Generic Audio Classification Using a Hybrid Model based on GMMs and HMMs

Menaka Rajapakse and Lonce Wyse
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{menaka,lonce}@i2r.a-star.edu.sg

Abstract

A hybrid model comprised of Gaussian Mixtures Models (GMMs) and Hidden Markov Models (HMMs) is used to model generic sounds with large intra class perceptual variations. Each class has variable number of mixture components in the GMM. The number of mixture components is derived using the Minimum Description Length (MDL) criterion. The overall performance of the hybrid model was compared against models based on HMMs and GMMs with a fixed number of mixture components across all classes. We show that a hybrid model outperforms both class-based GMMs, HMMs, and GMMs based on fixed number of components. Further, our experiments revealed that the contribution of transitions between states in HMMs has no significant effect on the overall classification performance of generic sounds when large intra class perceptual variations are present among sounds in the training and test datasets. Sounds that show multi-event structure with events that tend to be similar (repetitive) indicated improved performance when modeled with HMMs that can be attributed to HMM's state transition property. Conversely, GMMs indicate better performance when the sound samples show subtle or no repetitive behavior. These results were validated using the MuscleFish sound database.

1. Introduction

Several studies have been carried out to analyze sounds in the context of discriminating speech and music, speaker variability analysis, musical instrument classification, similar song search, however, only a few studies on generic sound classification have been performed to date. One major bottleneck in coming up with a robust sound classification framework is the difficulty in retrieving a single set of acoustic features that can optimally describe a certain sound class. This is due to the multiplicity of the sound sources that can comprise a single sound class (perceptual variations). For instance, a sound class categorized as *manufacturing* may contain manufacturing of steel, textile, printing, bottling etc. making it difficult to model and eventually to classify. Furthermore, the limited feature extraction capabilities may also undermine the separation of subtly distinct sound sources that generate sounds of different classes. In general, human sound labeling tends to correspond to physical or visual attributes of a sound source rather than directly to acoustic attributes. This causes sounds with significant variations to be grouped under a single category and makes it difficult to find a suitable feature sets that can invariably describe all sounds encompassed by a single class label.

A recent study carried out by Lie Lu [4] reported an impressive performance on the classification of 5 audio classes containing silence, music, background sound, pure speech and non-pure speech using Support Vector Machines (SVM). Another study [2] based on SVM had been carried out on the MuscleFish database and had reported lower classification error rate compared to the error rate reported by Wold [16]. According to the authors in [2], using SVMs is a trade-off between the accuracy and the computational complexity, and a trial and error approach has been adopted in choosing a kernel function and the parameters therein.

Moreover, techniques such as Hidden Markov Models (HMM) have been applied for speech recognition and sound related applications for several decades. The ability of HMMs to model speech sounds together with temporal sequences has made it a promising choice in modeling other sounds. The ability to model temporal sequences is an advantage when modeling text-dependent tasks, but the sequencing of sounds found in the training data does not always reflect the sounds found in the test dataset [7]. In [13] [5], it was shown that removing transition probabilities in HMM speaker models had no effect on the performance of text-independent tasks. This in turn raises the question about the appropriateness of temporal sequencing with HMM when modeling generic sounds, which can be considered as a text-independent task. Reyse-Gomez and

1550-5502/05 $20.00 © 2005 **IEEE**

Ellis [8] have used HMMs for generic acoustic modeling, and the model parameters were estimated using number of approaches. The overall best performance was achieved with low entropy criterion where a mixture of Gaussians has been used to estimate the number of states and the initial transition matrix for each class model.

In this study, we examine the effectiveness of using HMMs and GMMs in modeling generic audio. GMMs provide a probabilistic model for the underlying sounds generated by sound effects and also have the ability to model arbitrary densities by using a number of Gaussian component functions. A multi-modal density function can be constructed by combining Gaussian components. Moreover, GMMs are good for classifying data consisting of subcategories. The application of GMMs in the modeling and classification of speech, audio, music and text-independent speaker identification can be found in literature [11] [7] [1]. A recent comparison study carried out between MFCC and MPEG-7 features for sports audio classification has achieved equally good results from both types of features [17] which motivated us to use Mel-Frequency Cepstral Coefficients(MFCC) as the audio features for our experiments. In this paper, we compare generic acoustic models based on a hybrid model with HMMs and a variety of GMMs for their performance against a database of generic sounds with large inherent variability.

