CLASSIFICATION OF MEETING-ROOM ACOUSTIC EVENTS WITH SUPPORT VECTOR MACHINES AND VARIABLE-FEATURE-SET CLUSTERING

Andrey Temko and Climent Nadeu

TALP Research Center, Universitat Politècnica de Catalunya
Barcelona, Spain

{temko,climent}@talp.upc.es

ABSTRACT

Acoustic events produced in meeting-room-like environments may carry information useful for perceptually aware interfaces. In this paper we focus on the problem of classifying 16 types of acoustic events, using and comparing several types of features and various classifiers based on either GMM or SVM. A variable-feature-set clustering scheme is developed and compared with an already reported binary tree scheme. In our experiments with event-level features, the proposed clustering scheme with SVM achieves a 31.5% relative error reduction with respect to the best result from a binary tree scheme.

1. INTRODUCTION

Activity detection and description is a key functionality of perceptually aware interfaces working in collaborative human communication environments like meeting-rooms or classrooms. In such types of environments the human activity is reflected in a rich variety of acoustic events, either produced by the human body or by objects handled by humans, so acoustic scene analysis [1] may help to detect and describe human activity as well as to increase the robustness of automatic speech recognition systems. Actually, automatic scene analysis includes several tasks that target at the acoustic sources: segregation, localization, identification... Previously reported works have considered the problem of segmenting audio streams using a small number of categories (e.g. [2][3]), or detecting a given acoustic event (e.g. [4]). Several other published papers aim at classifying acoustic events, each one focusing on a given environment or a type of sounds: e.g. telemedicine [5], sports [6], animals [7], etc.

In this paper we focus on acoustic events that may take place in meeting-rooms or classrooms and on the preliminary task of classifying isolated sounds. The number of sounds encountered in such environments may be large, but in this initial work we have chosen 16 different acoustic events, including speech and music, and a database has been defined for training and testing. While in [8] the authors looked at the acoustic event classification (AEC) problem from the point of view of speech recognition, applying the usual automatic speech recognition strategy (cepstral features, hidden Markov model (HMM) classifiers), in our work we consider, develop and compare several feature sets and classification techniques, aiming at finding the ones which are most appropriate for the problem we are tackling. In this way, not only the parameters that model the short-time spectral envelope of the signals and its time derivatives are considered, but also other perceptual features, which may be more fitted to non-speech sounds.

HMMs require relatively large amount of data to accurately train the models, something that is not realistic in our task. That is the reason why we have tried Support Vector Machine (SVM) [9], a classifier that discriminates the data by creating boundaries between classes rather than estimating class conditional densities, so it may need considerably less data to perform accurate classification. In fact, SVMs have already been used for audio classification [10]. In this work we use SVM classifiers and compare them with Gaussian Mixture Model (GMM) classifiers.

As SVMs are binary classifiers, some type of strategy must be employed to extend them to the multi-class problem. In [10], the authors used the binary tree classification scheme to cope with several classes. In our work, we propose and develop a tree clustering technique using a specific feature set at each node. Relying on a given set of confusion matrices, that technique chooses the most discriminative feature set at each step of classification and, unlike the binary tree, it works for any number of classes.

Comparative tests have been carried out using the two basic classifiers (GMM and SVM) and several classification schemes. The effect of a confusion-based modification of the generalization parameters of the SVM classifier is also investigated in this work. The best results have been obtained using our proposed clustering scheme with SVM. Actually, it achieves a 31.5% relative average error reduction with respect to the best result from the binary tree scheme with SVM, and an even larger reduction with respect to the basic GMM-based classifier.

2. DATABASE

There is a lack of data for this classification problem, so acoustic event samples used in our work have been collected from several places, many of them from several websites. For four types of sounds we use 100 samples taken from the RWCP (Real World Computing Partnership) sound scene database [8]. Speech samples were taken from the ShATR Multiple Simultaneous Speaker Corpus [11] and short fragments from both close-talk and omnidirectional microphones are included. The database, which is

	Event	Source	Number
1	Chair moving	I	12
2	Clapping	RWCP + I	100+7
3	Cough	I	47
4	Door slam	I	80
5	Keyboard	I	45
6	Laugh	I	26
7	Music	I	38
8	Paper crumple	RWCP	100
9	Paper tear	RWCP	100
10	Pen/pencil handwriting	I	30
11	Liquid pouring	I	40
12	Puncher/Stapler	RWCP	200
13	Sneeze	I	40
14	Sniffing	I	13
15	Speech	ShATR	52
16	Yawn	I	12

Table 1. The sixteen acoustical events considered in our database, including number of samples and their sources (I means Internet).

