

Combining frame and segment based models for environmental sound classification

Pengfei Hu, Wenju Liu, Wei Jiang

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

pfhu@nlpr.ia.ac.cn, lwj@nlpr.ia.ac.cn, wei.jiang@ia.ac.cn

Abstract

The paper considers the task of recognizing environmental sounds, which plays a critical role in human's perception of an auditory context in audiovisual materials. A variety of features have been proposed for audio recognition, either frame-based or segmental. Here, we propose a two-stage framework to combine modeling in these two levels. First, the Gaussian Mixture Models (GMMs) are built based on short-term features and pre-classification are performed. Then, in the event that the GMMs are not certain about the result, the system engages Support Vector Machines (SVMs) to refine the output hypothesis. In the next stage, the features are combined by taking posterior estimates of GMMs along with segmental features as SVMs' input features. Experiments on the sound dataset show that the proposed framework makes an improvement over the traditional methods.

Index Terms: environmental sound classification, model combination, GMMs, SVMs

1. Introduction

Audio data is an integral part of many modern computer and multimedia applications. A typical multimedia database often contains millions of audio clips, including environmental sounds, music, speech sounds and other non-speech utterances. Unlike speech and music, most environmental sounds are unstructured, but they also provide a significant wealth of information, which is important and helpful in many applications, such as context-aware computing [1], video content parsing [2], and home automation [3].

Research on environmental sound analysis has received some interest in the last few years. Firstly, Wold *et al.* [4] presented a sound classification and retrieval system, in which pitch, loudness, brightness, and bandwidth, were used as audio features and the nearest neighbor (NN) rule was adopted to classify the query audio into one of the predefined sound classes, Li *et al.* [5] concatenated the perceptual and cepstral feature sets for sound recognition. A new classifier named nearest feature line was presented and produced better results than the NN-based methods. In the later work, Guo and Li [6] improved the classification performance by using SVMs with binary tree structure based on the same feature set. The SVM approach was also adopted by Lin *et al.* [7]. They applied wavelet transform to extract acoustic features such as subband power and pitch information.

In the work of [4] [5] [6] [7], the means and standard deviations of all short-term features in a sound file were computed to form a feature vector of the whole sound. This way of feature processing makes it convenient to apply the common clas-

sifier like SVM, but it discards much information. In [8], Lu *et al.* select more discriminative long-term features such as high zero-crossing rate ratio and spectrum flux, which are proved to be useful in characterizing different audio signals.

Other techniques such as GMM and HMM are modeling on frame level directly instead of representing an audio clip by a vector. For example, Aucouturier *et al.* [9] used 50-component GMMs to model the distribution of Mel-Frequency Cepstral Coefficients (MFCCs) and studied the differences between urban environment and polyphonic music. In the work of Selina *et al.* [10], they introduced a set of matching pursuit-based features as a supplement to MFCCs and built GMMs for fourteen types of environmental sounds. The average recognition accuracy is approximately 83.4%. Eronen [1] extended the work to audio-based context recognition. They classified twenty-four individual contexts and six higher-level categories by using discriminative HMMs. The reported average recognition rates are 66% and 79%. HMM has been applied for speech recognition and sound related applications for several decades. Its ability to model temporal sequences is an advantage when modeling text-dependent tasks. However in the task-independent task, [11] showed that removing transition probabilities in HMM models had little effect on performance of text-independent speaker recognition.

In this study, we first examine the effectiveness of using HMMs and GMMs in modeling environmental sounds. Furthermore, we propose a two-stage framework based on GMMs and SVMs to combine the features on frame and segment levels. In the first stage, the short-term features such as MFCCs and spectral centroid are extracted and GMMs are employed as models on the frame level. In the second stage, we perform reclassification by using SVMs as classifier based on segmental representation. The major contribution of our work is combining models by concatenating the segmental features and the posterior likelihoods of GMMs to form a high dimension vector as the input of SVMs. Fig 1. illustrates the overall block diagram of the proposed framework.