2. Acoustic Modeling

Features of a given sound class are extracted by subjecting sound files in the database to Mel-Frequency Cepstral Coefficients (MFCC) extraction. Fig. 1(a) shows the 2-D projection of these multidimensional features for 3 different sound classes. As we can see, due to the overlapping among classes in the feature space, finding the boundaries that optimally separate among these classes is challenging. As Fig.1(b) shows, the distribution of a typical sound class can be represented as a mixture of many sub classes that can be approximated by several Gaussian distributions.

2.1. Gaussian Mixture Model (GMM)

The motivation for using Gaussian densities as the representation of audio features [18], and speaker identification [7] is the potential of GMMs to represent an underlying set of acoustic classes by individual Gaussian components in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix. Also, GMMs have the ability to form a smooth approximation to the arbitrarily-shaped observation densities in the absence of other information.

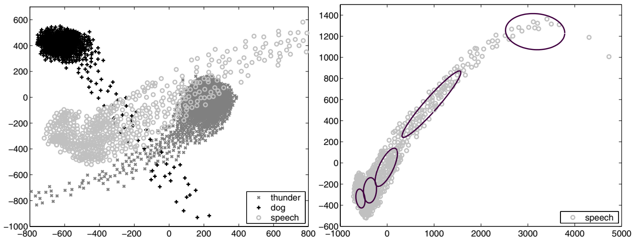

Figure 1. (a) 2-D projection of 3 different sounds (b) Fitted distributions of multiple Gaussian of a single class of speech data in (a).

GMMs have been extensively used for speaker identification [7] [12], music-speech discrimination [11]. With GMMs, each sound class is modeled as a mixture of several Gaussian clusters in the feature space; each sound cluster in the feature space is represented by a mean vector and a covariance matrix. Let $X = \{X_1, X_2, \cdots, X_n, \cdots, X_k\}$ be a set of sound classes where k denotes the number of classes. For a given sound $\mathbf{x} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_{\mathrm{T}}]$, we model the class conditional probability, $p(\mathbf{x}|X_k)$ with a GMM:

$$p(\mathbf{x}|X_k) = \sum_{j=1}^{m_k} p(\mathbf{x}|c_{kj}) b_{kj}$$

where each mixture component density $p(\mathbf{x}|c_{kj})$ is modeled by a d-variate Gaussian distribution where d represents the number of features, with mean μ_{kj} and covariance matrix $\sum_{kj}$. and the mixing parameter b_{kj} corresponds to the prior probability that the sound $\mathbf{x}$ was generated by component c_{kj}; $\sum_{j=1}^{m_k} b_{kj} = 1$; m_k denotes the number of components in the sound class X_k. By assuming that the given data are uncorrelated, we can use a diagonal covariance matrix, and write the distribution of the i^{th} component c_{kj} as multivariate normal with parameters, mean and auto correlations:

$$p(\mathbf{x}|c_{kj}) \sim (\mu_{kj}, \sigma_{kj})$$

with μ_{kj} denoting the mean, and σ_{kj} denoting the standard deviation of each component. And each mixture model is represented by a triplet; $M_k = (\mu_{kj}, \sigma_{kj}, b_{kj})$.

2.1.1. Parameter Estimation Let a set of sound classes, X, be represented by a set of GMMs, $M = \{M_1, M_2, \cdots, M_n, \cdots, M_k\}$ where

$$M_k = \{(\mu_{kj}, \sigma_{kj}, b_{kj}); j = 1 \cdots m_k\}.$$

The EM algorithm [6] is used to find the maximum likelihood parameters of each class. The aim here is to find a sound model, M_k, that optimizes the posteriori probability given the sound, $\mathbf{x}$. ie; $p(M_k|\mathbf{x})$. With parameters of the estimated sound models, the marginal posterior probabilities

$p(M_k|\mathbf{x})$ are evaluated.

$$M^* = \arg\max_k \quad p(M_k|\mathbf{x}) \tag{1}$$

Assuming equally likely prior probabilities $P(M_k)$ for all classes, we can rewrite Eq.(1) as the maximum likelihood criteria:

$$M^* = \arg\max_k \quad p(\mathbf{x}|M_k)$$

2.2. Hidden Markov Model (HMM)

HMMs are widely used for modeling time series data. A discrete-time HMM can be viewed as a Markov model whose states are not directly observable. Each of these hidden states is associated with a probability distribution function which models the symbol emission probability from that state. A HMM can be fully described by 3 entities: the state transition probability distribution A, the observation probability distribution, B, and the initial state distribution π. By using the compact notation, $\lambda = (A, B, \pi)$, we can represent the complete parameter set of the model. This parameter set defines a probability measure for an observation sequence O, which can be given as $P(O|\lambda)$.