	Name	*Content*	*Size*
1	*Perc*	Perceptual-spectral	11
2	*Ceps+der*	E+MFCC+d+dd	39
3	*Ceps*	E+MFCC	13
4	*FF+der*	FFBE+d+dd	39
5	*FF*	FFBE	13
6	*Perc+ceps+der*	"Perc" + "Ceps+der"	48
7	*Perc+ceps*	"Perc" + "Ceps"	24
8	*Perc+FF+der*	"Perc" + "FF+der"	50
9	*Perc+FF*	"Perc" + "FF"	24

Table 2. Feature sets that were used in this work, the way they were constructed from the basic features, and their size. d and dd denote first and second time derivatives, respectively. E means frame energy, and "+" means concatenation of features.

specified in Table 1, consists of 53 min of audio (942 files). The fact that sounds were taken from different sources makes the task more complicated due to the presence of several (at times even unknown) environments and recording conditions. An additional problem is the diversity in the number of samples per class.

3. FEATURES

All sounds were downsampled to 8 kHz, normalized to be in the range [-1 1], and framed (frame length=128, overlapping 50%, Hamming window). The silence portions were removed using an energy threshold. We used the following types of features:

1. Perceptual-spectral features

- Zero-crossing rate
- Short time energy
- Subband energies: 4 subbands equally distributed along 20 mel-scaled logarithmic filter-bank energies (FBE).
- Spectral flux: difference spectrum values between two adjacent frames and for the above-defined 4 subbands.
- Pitch. A simple cepstrum-based method was used to determine pitch in the range [70, 500] Hz.

2. Cepstral-based spectral parameters

12 mel-frequency cepstral coefficients (MFCC) were extracted from 20 bands. The zero-th coefficient was removed, but the frame energy was added to the set.

3. FF-based spectral parameters

Parameters based on filtering the frequency sequence of log FBEs (FFBE) have recently been reported [12]. We have used the usual second-order filter $H(z)=z-z^{-1}$, which implies a subtraction of the log FBEs of each two adjacent bands.

First and second time derivatives were also calculated for the last two types of features. The three types of features were combined to build 9 feature sets as shown in Table 2. The mean and standard deviation of the vectors estimated over the whole event signal were taken for classification, thus forming one vector per audio event with a number of elements which doubles the length of the feature set.

4. EXPERIMENTS

After randomly permuting the event samples within each class and indexing them, odd index numbers were taken for training and even index numbers for testing. 20 permutations in each experiment were performed. Because of unevenness in the number of representatives of the various classes, the overall performance is computed as an average of the individual class performances.

As SVMs with RBF kernels are used, there are two main parameters (hyperparameters) that are to be specified: the width σ of the Gaussian function, and the regularization parameter C, which controls the total distances from the misclassified points in the training phase to the bounding planes and it is considered as a tradeoff between minimizing the training error and maximizing the generalization capability of the classifier [9]. 5-fold cross-validation was applied to obtain the best setting of the kernel parameter σ. After the best kernel parameter is found, the whole training set is used again to generate the final classifier.

4.1. Binary tree scheme

First of all, the scheme proposed in [10], namely a binary tree with a SVM at each node, was applied to our acoustic event classification problem. In that reported work, each SVM was trained using a one against one strategy and C=200. In our work, we tried the same strategy, but with C=1, since this value yielded better results in our experiments, a fact that may indicate that our data are more noisy (contains more outliers) than data used in [10]. This SVM-based binary tree classification system was compared with a GMM classifier, which has one model per class and, for each test pattern, the model with maximal likelihood is chosen. Both a fixed and a variable number of Gaussians per class were tried, and the best accuracy was achieved by using a variable number that depends on the amount of data per class.

	1	2	3	4	5	6	7	8	9
SVM	78.7	72.1	77.9	77.5	77.6	81.2	82.3	82.9	82.4
GMM	71.9	67.3	72	69.9	73.4	73	77.4	75.1	78.9

Table 3. Percentage of classification rate for the SVM-based binary tree and the GMM classifiers on the defined feature sets.

Table 3 shows results for both classifiers. The best feature set in combination with the GMM classifier was the set number 9 (Perc + FF), with classification rate 78,9%, whereas for the SVM classifier was the set number 8 (Perc + FF + der), with 82,9% classification rate. Note that, in our experiments, the SVM approach shows a higher performance than the GMM one across all types of feature sets.

