# COMPARISON OF MPEG-7 AUDIO SPECTRUM PROJECTION FEATURES AND MFCC APPLIED TO SPEAKER RECOGNITION, SOUND CLASSIFICATION AND AUDIO SEGMENTATION

*Hyoung-Gook Kim, Thomas Sikora*

Communication Systems Group, Technical University of Berlin
{kim, sikora}@nue.tu-berlin.de

## ABSTRACT

Our purpose is to evaluate the MPEG-7 Audio Spectrum Projection (ASP) features for general sound recognition performance vs. well established MFCC. The recognition tasks of interest are speaker recognition, sound classification, and segmentation of audio using sound/speaker identification. For the sound classification we use three approaches: the direct approach, the hierarchical approach without hints, and the hierarchical approach with hints. For audio segmentation the MPEG-7 ASP features and MFCCs are used to train hidden Markov models (HMM) for individual speakers and sounds. The trained sound/speaker models are then used to segment conversational speech involving a given subset of people in panel discussion television programs. Results show that MFCC approach yields sound/speaker recognition rate superior to MPEG-7 implementations.

## 1. INTRODUCTION

Our challenge is to analyze/classify video sound track content for indexing purposes. To this end we compared the performance of MPEG-7 standard implementations vs. MFCC approach.

The MPEG-7 [1] standard, formally named "Multimedia Content Description Interface", focuses on describing the content for indexing, and retrieval of digital sounds, images and video. For sound classification the MPEG-7 audio standard group [2][3] has adopted a feature extraction method based on the projection of a spectrum onto a low-dimensional representation using decorrelated basis functions.

A comparison between MPEG-7 Audio Spectrum Projection (ASP) based on Principal Component Analysis (PCA) basis and MFCC has been performed in [4] for sports audio classification with 6 sound classes. For the classification Maximum Likelihood hidden Markov models (ML-HMM) and Entropic Prior HMM (EP-HMM) are used. Results indicate that they are comparable in performance with the best and the second best being MPEG-7 features with EP-HMM and MFCC with ML-HMM. In [5], we implemented and analyzed the MPEG-7 ASP features for the purpose of a speaker recognition system.

In this paper we focus on MPEG-7 ASP vs. MFCC speaker recognition, sound classification and audio segmentation.

## 2. MPEG-7 AUDIO SPECTRUM PROJECTION (ASP) FEATURES AND MFCC

In [3][4][5], the MPEG-7 ASP feature extraction is very well described. 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. For the basis decomposition step, we combined a basis dimension-reduction by PCA algorithm with a basis information maximization by FastICA [6].

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 normalized log-power version of the ASE called NASE. For each audio class, the spectral basis is extracted by computing the PCA for dimension reduction and FastICA for information maximization. The resulting spectrum projection is the product of the NASE matrix, the dimension-reduced PCA basis functions and the FastICA transformation matrix. The spectrum projection features and RMS-norm gain values are input to the HMM pattern classification.

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.

## 3. SPEAKER RECOGNITION

The speaker recognition/classification is useful for radio and television broadcast indexing.

For speaker recognition we performed experiments where 25 speakers (11 male and 14 female) were used. Each speaker was instructed to read 15 different sentences. After recording of the sentences spoken by each speaker, we cut the recordings into smaller clips: 21 training clips (about 3 minutes long), and 10 test clips (50 s.) per speaker.

## 4. SOUND CLASSIFICATION USING THREE AUDIO TAXONOMY METHODS

Our goal was to identify classes of sound based on MPEG-7 ASP and MFCC.

### 4.1. Building the Sound Libraries

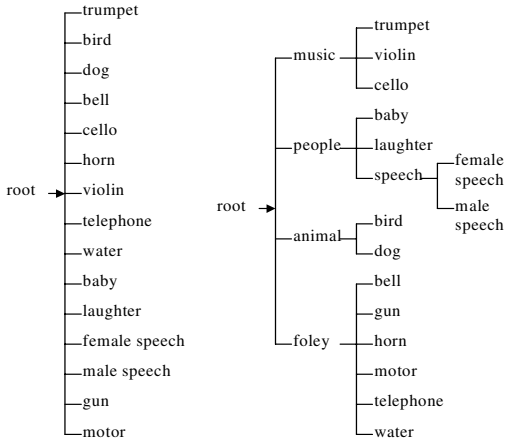
To test the sound classification system, we built sound libraries from various sources. This includes the speech database, that we collected for speaker recognition, and the "Sound Ideas" general sound effects library [7]. We created 13 sound classes from the sound effects library and 2 sound classes from the collected speech database. 70% of the data was used for training and the other 30% for testing.

### 4.2. Three Audio Taxonomy Methods

For sound classification, we use three different taxonomy methods: a direct approach, a hierarchical approach without hints and a 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. This approach is illustrated in Figure 1 (a). For the direct approach, we used a simple sound recognition system to generate the classification results. Each input sound is tested on all of the sound models, and the highest maximum likelihood score is used to determine the test clip's recognized sound class. This method is most straightforward, but would cause problems when there are too many classes.

