COMPARING MFCC AND MPEG-7 AUDIO FEATURES FOR FEATURE EXTRACTION, MAXIMUM LIKELIHOOD HMM AND ENTROPIC PRIOR HMM FOR SPORTS AUDIO CLASSIFICATION

Ziyou Xiong†, Regunathan Radhakrishnan‡, Ajay Divakaran‡ and Thomas S. Huang†

†Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801
‡Mitsubishi Electric Research Laboratory at Murray Hill,
Murray Hill, NJ 07974
E-mail: {zxiong, huang}@ifp.uiuc.edu, {regu, ajayd}@merl.com

ABSTRACT

We present a comparison of 6 methods for classification of sports audio. For the feature extraction we have two choices: MPEG-7 audio features and Mel-scale Frequency Cepstrum Coefficients(MFCC). For the classification we also have two choices: Maximum Likelihood Hidden Markov Models(ML-HMM) and Entropic Prior HMM(EP-HMM). EP-HMM, in turn, have two variations: with and without trimming of the model parameters. We thus have 6 possible methods, each of which corresponds to a combination. Our results show that all the combinations achieve classification accuracy of around 90% with the best and the second best being MPEG-7 features with EP-HMM and MFCC with ML-HMM.

Keywords: Sports Audio Classification, MFCC, MPEG-7 Audio Feature, HMM

1. INTRODUCTION AND RELATED WORK

Most of the audio features proposed so far have fallen into three categories: energy-based, spectrum-based and perceptual-based. Examples of the first category are 4Hz modulation energy used by Scheirer et al[1] for speech/music classification. Examples of the second category are roll-off of the spectrum, spectral flux, MFCC by Scheirer et al[1] and linear spectrum pair, band periodicity in [2]. Examples of the third category include pitch estimated by Zhang et al[3] to discriminate more classes such as songs, speech over music.

Although there are comparative studies of audio features for speech/music discrimination[4], there are few studies on this topic for general sound classification. Recently the MPEG-7 international standard has adopted the new, dimension-reduced, de-correlated spectral features[5] for general sound classification[6]. This motivates us to compare it with other widely used features.

In addition to numerous audio features, a broad spectrum of classifiers has been studied for audio classification such as Nearest Neighbor, Neural Networks, Gaussian Mixture Models(GMM), Hidden Markov Models(HMM), Nearest Feature Line, Adaboost and Support Vector Machines(SVM).

Among all the classifiers listed above, HMM have their advantage of better modelling the temporal evolution of dynamic sounds. Other forms of HMM have been studied as well, such as continuously-variable duration HMM, EP-HMM[7]. Their classification results have been reported to be better than those with ML-HMM on various "clean" databases. This motivates us to compare some of them with ML-HMM using our "noisy" sports audio database. For an explanation of "clean" and "noisy" databases, please see Section 4.

The rest of the paper is organized as follows. In Section 2 and Section 3 a brief overview is given on two different audio features and two different HMM. The experiments are described in Section 4. The experimental results and discussions are in Section 5 and Section 6.

2. MPEG-7 AUDIO FEATURES AND MFCC

We compare the MPEG-7 standardized features for sound recognition with MFCC. Although both are spectrum-based features, MPEG-7 audio features are new to the audio features family while MFCC have been widely used in speech recognition and audio classification. The extraction processes are summarized in Figure 1 and Figure 2 respectively.

The MPEG-7 features consist of dimension-reduced spectral vectors obtained using a linear transformation of a spectrogram. They are the basis projection features based on Principal Component Analysis(PCA) and an *optional* Independent Component Analysis(ICA). For each audio class, PCA is performed on the normalized log subband energy of

