Performance of MPEG-7 spectral basis representations for retrieval of home video abstract

Hyoung-Gook Kim, Thomas Sikora
Communication Systems Group, Technical University of Berlin
{kim, sikora}@nue.tu-berlin.de

ABSTRACT

In this paper, we present a classification and retrieval technique targeted for retrieval of home video abstract using dimension-reduced, decorrelated spectral features of audio content. The feature extraction based on MPEG-7 descriptors consists of three main stages: Normalized Audio Spectrum Envelope (NASE), basis decomposition algorithm and basis projection, obtained by multiplying the NASE with a set of extracted basis functions. A classifier based on continuous hidden Markov models is applied. For retrieval with accurate performance the system consists of a two-level hierarchy method using speech recognition and sound classification. For the measure of the performance we compare the classification results of MPEG-7 standardized features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC). Results show that the MFCC features yield better performance compared to MPEG-7 features.

Keywords: MPEG-7, Normalized Audio Spectrum Envelope (NASE), basis decomposition algorithm, Mel-scale Frequency Cepstrum Coefficients (MFCC)

1. INTRODUCTION

With the increasing number of people who can afford to make videos to record their lives, home videos play a more and more important role in multimedia.

Audio contents are very important clues for the retrieval of home videos, because different sounds can indicate different important events. In most cases it is easier to detect most important events and appealing things using audio features than using video features. For example, when interesting event occurs, people are likely to talk or laugh or cry out. So these events can be detected by audio content easily, while it is very difficult or even impossible using visual content.
For these reasons, effective video retrieval techniques using audio features are investigated by many researchers in the literature [1][2][3][4].

The standard MPEG-7 [5], formally named “Multimedia Content Description Interface”, focuses on describing the content for indexing, searching, and retrieval of audiovisual information. MPEG-7 audio standard group [6] [7] adapted a feature extraction method based on the projection of a spectrum onto a low-dimensional subspace via reduced-rank spectral basis functions. The dimension-reduced decorrelated features, called Audio Spectrum Projection (ASP), are used to train hidden Markov models (HMM) [8] in order to apply uniformly to diverse source classification tasks with accurate performance.

In [9][10], we applied MPEG-7 ASP features to sound classification using three different taxonomy methods: a direct approach, a hierarchical approach without hints and the hierarchical approach with hints. In the direct classification scheme, only one decision step is taken to classify the input audio into one of the various classes of the taxonomy. For the hierarchical approach the taxonomy consists of a number of sound categories into a hierarchical tree. Each of these classes is broken down further into more detailed classes. In hint experiments, each sound clip is assigned a hint, so that only one decision per clip needed to be made by the sound recognition program.

Results indicate that the recognition rate of the hierarchical approach without hints is lower compared to the direct approach. Many of the errors were due to problems with recognition in the highest layer that sound samples in different branches of the tree were too similar. For example, some bird sounds and horn sounds were difficult to tell apart with the human ear. Thus, the hierarchical structure for sound recognition does not necessarily improve recognition rates if sounds in different general classes are too similar unless some sort of additional information (e.g., a hint) is available. Thus the hierarchical classification with hints yields overall the highest recognition rate compared to one level structure or hierarchical classification without hints.

Effective video retrieval can be achieved using speech recognition taken from titles of the hierarchical approach. In order to achieve both the accuracy and flexibility of the system, a two-level hierarchy method using speech recognition and audio classification is used. The speech recognition is applied to selecting the subtitles as hints of the hierarchical classification.

In this paper, we focus on classification and retrieval technique targeted for retrieval of home video abstract using MPEG-7 standard ASP features. For the measure of the performance we compare the classification results of the MPEG-7 standardized features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC).

2. SYSTEM OVERVIEW

For retrieval of home video abstracts the system consists of a two-level hierarchy method using speech recognition and sound classification. Figure 1 depicts the block diagram of the system.