For the hierarchical approach we organize the database of sound classes on the hard disk using the hierarchy shown in Figure 1 (b). Because we modelled the database in this fashion, we decided to use the same hierarchy for recognition. That is, we create additional bases and hidden Markov models for the more general classes *animal*, *foley*, *people,* and *music*. For each test sound, a path is found from the root down to a leaf node with testing occurring at each level in the hierarchy.




(a) direct approach (b) hierarchical approach

Figure 1: Classification using a direct and hierarchical approach

In certain systems, such as hierarchical classification with hints, it would be feasible to assume that additional information is available. For instance, it would be possible to have a recording of human speech but not be able to tell the gender of the speaker by ear. The hint *speech* can be given, so that the program can determine what gender the speaker is with possibly higher accuracy. In our 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.

## 5. AUDIO SEGMENTATION

Our goal for audio segmentation was to separate audio into sound events. More specification, we were interested in identifying whenever particular speakers appeared in an audio event.

### 5.1. Data Set

We used two audio tracks from television panel discussions for our purpose.

Discussion 1 contained only four speakers. Each speaker model was trained with between 1 and 2 minutes of audio. Discussion 2, which was 60 minutes long, was much more challenging. 7 main speaker models were trained (5 male and 2 female), and an applause model was

also used as the studio audience often responded to comments with applause. The speakers themselves are mostly German politicians arguing about tax reforms, so they interrupt each other.

### 5.2. Segmentation Using Sound/Speaker Identification

For our test data, an audio track of a television panel discussion is used as input, but other kinds of audio input could be segmented in the same manner. The track was cut into 1.5 second sub-segments, which overlapped by 33%. That is, the "hop size" was 1 second. Overlapping increased the input data to be classified by 50% but yielded more robust sound/speaker segmentation results due to the filtering technique described below. We assumed that there is no speaker change within each sub-segment. Therefore, speaker detection can be performed at the sub-segment level. Given a 1.5 second long sub-segment as input, the NASE features were extracted and projected against each sound model's set of basis functions in the database. Then, the Viterbi algorithm was applied to align each projection on its corresponding sound class HMM. The Viterbi algorithm finds the maximum likelihood sequence of states through the recognition classifier and returns the most likely classification label for the sub-segment. Invalid input, such as pauses or heated discussions with multiple people speaking at the same time, cause sub-segments to sometimes be classified incorrectly when there are no appropriate models for the input. As a result, the sub-segment labels needed to be smoothed out. To this end, we used a low-pass filter to enable more robust segmentation by correcting errors. The filter waits for $A$ adjacent sub-segments of the same label before declaring the beginning of a segment. Errors can be tolerated within a segment, but once $B$ adjacent classifications of any other models are found, the segment is ended. For our data, the optimum values were $A = 3$ and $B = 3$. In a real-time system, this would imply at least a 3.5 second latency before detecting a new segment.

## 6. EXPERIMENTAL RESULTS

The audio data used throughout the paper were digitized at 22.05 kHz using 16 bits per sample. The features were derived from speech frames of length 25ms with a frame rate of 15ms. Each frame was windowed using a Hamming window function.

The features were used to train hidden Markov models (HMM) using conventional maximum likelihood estimation for individual audio classes. For speaker recognition and sound classification a 7-state left-right model were applied. However, in the case of the segmentation with a long panel discussion, parts of the temporal structure can be repeated in the video sequence, but not necessarily the whole temporal structure. Such temporal structures of video sequences require the use of an ergodic topology, where each state can be reached from any other state and can be revisited after leaving. Therefore, we built a 7-state ergodic model for the segmentation of audio.

### 6.1. Results of Speaker Recognition and Sound Classification

We performed experiments with different feature dimensions of the different feature extraction methods. The results of speaker recognition and sound classification for the direct approach are shown in Table 1.

| Feature Extraction | FD for Speaker Recognition | | | FD for Sound Recognition | | |
|---|---|---|---|---|---|---|
| | 7 | 13 | 23 | 7 | 13 | 23 |
| PCA-ASP | 58.4 | 85.1 | 88.9 | 83.3 | 90.4 | 95.0 |
| ICA-ASP | 65.7 | 84.9 | 93.6 | 82.5 | 91.7 | 94.6 |
| MFCC | 78.5 | 93.8 | 93.1 | 90.8 | 93.2 | 94.2 |

Table 1: Comparison of speaker recognition and sound classification accuracies (%). FD: feature dimension, PCA-ASP: MPEG-7 audio spectrum projection (ASP) based on PCA basis, ICA-ASP: MPEG-7 ASP based on ICA basis.

Regarding the recognition of 25 speakers MPEG-7 ASP onto ICA basis yields better performance than ASP onto PCA basis. The recognition rates using MPEG-7 conform ASP results appear to be significantly lower than the recognition rate of MFCC with the dimension 7 und 13. We achieve the better recognition rate with the MFEG-7 ASP features onto ICA vs. MFCC with the dimension 23.