The remainder of the paper is organized as follows. In section 2 the classification based on short-term features is described. In section 3, the SVM classifier on the segmental level and combination of the two models are presented in detail. The experiments are carried out in section 4. Finally the discussions and conclusions are presented in section 5.

2. Frame-based modeling

In the audio signal processing, the waveform of a sound is usually divided into separate frames. Passing through the procedure of feature extraction, each frame will be transformed into

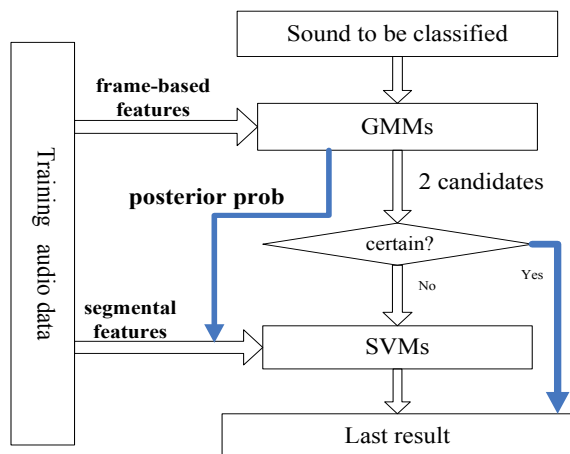


Figure 1: the overall block diagram of the proposed framework.

a feature vector. Based on the short-term features, the models are then built for the sound classification. The GMM and continuous HMM are the most common frame-based classifiers in sound related applications and we examine their effectiveness in modeling environmental sounds.

2.1. Feature Extraction

For the discriminant features that can capture the inherent characteristics of an audio signal, we employ two types of short-term features: perceptual features and Mel-Frequency Cepstral Coefficients (MFCCs). The perceptual features include spectral centroid, spectral flatness, spectral roll-off, and zero crossing rate [5].

- Spectral centroid measures the brightness of a sound. The higher the centroid, the brighter the sound.
- Spectral flatness indicates how flat the spectrum of a sound is.
- Spectral roll-off quantifies the frequency point at which the accumulative value of the frequency magnitude reaches 85% of the total magnitude.
- Zero crossing rate is defined as the number of time-domain zero-crossing within a frame. It is a simple measure of frequency content of a signal.
- The first 13 MFCCs and their derivatives are calculated for the salient features (we include log energy in the calculation)

Finally each analysis frame is thus represented by a 30-dimensional vector.

2.2. Classifier Design

In the first stage of the proposed framework, we select GMM as the classifier. GMM provides an effective way to describe the audio characters, and one of its powerful attributes is the capability to form smooth approximations to arbitrarily shaped densities. When making classification, the audio segment is labeled according to category that has the maximum *a posteriori* score. Similar to GMM, HMM is also a popular method based on short-term features. The HMM contains several states with a GMM in each state and transition probabilities between states. Thus the biggest difference between these two models

is that HMM has the ability to model temporal sequences by using transition probabilities. However, in the task of environmental sound classification, the audio data in the dataset are not usually aligned in advance. The sequencing of sounds in the training data does not always reflect the test sounds. This raises the question about the appropriateness of temporal sequencing with HMM when modeling environmental sound. In our work, we compare the performance of GMM and HMM with different state numbers and component numbers. The Expectation-Maximization and Baum-Welch algorithms are respectively used for their parameters estimation.

2.3. Confidence Measure

For each sound clip in the testing data, GMM that yields the highest log likelihood is selected as the output class. The average likelihood over all of frames from the two highest scoring GMMs can be used to form a confidence measure. Let L_1 be the log likelihood of the best matching model, and L_2 be the log likelihood of the second best model. The confidence measure for the classification can then be computed as

$$conf = \left| \frac{L_1 - L_2}{L_1} \right| \quad (1)$$

By selecting an appropriate threshold, the computed confidence measure can be used. Moreover, we investigate the classification result of GMMs and find which categories are easily confused. If the confidence is below the threshold and the two candidate categories are easily confused, the result can be labeled as uncertain and the system performs reclassification by SVM in the second stage.