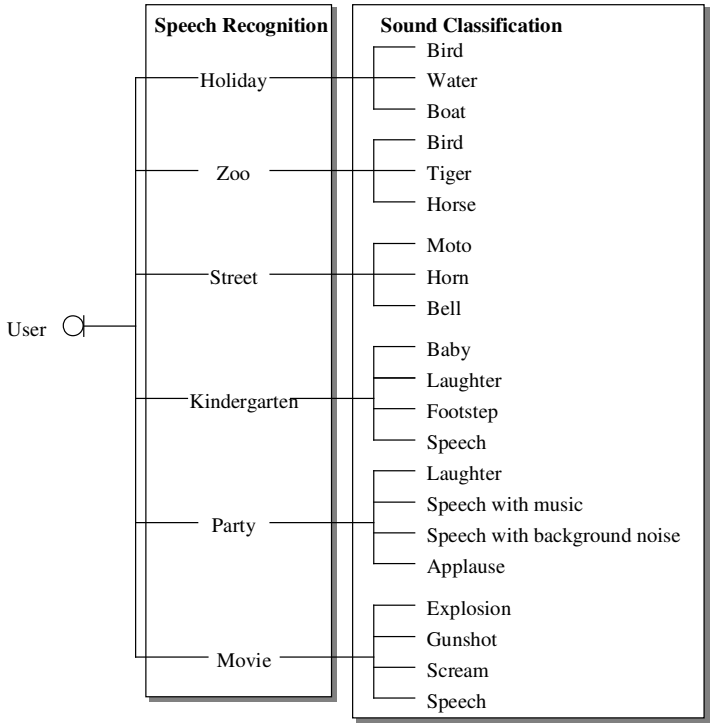

Figure 1: Block diagram of a two-level hierarchy method using speech recognition and sound classification

At the first/top level, a spoken annotation is recorded for each picture of the database. A speech recogniser extracts a spoken content descriptor from each spoken annotation. Each abstracted home video is annotated with one word of the five keyword lexicon: holiday, zoo, kindergarten, movie, party and street. Each keyword of the description vocabulary is modelled by concatenating the phone models. By uttering keywords, the user can automatically retrieve the corresponding pictures.

In the second/bottom level, each picture of the database includes it's own sound. Audio segments of keyword-category "holiday" is classified into bird, water, boat. Category "zoo" is classified into audio segments of bird, tiger, horse. Category "street" is classified into audio segments of motor, horn, bell. Category "kindergarten" is classified into audio segments of baby, laughter, footstep, speech. Category "movie" is classified into audio segments of explosion, gun, scream, speech. Category "party" is classified into audio segments of laughter, speech with music or background noise and applause.

Sounds are modelled according to category labels and presented in a set of model parameters. The sound retrieval is achieved based on sound classification. Given a selected query sound, the extracted sound features are used to run the sound classifier, which compares the pre-indexed sounds in the sound database with the audio query and outputs the classification results.

For the feature extraction in sound classification we use MPEG-7 basis projection features, while Mel-scale Frequency Cepstrum Coefficients (MFCC) are used in speech recognition. A continuous hidden Markov model (HMM) is used as the classifier for speech recognition and sound classification.

2.1 MPEG-7 Sound Classification System

The sound classification task is further divided into three steps: audio feature extraction, training of sound models, and decoding. Figure 2 depicts the procedure using a sound recognition classifier.

The audio feature extraction module extracts audio information from a given audio sound. The MPEG-7 ASP feature extraction mainly consists of a Normalized Audio Spectrum Envelope (NASE), basis decomposition algorithm –such as Singular Value Decomposition (SVD) or Independent Component Analysis (ICA)– and a spectrum basis projection, obtained by multiplying the NASE with a set of extracted basis functions:

- *Normalized Audio Spectrum Envelope* (NASE)
  The observed audio signal is divided into overlapping frames by the application of a hamming window function and analyzed using the short-time Fourier transform (STFT). To extract reduced-rank spectral features, the spectral coefficients are grouped in logarithmic sub-bands. The output of the logarithmic frequency range is the sum of the power spectrum in each logarithmic sub-band. The spectrum according to a logarithmic frequency scale, which the MPEG-7 standard refers to as Audio Spectrum Envelope (ASE), consists of one coefficient representing power between 0 Hz and "low edge", a series of coefficients representing power in logarithmically spaced bands between "low edge" and "high edge", and a coefficient representing power above "high edge". The resulting log-frequency power spectrum is converted to the decibel scale. Each decibel-scale spectral vector is normalized with the RMS (root mean square) energy envelope, thus yielding a normalized log-power version of the ASE called NASE. The full-rank features for each frame consist of both the RMS-norm gain value and the NASE vector.

