AUDIO SPECTRUM PROJECTION BASED ON SEVERAL BASIS DECOMPOSITION ALGORITHMS APPLIED TO GENERAL SOUND RECOGNITION AND AUDIO SEGMENTATION

Hyoung-Gook Kim, and Thomas Sikora

Communication Systems Group, Technical University of Berlin
Einsteinufer 17,D-10587 Berlin, Germany (Europe)
phone: +49 30 314-25799, fax: +49 30 314-22514, email: [kim, sikora]@tu-berlin.de
web: www.nue.tu-berlin.de

ABSTRACT

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 basis decomposition vs. Mel-scale Frequency Cepstrum Coefficients (MFCC). For basis decomposition in the feature extraction we have three choices: 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 hidden Markov model classifier. Experimental results show that the MFCC features yield better performance compared to MPEG-7 ASP in the sound recognition, and audio segmentation.

1. INTRUDUCTION

Our challenge is to analyze/classify video sound track content for indexing purposes.

Recently, audio contents become more and more important clues for effective video indexing, 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.

Toward this end, the MPEG-7 sound-recognition tools [1][2] provide a unified interface for automatic indexing of audio using trained sound classes in a pattern recognition framework. 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.

A feature extraction method of the MPEG-7 sound recognition framework is 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) [3] in order to apply uniformly to diverse source classification tasks with accurate performance.

In this paper, the MPEG-7 ASP features based on several basis decomposition algorithms are applied to sound recognition and to segment conversational speech of panel discussion television programs. For the measure of the performance we compare the classification and segmentation results of MPEG-7 standardized features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC).

2. MPEG-7 SOUND CLASSIFICATION SYSTEM

The sound classification task is performed using three steps: audio feature extraction, training of sound models, and decoding. Figure 1 depicts the procedure of sound recognition classifier.

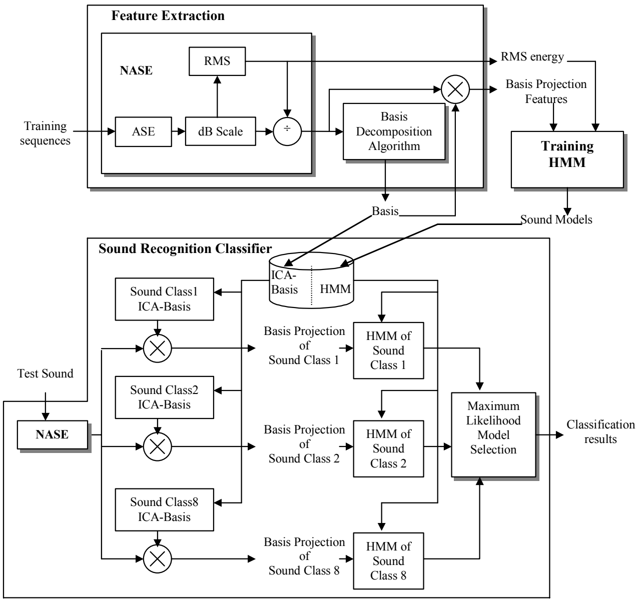

Figure 1: Block diagram of sound classification

2.1 Feature Extraction Using Basis Projection

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), a basis decomposition algorithm and a spectrum basis projection.

2.1.1 Normalized Audio Spectrum Envelope

The observed audio signal is divided into overlapping frames by 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 $m \times n$ NASE matrix X.

2.1.2 Decomposition Algorithm

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) [4], Independent Component Analysis (ICA) [5], and Non-negative Matrix Factorization (NMF) [6].

- Principal Component Analysis (PCA):
  PCA aims to reduce the dimensionality of a data set by only keeping the components of the sample vectors with large variance. PCA decorrelates the second order moments corresponding to low frequency properties and extracts orthogonal principal components of variations. By projecting onto these highly varying subspace, the relevant statistics can be approximated by a smaller dimension system.
- 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. Thus the independent components of a NASE matrix X can be thought of a collection of statistically independent sources for the rows (or columns) of X. The $m \times n$ matrix X is decomposed as

$$X = WS + N \tag{1}$$

  where S is the $r \times n$ source signal matrix, W is the $n \times r$ matrix mixing matrix or the matrix of spectral basis functions, and N is the matrix of noise signals. Here r 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. We use the Fast ICA algorithm [5] for performing the decomposition.
- Non-negative Matrix Factorization (NMF)
  On the other hand, NMF attempts a matrix factorization in which the factors have non-negative elements by performing a simple multiplicative updating. The NMF of X is given by

$$X = GF \tag{2}$$

where the factor G and F contain only non-negative entries. The columns of the $m \times n$ matrix X are the signals, the columns of the $m \times r$ matrix G are the basis signals, and the $r \times n$ matrix F is the mixing matrix. Here r is the number of non-negative components. The multiplicative divergence update rules are as the following:

$$F_{a\mu} = F_{a\mu} \frac{\sum_i G_{ia} X_{i\mu} / (GF)_{i\mu}}{\sum_k G_{ka}} \tag{3}$$

$$G_{ia} = G_{ia} \frac{\sum_\mu F_{a\mu} X_{i\mu} / (GF)_{i\mu}}{\sum_v F_{av}} \tag{4}$$

We can use the columns of G as the new basis.

