Chapter 13

Intelligent Content-Based Audio Classification and Retrieval for Web Applications

Mingchun Liu, Chunru Wan, Lipo Wang
School of Electrical and Electronic Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
E-mail: {p147508078,ecrwan,elpwang}@ntu.edu.sg

Content-based technology has emerged from the development of multimedia signal processing and wide spread of web application. In this chapter, we discuss the issues involved in the content-based audio classification and retrieval, including spoken document retrieval and music information retrieval. Further, along this direction, we conclude that the emerging audio ontology can be applied in fast growing Internet, digital libraries, and other multimedia systems.

13.1 Introduction

One booming technology today is the Internet, due to its fast growing number of users and rich contents. With huge data storage and speedy networks becoming available, multimedia contents like image, video, and audio are fast increasing. Although there are powerful text search engines, such as Google, which is frequently resorted by users to search for their interested text webpages, their multimedia search ability is limited or problematic. This is because, unlike text documents, most of these audio-visual documents are not well organized and structured for machine processing. Normally, they can only be accessed by external properties, such as file names, authors, publishers, formats and etc, rather than their intrinsic contents directly, such as the genres of a music audio, the scenes of a movie video and so on.

Although the content characteristics can be annotated or indexed manually, this kind of job is tedious and time-consuming, as we often heard, one picture is worth a thousand words. Also due to the annotation, some emotional and environmental information are lost. For example, listening to speech can collect much more information than reading from the transcription. In addition, due to the ambiguity of the nature of multimedia contents, sometimes, it is not easy to describe them using words precisely. Thus, it may cause fatal problems during searching and retrieval.

In order to search and index these media effectively, various automatic content-based multimedia retrieval systems have been studied. Compared with its counterparts, such as image and video, there has been less work done for the content-based audio processing, which is partly due to the difficulties involved in representing and classifying non-speech audio. However, as audio is a compulsory part in an integrated multimedia scenarios like the MPEG, digital library and entertainment industry, more efforts need to be placed in the audio field for a well-balanced multimedia system or for a full-fledged audio database system alone. On the other hand, some of the existing techniques derived for image and video processing can be utilized for audio with necessary changes. Therefore, starting from the early 1990s, the content-based audio signal processing has raised great interests in the research and industry communities. The audio objects being studied include speech, music and general sound effects.

In this chapter, we focus on the automatic classification and retrieval of audio and construction of audio ontology for machine processing. Since the speech and music are the two major audio sources, we give a brief literature review of spoken document retrieval and music information retrieval in Section 2 and 3. Next, we consider the issues involved in general audio classification and retrieval, including audio feature extraction, relevance feedback techniques in Section 4 and 5 respectively. Then, based on increasing audio concepts adopted and agreed upon, we present audio ontology for machine processing and inter-operation in Section 6. Finally, we conclude intelligence audio retrieval in Section 7.

13.2 Spoken Document Retrieval and Indexing

Speech signal is the widest studied audio signal in the literature. With the advance of ASR (automatic speech recognition) and IR (information retrieval) techniques, various spoken document retrieval systems have been developed. The Cambridge university spoken document retrieval system was described by [Johnson et al., (1999)]. The retrieval performance over a wide range of speech transcription error rates was presented and a num-

ber of recognition error metrics that more accurately reflecting the impact of transcription errors on retrieval accuracy were defined and computed. [Viswanathan et al., (1999)] proposed another spoken documents retrieval system utilized both content and speaker information together in retrieval by combining the results. Instead of speech transcription in normal spoken document retrieval system, [Bai et al., (1996)] represented a very-large-vocabulary Mandarin voice message file retrieval using speech queries.

The index has been found very beneficial for retrieval, in which it makes the search process cost less time and produce more meaningful results. [Kurimo, (2002)] presented a method to provide a useful searchable index for spoken audio documents. The idea was to take advantage of the large size of the database and select the best index terms for each document with the help of the other documents close to it using a semantic vector space determined from the training of self-organizing map. [Makhoul et al., (2000)] described a system integrating the requisite speech and language technologies, called Rough'n'Ready, which indexed speech data, created a structural summarization, and provided tools for browsing the stored data.

13.3 Music Information Retrieval, Indexing and Content Understanding

Besides speech, music is another type of audio being extensively studied. The major music information processing includes music index, music retrieval, music understanding and music instrument classification as illustrated in Figure 13.1. The major issue is regarded as music retrieval by string matching using query-by-humming technique.