3. Segment-based Classification

On the segment level, one audio segment is represented by a vector and traditional classifiers such as SVMs are applied based on segment representation. Here, we integrate two kinds of features as the segmental features of baseline system: the mean and variance of the short-term features and the long-term features. The long-term features include spectrum flux, high zero crossing rate ratio (HZCRR), and low short-time energy ratio (LSTER) [8]. In detail, spectrum flux reflects the average variation of spectrum between the adjacent two frames. HZCRR is defined as the ratio of the number of frames with higher zero crossing rate while LSTER is the ratio of number of frames whose energy are less than half time of average short-time energy. They describe the characteristic of energy and spectrum in the whole audio segment.

3.1. SVM

Compared with the generative model GMM, the Support Vector Machine (SVM) is a discriminative model with great generalization and more discriminative ability. The main idea of this model is to project the input vectors into a high dimension space in order to find a hyperplane which allows making a linear separation between two classes. In practice, SVM performs the computation of scalar product in the feature space by using kernel functions. These functions are also used to calculate the optimal hyperplane in the feature space without using directly the mapping function $\phi(x)$. The separating hyperplane is chosen in order to maximize the distance between the hyperplane and the training vectors closest to the border. These training vectors are called support vectors. An SVM classifier is given

by the following formula:

$$f(x) = \sum_{i=1}^N a_i y_i K(x, x_i) + b \quad (2)$$

where x is the vector to classify, x_i are the training examples, $K(x, x_i)$ is the kernel function and y_i correspond to the class label $y_i \in \{-1, +1\}$; a_i and b are the parameters of the model obtained in the training phase of SVM.

For SVM classification, the selection of kernel function is very important. In this paper we apply two different kernel types: the polynomial kernel and the radial basis function (RBF) kernel, which are very popular in many applications.

3.2. Model Combination

In our proposed framework, the posterior probabilities of GMMs are taken into account to form an improved more discriminant feature vector that is fed into the SVM classifier. This framework allows two ways of interpretation; 1.as a hybrid system that combines GMMs and SVMs, 2.as a feature extraction method, which considers the GMMs as the feature transformation. We hope that, in this way, more discriminative information on the frame level is inherited by concatenating the posterior probabilities and segmental features. When computing the posterior probabilities, we use the average value of log likelihood over all frames for each audio segment.

Next, the most direct implementation is feeding the new segmental representation into SVMs and performing multi-class classification. Prior to doing it, we introduce confidence measure to make judgement about the classification result of GMMs. As described in Fig.1, only when the GMMs' result is uncertain, the system engages SVMs for reclassification. In practice, we implement two schemes within this framework. In the first case, a multi-class classification task among all of categories is performed by using SVMs. The SVM is a binary classifier, so we need adopt the one-against-one strategy to construct the multi-class classifier; namely, $K(K-1)/2$ classifiers are built, where K is the number of classes. Furthermore, we consider another scheme to reduce the number of SVMs. The easily confused classes are found out by investigating the confusion matrix of GMMs' classification result. Only the classifiers among those easily confused categories will be trained. In implementation, a multi-class problem is reduced to a two-class problem between classes with two highest scoring GMMs. By doing so, the number of SVMs can be greatly reduced.

4. Experimental Evaluation

4.1. Experiment Setup

To evaluate the performance of the proposed system, we collect 1319 environmental sound clips from [13] and [14]. All of the sounds belong to 24 different categories which are commonly encountered in daily life such as alarm, bell, car-racing, car-crash, rain, stream, gun-shot, cry, laughter, bird, dog, crow, and so on. The length of sound files ranges from one second to less than one minute.

For experimental convenience, all sounds are unified to 16 KHz and 16 bit linear bit. The short-term features are extracted by using a window of 30ms with a 50% overlap. The extraction of MFCCs and the training of GMM and HMM are implemented by HTK 3.4 [15]. In the experiments, 5 fold cross validation is performed for accuracy estimation. More precisely, the sound database is equally divided into 5 disjoint subsets,

and classifiers were trained five times, each time with a different subset held out as a testing set. The estimated classification result is the average of these five accuracy rate for the testing data.