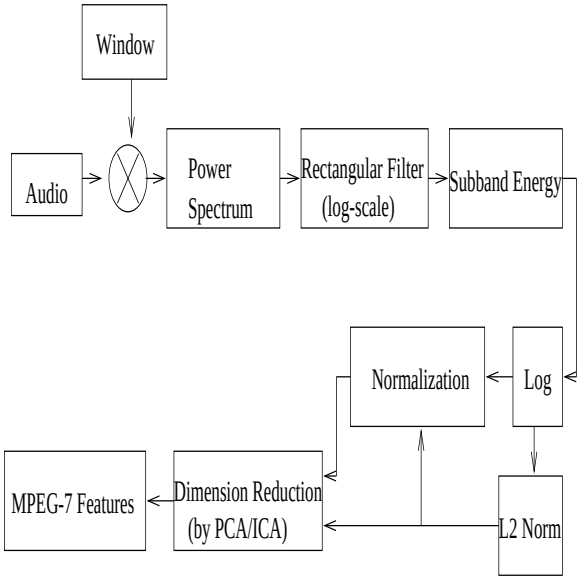

Fig. 1. Extraction Method for MPEG-7 Audio Features

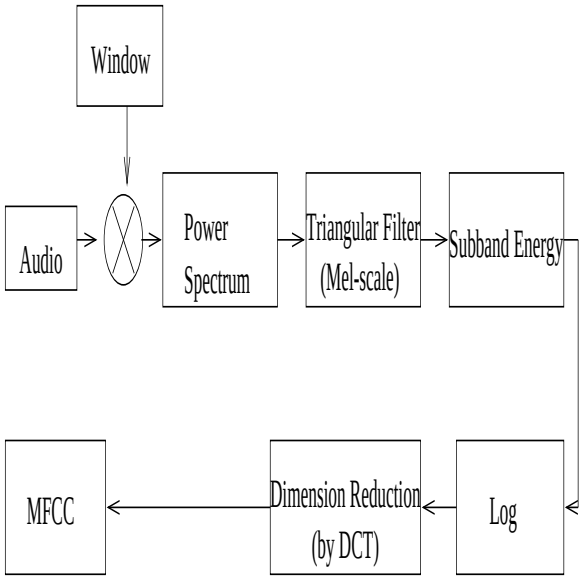

Fig. 2. Extraction Method for MFCC

all the audio frames from all the training examples in the class. The frequency bands are decided using the logarithmic scale(e.g. an octave scale).

MFCC are based on discrete cosine transform(DCT). They are defined as: $c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_k \times \cos[n(k - \frac{1}{2})\frac{\pi}{K}])$, $n = 1, \cdots, L$ where K is the number of the subbands and L is the desired length of the cepstrum. Usually $L \ll K$ for the dimension reduction purpose. $S'_k s, 0 \leq k < K$ are the filter bank energy after passing the k_{th} triangular band-pass filter. The frequency bands are decided using the Mel-frequency scale(linear scale below 1kHz and logarithmic scale above 1kHz).

Their differences are summarized as follows:

1. The Mel-frequency scale used for MFCC has been shown to be better than the logarithmic scale for speech recognition. MPEG-7 audio features use the logarithmic scale because of its simplicity.

2. MFCC of a testing audio example are the same for all the audio classes. This is because the DCT bases are the same. However, MPEG-7 audio features of the same example are different. Since each PCA space is derived from the training examples of each training class, each class has its distinct PCA space.

3. During training, the extraction of the MPEG-7 audio features requires more memory to buffer the features of all the training examples of the audio classes. During testing, the PCA projection needs to be performed for each class. The extraction of MFCC only requires buffering the features of one training example. The features are the same for different classes. Thus the extraction of the MPEG-7 audio features takes more time and memory.

3. ML-HMM[8] AND EP-HMM[7]

We compare EP-HMM with ML-HMM for classification of sports audio. Let's denote λ as the model parameters, O as the observation. The Maximum A Posteriori(MAP) test is the following: O is classified to be of class j if $P(\lambda_j|O) \geq P(\lambda_i|O), \forall i$.

When we don't have any bias towards any prior model λ_i, i.e, we assume $P(\lambda_i) = P(\lambda_j), \forall i, j$, the MAP test is equivalent to the ML test: O is classified to be of class j if $P(O|\lambda_j) \geq P(O|\lambda_i), \forall i$ due to the Bayes rule: $P(\lambda|O) = \frac{P(O|\lambda)P(\lambda)}{P(O)}$

However, if we assume the following biased probabilistic model $P(\lambda|O) = \frac{P(O|\lambda)P_e(\lambda)}{P(O)}$, where $P_e(\lambda) = e^{-H(P(\lambda))}$ and H denotes entropy, i.e, the smaller the entropy, the more likely the parameter, then we must use the MAP test and compare $\frac{P(O|\lambda_i)e^{-H(P(\lambda_i))}}{P(O|\lambda_j)e^{-H(P(\lambda_j))}}$ with 1 to see whether O should be classified to be of class i or j.