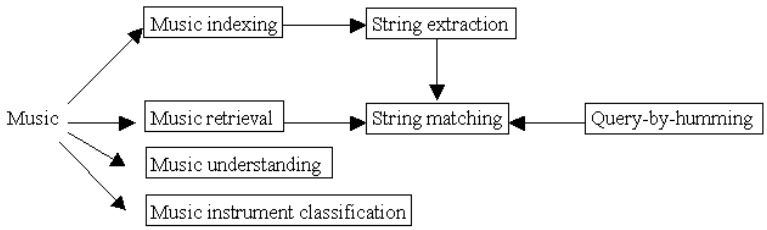


Fig. 13.1 The content-based music information processing

Usually, the extracted melody strings from original music data are adopted to represent the music itself. Hence, the problem of content-based music retrieval is transformed into a string-matching problem. [Tseng, (1999)] articulated an algorithm to extract the melody from MIDI files for retrieval. In that system, text strings were input as queries. The MIDI files were converted into text strings and compared to look for certain patterns in the query strings. [Liu et al., (1999)] extracted thematic feature strings, such as melody strings, rhythm strings, and chord strings, from the original music objects and treated them as the meta-data to represent their contents. A new approximate string-matching algorithm was proposed which provided fault tolerance ability according to the music characteristics.

In order to improve the music retrieval performance, several music index techniques have been proposed. [Chou et al., (1996)] implemented a music database system based on the chord-representation model and the PAT-tree index structure with "unstructured search" ability, where a chord was a combination of three or more notes which sound together in harmony and a PAT-tree was a Patricia-like tree constructed over all possible substrings of the chord. [Lap et al., (2000)] studied a number of issues regarding n-gram indexing of musical features using simulated queries.

Other relevant issues include music understanding and segmentation [Scheirer, (1999)], musical segmentation using hidden Markov models [Raphael, (1999)], and (MPEG layer III) digital music management [Pye, (2000)]. One particular area, the music instrument classification, has been raised notable interests and discussed in several papers [Kaminsky, (1995); Martin et al., (1998); Eronen et al., (2000); Miiva et al., (1999); Liu et al., (2001)]

13.4 Content-based Audio Classification and Indexing

Manually, it is natural to classify audio into hierarchical directory like the one shown in Figure 13.2. Firstly, the audio can be categorized into three broad classes which are speech, music and sound. Then, speech can be further classified to male and female speech or voice and unvoice speech according to different criteria. Music can be grouped into different genres and sound can be sorted into different environmental sounds, sound effects and so on. Researches have been conducted towards automatically building such tree-structure audio directory into different levels and branches according to applications.

On the first level of the directory, one simple yet important task of audio classification is to discriminate speech and music in audio clips. Different processings such as speech recognition and string matching can be further

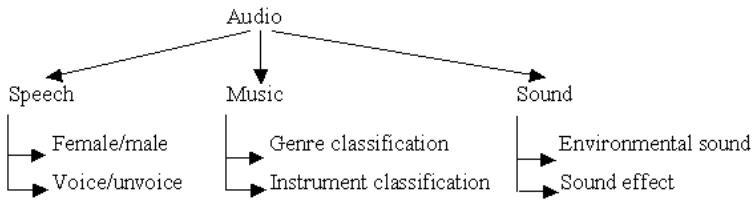


Fig. 13.2 A hierarchical audio classification

applied to the distinguished speech and music segments separately. For the rest of directory, the coarse to detail audio classes can be identified using corresponding features accordingly. For example, [Qi et al., (1999)] adopted a multi-layer feed-forward network based on hybrid features to conduct voiced-unvoiced-silence classification for speech signal. [Lambrou et al., (1998)] carried out a study to distinguish three different musical styles of rock, piano and jazz, using wavelet transform analysis in conjunction with statistical pattern recognition techniques. The environmental sound as one typical kind of general sound has raised attention because it can provide many contextual cues that enable us to recognize important aspects of our surroundings. A simple classification of five pre-defined classes of environmental sounds based on several extracted discriminating features was reported [Sawhney, (1997)]. [Zhang et al., (1998); Zhang et al., (1999)] utilized a hidden Markov model (HMM) to classify environmental sound into applause, explosion, bird's sound, and so on, in a hierarchical system based on the time-frequency analysis of audio. More recently, [Liu et al., (2002)] applied a fuzzy logic system to classify and retrieve audio clips, and further presented a framework to handle audio boolean search with multi-queries using concept adapted from fuzzy logic [Liu and Wan, (2002)]. Among the research conducted, we find that there are several works based on one common audio databases [Keislar et al., (1995); Wold et al., (1996); Li, (2000)]. We carry out our experiments of audio classification based on the same database which is described in more details below.

13.4.1 *The Audio Databases*

The database has two hierarchies of 16 classes. Among the 16 classes, there are two from speech (female and male speech), seven from music (percussion, oboe, trombone, cello, tubular-bell, violin-bowed, violin-pizzicato), and seven from other environmental sounds and sound effects. Thus, all

files have two labels, a coarse label from the three major classes: speech, music, and sound; a fine label from more specific classes.

Table 13.1 The audio database structure.

