COMPARISON OF SEQUENCE DISCRIMINANT SUPPORT VECTOR MACHINES FOR ACOUSTIC EVENT CLASSIFICATION

Andrey Temko, Enric Monte, and Climent Nadeu

TALP Research Center, Universitat Politècnica de Catalunya
Barcelona, Spain

{temko,enric,climent}@talp.upc.es

ABSTRACT

In a previously reported work, classification techniques based on Support Vector Machines (SVM) showed a good performance in the task of acoustic event classification. SVM are discriminant classifiers, but they cannot easily deal with the dynamic time structure of sounds, since they are constrained to work with fixed-length vectors. Several methods that adapt SVM to sequence processing have been reported in the literature. In this paper, they are reviewed and applied to the classification of 16 types of sounds from the meeting room environment. With our database, we have observed that the dynamic time warping kernels work well for sounds that show a temporal structure, but the best average score is obtained with the Fisher kernel.

1. INTRODUCTION

Recent works on statistical machine learning have shown the advantages of discriminative classifiers like SVM [1] in a range of applications, including audio classification [2]. In [3], we applied SVMs to the task of classifying a set of 16 types of acoustic events that may take place in a meeting room environment. In that work, the SVM-based techniques showed a higher classification capability than the Gaussian Mixture Models (GMM) based techniques, and the best results were consistently obtained with a confusion-based variable-feature-set clustering scheme, arriving with SVM to a 88,29 % classification rate. HMMs could not be considered in that task since the amount of available data was not large enough to accurately train the models.

A drawback of SVMs when dealing with audio data is their restriction to work with fixed-length vectors. Both in the kernel evaluation and in the simple input space dot product, the units under processing are vectors of constant size. However, when working with audio signals, although each signal frame is converted into a feature vector of a given size, the whole acoustic event is represented by a sequence of feature vectors, which shows variable length. In order to apply a SVM to this kind of data, one needs either to somehow normalize the size of the sequence of input space feature vectors or to find a suitable kernel function that can deal with sequential data.

Several methods have been explored to adapt SVMs to sequence processing [4]. The most common approach is to extract some statistical parameters from the sequence of vectors and thus transform the problem into that of fixed-length vector spaces. For example, the mean and the standard deviation of the features extracted from every frame of an audio clip were taken as feature vector for audio analysis in [2]. Despite the good results we obtained using this approach for acoustic event classification (AEC) [3], when frame-level features are transformed into statistical event-level features there exists an unavoidable loss of information.

In the work reported in this paper, we aim at using SVMs for AEC but preserving the information contained in the sequentiality of data, i.e. the temporal structure of the acoustic events. For that purpose, after choosing a set of meaningful reported techniques, we have compared their performance in the framework of our meeting-room AEC task. The fact that the used set of acoustic event types includes time structured sounds (e.g. music) but also sounds whose time evolution is not relevant (e.g. liquid pouring), allows us to investigate the appropriateness of the various techniques to classify the different types of sounds.

While in our previous work we tested several feature sets and several multi-class schemes for SVM, here we use only the best feature set from [3] and a Directed Acyclic Graph (DAG) [5] classification scheme. Moreover, the influence of the generative model parameters' estimation error on the Fisher score derivative is investigated.

The paper is organized as follows: Section 2 quickly reviews the SVM-based methods used in the work, Section 3 presents experimental results and discussions, and Section 4 concludes the work.

2. SVM-BASED SEQUENCE DISCRIMINANT TECHNIQUES

We have chosen three different SVM kernels techniques that make use of dynamic time warping (DTW), namely: dynamic time-alignment kernel (DTAK)[6], Gaussian dynamic time warping (GDTW) kernel [7], and the recent polynomial dynamic time warping kernel (PolyDTW) [8]. Additionally, we included in the comparison the Fisher score kernel [9] and the Fisher-ratio kernel [10][11], which aim at using generative model classifiers like GMM in the discriminative framework, and have been applied for speech/speaker recognition using SVM [10][11]. On the other hand, among the algorithms reported in the literature that normalize the size of the vector sequences [12], we have chosen the simple outerproduct of trajectory matrix method, which was the winner in [12]. As references for comparison, we also use a standard GMM classifier, and an SVM classifier with statistical event-level features.