EP-HMM have been shown to improve the classification accuracy over ML-HMM on melody, text[7] and general sound classification[6]. Moreover, in EP-HMM, it is possible to trim the parameters of their graphical structure, thus obtaining more compact graphical models. For further details on EP-HMM, please see [7].

4. SPORTS AUDIO CLASSIFICATION

4.1. Data Set

Unlike the data-set used in [9], our database of sports audio is not composed of relatively clean audio such as CD recordings and TIMIT database. It is from broadcast TV which is an un-controlled audio environment with high background interference that makes the audio classification more difficult.

We've collected 814 audio clips from TV broadcasting of golf, baseball and soccer games. Each of them is hand-labelled into one of the six classes as ground truth: applause, ball-hit, cheering, music, speech, speech with music. Their corresponding numbers of clips are 105, 135, 82, 185, 168, 139. Their duration differs from around 0.5 seconds(for ball hit) to more than 10 seconds(for music segments. The total duration is approximately 1 hour and 12 minutes. The database is partitioned into a 90%/10% training/testing set.

4.2. Feature Extraction

In our feature extraction, an audio signal is divided into overlapping frames of duration 30ms with 10ms overlapping for a pair of consecutive frames. Each frame is multiplied by a hamming-window function.

MFCCs are calculated from 40 subbands(17 linear bands between 62.5Hz and 1kHz, 23 logarithmic bands between 1kHz and 8kHz). 10th-order MFCCs are used as audio features. That is, $L = 10$ and $K = 40$.

The lower and upper boundary of the frequency bands for MPEG-7 features are also 62.5Hz and 8kHz that are over a spectrum of 7 octaves. Each subband spans a quarter of an octave so there are 28 subbands in between. Those frequencies that are below 62.5Hz are group into 1 extra subband. After normalization of the 29 log subband energy, a 30-element vector represents the frame. This vector is then projected onto the first 10 principal components of the PCA space of every class. Notice 10 principal components are used so that the number of MPEG-7 features is the same as that of the MFCC features.

5. EXPERIMENTAL RESULTS

The number of states for both HMM is selected to be 10 initially. The distribution of the observations is modelled as a single component multi-variate Gaussian. In order to compare the two HMM fairly, we first compare ML-HMM with EP-HMM without trimming states or parameters. We then compare ML-HMM with EP-HMM with trimming.

The results on classification accuracy performed on the 10-fold cross-validation data set are organized into Table 1 for a selected combination of features with classifiers. Because of the limited space, we omit those for the other 5 combinations. The average recognition rates for the 6 methods are summarized into Table 2. The following observations can be made based on these tables:

1. For our sports audio database, the best combination, on the average, is MPEG-7 features with EP-HMM with trimming of states and model parameters. The improvement of classification accuracy from Combination 2 to Combination 3 is solely due to the trimming of states and model parameters, especially so for the "ball-hit" class. We found that the most of the trimming was done for this class of short-duration impulse-like signals. The number of states needed was smaller than 10 and many of the state transitions were trimmed. With a more compact model, there were more training data per state per parameter to converge more closely to the global maximum.

2. Either with MFCC or MPEG-7 audio features, trimming of states and parameters for EP-HMM improves classification accuracy. This can be observed from the improvement from Combination 2 to Combination 3 and from Combination 5 to Combination 6.

3. Using ML-HMM as classifier, MFCC out-perform MPEG-7 audio features. This can be seen from Combination 4 and Combination 1. However, when EP-HMM is used as the classifier, either with or without trimming of state or parameters, the performance of MFCC drops by as much as 5% when compared with MPEG-7 features. The observation is from Combination 2, 3, 5 and 6.

4. For the 6 combinations, all of them perform well and are comparable in performance. No method seems to enjoy a significant advantage over the others. The choice of a particular combination would be governed by the computational complexity and memory requirement of the application.