- *Basis Decomposition*
  For the basis decomposition step, we combined a basis dimension-reduction by Principal Component Analysis (PCA) [11] algorithm with a basis information maximization by FastICA [12].
  a. Principal Component Analysis (PCA)
     In order to achieve a trade-off between further dimensionality reduction and information loss, PCA de-correlates the second order moments corresponding to low frequency properties and extracts orthogonal principal components of variations.
  b. Independent Component Analysis (ICA)
     ICA not only decorrelates the second order statistics but also reduces higher-order statistical dependencies. After extracting the reduced PCA basis, a further step consisting of basis rotation in the directions of maximal statistical independence is required for applications that require maximum decorrelation of features. The NASE is then fed to the FastICA algorithm, which maximizes the information. The basis function obtained by PCA and ICA is stored in the MPEG-7 basis function database for the classification scheme.

- *Basis Projection*
  The resulting spectrum projection is the product of the NASE matrix, the dimension-reduced PCA basis functions, and the ICA transformation matrix.

This spectrum projection is used to represent low-dimensional features of a spectrum after projection onto a reduced rank basis. The spectrum projection features and RMS-norm gain values are used as input to the HMM training module.

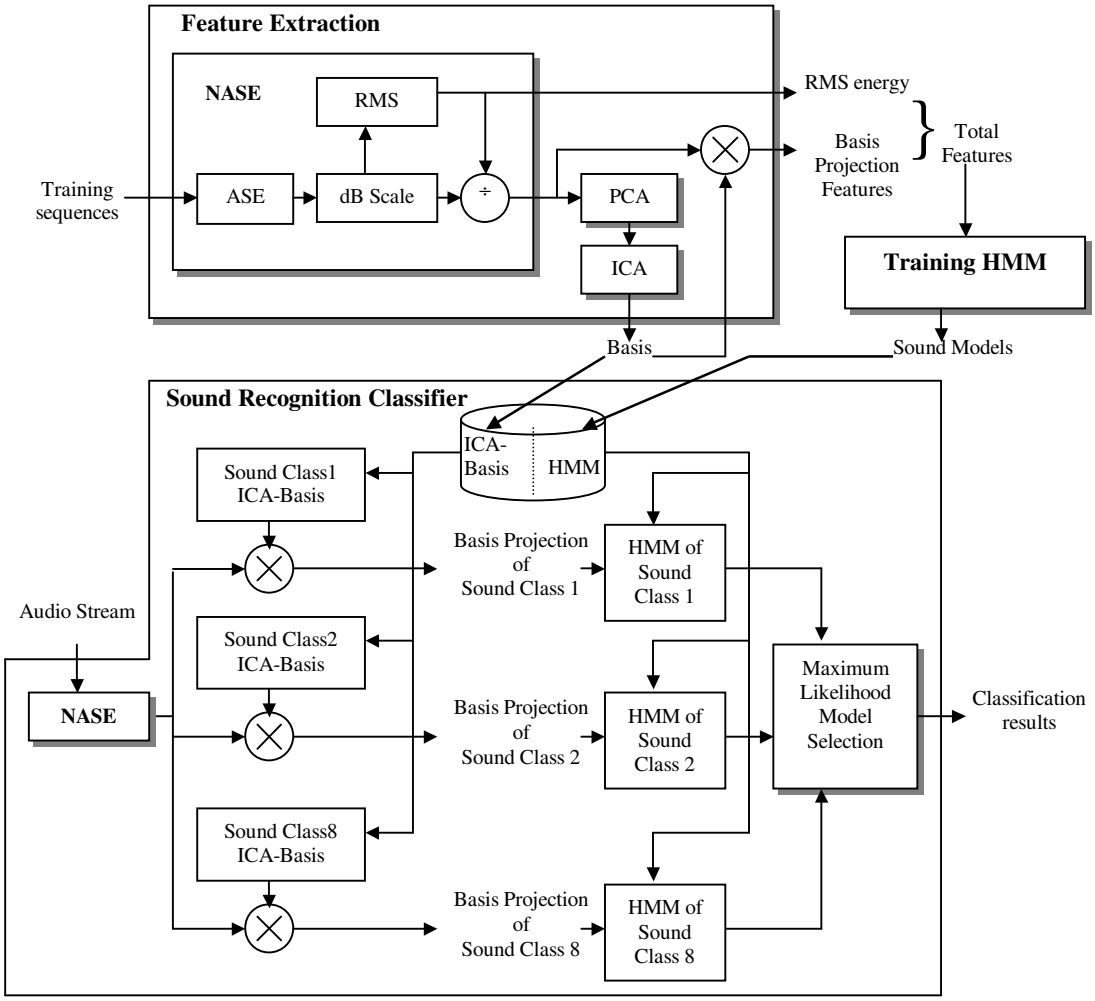

Figure 2: Block diagram of sound classification

For each pre-defined sound class of home video abstract, the training module builds a model from a set of training sounds using hidden Markov models (HMM) and continuous parametric distributions. Ergodic HMM parameters with 7 states are then obtained using a maximum likelihood estimation procedure known as the Baum-Welch algorithm. The procedure starts with random initial values for all of the model parameters and optimizes them by iterative re-estimation. Each iteration runs through the entire set of training data in a process that is repeated until the model parameters converge to satisfactory values. When the training process is complete, the statistical basis and HMM model of each sound class are stored in the sound model database of the sound recognition classifier.

At the end of the decoding process, the HMM with the highest maximum likelihood score is selected as the representative class for the sound. The algorithm also generates the optimal HMM state path for each model given the input sound. The state path corresponding to the most likely class is stored in the MPEG-7 sound database and used as an index for further query applications.