4.2. Initial Classification

We implement initial classification based on short-term features. To compare the performances of frame-based models, we examine the GMMs with 16,32,48 or 64 mixtures and HMMs with two or three states when the topology is fully-connected or left-right with skips. The overall recognition rates of GMMs are given in Table 1. We see that the classification performance of GMMs grows as the number of mixtures increases and become almost stable when the number is more than 48. Table 2 gives the result of HMMs. In general, the fully-connected HMMs outperform the left-right HMMs. Simultaneously, the overall performance of GMMs exceeds the results achieved with HMMs with same number of mixtures. This reveals that the HMM's transition property has little or no effect on the overall classification performance for environmental sound classification. This is because that the training data in the same class is not aligned beforehand and the sequencing information of sound clips in the same class is inconsistent. As the result, the transition probabilities in trained HMMs possibly reflect the transitions across different sound clips more than internal changes.

Table 1: Results of GMMs with different number of mixtures

Mixtures number	16	32	48	64
Accuracy	79.8%	80.8%	81.5%	81.6%

Table 2: Results of HMMs with varying topology, different number of states and mixtures

Mixtures number per state	Left-right with skips		Fully-connected	
	two states	three	two	three
8	74.3%	76.6%	75.9%	77.1%
16	76.4%	79.0%	78.1%	79.5%
32	78.3%	80.3%	79.2%	81.4%

By selecting GMMs with 48 mixtures as frame-based classifier, we investigate the confusion matrix of classification results to find out the easily confused classes. Table.3 gives a list of misclassification of some categories. In most situations, the testing sound from one categories are usually misclassified as specific classes. There are no overlap between results of most categories. In the next stage, we can reduce the number of the SVM classifiers according to this finding.

Table 3: List of some misclassification

Correct category	Misclassified as
alarm	car-racing ,cat,scream
bird	cough,wind,moto
cry	cat,laugh
thunder	explode,car-crash

4.3. SVM for classification

After the new segmental features are obtained by concatenating the posterior probabilities of 24 GMMs and basic segmental

features, the SVM will be used to perform classification, which is the essential part of the proposed framework. We conduct experiments for SVM classification in different schemes, which include the baseline system. The baseline system is multi-class classification based on basic segmental features and SVM. All of the rest three ones are based on new segmental features. The system I performs multi-class classification based on new features by using SVM directly without considering confidence measure (CM), in which all the results of GMMs will be relabeled by SVMs. By judging the certainty of the GMMs' results with CM, we propose the system II and III. The former handles multi-class classification by using SVM when the results of the first stage is uncertain. In experiments using multi-class SVM, we need train $24 * (24 - 1) / 2 = 276$ SVM classifiers. To reduce the number of classifiers, we only train the SVM classifiers for easy-confused categories in system III and perform two-class classification between categories with two highest scoring GMMs. In fact, the number of SVMs in system III is reduced to less than 70. Table 4 shows the performances of four systems. For all systems, the polynomial kernel and RBF kernel are used and the best parameters are searched in experiments.

Table 4: the result of SVM classification in different schemes

	Polynomial	RBF Kernel
SVM Baseline System	79.4%	80.2%
GMM+multi-SVM (I)	82.8%	83.5%
GMM+CM+multi-SVM (II)	83.3%	83.9%
GMM+CM+two-SVM (III)	84.7%	85.0%

From the Table 4, we can see that the introduction of posterior probabilities as a supplement to basic segmental features is significative, which makes an improvement over the traditional methods. The system I based on new segmental features outperforms the baseline by 3% and improve the accurate rate of the GMMs by 2%. With considering the confidence measure, the system II produces a better result than the system I. To find out the reason we track the results of two stages. For some sounds, the SVMs make a misclassification while the GMMs give a right label in the first stage. Checking the results of GMMs by confidence measure can prevent undesired things like that to some extent. In experiments, when the threshold of confidence measure is set to be 0.09, the optimal results are reached. Besides, it is also suggested that the performance is enhanced in system III, in spite of fewer SVM classifiers. The reason for this result is that the SVM is more suitable for two-class classification than multi-class classification. It demonstrates that the work about investigation of easily confused categories is meaningful. In sum, the best method in this two-stage framework can achieve an improvement over GMM method by more than 3%.