4.2. Variable-feature-set clustering scheme

Due to the acoustic variety of events for classification, it is reasonable to assume that different classes are better separated using specific feature sets (see Figure 1), the performance can improve by using a classification tree such that each of its nodes is associated to a different feature set. The tree is obtained by a clustering procedure based on the confusion matrices, one for each feature set, which result from the experiments reported in Section 4.1 by averaging over the 20 permutations. Those confusion matrices are used to find the best way of splitting the classes at a given node into two clusters and assigning a feature set with the least mutual confusion. For the sake of homogeneity, we use confusion matrices obtained by SVM classifiers for SVM clustering, and GMM matrices for GMM clustering. The confusion measures are normalized by the corresponding accuracies to cope with the dispersion of performance rates among the classes. As we have a relatively small number of classes, we can perform exhaustive search and get the global minimum. At the first step, all possible combinations of grouping 16 classes into two clusters (i.e. grouping 6 and 10, 8 and 8, etc) are searched over the available 9 confusions matrices that correspond to the 9 considered feature sets. For example, for the SVM clustering, we found that the 16 classes were best separated choosing the clusters {9} and {1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16}, and the 6th feature set. That process is carried out until we have single event clusters. Regarding the GMM classifier, the algorithm also groups the classes into two clusters, but in this case two models are generated at each step, one for each cluster. The algorithm is described with detail in [14].

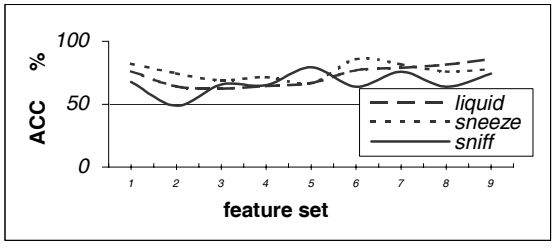

Figure 1. Dependence of performance of classifying "liquid_pouring", "sneeze" and "sniff" sounds upon the feature sets, using SVMs.

In our experiments, we have tried several ways of alleviating the problem of having a too much different amount of training data between the two clusters at a given tree node. The first way is to restrict the exhaustive search to look for an equal number of classes at each cluster. Hereafter, we will refer to this variant as restricted clustering.

Figure 2 shows the trees obtained by the normal (unrestricted) and restricted clustering algorithms in the SVM case. Note that the two trees show a very different structure but, denoting with N the number of classes (N=16), they have the same number of nodes (N-1), that is the same number of trained SVM classifiers. As observed in Figure 2, in normal clustering we mostly have only one class separated on each clustering step while the restricted tree shows a balanced structure. Regarding the GMM-based techniques, since each class model is trained without using information about the other classes it is not so much influenced by the problem of data unbalance. We also consider both clustering schemes for the GMM case. The resulting GMM schemes are similar to those in Figure 2

The second proposed way of coping with data unbalance is to introduce different C values for positively- and negatively-labeled training samples [13]. Additionally, since for each tree node an estimate of the number of confusions can be obtained as a byproduct of the clustering algorithm, we have used this estimate values to adapt the regularization parameters. Thus, apart from the normal way of using the regularization parameter C in the SVM-based classifiers, three other different methods are considered in this work: 1) only one C parameter computed as K times the inverse of the number of confusions at the given tree node, 2) two different C values (C^+ and C^-) whose values are K times the ratio between the number of negative (positive) training samples and the number of positive (negative) training samples, and 3) the effect of doing both adaptations simultaneously. The coefficient of proportionality K was set to 10 for all adaptations.

Results are shown in Table 4. Note that the highest performances are obtained by using method 3, and again SVM performs better than GMM.

5. DISCUSSION AND CONCLUSIONS

Regarding the features, we have observed from the experimental results that the best separating feature sets for the most confused classes mostly are FFBE-based features (sets 4,5,8,9), while it appears that the least confused classes mostly are MFCC-based (sets 2,3,6,7). This fact may indicate that the FFBE-based features are more discriminative than the MFCC features for highly overlapped data distributions, while MFCC features appear to show the best performance when there is a clearer separation between classes. However, for the most confused classes in the GMM case, the average best feature set is the one we have called perceptual set. This may be due to the relatively low size of that feature set, which facilitates the estimation problem.

The proposed clustering schemes (both normal and restricted) show two computational advantages in front of the binary tree classifier. First, the required number of trained SVMs is N-1, where N is the number of classes, while for the binary tree (N-1)N/2 trained SVMs are needed. Second, the proposed schemes involve a smaller number of classification steps, 4 for restricted clustering and between 1 and 14, depending on the input pattern,

	C=K	METHOD 1	METHOD 2	METHOD 3
SVM-N	84.67 ± 2.5	84.05 ± 1.7	86.71 ± 1.4	88.29 ± 2.1
SVM-R	84.72 ± 2.6	84.88 ± 2.7	84.95 ± 2.2	87.20 ± 1.5
GMM-N	83.6 ± 2.2			
GMM-R	82.15 ± 2.3			

Table 4. Performances of variable-feature-set clustering classifiers using different adaptations of the regularization parameters for the SVM case. -N and -R, denote normal and restricted clustering scheme, respectively. Standard deviations σ estimated over 20 repetitions are denoted with ± σ

for normal clustering in our case (see Figure 1), whereas the binary tree requires 15. However, the proposed variable-feature-set scheme has an obvious disadvantage: with our choice of feature sets (see Table 2) up to 9 feature sets can be involved in testing, 7 in our case (numbers 3 4 5 6 7 8 9).