Class name	No of files	Class name	No of files
1.Speech	53	Violin-pizzicato(9)	40
Female(1)	36	3.Sound	62
Male(2)	17	Animal(10)	9
2.Music	299	Bell(11)	7
Trombone(3)	13	Crowds(12)	4
Cello(4)	47	Laughter(13)	7
Oboe(5)	32	Machines(14)	11
Percussion(6)	102	Telephone(15)	17
Tubular-bell(7)	20	Water(16)	7
Violin-bowed(8)	45	Total	414

13.4.2 *Audio Feature Extraction, Normalization and Selection*

In our databases, all audio files are in 'au' format, the sample rate of individual file is 8000Hz. The lengths of the sound files range from half second to less than ten seconds, in that short period, segmentation is omitted. During feature extraction process, each audio file is divided into frames of 256 samples, with 50% overlap at the two adjacent frames. If the energy of an individual frame is below a predefined threshold, the whole frame is marked as silence frame and is ignored for further processing. After silence reduction, the audio frames are hamming-windowed. Then, the mean and standard deviation of frame-level characteristics are calculated and features are extracted from time, frequency and coefficient domains and combined to form the feature vector to represent the individual audio file.

Time domain features include RMS (root mean square), ZCR (zero-crossing ratio), VDR (volume dynamic ratio), frame energy, total energy and silence ratio. Frequency domain features include frequency centroid, bandwidth, four sub-band energy ratios, pitch, salience of pitch, spectrogram, first two formant frequencies, and formant amplitudes. The first 13 orders of MFCCs (Mel-Frequency Cepstral Coefficients) and LPCs(Linear Prediction Coefficients) are adopted as coefficient features. A summary of the features are list in Table 13.2. After feature extraction, the extracted feature vectors are then normalized and ready for selection in classification and indexing. The details of feature extraction can be found in [Liu et al., (2001)].

Each audio feature is normalized over entire files in the database by

Table 13.2 The structure of extracted features.

1.Time domain (9 features)	Mean and standard deviation of volume root mean square (RMS), zero-crossing ratio (ZCR), frame energy; volume dynamic ratio (VDR), total energy and silence ratio.
2.Frequency domain(26 features)	Mean and standard deviation of frequency centroid, bandwidth, four sub-band energy ratios, pitch, salience of pitch, spectrogram, first two formant frequencies and amplitudes.
3.Coefficient domain(52 features)	Mean and standard deviation of first 13 orders of MFCCs (Mel-Frequency Cepstral Coefficients) and LPCs(Linear Prediction Coefficients).

subtracting its mean and dividing by its standard deviation. After normalization, different features have similar distribution over the entire files in the database. This normalization process will ensure more accurate results during classification. Then, each audio file is fully represented by its normalized feature vector.

Theoretically, we can use the exhausted combination method to pick up the best feature vector, but the computation complexity is huge. A suboptimum method, the sequential forward selection (SFS) method is adopted to select the best feature set. The process is as follows: select the best single feature and then add one feature at a time which in combination with the already selected features that minimize the classification error rate. We continue to do this until all the features are selected.

13.4.3 *Audio Classification Experimental Results*

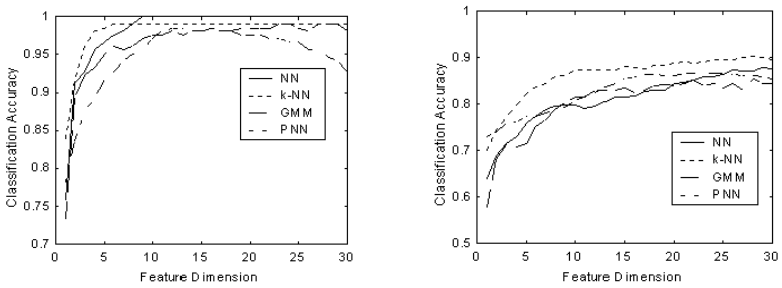
The database is split into two equal parts: one for training, and the other for testing. We conduct our experiments from various approaches including three statistical classifiers: Nearest Neighbor, modified k-Nearest Neighbor, Gaussian Mixture Model and one neural network classifier: the probabilistic neural network for audio classification.

- **Experiment 1:** Classifying the database into three major classes
- **Experiment 2:** Classifying the database into 16 classes

The most straightforward nearest neighbor rule can be conveniently used as a benchmark for all the other classifiers since it appears to always provide a reasonable classification performance in most applications. A variation of NN is the k-NN. Normally, the k samples in training set that are nearest to feature vector p are determined. The assignment of label to p is based on the majority of the labels of the k samples. In the modified k-NN, we firstly find the k nearest neighbors from each class instead of whole training set. Their means are calculated and compared, then assign the

testing feature vector with the class corresponding to the smallest mean. We set k to 4 in experiment 1, and set k to 2 in experiment 2. Usually, the pattern classification problem can be reduced to an estimation problem of a probability density function (*pdf*), since the classification can be performed according to the Bayes decision rule if a *posteriori* probability of the input pattern is obtained. The Gaussian mixture model has been proposed as a general model for estimating an unknown probability density function. While the PNN can be treated as a feed-forward network that implements a Gaussian mixture [Vlassis et al., (1999)]. In each experiment, the NN, k-NN, GMM, and PNN classifiers together with the SFS feature selection scheme are used to perform the classification task.

The classification accuracy versus feature dimension for the two classifiers in the experiments are shown in Figure 13.3. The overall and individual classification performances of these classifiers in each of the two experiments are given in Tables 13.3, and 13.4 respectively.



