THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.25, NO.3E 2006. 9 pp. 89-94

Dimension-Reduced Audio Spectrum Projection Features for Classifying Video Sound Clips

Hyoung-Gook Kim*
*Samsung Advanced Institute of Technology
(Received May 6 2006; Revised September 5 2006; accepted September 15 2006)

Abstract

For audio indexing and targeted search of specific audio or corresponding visual contents, the MPEG-7 standard has adopted a sound classification framework, in which dimension-reduced Audio Spectrum Projection (ASP) features are used to train continuous hidden Markov models (HMMs) for classification of various sounds. The MPEG-7 employs Principal Component Analysis (PCA) or Independent Component Analysis (ICA) for the dimensional reduction. Other well-established techniques include Non-negative Matrix Factorization (NMF), Linear Discriminant Analysis (LDA) and Discrete Cosine Transformation (DCT). In this paper we compare the performance of different dimensional reduction methods with Gaussian mixture models (GMMs) and HMMs in the classifying video sound clips.

Keywords: *MPEG-7, Audio spectrum projection (ASP), PCA, ICA, LDA, NMF, DCT*

I. Introduction

Among multimedia documents that are today available in profusion on the Internet or in private archives, many documents contain an audio part. In the field of application of content-based audio indexing and retrieval as well as in the field of audio analysis for video indexing, it is important to obtain a small number of feature dimensions after the extraction process, so that the training and the classification task of pattern recognition can be done faster and with less required storage space for the trained statistical models since the number of dimensional complexity of models corresponds with the quantity of data of models. So the computational and storage costs should be observed in connection with achievable classification results. The curse of dimensionality, which states that for every additional feature dimension, extensively more data are needed for the training in order to obtain good models, is a good reason to apply dimension reduction methods to the high dimensional feature data. The avoidance of this problem with a small number of dimensions is also beneficial considering the fact that often the costs for the creation of a set of audio classes are underestimated and a sufficient amount of audio data for training purposes is not available at all times.

The features, which are useful for the classification of an audio pattern, are often obtained from an auditory spectrum, where empirically adjacent bands are highly correlated. This is the initial point for decorrelation and multivariate techniques of linear transforms for an efficient representation of the dimension-reduced feature space. To obtain a more compact representation of the auditory spectrum, transforms can be applied so that only a few transform coefficients represent the auditory spectrum of one analysis frame with a loss of information. Hence, it is the purpose to keep the shape of the auditory spectrum lossy but sufficient in a few transform coefficients.

In order to provide a unified interface for automatic indexing of audio, the MPEG-7 standard has adopted a sound classification framework [1-2] using dimension-reduced, decorrelated spectral features called Audio Spectrum Projections (ASP) and continuous hidden Markov models (CHMM) [3] as classifier. The MPEG-7 employs Principal Component Analysis (PCA) or Independent Component Analysis (ICA) for the dimensional reduction. Other well-established techniques include Non-negative Matrix

Corresponding author: Hyoung-Gook Kim
(hyounggook.kim@samsung.com)
SAIT.MT. 14-1 Nongseo-dong, Giheung-gu, Younggin-si
Gyunggi-do, 449-712

Factorization (NMF), Linear Discriminant Analysis (LDA) and the Discrete Cosine Transform (DCT).

Two widely used statistical models for classification are Gaussian mixture models (GMMs) and HMMs. An HMM models the time-dependent short-time observation probability densities as well as their probabilities occurrence, while a GMM models the long term observation probability density of an observable stochastic process. Thereby the short-time observation probability densities are approximated coarser with uni-modal Gaussian densities compared with the approximation of the long-term observation probability with Gaussian mixtures.

In this paper, the ASP features based on different basis decomposition algorithms are tested with GMMs and HMMs for the classifying of video sound clips. For the measure of the performance we compare the classification results and the computational costs.

II. MPEG-7 Sound Recognition Structure

The MPEG-7 sound recognition classifier is performed using three steps: audio feature extraction, training of sound models, and classification. Figure 1 depicts the procedure of the MPEG-7 sound recognition classifier.

To extract a reduced-rank spectral feature called Audio Spectrum Envelope (ASE) ASE(l,f), 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:

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

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.