2.1 Fisher kernel

Fisher kernel is one of the most successful approaches that enable SVM to classify whole sequences. Inspired by using statistical modeling method, Fisher kernel recently has become very popular in the areas that involve time-series recognition. The generalized idea of Fisher kernel the score-space kernel was applied to speech recognition in [11]. Modification of likelihood score space kernel (i.e. Fisher kernel) known as likelihood ratio score-space kernel has shown comparative results in the sphere of speaker verification [10].

The idea of Fisher kernel includes in mapping the variable length sequence to a single point in fixed-dimension space, the so-called *score-space*. To perform such a mapping, Fisher kernel applies the first derivative operator to the likelihood score of the generative model. Given an input sequence X, and a model M, parameterized by θ, the Fisher score is defined as

$$\psi_{fisher}(X) = \nabla_\theta \log P(X|M,\theta) \qquad (2.1.1)$$

The Fisher score can be interpreted in the following way. When a generative model is trained by ML (maximum likelihood) criterion, it uses the same set of derivatives to compute how close it is to the local extreme. Another motivation of using Fisher score is that the gradient of the log-likelihood can capture the generative process of the whole sequence better than just a posterior probability. Furthermore, in [9] it was shown that, under the condition that the class variable is a latent variable in the probability model, the learning machines, that use Fisher kernel, are asymptotically at least as good as making decision based on the generative model itself (maximum a posteriori). In [9] applied to bio-sequences recognition Fisher kernel performed significantly better than HMM.

2.2 Outerproduct of trajectory matrix

The time analysis of the data gives a sequence of *l*-dimensional parametric vectors. The sequence is considered as a trajectory in the *l*-dimensional space. If we define the *l-by-m* trajectory matrix as $X = [x_1, x_2, \ldots x_m]$, the outerproduct matrix Z [12]is defined as

$$Z = X^T X \qquad (2.2.1)$$

Thus the outerproduct matrix Z is *l*-by-*l* and no longer depends on the length of the sequence. The vectorized outerproduct thus can feed the SVM classifier directly. It is obvious that this method explicitly considers sequence duration information. Despite the simplicity of the given approach, it showed considerably better results than *Compaction and Elongation* method in the task of spoken letters recognition [12].

2.3 Gaussian dynamic time warping (GDTW)

This approach as well as a previous one does not assume a model for the generative class conditional densities. The GDTW [7] method addresses the problem of variable length sequences classification by introducing the DTW technique to SVM kernel. Recalling the standard RBF kernel for SVM

$$K(T,R) = \exp\left(-\gamma \|T-R\|^2\right) \qquad (2.3.1)$$

where *T, R* denote two patterns to compare. As mentioned in Section 1, if the two patterns are sequences of different length, they cannot be compared in the kernel evaluation directly. An obvious modification of (2.3.1) is to substitute the squared Euclidian distance computation with the equivalent that can cope with temporally distorted, variable length sequences. Thus, in [7] GDTW kernel was defined as

$$K(T,R) = \exp(-\gamma D(T,R)) \qquad (2.3.2)$$

where $D(T,R)$ is a DTW distance between sequences *T* and *R*.

The proposed method was successfully applied to handwriting recognition and showed comparative and at times superior results to HMM and MLP in [7].

2.4 Dynamic time alignment kernel (DTAK)

The approach proposed in [6] also deals with DTW. Instead of substituting the Euclidian distance in Gaussian kernel (2.3.1) by DTW distance, it substitutes the Euclidian distance in definition of DTW local distance by a kernel.

$$K(T,R) = D_\phi(T,R) = \frac{1}{N}\sum_{n=1}^{N} k\left(t_{\phi_{T(n)}}, r_{\phi_{R(n)}}\right) \qquad (2.4.1)$$

where *k(.)* is a kernel function that can be either a simple dot product or any conventional SVM kernel and Φ is the optimal DTW path. Actually, DTAK performs DTW in the feature space. Unlike the original DTW, which finds the optimal path that minimizes the accumulated distance/distortion, the DTAK algorithm maximizes the similarity. In the task of phoneme recognition, the proposed DTAK method outperformed HMM with a small or medium amount of training data and it got comparable results with a larger dataset [6].