(a) Performance of experiment 1 (b) Performance of experiment 2

Fig. 13.3 The performances of audio classification

Table 13.3 The classification accuracy of experiment 1.

Class name	Test	NN	k-NN	GMM	PNN
1.Speech	26	26(100)	26(100)	26(100)	23(88.5)
2.Music	150	150(100)	150(100)	149(99.3)	150(100)
3.Sound	31	31(100)	29(93.5)	29(93.5)	29(93.5)
Total	207	207(100)	205(99.0)	204(98.6)	202(97.6)

From the figures, we can see that generally the classification performance increases rapidly with the increase of features at the beginning. After reaching a peak value, it remains more or less constant or may even decrease for certain classifiers. All classifiers reach their best performance at approxi-

Table 13.4 The classification accuracy of experiment 2.

Class name	Test	NN	k-NN	GMM	PNN
1.Speech	26	21(80.8)	22(84.6)	20(76.9)	18(69.2)
Female	18	14(77.8)	16(88.9)	15(83.3)	15(83.3)
Male	8	7(87.5)	6(87.5)	5(62.5)	3(37.5)
2.Music	150	141(94.0)	145(96.7)	133(88.7)	142(94.7)
Trombone	7	6(85.7)	7(100)	5(71.4)	5(71.4)
Cello	23	23(100)	23(100)	23(100)	23(100)
Oboe	16	14(87.5)	15(93.8)	13(81.3)	14(87.5)
Percussion	51	49(96.1)	49(96.1)	46(90.2)	49(96.1)
Tubular-bell	10	10(100)	10(100)	10(100)	9(90.0)
Violin-bowed	23	20(87.0)	22(95.7)	18(78.3)	23(100)
Violin-piz	20	19(95.0)	19(95.0)	18(90.0)	19(95.0)
3.Sound	31	23(74.2)	25(80.6)	21(67.7)	25(80.6)
Animal	4	3(75)	4(100)	4(100)	4(100)
Bell	4	3(75)	2(50)	2(50.0)	3(75.0)
Crowds	2	2(100)	2(100)	2(100)	2(100)
Laughter	3	3(100)	3(100)	1(33.3)	3(100)
Machines	6	2(33.3)	4(66.7)	2(33.3)	3(50.0)
Telephone	8	8(100)	8(100)	8(100)	8(100)
Water	4	2(50.0)	2(50.0)	2(50.0)	2(50.0)
Total	207	185(89.4)	192(92.8)	174(84.1)	185(89.4)

mately 20 features. This shows that the SFS feature selection procedure is an efficient method to quickly find out a small set of features to yield a satisfactory result among a large set. Thus, for simplicity and fair comparison, all the listed classification accuracies in the two tables are achieved by their corresponding classifiers using 20 features selected from SFS method. Note that, the best 20 feature sets for different classifiers are different. In particular, during the second experiment, our k-NN classifier with 28 features selected by the SFS method, yields 93.72% accuracy, as compared to that of 90.4% by nearest feature line (NFL) using 34 features [Li, (2000)].

13.5 Content-based Audio Retrieval

The goal of content-based audio retrieval is to find documents from audio database which satisfy certain user's requirements regarding to his/her query. A typical situation is to search for audios sound similar to the proposed query example based on distance measurement of their extracted features. The best audio search engine would retrieve similar sounds on top of the similarity ranking list while leave the dissimilar ones at the bottom.

The pioneer work for retrieval of general audio database was done by [Keislar et al., (1995)], where they claimed that many audio and multi-

media applications would benefit from the ability to classify and search for audio based on the characteristics of the audio rather than by resorting exclusively to keywords. They built such a prototype audio classification and retrieval system which led the research along this direction [Wold et al., (1996)]. In that system, sounds were reduced to perceptual and acoustical features, which let users search or retrieve sounds by different kinds of query. [Li, (2000)] presented a new pattern classification method called the nearest feature line (NFL) for equivalent task. Experiments were carried out based on the same database with lower error rate achieved. Other works in content-based audio retrieval can be found in the literature [Liu et al., (2000); Foote, (1997); Smith et al., (1998); Kashino et al., (1999); Kashino et al., (2000); Johnson et al., (2000); Piamsa-Nga et al., (1999); Zhang et al., (1999); Melih et al., (1998); Melih et al., (1998)].

A full-fledged procedure of an integrated audio retrieval system is illustrated in Figure 13.4. Raw audio recordings are analyzed and segmented based on abrupt changes of features. Then audio segments are classified and indexed. They are stored in corresponding archives. The audio archives can be organized in a hierarchical way for the ease of the storage and retrieval of audio clips. When a user wants to browse the audio samples in the archives, he/she may put a set of features or a query sound into the computer. The search engine will then find the best matched sounds and present them to the user. The user may also give feedbacks to get more audio material relevant to his/her interest.

