HOW EFFICIENT IS MPEG-7 FOR GENERAL SOUND RECOGNITION?

HYOUNG-GOOK KIM, JUAN JOSÉ BURRED, AND THOMAS SIKORA

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

Our challenge is to analyze/classify video sound track content for indexing purposes. To this end we compare the performance of MPEG-7 Audio Spectrum Projection (ASP) features based on several basis decomposition algorithms vs. Mel-scale Frequency Cepstrum Coefficients (MFCC). For basis decomposition in the feature extraction we evaluate three approaches: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF). Audio features are computed from these reduced vectors and are fed into a continuous hidden Markov model (CHMM) classifier. Our conclusion is that established MFCC features yield better performance compared to MPEG-7 ASP in the general sound recognition under practical constraints.

INTRODUCTION

Due to the advances of information technology, more and more digital audio, images, and video are being captured, produced and stored. As a result, there are strong research and development interests in multimedia databases regarding the efficient use of the information stored in these media types . For these reasons, the MPEG-7 standard [1], formally named "Multimedia Content Description Interface", has been focusing on a standardized set of technologies for describing multimedia content in a wide range of multimedia applications such as data indexing, data filtering or data retrieval systems.

Among multimedia documents that are today available in profusion on the Internet or in private archives, many contain an audio part. These audio signals enclose information that can be used to index and retrieve the documents they belong to. Recently, audio classification has become more and more important for effective indexing of multimedia documents because different audio types should be processed differently, and the searching space after classification is reduced to a particular subclass during the retrieval process.

In order to provide a unified interface for automatic indexing of audio, the MPEG-7 sound-recognition tools [2][3] use dimension-reduced, decorrelated spectral features, called Audio Spectrum Projections (ASP), based on Audio Spectrum Basis (ASB) as feature extraction and continuous hidden Markov models (CHMM) [4] as classifier. Each classified audio piece will be individually processed and indexed so as to be suitable for efficient comparison and retrieval by the sound recognition system.

In this paper, the MPEG-7 ASP features based on several basis decomposition algorithms are applied to sound recognition. For the measure of the performance we compare the classification results of MPEG-7 standardized features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC) [5], which have been widely used in speech recognition and audio classification.

1 MPEG-7 COMPLIANT FEATURES

The MPEG-7 ASP feature extraction mainly consists of a Normalized Audio Spectrum Envelope (NASE), a basis decomposition algorithm and a spectrum basis projection, obtained by multiplying the NASE with a set of extracted basis functions. Figure 1 depicts the procedure of MPEG-7 ASP feature extraction.

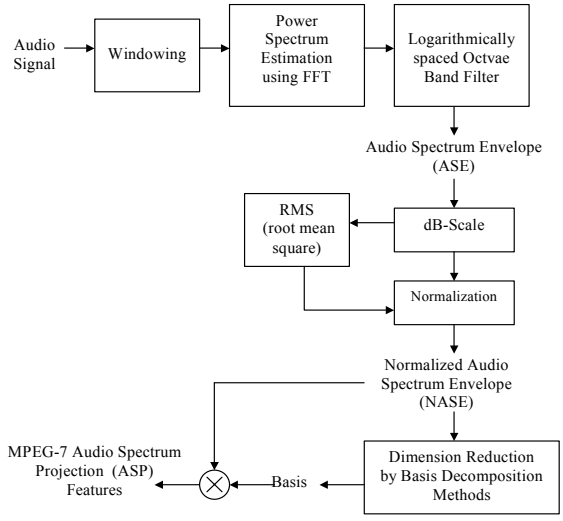


Figure 1: Block diagram of MPEG-7 ASP features

1.1 Normalized Audio Spectrum Envelope (NASE)

To extract a reduced-rank spectral feature called Audio Spectrum Envelope (ASE), the observed audio signal is analyzed using a FFT. The power spectral coefficients are grouped in logarithmic sub-bands spaced in octave bands spanning between the low edge and high edge parameters. The resulting ASE features are 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 and represented by the $L \times F$ matrix $X(l,f)$. It is defined as:

$$X(l,f) = \frac{10\log_{10}\left(ASE(l,f)\right)}{\sqrt{\sum_{f=1}^{F}\left\{10\log_{10}\left(ASE(l,f)\right)\right\}^2}} \quad (1)$$

where l ($1 \le l \le L$) is the time frame index, f ($1 \le f \le F$) is the logarithmic frequency range, L is the total number of frames and F is the number of ASE spectral coefficients.