2.5 Polynomial dynamic time warping (PolyDTW)

The method shares the same idea of performing DTW in transformed feature space. After spherical normalization [10] each vector *t* of a sequence is projected onto the sphere surface as

$$\hat{t} = \frac{1}{\sqrt{t^2+\alpha^2}}\begin{bmatrix} t \\ \alpha \end{bmatrix} \qquad (2.4.1)$$

Then the arcos of the dot product between normalized vectors can be taken as a local distance for DTW. Thus, the kernel is given as

$$K(T,R) = \cos^m\left(\frac{1}{N}\sum_{n=1}^{N}\arccos\left(\hat{t}_{\phi_{T(n)}} \cdot \hat{r}_{\phi_{R(n)}}\right)\right) \qquad (2.4.2)$$

This method has been successfully applied to the task with high intra-class variation such as dysarthric speech recognition and showed superior results to HMM [8].

3. EXPERIMENTS AND DISCUSSION

3.1 Experimental setup

Our previous efforts in [3] were focused on developing a variable-feature-set clustering scheme and using SVM with statistical event-level features. In this work, for simplicity, we use DAG [5] multi-class scheme, and only one feature set, the one that showed best results in [3], namely, a concatenation of perceptual features (ZCR, Spectral Flux, etc) and frequency filtering features [13] (plus their first and second derivatives). The number of features per frame is 50 and there is a frame each 10ms.

1-chair moving	2-clapping	3-cough
4-door	5-keyboard	6-laugh
7-music	8-paper crumple	9-paper tearing
10 pen writing	11-liquid pouring	12-puncher
13-sneeze	14-sniffing	15-speech
16-yawn		

Table 1. Database of the sixteen classes of acoustic events

In all the experiments we use the databases of acoustic events described in [3]. The database contains the 16 classes of meeting-room acoustic events that are summarized in Table 1.

For the outerproduct, DTAK, and GDTW methods we use a Gaussian kernel, and a 5-fold cross-validation on the training database was applied to find the optimal kernel parameter. The techniques that exploit DTW required some optimization steps to be feasible in practice (beam search strategy, kernel caching). For PolyDTW, a polynomial of third degree was chosen with α=1, as suggested in [8]. Also, we chose the linear SVM kernel for the Fisher score and the likelihood ratio methods, since it performed better than RBF.

The mean of individual class accuracies was chosen as a metric as in [3].

3.2 Comparison results

Figure 1 shows the results of the 8 considered techniques when applied to the database of acoustic events. The best average result is obtained with the Fisher kernel, 88.13%, and it is followed by the results from PolyDTW, likelihood ratio kernel and GMM. All mentioned results are better than 83.1%, the score of the non-sequential SVM technique that uses statistical event-level features (SVM stat). A similar result was observed in [3] using a binary tree instead of a DAG scheme: 82.9%.

It is also worth noticing that the result with the Fisher kernel (88.13%) is comparable to the best result in [3] using non-sequential SVM techniques: 88.29%. However, the latter result was obtained by using a variable-feature-set clustering, a classification scheme that is more developed than DAG, and by using the most discriminative feature set on each step of classification.

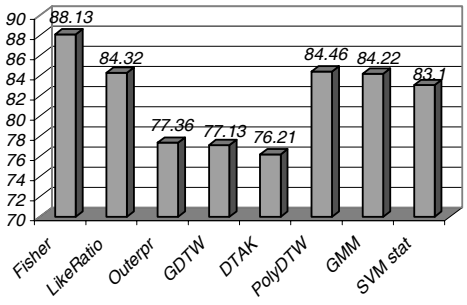

Figure 1. Classification accuracy for the 8 techniques

3.3 Influence of the number of Gaussians on the derivatives of the generative model

Interesting enough that the best results for GMM were obtained with 8 Gaussians, while for Fisher kernel the appropriate generative model that leaded to the best performance was 1-Gaussian GMM. Figure 2 shows the dependence of performance of Fisher kernel, Likelihood ratio kernel and GMM on the number of Gaussians.