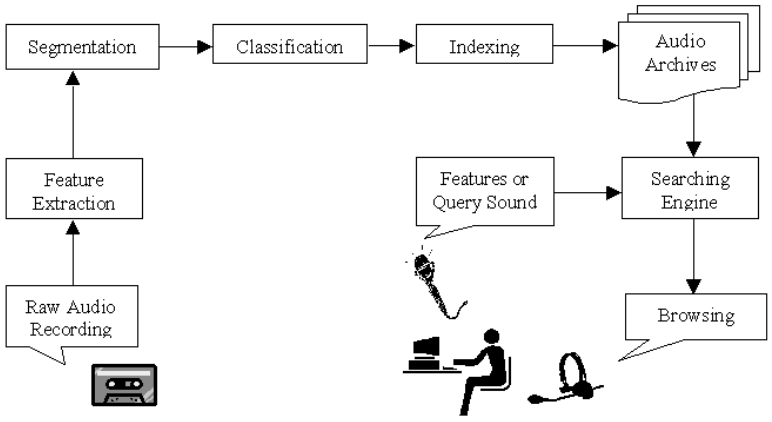


Fig. 13.4 The procedure of content-based audio retrieval

13.5.1 *Performance Measurement*

The performance of retrieval is measured by precision and recall [Grossman et al., (1998)]. Precision and recall are defined as follows:

$$Precision = \frac{Relevant\ Retrieved}{Total\ Retrieved} \tag{13.1}$$

$$Recall = \frac{Relevant\ Retrieved}{Total\ Relevant} \tag{13.2}$$

Another measurement is the average precision (AP), which is used as one single indicator of retrieval performance, which refers to an average of precision at various points of recall. We assume that the files in a same class are relevant, and otherwise they are non-relevant. Hence, the performance can be measured automatically without hearing the sound.

13.5.2 *Audio Retrieval Experimental Results*

Simple retrieval scheme directly uses the distance between the query example and individual samples, and the retrieval list is given based on the distance measurement. It is searched through the whole database, many non-relevant sounds are retrieved and the speed is slow, especially when the database is growing large. To increase the search speed and retrieve more relevant files to the query, a probabilistic neural network (PNN) classification combined with Euclidean distance measurement scheme for retrieval is proposed.

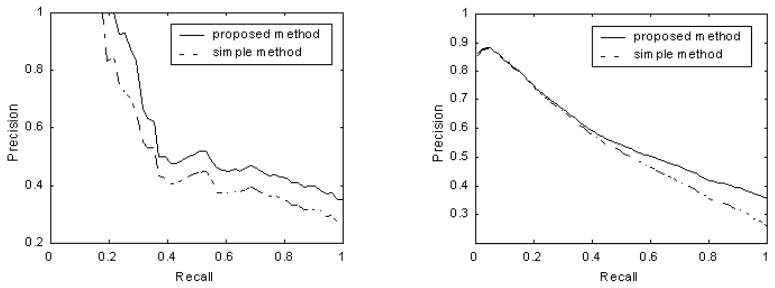
With this hierarchical retrieval strategy, query sound is firstly classified into one of three major classes, namely music, speech and sound by the calculated probabilities from PNN. Then the distances between query and samples in that class instead of whole database are measured and an ascending distance list is given as retrieval result. With this approach, many non-relevant documents are avoid before the search begins.

Key advantages of PNN are that training requires only a single pass and the decision surfaces are guaranteed to approach the Bayes optimal decision boundaries as the number of training samples grows. PNN is easy to use and extremely fast for moderate-sized database. For very large databases and for mature applications in which classification speed is more important than training speed, a complimentary probabilistic neural network, polynomial adaline can be used instead [Johnson et al., (2000)]. Based on these facts, PNN is chosen as the classifier for the first stage of retrieval. Following experiments are conducted by the simple distance method and the

proposed scheme. The PNN classification used in this section is the same as introduced in previous section with first 20 best features selected.

Same as the experiment in audio classification, the database is split into two equal parts: one for training, and the other for testing. Every file in testing sets are used as query sound and submit to the search engine one by one. One typical and mean precision-recall curve by the proposed method and direct method are shown in Figures 13.5. The curve of proposed method is above the curve of the direct distance method, which means if we recall same number of documents, the proposed method can retrieval less irrelevant files. It also means if we retrieval same number of total documents using both methods, among them, more relevant files are retrieved by proposed method than by direct method. The average precision retrieved by proposed scheme and the simply distance method are 0.57, and 0.53. These results show that the proposed retrieval scheme yields better overall performance than direct distance method in both recall-precision relation and average precision.

Most often, people only browse the files rank in the top list. For this concern, top ten retrieved files for several queries are listed in Table 13.5. Search method 'A' means the direct distance method, while search method 'B' means the proposed method. As we can see in the table, shown in the second and third column of Table 13.5, the top ten searching results by direct distance method for a male speech query, are from the classes of male speech, percussion, female speech, machine, cello, violin-pizzicato, and so on according to similarity. The top ten results by proposed method for the same query are from the classes of male speech and female speech only. There is only 1 relevant retrieved by direct method. While there are 2 files retrieved from the same class of query using proposed method, and all the results are in the query's up-level coarse class "speech".



(a) One typical P-R curve (b) The mean P-R curve

Fig. 13.5 Precision-recall (P-R) curve of audio retrieval

Table 13.5 Experimental result of audio retrieval.