1.2 Basis Decomposition Algorithms

In general, removing statistical dependence of observations is used in practice to dimensionally reduce the size of datasets while retaining as much important perceptual information as possible. For such a basis decomposition step, we can choose one of the following methods: Principal Component Analysis (PCA) [6], Independent Component Analysis (ICA) [7], and Non-negative Matrix Factorization (NMF) [8].

1.2.1 Principal Component Analysis (PCA)

PCA aims to decorrelate variables or signals, in order to find orthogonal directions with maximal variance. The first step of PCA consists of removing the sample mean of each signal:

$$\hat{X}(l,f) = X(l,f) - \frac{1}{L}\sum_{l=1}^{L} X(l,f), \quad (2)$$

where L is the number of frames and X is the NASE matrix.

The second step consists of applying a linear transformation on $\hat{X}$. This transformation rotates the coordinate system in such a way that the first new axis points in the direction of maximal variance, the second axis, orthogonal to the first one, collects the largest part of the remaining variance, and so on.

The new axes are determined by a spectral decomposition of the sample covariance matrix

$$C_X = \left(\hat{X}\hat{X}^T\right)/L = V\Sigma V^T \quad (3)$$

where V is an orthonormal matrix and Σ a diagonal one. As C_X is symmetric and semipositive definite, all eigenvalues λ_i (the diagonal entries of Σ) are real and non-negative. The variance along each of the new axes V_i is simply given by its associated eigenvalue λ_i. The projection gives the decorrelated signals Y according to

$$Y = V^T\hat{X}, \quad (4)$$

where V gathers the m eigenvectors associated to the m largest eigenvalues.

Sphered signals can be obtained with a slight modification of PCA as

$$C_P = \sqrt{\Sigma^{-1}}V^T, \quad (5)$$

In order to perform dimensionality reduction, we reduce the size of the matrix C_P by throwing away $F - E$ of the columns of C_P corresponding to the smallest eigenvalues of D. The resulting matrix C_E has the dimensions $F \times E$.

The projection is given by

$$Y_E = XC_E \quad (6)$$

yielding decorrelated signals with unit variance.

1.2.2 Independent Component Analysis (ICA)

ICA is a statistical method which not only decorrelates the second order statistics but also reduces higher-order statistical dependencies. Thus, ICA produces mutually uncorrelated basis. The independent components of a NASE matrix X can be thought of as a collection of statistically independent bases for the rows (or columns) of X. The $L \times F$ matrix X is decomposed as

$$X = WS + N \quad (7)$$

where S is the $P \times F$ source signal matrix, W is the $L \times P$ mixing matrix or the matrix of spectral basis functions, and N is the $L \times F$ matrix of noise signals. Here P is the number of independent sources. The above decomposition can be performed for any number of independent components and the sizes of W and S vary accordingly.

From the several ICA algorithms we use a combination of PCA and FastICA algorithm [7] for performing the decomposition. After extracting the reduced PCA basis C_E, a further step consisting of basis rotation in the directions of maximal statistical independence is needed for applications that require maximum decorrelation of features. The whitening closely related to PCA is done by multiplying the $F \times E$ transformation matrix C_E with the $L \times E$ matrix Y_E. The input Y_E is then fed to the FastICA algorithm

based on a Gram-Schmidt-like decorrelation. When we have estimated E independent components, or E vectors $W_1,\ldots,W_E$, we run the one-unit fixed-point algorithm for W_{E+1}, and after every iteration step, subtract from W_{E+1} the projections $W_i^T W_j W_j$, j=1,…,E of the previously estimated E vectors according to

$$W_{E+1} \leftarrow W_{E+1} - \sum_{j=1}^{E} W_{E+1}^T W_j W_j \tag{8}$$

and then renormalize W_{E+1}:

