}|\boldsymbol{\theta}_j)) \tag{4}$$

To compute the $\boldsymbol{\theta}_i$ parameters for a given utterance X_i we use a maximum likelihood approach. In the case of diagonal mixture models there is no analytical solution for $\boldsymbol{\theta}_i$ and we use the Expectation Maximization algorithm. In the case of single full covariance Gaussian model there is a simple analytical solution for the mean vector and covariance matrix. Effectively we are proposing to map the input space X_i to a high dimensional feature space $\boldsymbol{\theta}_i$. Notice that if the number of vector in the X_i utterance is small and there is not enough data to accurately estimate $\boldsymbol{\theta}_i$ we can use regularization methods, or even replace the maximum likelihood solution for $\boldsymbol{\theta}_i$ by a maximum a posteriori solution. Other solutions like adapting the $\boldsymbol{\theta}_i$ parameters are possible and will be described in publications in the future.
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

$$D(p(\mathbf{x}|\boldsymbol{\theta}_i), p(\mathbf{x}|\boldsymbol{\theta}_j)) = \int_{-\infty}^{\infty} p(\mathbf{x}|\boldsymbol{\theta}_i)\, log(\frac{p(\mathbf{x}|\boldsymbol{\theta}_i)}{p(\mathbf{x}|\boldsymbol{\theta}_j)})\, d\mathbf{x} + \int_{-\infty}^{\infty} p(\mathbf{x}|\boldsymbol{\theta}_j)\, log(\frac{p(\mathbf{x}|\boldsymbol{\theta}_j)}{p(\mathbf{x}|\boldsymbol{\theta}_i)})\, d\mathbf{x} \tag{5}$$

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 definite matrix, we need a further step to generate a valid kernel. Among many posibilities we simply exponentiate the symmetric KL divergence, scale, and shift (A and B factors below) it for numerical stability reasons

$$\begin{aligned} K(X_i, X_j) &\Longrightarrow K(p(\mathbf{x}|\boldsymbol{\theta}_i), p(\mathbf{x}|\boldsymbol{\theta}_j)) \\ &\Longrightarrow e^{-A\, D(p(\mathbf{x}|\boldsymbol{\theta}_i), p(\mathbf{x}|\boldsymbol{\theta}_j)) + B} \end{aligned} \tag{6}$$

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 5 and we have to resort to Monte Carlo methods

or numerical approximations. In the case of single full covariance models the KL divergence has an analytical solution

$$D(p(\mathbf{x}|\boldsymbol{\theta}_i), p(\mathbf{x}|\boldsymbol{\theta}_j)) = tr(\Sigma_i \, \Sigma_j^{-1}) + tr(\Sigma_j \, \Sigma_i^{-1}) - 2\,S + tr((\Sigma_i^{-1} + \Sigma_j^{-1})\,(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T) \quad (7)$$

where S is the dimensionality of the original feature data $\mathbf{x}$. Notice that this similar to the Arithmetic harmonic sphericity (AHS) distance quite popular in the speaker identification and verification research community [8].

Our approach, while independently derived, shows remarkable similarity to the Information Diffusion kernel proposed in [9]. There are however some differences. Among others, our approach is conceptually much simpler and is applied to continuous data sets as opposed to discrete data sets such as text corpora.

4 Experiments and Results

We chose the HUB4-96 [10] News Broadcasting corpus and the Narrowband version of the KING corpus [11] 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.

The HUB4 corpus has over 2000 speakers. However, we only used the 50 speakers who 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 commonly used for speaker identification and verification in the speech community. 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 (excluded from the training set) from each speaker. This produced a total of 200 training utterances and 300 testing utterances. We use standard Mel-Frequency Cepstral Coefficients (MFCC's) and their first and second derivatives to compose a 39 dimensional feature vector in all our experiments.

Two types of probabilistic distance kernels were explored: the GMM/KL divergence and the full-covariance/AHS distance. In the first kernel a sequence of feature vectors from each utterance was modeled by a single GMM of diagonal covariances with 16 mixtures. Then the KL divergences between each of these GMM's were computed, this formed a 1198x1198 training matrix and a 15325x1198 test matrix for the HUB4 corpus; and a 200x200 training matrix and a 300x200 test matrix for the KING corpus. For the full-covariance/AHS distance based kernel a full covariance Gaussian was computed for each speaker, then the AHS distances between each of these full covariances were computed.