Query	Male		Oboe		Animal	
Method	A	B	A	B	A	B
No.1	Male	Male	Oboe	Oboe	Animal	Animal
No.2	Perc	Female	Oboe	Oboe	Telephone	Telephone
No.3	Perc	Female	Oboe	Oboe	Telephone	Telephone
No.4	Female	Female	Telephone	Oboe	Telephone	Telephone
No.5	Female	Female	Oboe	Trombone	Violin-piz	Animal
No.6	Machines	Female	Trombone	Oboe	Trombone	Telephone
No.7	Cello	Female	Telephone	Oboe	Animal	Laughter
No.8	Perc	Female	Telephone	Oboe	Telephone	Telephone
No.9	Violin-piz	Male	Oboe	Trombone	Trombone	Machines
No.10	Perc	Female	Oboe	Oboe	Perc	Animal
Relevant No	1	2	6	8	2	3

13.5.3 *Content-based Audio Retrieval With Relevance Feedback*

In both the proposed and direct audio retrieval schemes introduced at previous section, the user interaction is not considered. However, the user involvement may be crucial to achieve a better performance. Basically, the purpose of relevance feedback is to move relevant files ranking to the top and irrelevant files ranking to the bottom progressively. In principle, there are two strategies to apply user's feedback information. One is to update the weights in the similarity measurement and the other is to refine the query [Liu et al., (2003)]. Here, we focus on the former approach.

13.5.3.1 *Proposed Relevance Feedback Algorithm*

Suppose that we have obtained the relevance audios set R_{rel}, which includes the query example q and relevance feedbacks $f^j, j = 1, \cdots, M$, where M is the number of relevant feedbacks. If we can decrease the sum of the square weighted $L2$ distance $\sum_{j \in R_{rel}} \rho^2(f^j, q : w)$ between relevance feedbacks and the query example, more relevant audios may emerge on the top of the retrieval list because of their similar feature characteristics. The weighted $L2$ distance is defined as follows.

$$\rho(\mathbf{f}^j, \mathbf{q} : \mathbf{w}) = (\sum_{i=1}^{N} w_i (f_i^j - q_i)^2)^{1/2} \tag{13.3}$$

where the subscript i refers to the ith feature element, the superscript j refers to the jth file in the relevant set. Based on this observation, we

consider minimizing the following objective function:

$$f(\mathbf{w}) = \mathbf{w}^T\mathbf{D}\mathbf{w} + \varepsilon\mathbf{w}^T\mathbf{w}, \text{subject to } \mathbf{c}^T\mathbf{w} = 1 \tag{13.4}$$

where ε is a positive constant, and

$$\mathbf{D} = diag\{d_1, \cdots, d_N\} = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_N \end{pmatrix}, \mathbf{c} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{13.5}$$

Here, $d_i = \sum_{j \in R_{rel}} d_{ji}^2$ and d_{ji} is the distance between ith feature component of the jth relevance feedback and query example, i.e. $d_{ji} = f_i^j - q_i$. The term $\varepsilon\mathbf{w}^T\mathbf{w}$ is introduced to avoid very large variation of w. This is a typical constrained optimization problem, which can be solved by the Lagrangian method. The solution of the constrained optimization problem is given by

$$\mathbf{w} = \frac{\mathbf{R}^{-1}\mathbf{c}}{\mathbf{c}^T\mathbf{R}^{-1}\mathbf{c}} = \frac{1}{(r_1^{-1} + \cdots + r_N^{-1})} \begin{pmatrix} r_1^{-1} \\ \vdots \\ r_N^{-1} \end{pmatrix} \tag{13.6}$$

which has an equivalent form as follows:

$$w_i = \frac{1}{(r_1^{-1} + \cdots + r_N^{-1})} r_i^{-1} \tag{13.7}$$

where $\mathbf{R} = \mathbf{D} + \varepsilon\mathbf{I}$ and $r_i = d_i + \varepsilon$. In the case of negative feedback, the objective function can be adjusted as follow:

$$f(\mathbf{w}) = \mathbf{w}^T\mathbf{D}\mathbf{w} - \beta\mathbf{w}^T\mathbf{D}'\mathbf{w} + \varepsilon\mathbf{w}^T\mathbf{w} + \lambda(\mathbf{c}^T\mathbf{w} - 1) \tag{13.8}$$

where β is a positive number and it is usually small to reduce the effect of negative feedback compared to positive feedback. $\mathbf{D}' = diag\{d_1', \cdots, d_N'\}$, where $d_i' = \sum_{j \in R_{irrel}} d_{ji}'^2$, and R_{irrel} is defined as the irrelevance or negative feedback audio set, $d_{ji}' = f_i'^j - q_i$ and f'^j is a negative feedback in the set R_{irrel}. In this case, the solution to Eq. 13.8 has the same form as in Eq. 13.6 and Eq. 13.7 with $\mathbf{R}$ being replaced by $\mathbf{R} = \mathbf{D} - \beta\mathbf{D}' + \varepsilon\mathbf{I}$. In our experiments, we set $\alpha = 0.5, \beta = 0.1, \varepsilon = 0.5$ empirically.