$$W_{E+1} \leftarrow \frac{W_{E+1}}{\sqrt{W_{E+1}^T W_{E+1}}} \tag{9}$$

The resulting spectrum projection Z is the product of the NASE matrix X, the dimension-reduced PCA basis functions C_E, and the $E \times E$ ICA transformation matrix W_E:

$$Z = XC_E W_E . \tag{10}$$

1.2.3 Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) has been recently proposed as a new method for dimensionality reduction. The NMF is a subspace method which finds a linear data representation with the non-negativity constraint. It is conceptually simpler than PCA or ICA, but not necessarily more computationally efficient. Within this context, NMF was first applied in generating parts-based representations from still images [9] and has later been evaluated in audio analysis tasks, such as general sound classification [10] and polyphonic music transcription [11].

Given a non-negative $m \times n$ matrix $|X|$, NMF consists of finding the non-negative matrices G ($m \times p$) and H ($p \times n$) such that $|X| \approx GH$, where $p < m$ and $p < n$. Several algorithms have been proposed to perform NMF. Here, the Divergence Update algorithm is used. The divergence of two matrices A and B is defined as

$$D(A \,||\, B) = \sum_{i,j} \left(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right) \tag{11}$$

The algorithm iterates updating the factor matrices in such a way that the divergence $D(\,|X| \,||\, GH\,)$ is minimized.

Such a factorization can be found using the update rules

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i G_{ia} X_{i\mu} / (GH)_{i\mu}}{\sum_k G_{ka}} \tag{12}$$

$$G_{ia} \leftarrow G_{ia} \frac{\sum_\mu H_{a\mu} X_{i\mu} / (GH)_{i\mu}}{\sum_v H_{av}} \tag{13}$$

More details about the algorithm can be found in [9].

In this case, X is the $L \times F$ NASE matrix, and thus factorization yields the matrices G and H with sizes $L \times E$ and $E \times F$, respectively, where E is the desired number of bases. In this way, H is the basis matrix, which is stored and used to obtain the ASP needed to perform classification. The projection is defined as:

$$Y = |X| H^T \tag{14}$$

2 MFCC FEATURES

For the feature extraction many researchers are interested in comparing of the performance of MPEG-7 ASP features vs. MFCCs according to reduced dimension.

Compared to MPEG-7 ASP features, MFCCs are not always able to express the domain's statistical structure, but they assume that all signals are infinitely stationary and that the probabilities of the basis functions are all equal.

Their processes are compared in Table 1.

	MFCCs	MPEG-7 ASP
1	Convert to Frames	Convert to Frames
2	For each frame, obtain the amplitude spectrum	For each frame, obtain the amplitude spectrum
3	Mel-scaling and smoothing	Log-scale octave bands
4	Take the logarithm	Normalization
5	Take the discrete cosine transform (DCT)	Perform basis decomposition using PCA, ICA, or NMF for projection features

Table 1: Comparison of MPEG-7 ASP and MFCCs.

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 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.

The components of the Mel-spectral vectors calculated for each frame are highly correlated. Features are typically modeled by mixtures of Gaussian densities. Therefore, in order to reduce the number of parameters

in the system, the last step of MFCC feature construction is to apply a transform to the Mel-spectral vectors which decorrelates their components. Theoretically, the Karhunen-Loève Transform (KLT, or equivalently, PCA) achieves this. In the speech community, the KLT is approximated by the Discrete Cosine Transform (DCT) [12][13].

The MFCCs are defined as:

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} \left((\log S_k) \times \cos\left[n \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \right)$$

$$n = 1,2,...,L \qquad (15)$$

where K is the number of critical bands and L is the desired length of the cepstrum. Usually $L<<K$ for the dimension reduction purpose. S_k, $0 \le k < K$, are the filter bank energies after passing each corresponding kth triangular band-pass filter.

3 EVALUATION

In order to evaluate the proposed feature sets, a left-right continuous HMM classifier with 7 states was used with a variety of different sound sources.