2.2.1. Parameter Estimation The training of the models for each sound class is carried out using the Forward-Backward algorithm [6], also known as Baum-Welch reestimation, which determines the HMM parameters, λ that maximize the probability $P(O|\lambda)$. For the evaluation of a sequence, forward algorithm was adopted.

3. Experiments And Results

3.1. Feature Selection

For our experiments, we use MFCCs which are short-term spectral features. The extracted MFCCs are d-dimensional feature vectors, where $d = 39$ containing the energy, cepstrum, first order derivative and the second order derivative terms. The MFCC feature vectors are extracted at 8kHz sampling rate with an overlapping window of size 128, and a DFT size of 256. Each sound file within a class is normalized across individual features so that the mean is zero and the variance is equal to one.

3.2. GMM Model Order Selection

For GM modeling, the MDL criterion [9] was used to determine the number of optimal mixture components per class. According to the experiments performed on synthetic and real data [10], MDL had been able to derive reliable estimates in most cases, and yielded sensible estimates even for the most difficult data partioning instances. The experiments in [10] have been carried out assuming the full covariance of data. In our case, we assume diagonal covariance of the features due to the highly uncorrelated nature of the MFCC based feature vectors [3]. The MDL cost function can be written as:

$$MDL(m_k, M_k) = -log(p(\mathbf{x}|m_k)) + \frac{1}{2}N(M_k)logT \tag{2}$$

The first term in Eq.(2) measures the model's entropy and the second term penalizes over complex models. The number of free parameters, $N(M_k)$, in the GM model can be estimated as in [14]: $N(M_k) = 2dm_k + (m_k - 1)$. The first term is for d dimensional mean and diagonal covariance, and the second term is introduced for the $(m_k - 1)$ adjustable mixture weights imposed by the constraint $\sum_{j=1}^{m_k} b_{kj} = 1$. Once several models per class are trained, the model selection is carried out based on the MDL principle by choosing the model with the minimum description length as:

$$\hat{M_k} = \arg_{M_k} \min(MDL(m_k, M_k)) \tag{3}$$

3.3. Hybrid Model

A hybrid model is constructed using GMMs based on class-based variable mixture component and HMMs. Hybrid models take advantage of the positive aspects of each model used in the combination. Each type of model is capable of capturing specific information embedded in the training data. For instance, GMMs are more likely to capture structural information while HMMs are inclined to extract transition information in sequences. The selection of a single model alone may discard these potentially relevant information which may hinder its classification performance. By pooling the outputs from both models one can improve the performance and reliability of the method before making the classification decision. With this in mind, we use a hybrid model that exploits the sum rule that combines the results from both models. The sum rule is known to be robust to errors in the estimated posterior probabilities of the each model [15]. Given a set of weights $\{w_i; i = 1..L\}$ for L number of classifiers, and k classes, a given sound, $\mathbf{x}$ is assigned to a model M_n, if

$$\sum_{i=1}^{L} w_i P(M_n|\mathbf{x}_i) > \sum_{i=1}^{L} w_i P(M_l|\mathbf{x}_i) \quad l = 1, \cdots, k; l \neq n$$

A smaller training set, which is a subset of the training set used in training the models was used to estimate the weights in order to minimize the combined classifier error rate. The devised model is illustrated in Figure 2.

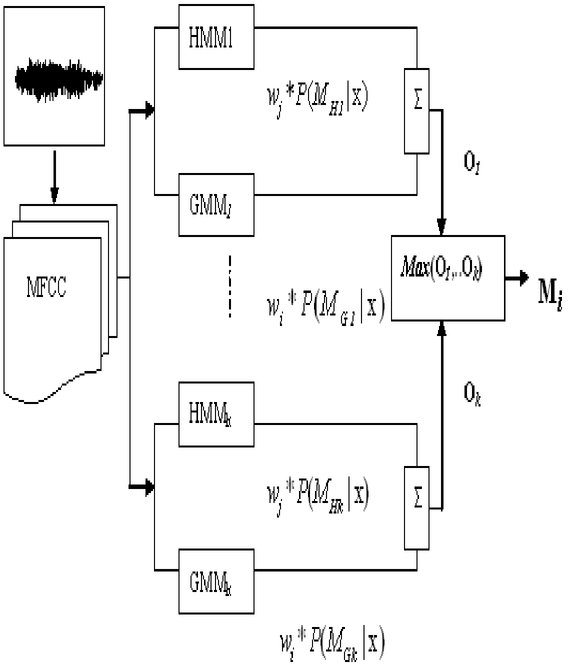

Figure 2. The hybrid generic audio classification system.