Our experiments trained and tested using five different types of classifiers: Baseline GMM, Baseline AHS, SVM using Fisher kernel, SVM using GMM/KL Divergence based kernels, and SVM using Full-Covariance/AHS distance based kernels. We compared the performance of all these classifiers. [12] and [8] describe in detail the first and second classification approaches. For the Fisher kernel experiments we used as $\boldsymbol{\theta}$ parameters the prior probabilities of each mixture Gaussian as described in section 2.
In order to identify the 50 speakers from HUB4, 50 SVM's 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 SVMFu [13] 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 or AHS probabilistic classifiers 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 Θ

$$Score_i - \frac{1}{49} \sum_{t=1,t\neq i}^{50} Score_t > \Theta \tag{8}$$

The Detection Error Tradeoff (DET) curve as shown in Fig. (1) is computed by varying Θ. DET's can be computed in two different ways based on the pool of speakers. The DET shown in Fig. (1) was computed by using all the 50 speakers in the HUB4 corpus. Each utterance was tested against all the 50 classifiers. However, the DET shown in Fig. (2) was computed by using only three cohort speakers and the target speaker. Cohort speakers are a subset of speakers who are highly confusable with the target one. The use of cohorts represents a worst case scenario for speaker verification. We only show DET curves on the HUB4 corpus. Results on the KING corpus are quite similar and are shown in Table 2.

Table 1: *Comparison of all the classifiers used on the HUB4 corpus. Both classification accuracy (Acc) and equal error rates (EER) are reported in percentage points.*

Type of Classifier	HUB4 Acc	HUB4 EER	HUB4 Cohort EER
GMM NG=256	87.4	8.1	13.8
AHS	81.7	9.1	16.8
SVM Fisher	62.4	14	20.8
SVM GMM/KL	83.8	7.8	10.8
SVM AHS	84.7	7.4	10.0

We tested our probabilistic kernels and compared their results with each other, as well as with the results of the baseline GMM classifier[1], baseline AHS classifier, and the Fisher kernel SVM in both speaker verification and identification. The following tables show the equal-error rates

[1]Experiments were done where we varied the number of Gaussians. We only report results on the best GMM configuration.

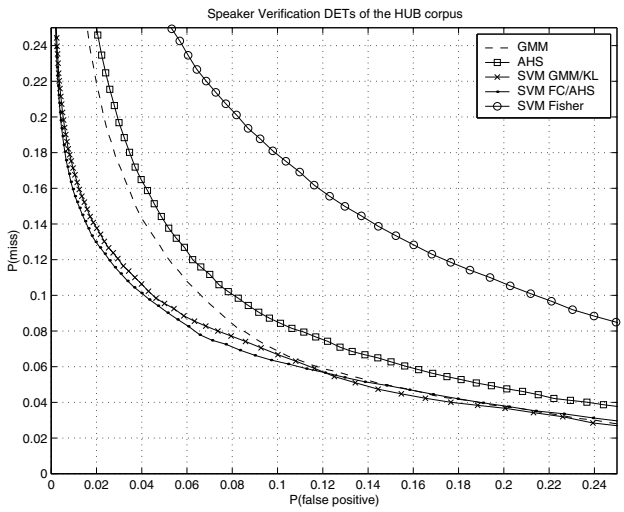

Figure 1: *Speaker verification detection error tradeoff (DET) curves for the HUB4 corpus, tested on all 50 speakers.*

(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.

Our approach using the probabilistic SVM kernels shows quite promising results. As we can see in the case of the HUB4 corpus all classifiers perform similarly in the speaker identification task with the exception of the SVM Fisher. This is mostly likely because of the availability of sufficient training data in HUB4 for the generative classifiers (GMM's and AHS). Similar performance is observed when we look at the speaker verification task and the DET plot in Fig. (1).

The results of the KING corpus are shown in Table 2. As we can see in both speaker identification and verification tasks, our probabilistic SVM methods outperform the generative classifiers significantly. This is because the amount of data is more limited and SVM methods can take better advantage of fewer data points.

Table 2: *Comparison of all the classifiers use on the KING corpus. Both classification accuracy (Acc) and equal error rates (EER) are reported in percentage points.*

Type of Classifier	KING Acc	KING EER	KING Cohort EER
GMM NG=256	70.7	16.1	25.2
AHS	48.3	26.8	28.0
SVM GMM/KL	72.7	7.9	11.1
SVM AHS	79.7	6.6	9.1

Looking at the HUB4 speaker verification DET in Fig. (2), we can see the different performance when cohorts are used. Naturally the EER's are worse when we look at a worse case scenario