The audio retrieval is based on audio classification results. Figure 2 depicts the recognition module used to classify an audio input based on pre-trained sound class models (HMMs). Sounds are read from a media source format, such as WAV files. Given an input sound, the NASE features are extracted and projected against each individual sound model's set of basis functions, producing a low-dimensional feature representation. Then, the Viterbi algorithm is applied to align each projection on its corresponding sound class HMM (each HMM has its own representation space). The HMM yielding the best maximum likelihood score is selected.

2.2 Sound Retrieval Using Histogram State Technique

The hidden Markov models (HMMs) consist of several states. During training, the parameters for each state of a speaker's model are estimated by analyzing the feature vectors of the training set. Each state represents a similar behaving portion of an observable symbol sequence process. At each instant in time, the observable symbol in each sequence either stays at the same state or moves to another state depending on a set of state transition probabilities associated with the states. Different state transitions may be more important for modeling different kinds of data. In this paper, we use the ergodic HMM topology, where each state can be reached from any other state and can be revisited after leaving.

An input sound is indexed by selecting the HMM yielding the maximum likelihood score and storing the corresponding optimal HMM state path, which was obtained using the Viterbi algorithm. This state path describes the evolution of a sound through time with a sequence of integer state indices.

The MPEG-7 standard proposes a method for computing the similarity between two state paths generated by the Viterbi algorithm. This method, based on sum-of-squared-differences between "state path histograms [6][7]", is explained in the following.
A normalized histogram can be generated from the state path obtained at the end of the classification procedure. Frequencies are normalized to values in the range [0-1] obtained by dividing the number of samples associated with each state of the HMM by the total number of samples in the state sequence:

$$hist_a(j) = \frac{N(j)}{\sum_{i=1}^{K} N(i)}, \qquad 1 \le j \le K \qquad (1)$$

K is the number of states in the HMM and $N(j)$ is the number of samples for state j for the given state path.

A similarity measure between two state paths a and b is computed as the absolute difference between each relative frequencies summed over state indices k ($1 \le k \le K$). This gives the Euclidian distance between the two sounds indexed by a and b:

$$\delta(a,b) = \sum_{j=1}^{k} \sqrt{(hist_a(j) - hist_b(j))^2} \qquad (2)$$

3. EXPERIMENTAL PROCEDURE AND RESULTS

3.1 Database

We collected 180 home video abstracts. Each of them is hand-labeled into one of 18 audio classes which includes a controlled audio environment with very low background interference. Their duration differs from around 7 seconds to more than 12 seconds long. The recorded audio signal of home video abstracts is sampled at a 22.05 kHz/16 bit PCM format which was used throughout all our experiments.
The video part of the signal was not used for the experiments described in this paper.