As can be seen from Figure 2 there is an apparent inconsistency in the results, in the sense that the recognition rate improves in the case of the GMM classifier as the number of Gaussians increases, but at the same time, the results degrade in the case of the Fisher kernel. There is a two-fold explanation of this phenomenon. The first is related to the fact that the likelihood of the observation given the model is computed by means of a linear combination of Gaussians. The weight of each Gaussian is proportional to the number of samples that are assigned to it. Therefore, the parameters estimated with a small number of samples (i.e. that have a higher estimation error), have a lower influence in the likelihood. In the case of the Fisher score, the derivative of the likelihood with respect to each parameter inherits the estimation error, and it is not concealed, as it is the case of the GMM. Furthermore this effect is augmented by the fact that the dimensionality of the Fisher kernel increases proportionally to the number of Gaussians, and the number of noisy coordinates can be majority [14]. The second explanation uses the concept of sensitivity, which is the percentage change of a function for a given percentage change of one of the parameters:

$$S = \frac{\Delta f(x) / f(x)}{\Delta x / x} \approx \frac{x}{f(x)} \frac{df(x)}{dx} \qquad (3.2.1)$$

We computed the sensitivity of the likelihood of a GMM, and the Fisher kernel associated to the GMM. The resulting expressions are highly complicated. Nevertheless, simulations for one Gaussian confirmed that the sensitivities to the mean and the weight of each Gaussian are similar for both GMM and Fisher kernel, but the sensitivity to the variance is at least three times higher in the case of the Fisher kernel.

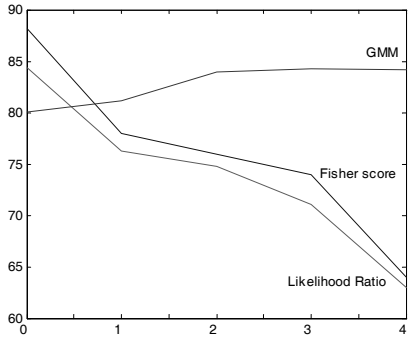

Figure 2. Dependence of the performance of the Fisher score kernel, likelihood ratio kernel and GMM on the number of Gaussians (log_2N_g)

3.4 Dependence of the classifier performance on the temporal structure of the acoustic event signals

The signals to be classified are quite heterogeneous, and have different temporal structures. Therefore, as was expected the performance of each classifier was biased to a given subset of the classes. For instance the DTW based classifiers behaved better with signals such as “music”, or “sneeze”, while classifiers that did not take into account the temporal structure of the signal did better with other signals that did not have that structure, such as

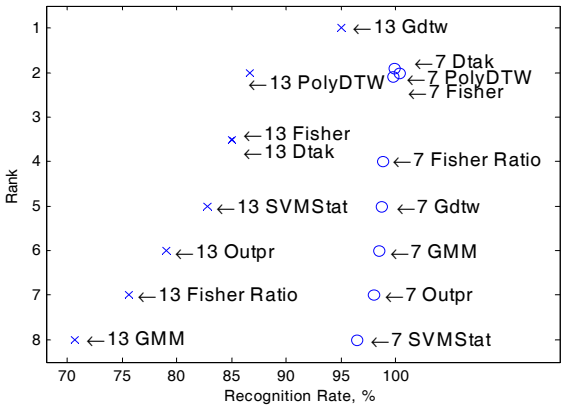

Figure 3. Comparison results for the classes "music" (7) and "sneeze" (13)

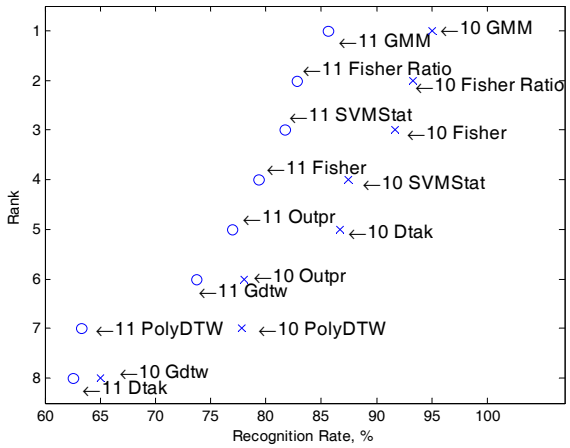

Figure 4. Comparison results for the classes "pen writing" (10) and " liquid pouring" (11)

"pen writing" or "liquid pouring". Ranking eight classifiers for a given class (giving the score 1 to the best one and 8 to the worst one) these properties can be summarized in Figure 3 and Figure 4, where we compare the 8 classifiers for above-mentioned pairs of classes.