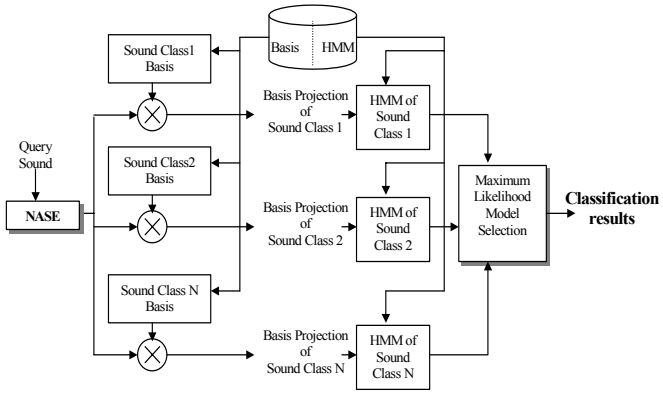
3.1 Classification

Given a training sequence for each pre-defined sound class, the HMM for that class is trained using a maximum likelihood estimation procedure known as the Baum-Welch algorithm. Classification is achieved by using the Viterbi algorithm to determine the maximum likelihood state sequence through the HMMs given an observed sequence of feature vectors.

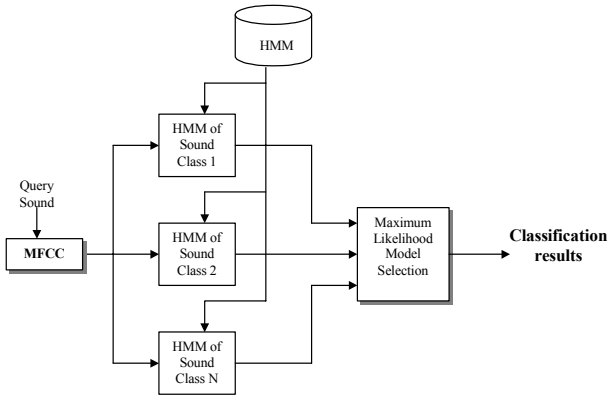
The procedures of the sound recognition classifier using MPEG-7 ASP features and MFCC are summarized in Figures 2a and 2b, respectively.

In the case of sound classification using MPEG-7 ASP features, the NASE features are extracted from the query sound and projected onto each individual sound basis functions. Then, the Viterbi algorithm is applied to align each projection on its corresponding sound class HMM. On the one hand, this process causes testing to last considerably longer, as each test clip has to be projected onto different bases, before it could be tested on the different HMM's to determine what it should be recognized as, but on the other hand, the performance due to the projection onto the well-chosen bases increased performance considerably. In order to obtain good results with the PCA and ICA algorithms, feature extraction parameters need to be selected with care.

In the case of MFCCs, the classification process is easy because there are no bases. Each query sound is simply matched against each of the HMMs (trained with MFCC features) via the Viterbi algorithm. The HMM yielding the best acoustic score determines the recognized sound.



(a): Sound recognition classifier using MPEG-7 ASP features



(b): Sound recognition classifier using MFCC features

Figure 2: Block diagram of sound recognition classifier

3.2 Datasets

To test the sound classification system, we built sound libraries from various sources. These include a speech database, and the "Sound Ideas" general sound effects library [14]. We created 10 sound classes (bird, dog, bell, horn, telephone, water, baby, laughter, gun, motor) from the sound effects library and 2 speech classes (male speech, female speech) from the speech database. 60 sound examples were collected for each class. 66% of the data was used for training and the other 33% for testing.

3.3 Feature Extraction

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 30ms with a frame rate of 15ms. Each frame was windowed using a Hamming window function and transformed into the frequency domain using a 512-point FFT. The low and high boundaries of the logarithmic frequency bands for MPEG-7 features are 62.5 Hz and 8 kHz, that are over a spectrum of 7 octaves. For each audio class, one of the PCA, FastICA, or NMF methods is performed on the NASE features of all the audio frames from all the training examples in the class.

For NMF of the audio signal we had two choices: (1) The NMF basis was extracted from the NASE matrix. The ASP projected onto the NMF basis were directly applied to the HMM sound classifier. (2) The audio signal was transformed to the spectrogram. NMF component parts were extracted from spectrogram image patches. Basis vectors computed by NMF were selected according to their discrimination capability. Sound features were computed from these reduced vectors and fed into HMM classifier. This process is well described in [10].