3.2 Feature Extraction

The input audio signal is divided into overlapping frames of duration 30ms with 15ms overlapping for consecutive frames. Each frame was windowed using a Hamming window function. To extract Audio Spectrum Envelope (ASE) features, the observed audio signal is analyzed using a 512-point FFT. The power spectral coefficients are grouped in logarithmic sub-bands spaced in non-overlapping 7-octave bands spanning between low boundary (62.5 Hz) and high boundary (8 kHz). The resulting 30-dimensional ASE is converted to the decibel scale. Each decibel-scale spectral vector is normalized with the RMS (root mean square) energy envelope, thus yielding a NASE. For each audio class, the spectral basis is extracted by computing the PCA for dimension reduction and FastICA for information maximization. The recognition task of interest is to test the reduced dimension of the PCA with 7, 13, and 23.

MFCCs are based on a short-term spectrum, where Fourier basis audio signals are decomposed into a superposition of a finite number of sinusoids. The power spectrum bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. Then the spectrum is segmented into 23-critical bands by means of a filter bank that typically consists of overlapping triangular filters. Finally, a discrete cosine transform applied to the logarithm of the filter bank outputs results in vectors of decorrelated MFCC features.

As feature parameters for speech recognition, the 13^{th} order MFCC coefficients plus delta and acceleration calculations were used. In order to compare the performance vs. MFCC approach for sound classification we used only MFCC coefficients without delta and acceleration calculations.

3.3 Experimental Results

For the speech recognition in the first level, the recognition rate is always very good because we use only 6 keywords. Therefore, our purpose in this paper was to evaluate the MPEG-7 ASP for sound classification performance vs. MFCC. Additionally, we performed experiments with different feature dimensions of the different feature extraction methods.

The results of sound classification are shown in Table 1. Results show that the sound classification system achieves high recognition rate, because three or four sound classes in one of 8 categories are tested.

In the averaged recognition rate, the MPEG-7 ASP based on ICA basis yields better performance than ASP based on PCA basis. However, the recognition rates using MPEG-7 conform ASP results appear to be significantly lower than the recognition rate of MFCC. Overall MFCC achieves best recognition rate.

The category "Movie" provides best recognition rate compared to other categories. The movie contains no background interference that makes the audio classification more difficult. On the other hand, the category "Party" yields lowest recognition rate vs. other categories. The reason is that the speech with background noise and the speech with music were too similar, it was difficult to be recognized, as two separate classes.

Feature Dimension	Feature Extraction	Holiday	Zoo	Street	Kinder-garten	Movie	Party	Average
7	PCA-ASP	92.5	95.5	92.1	91.3	96.5	75.1	90.05
	ICA-ASP	91.3	96.2	90.7	90.5	96.9	82.3	91.32
	MFCC	97.08	97.6	95.3	96.3	97.6	94	96.31
13	PCA-ASP	96.3	97.6	95.7	95.8	98	82.4	94.3
	ICA-ASP	97.9	94.3	96.6	96.6	98.7	93.9	96.33
	MFCC	100	99	96.6	99	100	90.1	97.45
23	PCA-ASP	100	98.8	98.5	98.5	100	88.2	97.33
	ICA-ASP	99	99.4	97.8	99	100	94	98.2
	MFCC	100	100	99	100	100	93.4	98.73
#	Average	97.12	97.6	95.81	96.28	98.63	88.15	95.56

Table 1: Comparison of sound classification accuracies (%)

3.4 Sound Retrieval Results

Once an input sound *a* was recognized as a sound of class *Cl*, the state paths of the sounds *b* in the MPEG-7 database (which belong to class *C*l) were compared to the state path of *a* using the Euclidean distance $\delta(a,b)$ as described in section 2.2. These sounds were sorted so that those corresponding to the smallest distances were at the top of the list. That is, the items which are the most similar to the query appeared at the top of the list and the most dissimilar ones at the bottom.
Such system basically acts as a search engine for similar sounds within a given sound class. In the Table 2, gunshot_15 was input as a test sound *a* and recognized as *Cl=gunshot*. The list of the retrieved items indexed with *gunshot*, sorted by similarity with query gunshot_15, is depicted in Table 2.