For general sound recognition of 15 audio classes MFCC performs superior at low dimension, while slightly inferior at high dimensions. The MFEG-7 ASP features onto PCA provides slightly better recognition rate than MFEG-7 ASP features onto ICA with the dimension 23.

Table 2 describes the recognition results of several sounds with different classification structures.

| Feature Extraction | Feature Dimension (13) | | |
|---|---|---|---|
| | a | b | c |
| PCA-ASP | 90.41 | 75.83 | 97.05 |
| ICA-ASP | 91.67 | 76.67 | 97.08 |
| MFCC | 93.24 | 86.25 | 96.25 |

Table 2: Comparison of sound classification accuracies (%) using several audio taxonomy. a: direct approach, b: hierarchical classification without hints, c: hierarchical classification with hints

The MPEG-7 ASP features yields 91.67% recognition rate in the classification using a direct approach. This recognition rate appears to be significantly lower than the 93.24% recognition rate obtained with MFCC.

In the classification using hierarchical approach without hints, the MFCC features gives a significant recognition improvement over the MPEG-7 ASP features. However, the recognition rate 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, a 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.

The hierarchical classification with hints yields thus overall the highest recognition rate compared to one level structure or hierarchical classification without hints. Specially, the recognition rate of the MPEG-7 ASP is slightly better than the recognition rate of the MFCC features, because some male and female speeches are better recognized by the MPEG-7 ASP than by the MFCC.

### 6.3 Segmentation Results

The results achieved with two panel discussion materials are summarized in Table 3.

| M | FD | FE | Reco. Rate (%) | R (%) | P (%) | F (%) |
|---|---|---|---|---|---|---|
| d 1 | 13 | ASP | 83.2 | 84.6 | 78.5 | 81.5 |
| | | MFCC | 87.7 | 92.3 | 92.3 | 92.3 |
| | 23 | ASP | 89.4 | 92.3 | 92.3 | 92.3 |
| | | MFCC | 95.8 | 1 | 92.8 | 96.2 |
| d 2 | 13 | ASP | 61.6 | 51.5 | 28.8 | 36.9 |
| | | MFCC | 89.2 | 63.6 | 61.7 | 62.6 |
| | 23 | ASP | 84.3 | 66.6 | 61.1 | 63.7 |
| | | MFCC | 91.6 | 71.2 | 73.8 | 73.4 |

Table 3: Performance of the segmentation using sound/speaker identifiers. M: TV materials, d1: *discussion 1*, d2: *discussion 2*, FD: feature dimension, FE: feature extraction methods, C: number of correctly found boundaries, B: total number of boundaries, H: number of hypothesized boundaries, R:recall=C/B, P: precision=C/H, F: F-measure=(2·recall·precision)/(recall+precision).

The segmentation results for *discussion 1* was quite good because there were only four speakers, and they rarely interrupted each other. The algorithm runs fast enough so that it was implemented on a real-time system. On the other hand, the results of the segmentation for *discussion 2* was not as good, but still impressive in view of the numerous interruptions. The training data also differed somewhat from the test data because the politicians did not raise their voices until later in the show. That is, we used their calm introductions as training data, while the test data sounded quite different because the politicians had become more excited.

The Table 3 shows that the recognition accuracy, recall, precision and F-measure of the MFCC features are better than MPEG-7 ASP features in the case of both 13 und 23 feature dimensions for discussion 1. For discussion 2 the MFCC features show a remarkable improvement over the MPEG-7 ASP features. Recall that the recognition system identifies speakers as part of the segmentation task.

## 7. CONCLUSION

Our results show that the MFCC features yield better performance compared to MPEG-7 ASP in the speaker/sound recognition, and audio segmentation. In the case of MFCC, the process of recognition, classification and segmentation is simple and fast because there are no bases used. On the other hand, the extraction of the MPEG-7 ASP is time and memory consuming compared to MFCC.

## 7. REFERENCES


[1] B. Manjunath, P. Salembier, and T. Sikora, "Introduction to MPEG-7," *Wiley,* NY, Sep, 2001.

[2] ISO., "ISO 15938-4:2001 (MPEG-7: Multimedia Content Description Interface, Part 4: Audio," ISO, 2001.

[3] M. Casey, "MPEG-7 Sound-Recognition Tools," *IEEE Transactions on Citcuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737-747, June, 2001.

[4] Z. Xiong, R. Radhakrishnan, A. Divakaran, T. Huang, "Comparing MFCC and MPEG-7 Audio Features for Feature Extraction, Maximum Likelihood HMM and Entropic Prior HMM for Sports Audio Classification," *ICASSP 2003*, vol. 5, pp. 628-631, April 2003.

[5] H.-G. Kim, E. Berdahl, N. Moreau, T. Sikora "Speaker Recognition Using MPEG-7 Descriptors," *EUROSPEECH 2003,* September 2003.

[6] A. Hyvarinen, E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks,* vol. 13, pp. 411-430, 2000.

[7] http://www.sound-ideas.com/