MFCCs are calculated from 40 subbands between 62.5 Hz and 8 kHz.

3.4 Results

We performed experiments with different feature dimensions for each of the feature extraction methods. Particularly, the recognition task was performed for a number of 7, 13 and 23 reduced dimensions from the basis vectors. The sound recognition results are shown in Table 2.

Feature Extraction	Feature Dimension		
	7	13	23
PCA-ASP	83.3	90.4	95.0
ICA-ASP	82.5	91.7	94.6
NMF-ASP	75.83	78.33	79.58
MFCC	90.8	93.2	94.2

Table 2: Sound Classification Accuracies (%)

Regarding the recognition of 12 sound classes MPEG-7 ASP projected onto FastICA basis provides slightly better recognition rate than ASP projected onto PCA basis with 7 and 23 dimensions, while slightly worse with 13 dimensions. Compared to MPEG-7 ASP based on PCA, NMF or FastICA, MFCC performs superior at dimension 7 and 13 while slightly inferior at dimension 23. Furthermore, the MFCC feature extraction is simpler and faster than ICA. On the other hand, the ASP projected onto NMF derived from NASE matrix $|X|$ yields lowest recognition rate, while NMF with 95 ordered basis according to the spectrogram image patches provides 95.8 % recognition rate. The NMF by divergence update algorithm converges very slowly in comparison with PCA or FastICA.

4 CONCLUSIONS

In this paper we compared the performance of MPEG-7 Audio Spectrum Projection (ASP) features based on three basis decomposition algorithms vs. MFCC.

For a basis decomposition step PCA decorrelates the second order moments corresponding to low frequency properties and extracts orthogonal principal components of variations. ICA is a statistical method which not only decorrelates the second order statistics but also reduces higher-order statistical dependencies. Thus, ICA produces mutually uncorrelated basis. On the other hand, NMF attempts a matrix factorization in which the factors have non-negative elements by performing a simple multiplicative updating. In the case of MFCCs, a discrete cosine transform is used to the logarithm of the filter bank outputs results in vectors of decorrelated MFCC features.

Our results show that the MFCC features yield better performance compared to MPEG-7 ASP basis functions in sound recognition. In the case of MFCC, the process of feature extraction is simple and fast because there are no bases used. On the other hand, the extraction of the MPEG-7 ASP is more time and memory consuming compared to MFCC. The NMF updating process is very slow compared to FastICA.

Future work should focus on a set of perceptually motivated features after examination in some detail MFCCs: a) the use of the Mel frequency scale to model the spectra; and b) the use of the DCT to decorrelate the Mel-spectral vectors, for noise robust sound classification system.

REFERENCES

[1] B. S. Manjunath, P. Salembier and T. Sikora, “Introduction to MPEG-7”, *Wiley* (2002).

[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 circuits and Systems for video Technology*, vol. 11, no.6, June 2001.

[4] L. Rabiner and B.-H. Juang, “Fundamentals of speech recognition,” *Prentice Hall, N.J.* (1993).

[5] S. Davis and P. Mermelstein, “Experiments in syllable-based recognition of continuous speech,” *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. 28, pp. 357—366 (1980).

[6] I. T. Jolliffe, "Principal component analysis," *Springer-Verlag* (1996).

[7] A. Hyvärinen, E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks,* vol.. 13, pp. 411-430 (2000)

[8] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization" *Nature*, no. 401, pp. 788—791 (1999).

[9] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization" *Adv. Neural Info. Proc. Syst.* 13, pp. 556—562 (2001).

[10] Y.-C. Cho, S. Choi and S.-Y. Bang, "Non-negative component parts of sound for classification" in *Proc. IEEE Int. Symp. Signal Processing and Information Technology,* Darmstadt, Germany (2003).

[11] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription" in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, USA (2003).

[12] N. Marhav and C.-H. Lee, "On the asymptotic statistical behavior of empirical ceptsral coefficients" in *IEEE Transactions on Signal Processing* 41, pp. 1990—1993 (1993).

[13] B. Logan, "Mel frequency cepstral coefficients for music modeling" in *Proc. International Symposium on Music Information Retrieval* (2000).

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