2.1.3 Audio Spectrum Projection

The resulting audio spectrum projection is obtained by multiplying the NASE matrix with a set of extracted basis functions. 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.

2.2 Training of Sound Models and Decoding

For each pre-defined sound class, the training module builds a model from a set of training sounds using hidden Markov models (HMM). When the training process is complete using a maximum likelihood estimation procedure known as the Baum-Welch algorithm, the statistical basis and HMM model of each sound class are stored in the sound model database of the sound recognition classifier.

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.

3. SEGMENTATION OF SPEAKERS USING MPEG-7 SOUND CLASSIFICATION SYSTEM

In this section, MPEG-7 dimension-reduced, decorrelated ASP features are applied to segment conversational speech of panel discussion television programs. Without a priori information about number of speakers, the speech stream is segmented by a hybrid metric-based [7] and model-based [8] segmentation algorithm.

3.1 Segmentation System

The hybrid segmentation using MPEG-7 ASP features is mainly composed of six modules: silence removal, feature extraction module using normalized audio spectrum envelope (NASE), speaker change detection, segment-level-clustering module, speaker model updating using MPEG-7

ASP features and the HMM re-segmentation module. Figure 2 depicts the algorithm flow chart.

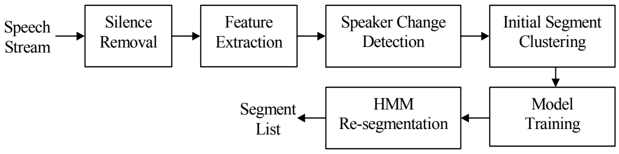

Figure 2: Segmentation procedures.

The speech stream is divided into sub-segments of length 3-second window with 2.5-second overlapping. That is, the temporal resolution of the segmentation is 0.5 second. In the sub-segments silence segments are detected by a simple energy-based algorithm and the detected silence segments are removed. Each non-silence 3s sub-segment is converted into NASE features. The speaker change detection step is performed using a NASE divergence distance [9] measure between two sub-segments and splits the conversation into smaller segments that are assumed to contain only one speaker. The segments created by the speaker change detection step are used to form initial clusters from similar segments. At this stage, we use a hierarchical agglomerative method [10] that computes the Generalized Likelihood Ratio (GLR) distance [11] between every pair of clusters and merges two clusters with the minimum distance at every step. The clustering procedure continues to aggregate clusters together until there is just one large cluster. The output from the procedure is a tree of clusters. At the end of the hierarchical classification algorithm a dendrogram is built in which each node corresponds to a cluster. The cutting of the dendrogram produces a partition composed of all the utterances. The initial clusters are used to train an initial set of speaker models from all segments of each respective cluster.

Given the NASE features of every cluster, the spectral basis is extracted by computing the several basis decomposition algorithms. The resulting spectral basis vectors are multiplied with the NASE matrix, thus yielding the dimension-reduced decorrelated ASP features.

The spectrum projection features and RMS-norm gain values are input to the HMM training process. In order to train a statistical model on the basis projection features and RMS-norm gain value of each cluster an ergodic HMM with 7 states is trained for each cluster. The trained speaker models are then used to resegment the speech stream.

Re-segmentation is achieved by using the Viterbi algorithm to determine the maximum likelihood state sequence through the sound recognition classifier shown in Figure 1, given an observed sequence of feature vectors.

4. EXPERIMENTAL PROCEDURE AND RESULTS

4.1 Database

To test the sound classification system, we built sound libraries from various sources. We created 13 sound classes of trumpet, bird, dog, bell, cello, horn, violin, telephone, water, baby, laughter, gun and motor from the "Sound Ideas" general sound effects library and 2 sound classes of male and female speech from the collected speech database. 70% of the data was used for training and the other 30% for testing.

For the segmentation we used two audio tracks from television talk- show programs. "Talk Show 1" is approximately 15 minutes long and contains only four speakers. "Talk Show 2", which is 60 minutes long, is much more challenging because they interrupt each other frequently. It contained 7 main speakers (5 male and 2 female), and an applause as the studio audience often responded to comments with applause.

4.2 Parameters used in the Implementation

The audio data used throughout the paper were digitized at 22.05 kHz using 16 bits per sample. The ASP features based on PCA/ICA basis were derived from sound/speech frames of length 25ms with a frame rate of 15ms. The lower and upper boundary of the logarithmic frequency bands are 62.5 Hz and 8 kHz that are over a spectrum of 7 octaves. For sound classification a 7-state left-right HMM model were applied, while we built a 7-state ergodic model for the segmentation of audio.

For NMF of the audio signal we did not use the spectrogram image patches, but computed the NMF basis from the NASE matrix. The ASP projected onto the NMF basis without further basis selecting was applied direct to sound classifier.