For each audio class, the spectral dimension-reduced basis is extracted by basis decomposition algorithms. The resulting spectrum projection is the product of the NASE matrix and the dimension-reduced basis. The spectrum projection features and RMS-norm gain values are used as input to the HMM training module. 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. Given a test sound as input, the NASE features are extracted and projected against each sound model's set of basis functions in the database. Then, the Viterbi algorithm is 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.

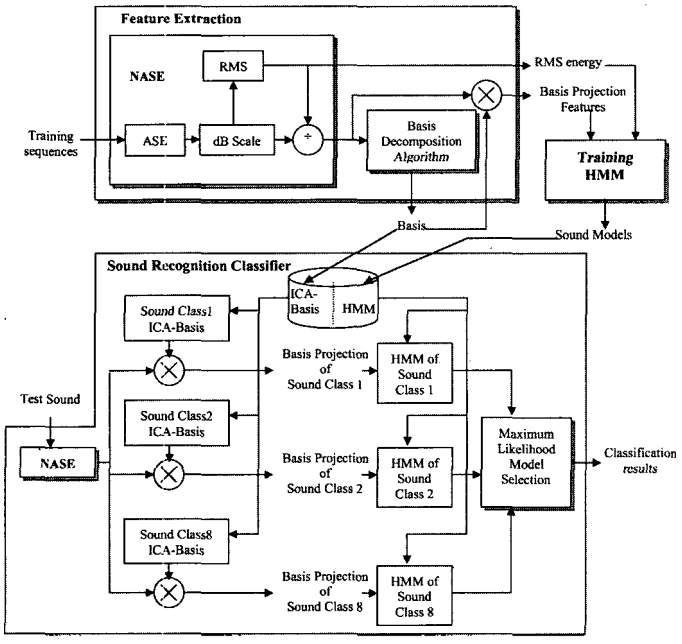

Figure 1. Block diagram of Sound Recognition Based on Basis Decomposition Algorithms.

III. Basis Decomposition Algorithms

Removing statistical dependence of observations is used in practice to dimensionally reduce the size of feature vectors while retaining as much important perceptual information as possible. We may choose one of the following methods: Principal Component Analysis (PCA) [5] or Independent Component Analysis (ICA) [6], Linear Discriminant Analysis (LDA) and Non-negative Matrix Factorization (NMF) [7].

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

$$C_P = \sqrt{\Sigma^{-1}}V^T, \quad (4)$$

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. The resulting matrix C_E has the dimensions $F \times E$.

The projection is given by

$$Y_{PCA,E} = \hat{X}C_E \quad (5)$$

yielding decorrelated signals with unit variance.

3.2. 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

$$X = WS + N_{noise} \quad (6)$$

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_{Noise} 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.

In this paper 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 $L \times F$ matrix $\hat{X}$. The input $Y_{PCA,E}$ is then fed to the FastICA algorithm based on a Gram-Schmidt-like decorrelation. When we have estimated E independent components, or E vectors $W_1, \cdots, W_E$, we run the one unit fixed point algorithm for W_{E+1}, and after every iteration step, subtract from W_{E+1} the projections $W_i^T W_j W_j$, $j=1,\cdots,E$ of the previously estimated E vectors according to

$$W_{E+1} \leftarrow W_{E+1} - \sum_{j=1}^{E} W_{E+1}^T W_j W_j \quad (7)$$

and then renormalize W_{E+1}:

$$W_{E+1} \leftarrow \frac{W_{E+1}}{\sqrt{W_{E+1}^T W_{E+1}}} \quad (8)$$

The resulting spectrum projection $Y_{ICA,E}$ 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:

$$Y_{ICE,E} = XC_E W_E. \quad (9)$$

3.3. Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (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.