The column C=K in Table 4 shows that, without any adaptation, SVM restricted clustering performs equally well as normal clustering. In that table, we can notice that SVM-N takes advantage of using different C values for each cluster. And SVM-R does not take any advantage, due presumably to the balancing average implied by the half-to-half constraint. Additionally, as we can see from Table 4, introducing prior knowledge (about confusions) with the generalization parameter C (method 1) does not have a positive influence on the classification performance, while introducing it along with different C values for positive and negative classes (method 3) leads to an improvement for both types of clustering trees. The gain in performance, however, is not much significant, so there is a need to have a more sophisticated algorithm of introducing prior knowledge about confusions in the regularization parameters. In restricted clustering we can obtain only the global minimum of error within the constraint so the final performance of the SVM-R technique is worse than that of the normal one (Table 4, method 3). We can also observe that normal clustering seems to perform slightly better than restricted clustering for GMM.

In summary, the best results were obtained with SVM and the proposed clustering scheme, arriving to a 88.29 % classification rate, which means a 31.5% relative average error reduction with respect to the best result from the binary tree scheme with SVM. That good performance is attributable to the clustering technique, and to the fact that SVM provides the user with the ability to introduce knowledge about class confusions.

6. ACKNOWLEDGEMENTS

The authors wish to thank Enric Monte for his valuable insights into pattern recognition and machine learning, and also Jaume Padrell for his support during the work. 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.

7. REFERENCES

[1]. A.Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge,1990

[2]. L. Lu et al "Content Analysis for Audio Classification and Segmentation", *IEEE Transactions on Speech and audio processing*, V.10, N. 7, pp. 504-516, 2002.

[3]. J. Pinquier et al "Robust Speech / Music Classification in Audio Documents", *Proc.ICSLP'02*, Vol.3, pp. 2005-2008, 2002.

[4]. L. Kennedy and D. Ellis, "Laughter Detection in Meetings", NIST Meeting Recognition Workshop, *Proc.ICASSP'04*, 2004.

[5]. M. Vacher et al, "Sound Detection and Classification through Transient Models using Wavelet Coefficient Trees", *EUSIPCO'04* pp. 1171-1174, 2004

[6]. Z. Xiong et al, "Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework", ICME'03, Vol. 3, pp. 401-404, 2003

[7]. M. Slaney, "Mixtures of Probability Experts for Audio Retrieval and Indexing", *Proc. ICME* , pp.345- 348, Vl, 2002.

[8]. T.Nishiura et al., "Environmental Sound Source Identification Based on Hidden Markov Model for Robust Speech Recognition" *Proc. Eurospeech'03,* Geneva, pp.2157-2160, 2003

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

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

[11]. ShATR Multiple Simultaneous Speaker Corpus *http://www.dcs.shef.ac.uk/research/groups/spandh/projects/shatrweb/index.html*

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

[13]. G. Karakoulas, J. Shawe-Taylor, "Optimizing classifiers for imbalanced training sets", *Proc. Neural Information Processing Workshop, NIPS'98*, pp.253-259, 1999.

[14]. A. Temko, C. Nadeu, "Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering", *Internal UPC-TALP report*, December 2004.

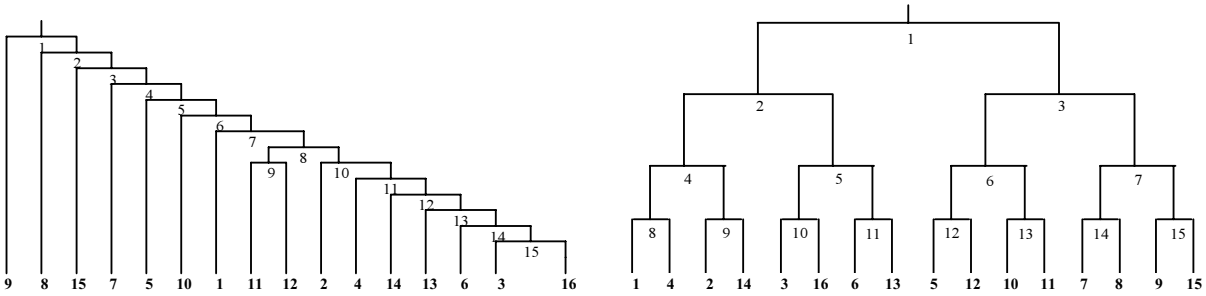

Figure 2. Normal and restricted clustering schemes for SVM classifiers