Similar Sound	Maximum Likelihood Score	Euclidean Distance
gunshot 27	27.8624	0.111023
gunshot 31	28.5342	0.111532
gunshot 46	28.4523	0.115978
gunshot 33	16.5835	0.138513
gunshot 41	29. 5342	0.162056
gunshot 50	26.3256	0.167023

Table 2: Results of sounds similarity

The maximum likelihood scores used for classification are also in Table 2. Note that calculating the similarity by comparing the state paths and by comparing the maximum likelihood scores produce different results.
If, however, the sound *a* had been incorrectly recognized as something else, such as *explosion*, we would have searched in the *explosion* class for similar training data and found irrelevant results. Thus, it would have been better to have also searched the second and/or third most likely classes while hoping that *gunshot* were one of these.

We used training data from the three classes with the highest maximum likelihood scores to produce a list sorted by the Euclidean distance between the query *a* and each of the retrieved items. This method seemed promising but often produced inconsistent lists including data from different classes mixed up with each other. We decided to compare the reliability of our retrieval technique using different methods for recognition.

To compare lists of similar items, we used our own measure called *consistency*. A list is consistent when the elements next to each other belong to the same class, and a list is inconsistent when any two adjacent elements always belong to different classes. We used the following method to calculate the consistency *C* of a retrieval method:

M sound clips are tested to produce M lists l_m of similar sounds, such that $1 \le m \le M$. Let L_m be the length of the list l_m, and let N_m be the number of times that two adjacent entries in the list l_m belong to the same class. Compute the consistency C according to:

$$C_m = \frac{N_m}{L_m - 1} \qquad (3)$$

$$C = E\{C_M\} \cong \frac{1}{M} \sum_{m=1}^{M} C_m \qquad (4)$$

Thus, the consistency is a real number between 0 and 1, with 0 is as inconsistent as possible and 1 as consistent as possible.

Using the same library of test sounds, we then measured the inconsistency for retrieval methods using ASP projections based on PCA basis, and ASP projections based on FastICA basis as input to the hidden Markov models. Results are depicted in Table 3.

Method	With the state paths	With maximum likelihood score
PCA-ASP	0.63	0.44
FastICA-ASP	0.66	0.51
MFCC	0.65	0.50

Table 3: Results of sounds similarity

The results show that the lists of similar sounds were more consistent if we used the state paths instead of the maximum likelihood scores for comparison. We attribute this result to the fact that the state paths contain more information because they are multi-dimensional, whereas the maximum likelihood scores are one-dimensional. Thus, our best technique for retrieving similar sounds is the FastICA method using the state paths for comparison

3.5 Graphical User Interface of the System

Figures 3, 4, 5 and 6 show the graphical interface of the system: Each home video abstract includes two MPEG-7 descriptors: spoken content descriptor and sound descriptor as shown in Figure 3.