4.3 Results of Sound Recognition

Our goal was to identify classes of sound using MPEG-7 ASP features based on three basis decomposition algorithm and MFCC.

We performed experiments with different feature dimensions of the different feature extraction methods. The results of sound classification are shown in Table 1.

Feature Dimension	Feature Extraction Method			
	ASP-PCA	ASP-ICA	ASP-NMF	MFCC
7	83.3	82.5	72.92	90.8
13	90.4	91.7	75	93.2

Table 1: Comparison of sound classification accuracies (%). ASP-PCA: MPEG-7 audio spectrum projection (ASP) based on PCA basis, ASP-ICA: MPEG-7 ASP based on FastICA basis, ASP-NMF: MPEG-7 ASP based on NMF basis.

Regarding the recognition of 15 sound classes MPEG-7 ASP projected onto FastICA basis provides slightly better recognition rate than ASP projected onto PCA basis at dimension 7, while slightly worse at dimension 13. 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. On the other hand, the ASP projected onto NMF yields lowest recognition rate vs. other feature extraction methods. The reason is that NMF basis matrix, which was produced without spectrogram image patches and basis ordering, reduced the data too much, and the HMM did not receive enough information.

4.4 Results of Segmentation of Speakers

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.

For the measure of the performance of the error correction the recognition rate and F-measure are used. The F-measure *F* is a combination of the recall (*RCL*) rate of correct boundaries and the precision (*PRC*) rate of the detected boundaries. When *RCL* and *PRC* errors are weighted as being equally detractive to the quality of segmentation, *F* is defined as

$$F = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL} \qquad (5)$$

The recall is defined by *RCL*= *ncf* / *tn*, while precision *PRC*=*ncf* / *nhb*, where *ncf* is the number of correctly found boundaries, *tn* is the total number of boundaries, and *nhb* is the number of hypothesized boundaries, meaning the total number of boundaries found by the segmentation module. *F* is bounded between 0 and 100, where *F*=100 is a perfect segmentation result and *F*=0 implies segmentation to be completely wrong.

Table 2 shows results for segmentation by the hybrid method.

System	M	FD	Feature Extraction	Reco. Rate (%)	F (%)
Hybrid	Talk 1	13	ASP-PCA	86.4	88.6
			ASP-ICA	86.2	88.5
			ASP-NMF	70.9	72.5
			MFCC	90.5	93.5
		24	ASP-PCA	87.5	91.8
			ASP-ICA	91.5	94.7
			ASP-NMF	77.5	79.3
			MFCC	96.8	98.1
	Talk 2	13	ASP-PCA	71.8	55.8
			ASP-ICA	72.1	56.1
			ASP-NMF	56.3	41.5
			MFCC	87.2	69.7
		24	ASP-PCA	84.6	72.9
			ASP-ICA	88.9	75.2
			ASP-NMF	72.5	57.1
			MFCC	93.2	82.7

Table 2: Comparison of Segmentation Results (%) based on several feature extraction methods. M: TV materials, FD: feature dimension

The segmentation results for Talks 1 was quite good because there were only four speakers, and they rarely interrupted each other. On the other hand, the results of the segmentation for Talks 2 was not as good, but still impressive in view of the numerous interruptions.

The recognition accuracy 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 "Talk Show 1". For "Talk Show 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. Overall MFCC achieves best recognition rate and F-measure rate.

5. CONCLUSIONS

In this paper we compare the performance of MPEG-7 Audio Spectrum Projection (ASP) features based on three basis decomposition algorithms vs. Mel-scale Frequency Cepstrum Coefficients (MFCC). Our results show that the MFCC features yield better performance compared to MPEG-7 ASP in sound recognition and audio segmentation. In the case of MFCC, the process of feature extarction 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.

REFERENCES

[1] M. Casey, "MPEG-7 sound recognition tools," IEEE Transactions on circuits and Systems for video Technology, vol. 11, no.6, June 2001.

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

[3] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," Prentice Hall, N.J., 1993.

[4] I. T. Jolliffe, "Principal component analysis," Springer-Verlag, 1986.

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

[6] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," NIPS 2000.

[7] P. Delacourt, C. j. Welekens, "DISTBIC: A speaker-based segmentation for audio data indexing," Speech Communication 32, 2000, pp. 111-126.

[8] L. Wilcox, F. Chen, d. Kimber, V. Balasubramanian, "Segmentation of speech using speaker Identification," Proceedings of ICASSP, 1994.

[9] L. Lu, H.-J. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," Proceedings of 10th ACM international conference on multimedia, Dec. 2002, pp. 602-610.

[10] D. A. Reynolds, E. Singer, B. A. Carlson, J.J. McLaughlin, G.C. O'Leary, and M.A. Zissman, "Blind clustering of speech utterances based on speaker and language charchteristics," ICASSP 1998, 1998.

[11] H. Gish, M.-H. Siu, R. Rohlicek, "Segration of speaker for speech recognition and speaker identification," Proceedings of ICASSP 1991, 1991, pp. 873-876.