13.5.3.2 *Query Refining and Other Issues*

As we have mentioned earlier, another way to conduct feedback is to modify query accordingly. It aims to adjust the query for better representation of the user's information request. Intuitively, we use the mean of the relevant retrieved files (including the query example) to form the new query in our

experiments. The number of relevant files to select is another issue for consideration during feedbacks. In the experiments, we chose 1 to 3 files. This is because we notice that normally users are only willing to provide a few feedbacks and too many feedbacks can't give much further performance improvement.

13.5.3.3 *Experimental Results*

In the retrieval system with feedback, different users may have different opinions and may choose different files as feedbacks or even determine the same file as relevance or irrelevance. Hopefully, since files in same class are already assumed as relevant, we can mark those files from most similar to least similar automatically. Therefore, the first 1-3 files are used as relevance feedback for weight updating. Thus, ambiguity of relevance judgment is avoided and experiments are conducted in a fully automatic way.

In most cases, however, users don't have patience to listen to all the possible retrieved files. Normally, they only interest in several files ranking at the top. Thus, AP is calculated again based on top T (T=15) retrieved files considered. We call it AP(15), and defined as follows:

$$AP(15) = \frac{1}{TopR} \sum_{i=1}^{TopR} Precision(i) \tag{13.9}$$

where $TopR$ is the number of relevant files ranking at Top 15 retrieved files, $Precision(i)$ is the precision at ith relevant files retrieved. This AP(15) may be a more accurate indicator for practical retrieval performance.

The mean APs and AP(15)s of the tests on the two databases are measured and listed in Tables 13.6. The retrieval performance without feedback is measured at the beginning. The original mean AP is 0.485, while mean AP(15)s is 0.852 when the top 15 files are considered only. From the Table, we can see that when first 3 relevant files are chosen as relevance feedback, the AP performance can increase to 0.59 and the mean AP(15) performance increase to 0.946 using our relevance feedback algorithm.

In order to show the overall performance improvement rather than particular one, the AP difference of the database after and before 1st iteration of feedbacks with 3 relevant files selected with query updating are shown in Figures 13.6 and 13.7. Figure 13.6 considers the whole retrieved files, while Figure 13.7 considers the top 15 retrieved files only. The bar above the horizontal zero line means that the AP after feedback is higher than the AP before feedback and vice versa. We can clearly see that in most cases, the performances after feedbacks are better.

Table 13.6 The AP and AP(15) performance of the relevance feedback.

	AP:0.485	AP(15):0.852
1 Files	0.52	0.892
2 Files	0.558	0.924
3 Files	0.59	0.946

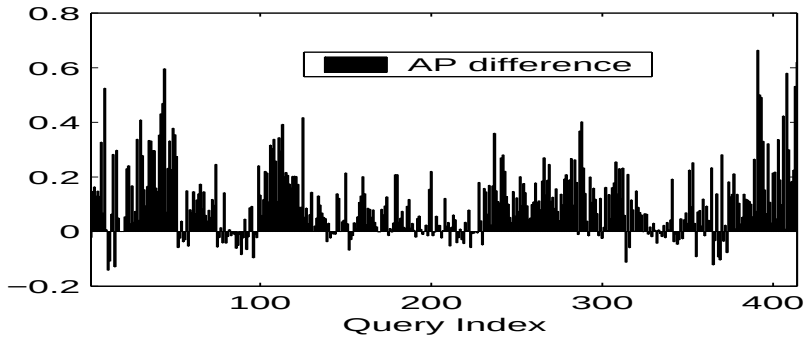


Fig. 13.6 The difference in AP of retrieval performance

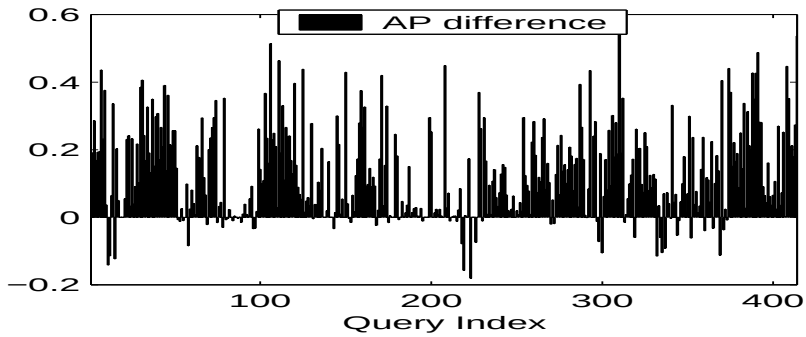


Fig. 13.7 The difference in AP(15) of retrieval performance

13.6 Audio Retrieval Based on the Concepts of Audio Ontology and Audio Item

There are growing audio databases becoming available from diversified web resources. Correspondingly, different application dependent and heuristic content-based audio systems have been built up. However, it is neither

possible to make the audio processing system always keep pace with its expanding data nor to manually implement new database systems individually from scratch in a short period. Therefore, the re-usability and inter-operability with existing domains become crucial and beneficial. The solution is dependent upon the establishment of a semantic web–an extended web of machine-readable information and automated services that extend far beyond current capabilities [Lee et al., (2001)]. The semantic web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding.