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

$$D(A\|B)=\sum_{i,j}(A_{ij}\log\frac{A_{ij}}{B_{ij}}-A_{ij}+B_{ij}) \quad (10)$$

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

$$H_{a\mu}\leftarrow H_{a\mu}\frac{\sum_i G_{ia}X_{i\mu}/(GH)_{i\mu}}{\sum_k G_{ka}} \quad (11)$$

$$G_{ia}\leftarrow G_{ia}\frac{\sum_\mu H_{a\mu}X_{i\mu}/(GH)_{i\mu}}{\sum_\nu H_{a\nu}} \quad (12)$$

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:

$$Y_{NMF,E}=|X|H^T \quad (13)$$

3.4. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [14] seeks for a global linear transform, so that the class separability of the projected and dimension-reduced feature space is increased. The within-class scatter matrix O_W are defined as

$$O_W=\sum_{i=1}^{C}\frac{N_i}{N}\Sigma_i \quad (14)$$

where C is the number of involved classes, N_i is the number of observed analysis frames in the i-th class, N is the number of all analysis frames of all classes and $\sum_i$ is the sample covariance matrix of the NASE matrix X of the i-the class.

The eigenvectors U and eigenvalues Λ of the within class scatter matrix O_W with

$$O_W=U\Lambda U^T \quad (15)$$

are determined. And all zero eigenvlaues and their associates are removed, so that Λ and U contain non-singular eigenvalues and eigenvectors.

The between-class scatter matrix O_B is computed and its eigenvectors is obtained as

$$O_B=V\Delta V^T \quad (16)$$

Only D-1 greatest eigenvectors of V is taken due to dimension reduction and the reduced eigenvector matrix V_C is built. Then the dimensional-reduced LDA transformation matrix is given by

$$W_{LDA}=U\Lambda^{-\frac{1}{2}}V_C \quad (17)$$

With the NASE matrix X, the LDA projected feature matrix is computed by the projection of X onto the LDA basis W_{LDA}

$$Y_{LDA}=W_{LDA}^T X \quad (18)$$

The projection W_{LDA} has then a dimension of $(D\text{-}1)\times N$. since the LDA basis is a global one, N is the number of all analysis frames of all classes.

IV. Evaluation

4.1. Datasets

To test the sound classification, we built sound libraries from various video sound tracks. These sounds are correlated with the background interference compared to general sound effects library. We created 13 sound classes (bird, sirens, bell, explosion, whistle, water, baby, laughter, guns, motor, cheering, music, applause) from the movie sound tracks and 2 speech classes (male speech, female speech) from the speech database of the broadcast news and TV panel discussion programs. 200 sound examples were collected for each class. 60% of the data was used for training and the other 40% for testing.

4.2. Feature Extraction and Classification

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 ASP 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, LDA or NMF methods is performed on the NASE features of all the audio frames from all the training examples in

the class. In order to evaluate the ASP feature sets, a left-right continuous HMM classifier with 7 states and a GMM classifier with 10 Gaussian densities were used with a variety of different sound sources.

4.3. 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 6, 12 and 23 reduced dimensions from the basis vectors. The sound recognition results are shown in Table 1.

Table 1. Sound Classification Accuracies (%).

Feature Extraction	Feature Dimension							
	HMM				GMM			
	6	12	23	Average	6	12	23	Average
PCA-ASP	73.54	83.67	92.32	83.17	87.32	89.17	96.2-	90.90
ICA-ASP	71.67	82.08	90.67	81.47	85.42	90.33	95.33	90.36
NMF-ASP	63.87	65.38	70.36	66.54	66.28	68.91	75.77	70.32
LDA-ASP	81.00	88.75	94.75	88.17	92.83	95.75	95.75	94.78
DCT-ASP	70.23	83.25	90.05	81.18	84.50	87.33	94.75	88.86

Regarding the classification of 15 sound classes GMM provides significantly better classification accuracies vs. HMM. Among the classification using different ASP feature extraction methods GMM using LDA-ASP features obtains the best recognition rates. With HMM best recognition rates are also reached by LDA-ASP features, PCA-ASP on the second ranking position. It is revealed clearly that with the same dimension of feature space LDA outperforms PCA by a great margin, because LDA aims at extracting the features which separate the classes to a maximal extent while PCA only focuses on the features which approximate the original feature space with the lowest mean-square-error and does not consider the discriminability of these features. The ASP projected onto NMF derived from NASE matrix yields the lowest recognition rate.