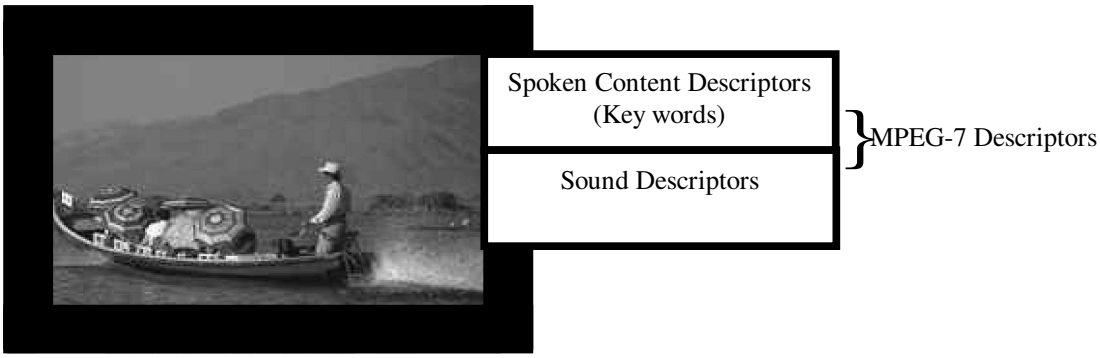

Figure 3: The home video abstract including spoken content and sound descriptors

Figure 4 illustrates the global view including all home video abstracts. If the user is looking for home video abstract of a holiday video, he starts with the global view in Figure 4 and refines the search by speech recognition. Figure 5 illustrates the result of a query by speech recognition. If the user is looking for home videos from category “holiday”, query sounds are read from a media source format, such as a WAV file, and projected into a separate classification space for each model. They are then passed through a Viterbi decoder that yields a sound class and the state path resulting from the alignment on the corresponding HMM. The matching module compares the state paths indexing sounds in the MPEG-7 sound database with the state path resulting from the decoding of the audio query. The matching results are presented within the class assigned to the query sound. Figure 6 depicts the results of an audio query belonging to the “water” class.

Figure 4: Global view of home video abstract database.

Figure 5: The results of a speech query “holiday” by speech recognition.

Figure 6: The results of a sound query “water” by sound recognition.

4. CONCLUSIONS

In this paper we presented a two-level hierarchy method using speech recognition and sound classification. Different feature dimensions of the different feature extraction methods for sound classification were compared on home video abstract test sets. Our results show that the MFCC features yield better performance compared to MPEG-7 ASP in the sound recognition. In the case of MFCC, the process of segmentation is simple and fast because there are no bases used. The extraction of the MPEG-7 ASP is also more time and memory consuming compared to MFCC.

The sound retrieval system relies on the performance of the classifier. Our retrieval system is based on a distance metric allowing to compare two state paths in conformance with the audio part of the MPEG-7 standard. The approach proposed in this paper consists in retaining the 3 most likely sound class hypotheses, in order to cope with possible recognition errors. Even in case of a classification error, most of the retrieved sounds may thus be part of the correct class. We also proposed a consistency measure to evaluate the homogeneity of such mixed result lists. Some tests reveal that ICA-ASP based classification yield more consistent retrieval results using MPEG-7 conform state path descriptors rather than HMM acoustic scores.

Our further research will include increasing the size of our database. We will focus on robust sound classification using home videos from an uncontrolled audio environment with high background interference.

5. REFERENCES

[1] E. M. Bakker, M. S. Lew "Semantic video retrieval using audio analysis," Proceedings of CIVR 2002, , pp. 271-277, July, 2002.

[2] S. Srinivasan, D. Petkovic, D. Ponceleon, "Towards robust features for classifying audio in the CueVideo system," Proceedings of 7th ACM International Conference on Multimedia, pp. 393-400, 1999.

[3] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using audio and visual Information," IEEE Signal Processing Magazine (invited paper), Nov. 2000.

[4] Z. Xiong, R. Radhakrishnan, A. Divakaran, T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," Proceedings of ICASSP 2003, Vol. 5, pp. 632-635, April 2003.

[5] B. S. Manjunath, P. Salembier, and T. Sikora, "Introduction to MPEG-7," Wiley, April, 2002.

[6] M. Casey, "MPEG-7 sound-recognition tools," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11. No.6, June 2001.

[7] ISO/IEC JTC 1/SC 29, "Information Technology Multimedia Content Description Interface-Part 4 : Audio," June, 2001.

[8] L. Rabiner, B.-H. Huang, "Fundamentals of speech recognition," Prentice Hall, N.J., 1993.

[9] H.-G. Kim, N. Moreau. T. Sikora, "Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation," submitted to ICASSP 2003, 2003.

[10] H.-G. Kim, N. Moreau. T. Sikora, "Audio classification based on MPEG-7 spectral basis representations," submitted to Special Issue of IEEE Transactions on Circuits and Systems for Video Technology, 2004.

[11] I. T. Jolliffe, "Principal components analysis," Springer-Verlag, 1986.

[12] A. Hyvarinen, E. Oja "Independent component analysis: algorithms and application," Neural Networks 13, pp. 411-430, 2000.