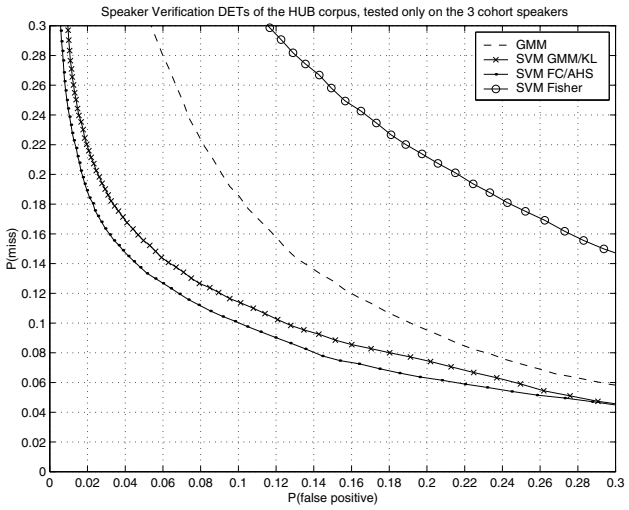

Figure 2: *Speaker verification DET curve for the HUB4 database tested only on 3 cohort speakers and the target speaker.*

such as cohorts. However, the relative performance of all classifiers remains the same with the new proposed probabilistic SVM systems outperforming all other classifiers.

5 Conclusion and Future Work

In this paper we have proposed two new methods of combining generative classifiers that maximize the likelihood of observed data under some model assumptions and discriminative classifiers (SVM's) that effectively minimize training error rates.

Our approach is extremely simple. For every utterance a PDF is learned using maximum likelihood approaches. In the case of GMM's we use the EM algorithm to learn the model parameters $\boldsymbol{\theta}$. In the case of a single full covariance Gaussian we directly estimate the full covariance. Then we introduce the idea of computing kernel distances via a direct comparison of PDF's. In effect we replace the traditional kernel distance on the original data $K(X_i, X_j)$ by a new kernel derived from the symmetric Kullback-Leibler (KL) divergence $K(X_i, X_j) \longrightarrow K(p(\mathbf{x}|\boldsymbol{\theta}_i), p(\mathbf{x}|\boldsymbol{\theta}_j))$. After that a kernel matrix is computed and a traditional SVM can be used.

In our experiments we have validated this new approach to speaker identification and verification comparing its performance with Fisher kernel SVM's and with other well-known speaker recognition algorithms: GMM and AHS methods. Our results show that these two new kernels always outperform the SVM Fisher kernel and the AHS methods, and they do equally well as the baseline GMM in the case of speaker identification when training with a large corpus (HUB4). These new kernels outperform both the baseline classifiers and the Fisher kernel SVM when training with a small corpus (KING). They also outperform all other classifiers in the case of speaker verification. All these encouraging results show that SVM's can be improved by paying careful attention to the nature of the data being model. In the case of speech signals we just take

advantage of previous years of research in generative methods.
The most remarkable result is the good results obtained using a full covariance single Gaussian probabilistic based SVM classifier. Its simplicity and similarity with the well known AHS method makes it a very attractive alternative to more complex methods of combining generative classifiers and discriminative methods such as Fisher SVM. Its performance is consistently good across both databases, is specially fast to compute, and it requires no tuning of system parameters.
We feel that this approach of combining generative classifiers via KL distances of derived PDF's is quite generic and can possibly be applied to other domains. We plan to explore its use in image classification tasks and other multimedia related tasks.

References

[1] Vapnik, V., *Statistical learning theory*, John Wiley and Sons, New York, 1998.

[2] Clarkson, P. and Moreno, P. J., "On the use of support vector machines for phonetic classification," *ICASSP*, 1999.

[3] Smith, N. and Niranjan, M., "Data-dependent kernels in svm classification of speech patterns," in *International Conference on Spoken Language Processing*, 2000.

[4] Ho, P. P., "A Handset Identifier Using Support Vector Machines," in *IEEE International Conference on Spoken Language Processing*, Denver, CO, USA, 2002.

[5] Wan, V. and Campbell, W., "Support vector machines for speaker verification and identification," *IEEE Proceeding*, 2000.

[6] Jaakkola, T., Diekhans, M. and Haussler, D., "Using the fisher kernel method to detect remote protein homologies," in *Proceedings of the Internation Conference on Intelligent Systems for Molecular Biology*, Aug. 1999.

[7] Moreno, P. J. and Rifkin, R., "Using the fisher kernel method for web audio classification," *ICASSP*, 2000.

[8] Bimbot, F., Magrin-Chagnolleau, I. and Mathan, L., "Second-order statistical measures for text-independent speaker identification," *Speech Communication*, vol. 17, pp. 177–192, 1995.

[9] J. Lafferty and G. Lebanon, "Information diffusion kernels," in *Neural Information Processing Systems 15*, 2002.

[10] Stern, R. M., "Specification of the 1996 HUB4 Broadcast News Evaluation," in *DARPA Speech Recognition Workshop*, 1997.

[11] "Brief Description of the KING Speech Database," http://www.ldc.upenn.edu/Catalog/docs/LDC95S22/ kingdb.txt.

[12] Reynolds, D. A., "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, pp. 91–108, August 1995.

[13] Rifkin, R., "SVMFu Documentation," http://five-percent-nation.mit.edu/SvmFu.