A comparison of the computational cost between the five feature extraction methods and between two statistical methods GMM and HMM is shown in the Table 2. With an impression of these costs, it is fairer to decide, which method is more suitable for particular real-world applications. For such decisions, it is often a trade-off between quality of the classification results, the complexity of the applied methods and the computational cost of these methods. To give practical values of computational cost, the time of execution is measured for feature extraction for training, for the training of GMM and HMM and for the classification. Exemplary, the data set sound for sound classification is used with 12 feature dimensions. The execution were measured on a personal computer with a Pentinum IV CPU, clocked with 2.53 GHZ and equipped with 512 MB RAM memory. To do a NMF, the divergence update algorithm was iterated 200 times. This algorithm converges very slowly in comparison with PCA, FastICA or LDA. Therefore, we did not measure the execution times of NMF-based classification.

Table 2. Execution times of feature extraction and statistical methods in second with 12 feature dimension (FeExtr-feature extraction, FeaTest: feature extraction of test audio and classification).

	PCA-ASP		ICA-ASP		LDA-ASP		DCT-ASP	
	HMM	GMM	HMM	GMM	HMM	GMM	HMM	GMM
FeExtr	31.47		42.67		34.42		25.08	
Trainging	39.56	846.38	45.34	903.2	44.78	785.67	40.06	789.80
FeaTest	58.53	26.41	62.31	26.41	42.94	24.41	41.89	23.56

As DCT-based ASP uses no signal dependent transform, the method is the fastest feature extraction methods of the four examined. PCA-based ASP has to compute per-class bases for dimension reduction, what is slightly less computational intensive as the computation of the global basis of LDA-based ASP. A few times more execution time is needed for training of GMMs than for training of HMMs as different numbers of model reestimation iterations are used and the training algorithms are implemented differently. However, the feature extraction of test audio data and classification with GMM is performed faster than with HMMs. Therefore, it could be interesting to use GMMs in applications, where training is done off-line and only the classification should be fast as possible for on line processing of audio data. The execution of the different feature extraction and classifications methods follows in general the results for the feature extraction of training. The PCA-based ASP with per-class projections is the slowest classification methods. The LDA-based ASP is approximately fast as the global transform of the feature space with the LDA basis is comparable with the DCT.

V. Conclusions

In this paper we compared the performance of Audio Spectrum Projection (ASP) features based on five basis decomposition algorithms.

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. On the other hand, NMF attempts a matrix factorization in which the factors have non-negative elements by performing a simple multiplicative updating. LDA finds the linear combination of features which best separate two or more classes of object or event.

Our results show that the ASP features based on LDA basis in general yield better performance compared to ASP projected on other basis functions in sound recognition. The NMF updating process is very slow compared to PCA, FastICA, LDA, DCT.

Further work focus on applying the audio segmentation to film indexing and the detection of sports video highlights using audio contents.

References

1. B. S. Manjunath, P. Salembier and T. Sikora, *Introduction to MPEG-7*, (Wiley 2002)
2. H.-G. Kim, N. Moreau, T. Sikora, *MPEG-7 Audio and beyond*, (Wiley 2005)
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 1996)
5. A. Hyvärinen, E. Oja, "Independent component analysis: algorithms and applications," Neural Networks, vol., 13, 411-430 (2000)
6. D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization" Adv. Neural Info. Proc. Syst. 13, 556-562 (2001).
7. N. Marhav and C.-H. Lee, "On the asymptotic statistical behavior of empirical ceptsral coefficients" in IEEE Transactions on Signal Processing 41, 1990-1993 (1993).
8. R. Duda, *Pattern classification*, (John Wiley 2001)

[Profile]

•Hyoung-Gook Kim

received the diploma degree in electrical engineering and the Ph. D. degree in computer science from the Technical University of Berlin, Berlin, Germany. From 1998 to 1999, he worked on speech recognition at Siemens AG. From 1999 to 2002, he was a Project Leader of the Speech Processing Laboratory at Cortologic AG, where he developed a noise reduction preprocessor and a 1.2 kb/s low-bit-rate speech coder for mobile voice communication. From 2002 to 2005, he served as Adjunct Professor of the Communication Systems Department, Technical University of Berlin. Since 2005 he joined Samsung Advanced Institute of Technology as a Project Leader. His research interests include audio signal processing, music information retrieval, audiovisual content indexing and retrieval, automatic segmentation, speech enhancement, and robust speech recognition.