In Figure 3 it can be seen that in the case of "music" and "sneeze" the best classifiers, i.e. highest ranking and recognition rate, are DTW-based such as GDTW, PolyDTW and DTAK. While the classifiers that do not take into account the temporal structure, give inferior results. In Figure 4 the ranking of classifiers is opposite, and the classifiers that specifically dismiss the temporal order fare better; the highest ranking corresponds to the GMM, and the Fisher Ratio. Another general feature that was detected, and that is reflected in these figures, is that there are signals that are easier to classify. It can be seen that systematically the results for a given class are better than for the others consistently for all the 8 classifiers, i.e. the distribution of the results for all classification systems are separated, although the order of the systems can be different for each signal.

As a general summary, we can assert that there was a correlation between the classes and the classifiers, which is masked in the mean values presented in Figure 1. For both types of signals, with time structure or without it, the overall best accuracy with Fisher kernel is usually in the middle offering a good balance between the two groups of classes.

4. CONCLUSIONS

Several methods that adapt SVMs to sequence processing have been reviewed and applied to the classification of sounds from the meeting room environment. We have seen that the dynamic time warping kernels work well for sounds that show a temporal structure, but due to the presence of less-time-structured sounds in the database the best average score is obtained with the Fisher kernel. Moreover, only one Gaussian is used in that method due to its high sensitivity to the variance parameters as a consequence of the scarcity of data. On the other hand, the observed bias of the classifiers to specific types of classes is a good condition for a successful application of fusion techniques.

5. ACKNOWLEDGEMENTS

This work has been partially sponsored by the EU-funded project IP506909 – CHIL: Computers in the Human Interaction Loop, and the Spanish Government-funded project ALIADO.

6. REFERENCES

[1]. B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002

[2]. G. Guo, Z. Li, "Content-based audio classification and retrieval using Support Vector Machines", *IEEE Transactions on Neural Networks*, Vol. 14, pp 209-215, 2003

[3]. A. Temko, C. Nadeu, "Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering*", Proc. ICASSP'05*, pp. 505-508, 2005

[4]. T. Dietterich, "Machine Learning for Sequential Data: A Review", *LNCS*, Vol. 2396, pp 15-30, 2002, Springler-Verlag

[5]. J. Platt et al., "Large Margin DAGs for Multiclass Classification", *Proc. Advances in Neural Information Processing Systems* 12, pp. 547-553, 2000

[6]. H. Shimodaira, et al., "Dynamic Time-Alignment Kernel in Support Vector Machine", *Proc. Advances in Neural Information Processing Systems*, 14, vol.2, pp.921-928, 2001

[7]. C. Bahlmann, et al., "On-line Handwriting Recognition using Support Vector Machines - A kernel approach". *Proc. Int. Workshop on Frontiers in Handwriting Recognition*, 2002

[8]. V. Wan, J. Carmichael, "Polynomial Dynamic Time Warping Kernel Support Vector Machines for Dysarthric Speech Recognition with Sparse Training Data", *Proc. Interspeech'05*, pp 3321-3324, 2005

[9]. T. Jaakkola, D. Haussler, "Exploiting generative models in discriminative classifiers", *Proc. Advances in Neural Information Processing Systems* II, pp.487-493, 1999

[10]. V. Wan, S. Renals, "Speaker Verification using Sequence Discriminant Support Vector Machines", *IEEE Transactions on Speech and Audio Processing*, V. 13, no. 2, pp. 203-210, 2005

[11]. N. Smith, M. Gales, "Using SVMs and Discriminative Models for Speech Recognition" *Proc. ICASSP'2002*, pp 77-80, 2002

[12]. R. Anita, et al., "Outerproduct of trajectory matrix for acoustic modelling using support vector machines", *Proc* IEEE Int. Workshop on Machine Learning for Signal Processing, 2004

[13]. C. Nadeu et al., "On the decorrelation of filter-bank energies in speech recognition", *Proc. Eurospeech'95*, pp. 1381-1384, 1995

[14]. K. Tsuda et al., "Clustering with the fisher score", *Proc. Advances in Neural Information Processing Systems*, 15, pp. 729-736, 2003