5. Conclusions

This paper has presented a novel approach to combine models on the frame and segment levels for environmental sound classification. On the frame level, GMMs are selected as models to gather most information about shot-term features. Then, the GMMs' likelihoods of sound clip are used as one part of segmental features. Based on the new segmental representation, SVMs are employed for reclassification on the segment level. The posterior probabilities of GMMs play the role of bridging GMM and SVM models. Therefore, this approach integrates the merits of both generative model and discriminative model. Ex-

perimental results show that the proposed system has achieved a significant improvement over the traditional methods.

In addition, we compare the effectiveness of GMM and HMM in the task of environmental sound classification. It is verified that the transition probabilities of HMM have no effect on modeling environmental sound. We also investigate the classification results of GMMs and introduce confidence measure to serve SVM classification. As a result, the number of SVMs to be trained is reduced greatly.

6. Acknowledgements

This work was supported in part by the China National Nature Science Foundation (No.91120303, No.90820011 and No.90820303), 863 China National High Technology Development Project (No.20060101Z4073, No.2006AA01Z194), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

7. References

- [1] Antti,J.E.,Vesa,T.P.,Juha,T.T., "Aduio-Based Context Recognition", IEEE Trans. on Audio, Speech, and Language Processing,14(1):321-328, 2006.
- [2] Simon,M.,Chitra,D., and Svetha,V., "Detecting indexical signs in film audio for scene interpretation", in Proceedings of IEEE international Conference on Multi and Expo,Tokyo,Japan, 2001.
- [3] Wang,J.C., Lee,H.P., Wang,J.F., and Lin,C.B, "Robust environmental sound recognition for home automation", IEEE Trans. on Automation Science and Engineering",5(1):25-30, 2008.
- [4] Wold,E., Blum,T., Keislar,D., and Wheaton,J., "Content-based classification, search and retrieval of aduio", IEEE Multimedia Magazine,3(3):27-36, 1996.
- [5] Li,S.Z., "Content-based classification and retrieval of audio using the nearest feature line method", IEEE Trans. on Speech and Audio processing,8(5):619-625, 2000.
- [6] Guo,D.G., and Li,S.Z., "Content-based audio classification and retrieval by support vector machines", IEEE Trans. on Neural Network,14(1):209-215, 2003.
- [7] Lin,C.C., Chen,S.H., Truong,T.K., and Chang,Y.K, "Audio classification and categorization based on wavelets and support vector machine", IEEE Trans. on Speech and Audio Processing,13(5):644-651, 2005.
- [8] Lu,L., Zhang,H.J., and Jiang,H., "Content analysis for audio classification and segmentation", IEEE Trans. on Speech and Audio Processing,10(7):504-515, 2002.
- [9] Aucouturier,J.J., Defreville,B., and Pachet,F., "The bag-of-frames approach to audio pattern recognition:A sufficient model for urban soundscapes but not for polyphonic music", The Journal of Acoustic Society of America,122(2):881-891, 2007.
- [10] Chu,S., Narayanan,S., and Jay Kuo,C.C., "Environment Sound recognition with time-frequency audio features",IEEE Trans. on Audio,Speech,and Language Processing,17(6):1142-1158, 2009.
- [11] Tishby,N.Z., "On the application of mixture ar hidden markov models to text independent speaker recognition", IEEE Trans. Signal Processing,39(3):563-570, 1991.
- [12] Vanick,V.N., "Statistical Learning Theory",Wiley, 1998.
- [13] "The BBC Sound Effects Library-Original Series", Online: <http://www.sound-ideas.com/bbc.html>.
- [14] Online: <http://www.freesound.org/>.
- [15] Young,S.,et al., "The HTK Book for version 3.4.1", Cambridge, 2000.