6. DISCUSSION

1. Casey[6] shows that MPEG-7 features with EP-HMM with trimming, which is also the best combination for our database, yielded significant better results(on average 6.5%) than MPEG-7 features with ML-HMM for a database of 1000 sounds ranging over 20 audio classes. Our gain is smaller(about 4.0%), possibly due to the noisy nature of our database.

2. The small marginal advantage of MPEG-7 features over MFCC and the result that MFCC with ML-HMM yielded better results than the ones in Table 1, 2, 5, 6 is a bit surprising to us. Based on the results in [6][7]

	[1]	[2]	[3]	[4]	[5]	[6]
[1]	1.00	0	0	0	0	0
[2]	0	0.923	0	0	0.077	0
[3]	0.125	0	0.875	0	0	0
[4]	0	0	0	0.944	0.056	0
[5]	0	0	0	0	0.941	0.059
[6]	0	0	0	0	0	1
	Average Recognition Rate: 94.728%					

Table 1. Recognition Matrix(or Confusion Matrix) on a 90%/10% training/testing split of a data set composed of 6 classes. [1]: Applause; [2]: Ball-Hit; [3]: Cheering; [4] Music; [5] Speech; [6] Speech with Music. The results here are based on MPEG-7 Audio Features and EP-HMM with trimming of states and parameters.

#	Combination	Accuracy Rate
1	MPEG-7+ML-HMM	90.974%
2	MPEG-7+EP-HMM No Trimming	90.974%
3	MPEG-7+EP-HMM+Trimming	94.728%
4	MFCC+ML-HMM	94.604%
5	MFCC+EP-HMM No Trimming	88.427%
6	MFCC+EP-HMM+Trimming	89.353%

Table 2. Comparison of the average recognition rates for the 6 methods.

we expected that the MPEG-7 features would beat MFCC and EP-HMM would beat ML-HMM. One factor we may have missed is that our audio signals' sampling rate has been set to be 16kHz, not 32kHz or 44.1kHz, hence the MPEG-7 features have not been able to capture the spectral information that is outside the speech spectrum range. We need further research to get a definitive comparison.

7. CONCLUSIONS AND FUTURE DIRECTIONS

We present a comparison of 6 methods for classification of sports audio. Our results show that all the combinations achieve classification accuracy of around 90%. They are comparable in performance with the best and the second best being MPEG-7 features with EP-HMM and MFCC with ML-HMM.

Possible further directions include increasing the size of our database by both increasing the number of samples per class as well as the number of classes in order to derive more robust audio models, studying the effect of the audio sampling rate on the classification results for both MFCC and MPEG-7 audio features and applying our findings to audio highlights extraction.

8. ACKNOWLEDGEMENT

We appreciate the help from Dr. Michael Casey and Dr. Matthew Brand of MERL Cambridge Research Lab and Dr. Vladimir Pavlovic and Dr. Malcolm Slaney.

9. REFERENCES

[1] E. Scheirer and S. Malcolm, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP-97*, April 1997, Munich, Germany.

[2] L. Lu, S. Li, and H.J. Zhang, "Content-based audio segmentation using support vector machines," *Proceeding of ICME 2001*, pp. 956 – 959, 2001, Tokyo, Japan.

[3] T. Zhang and C. Kuo, "Content-based classification and retrieval of audio," *Proceeding of the SPIE 43rd Annual Conference on Advanced Signal Processing Algorithms, Architectures and Implemenations*, vol. VIII, 1998, San Diego, CA.

[4] M.J. Carey, E.S. Parris, and H.L. Thomas, "A comparison of features for speech, music discrimination," *Proceedings of ICASSP 1999*, pp. 149–152, 1999.

[5] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737–747, June 2001.

[6] M. Casey, "Reduced-rank spectra and entropic priors as consistent and reliable cues for general sound recognition," *Proceeding of the Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis*, 2001.

[7] M. Brand, "Structure discovery in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1183, 1999.

[8] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–86, February 1989.

[9] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio," *IEEE Multimedia Magazine*, vol. 3, no. 3, pp. 27–36, 1996.