3.4. Database

For our experiments MuscleFish database [16] was used. The MuscleFish database used consists of 414 sound files belonging to $k = 16$ sound classes. Table 1 illustrates the sample distribution of sound categories in the MuscleFish database.

Class(k)	N	m	Class	N	m
Female	36	8	Violin-pizz	40	6
Male	17	6	Animal	9	24
Altrombone	13	14	Bell	7	14
Cello	47	72	Crowds	4	8
Oboe	32	28	Laughter	7	16
Percussion	102	46	Machines	11	24
Tubular-bell	20	40	Telephone	17	14
Violin-bowed	45	50	Water	7	14

Table 1. Class distribution(N) of MuscleFish Database together with the Gaussian components (m) selected using the MDL criterion for each respective training dataset.

3.5. Classification Performance

We divided the database in to 2 mutually exclusive sets; training, and test sets with the training set having twice as much data as in the test set. During the testing, each frame of the query sound is evaluated with all class models, and the class model which yields the highest likelihood is assigned as the respective class of the tested frame. Once all the frames of the query sound are evaluated, the class model, which claims the majority of the frames as its own is assigned as the *winner* model for the tested sound. Table 2 illustrates the results of modeling sounds with HMMs using varying Gaussian mixture components and states. The best performance is achieved with 2 states with 2 mixture components. Further increase of the number of states in the models deteriorates the overall performance. With GMM, the best performance was achieved with 10 Gaussian mixtures as illustrated in Table 3. The overall classification performance of the GMM exceeded the results achieved with 2-state, 2-mixture HMM indicating GMM's potential of capturing the structure of generic sounds without utilizing the transition information encoded in sequences. The individual class performances from GMM and HMM are illustrated in the Figure 3. The class *crowd* was not recognized by any of the models. This can be attributed to the fewer number of samples available in the database, which is only 4, the lowest of all the class samples. Classes such as laughter, machines, violinpiazz, female, and telephone performed better with HMM. These classes show strong repetitive behavior compared to the ones which scored high with GMMs - bells, altrombone, violinbowed that show no obvious repetitive nature. This suggests that when HMMs are used to model such non-repetitive sounds, the HMM's transition property has little or no effect on the overall classification performance. The best performance was achieved with the hybrid model as shown in Table 4.

4. Conclusions

In this paper we have modeled and classified generic sounds using a hybrid model comprised of a class-dependent variable component GMM and HMM. Overall performance was compared against class-based GMMs, GMMs with fixed number of Gaussian components across all classes and HMMs. Many of the classes in the database did not have sufficient data samples for a better representation of a class. The performance decreases as the number of states and Gaussian mixtures increased in HMMs. With GMMs, individual sound classes were modeled using class-dependent variable Gaussian components determined by the MDL principle. According to these results, Gaussian Mixture Models with varying number

Q	m	top 1	top 2	top 3
1	2	63.28	78.91	87.50
	4	61.70	77.34	87.50
	6	58.59	74.22	82.81
2	2	67.97	80.47	88.28
	4	54.69	70.31	83.38
	6	50.78	66.41	80.47
4	2	57.81	68.75	82.81
	4	47.70	64.10	75.0
	6	43.75	55.47	64.84
6	2	53.91	67.97	80.47
	4	42.97	53.91	62.50
	6	35.94	50.78	60.16

Table 2. Overall Performance(%) with varying number of Gaussian mixtures(m) and states(Q) per class using HMMs. The results are given for top 1,top 2,and top 3 retrievals.

Number of mixtures	top 1	top 2	top 3
m=4	44.53	63.89	71.09
m=8	70.31	83.59	92.19
m=10	73.44	82.81	90.62
m=12	70.31	81.25	89.84
m=16	67.97	79.69	90.62
Class-dependent	81.23	85.21	93.62

Table 3. GMM Performance(%) for top 1,top 2, top 3.

of mixtures in each respective class are better in capturing the structure of sounds and thereby modeling generic sounds when there are large perceptual variations in the training and test sets in which samples lack repetition. Similarly, sounds which are more repetitive in nature performed well with HMMs. The hybrid model outperformed all the other derived models by exploiting the strengths of individual models and evinced the suitability of such a structure for modeling generic sounds that exhibit large intra-class variations.

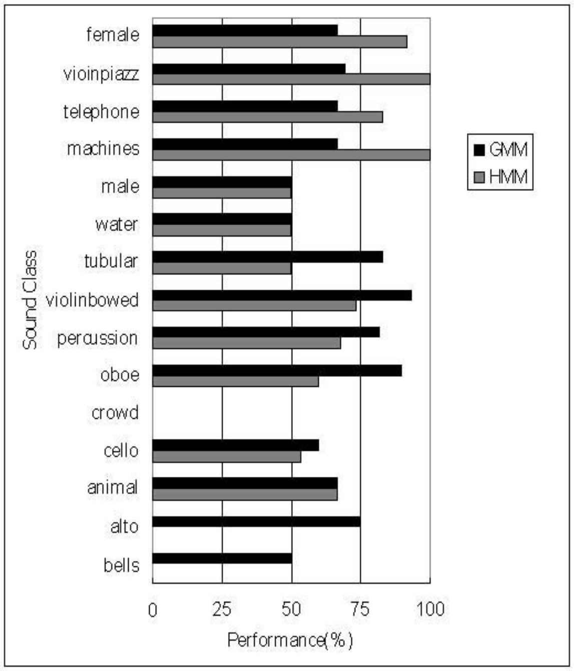

Figure 3. Individual class classification performance from GMM, and HMM

Method	top 1	top 2	top 3
HMM	67.97	80.47	88.28
GMM	73.44	82.81	90.62
GMM-Class dependent	81.23	85.21	93.62
Hybrid Model	83.14	89.06	96.88

Table 4. Overall Performance(%) Summary for top 1,top 2,top 3 retrievals.

References

[1] M. Casey. Reduced-rank spectra and minimum entropy priors for generalized sound recognition. In *Proc. Workshop on Consistent and Reliable Cues for Sound Analysis, EUROSPEECH*, 2001.

[2] G. D. Li Stan, Guo. Content-based audio classification and retrieval using svm. In *IEEE Pacific-Rim Conference on Multimedia Proccedings, Sydney, Australia*, 2000.

[3] B. Logan. Mel frequency cepstral coefficients for music modelling. In *International Symposium on Music Information Retrieval*. MusicIR2000, 2000.

[4] L. Lu and et al. Content-based audio classification and segmentation by using support vector machines. In *Multimedia Systems*, pages 482–492, 2003.

[5] T. Matsui and S. Furui. Comparison of text-independent speaker recognition method using vq-distortion and discrete/continuous hmm. In *Proc. IEEE ICASSP*, pages 157–164, 1992.

[6] K. P. Murphy. The bayes net toolbox for matlab. www.ai.mit.edu/murphyk/Software/BNT/usage.html.

[7] D. Reynolds and A. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–82, 1995.

[8] M. J. Reyse-Gomez and D. P. W. Ellis. Selection, parameter estimation and discriminative training of hmms for generic acoustic modeling. In *Proc. IEEE ICME*, 2003.

[9] J. Rissanen. Modelling by shortest data descritption. *Automatica*, 14:465–471, 1978.

[10] S. J. Roberts and D. Husmeier. Bayesian approaches to gaussian mixture modeling. *IEEE Trans. on Pattern and Machine Intelligence*, 20(11):1133–1141, November 1998.

[11] E. Scheirer and M. Slaney. Construction and evaluation of a robust multi-feature speech/music discriminator. In *Proc. IEEE ICASSP*, 1997.

[12] A. P. Schmidt and et al. Reduced-rank spectra and minimum entropy priors for generalized sound recognition. In *Music Classification and Identification System*. www.trevorstone.org/school/ MusicRecognitionDatabase.pdf.

[13] N. Z. Tishby. On the application of mixture ar hidden markov models to text independent speaker recognition. *IEEE Transactions Signal Processing*, 39:563–570, 1991.

[14] M. Walter and et al. Data driven gesture model acquisition using minimum description length. In *Proc. BMVC*, 2001.

[15] A. Webb. *Statistical Pattern Recognition*. Willy Publishing Company, 2002.

[16] E. Wold and et al. Content-based classification, search and retrieval of audio. In *Proc. IEEE Multimedia*, pages 27–36., 1996.

[17] Z. Xiong and et al. Comparing mfcc and mpeg-7 audio features for feature extraction, maximum likelihood hmm and entropic prior hmm for sports audio classification. In *Proc. IEEE ICME)*, pages 397–400, 2003.

[18] T. Zhang and C. Kuo. *Content-Based Audio Classification and Retrieval for AudioVisual Data Parsing*. Kluwer Publishing Company, 2001.