13.6.1 *Ontology and Audio Ontology*

Ontology is a formal, explicit specification of a shared conceptualization [Gruber, (1993)]. In this context, "conceptualization" refers to an abstract model of some phenomenon in the world that identifies that phenomenon's relevant concepts. "Formal" means that the ontology should be machine understandable. "Explicit" means that the type of concepts used and the constraints on their use are explicitly defined, and "shared" reflects the notion that an ontology captures consensual knowledge. Ontology also specifies the relations among concepts, which provides a formal vocabulary for information exchange. Specific instances of the concepts defined in the ontologies paired with ontologies constitute the basis of the semantic web.

With in-depth manual audio catalog and corresponding automatic audio classification, it is feasible to represent semantics of audio using audio ontology because there already exists some common audio semantics with high-level meanings derived from various content-based audio systems. Although it is still unlikely to create a universal acceptable machine-processable audio retrieval system in the near future, audio domain ontology can be constructed towards such direction with an open architecture. In particular, an audio domain-specific ontology was utilized to improve the accuracy (precision and recall) and communication effectiveness of a database system response to a user information request in [Khan et al., (2000)]. The ontology was employed along with user profile information, to automatically select and deliver appropriate information units from a multimedia audio databases.

An audio ontology is a formal explicit description of concepts in audio domain(classes are sometimes called concepts), properties of each concept describing various features and attributes of the concept (slots are sometimes called roles or properties), and restrictions on slots (facets are sometimes called role restrictions). The audio ontology together with a set of individual instances of classes constitutes an audio knowledge base.

Classes are the focus of most ontologies. Classes describe concepts in

the domain. For example, a class of **audios** represents all audios. Specific audios are instances of this class. The music audio streaming from the Internet is an instance of the class of **music audio**. A class can have subclasses representing concepts that are more specific than the superclass. For example, we can divide the class of all audios into speech, music, and sound yet speech can be further divided into female and male speech.

Slots describe properties of classes and instances. For a particular female speaker such as Mary with the speech "univerisity" in the audio database, it has a female gender and is spoken by a female author named Mary. We have two slots describing the female speech in this example: the slot **gender** with the value female and the slot **speaker** with the value Mary. At the class level, we can say that instances of the class female speech will have slots describing their name, copyright, encoding, length, the speaker of the female speech and so on.

All instances of the class speech, and its subclass female speech, have a slot **speaker**, whose value "Mary" is an instance of the class **author** as shown in Figure 13.8. All instances of the class **author** have a slot **produces** that refers to all the speeches (instances of the class speech and its subclasses) that the speaker produces.

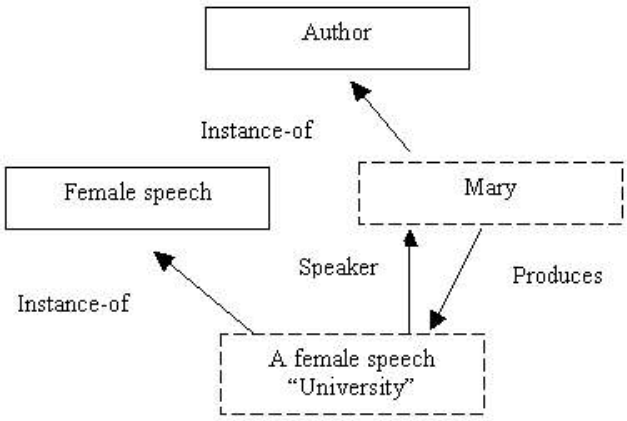


Fig. 13.8 A female speech ontology example in the audio domain

13.6.2 *MPEG-21 and Audio Item*

Many elements (standards) exist for delivery and consumption of multimedia contents, but there was no "big picture" to describe how these elements relate to each other. MPEG-21, the newest standard in MPEG family, will fill the gaps and allow existing components to be used together, thereby increasing interoperability. MPEG-21 defines a multimedia framework to enable electronic creation, delivery, and trade of digital multimedia content by providing access to information and services from anywhere at anytime. The fundamental concept in MPEG-21 framework is the 'digital item', a structured digital object with a standard representation, identification and metadata. The digital item incudes three parts, resource (individual asset), metadata (data about or pertaining to the item), and structure (relationships among the parts of the item). For example, a digital item called 'audio music album' can has its mp3 songs (possibly encoded for different bit-rates as its resource), text lyrics and artist biography, intellectual property information as its metadata, and links to web page to purchase the album as its structure. Using the available audio ontology and the concept of digital item, we can define the so-called 'audio item' for audio retrieval.

13.6.3 *Audio Retrieval using Audio Item with Ontology*

Here, we illustrate a framework for audio retrieval using the concept of "audio item" with ontology and all the available content-based audio processing techniques, shown in Figure 13.9. The audio item is built by segmentation, classification (speech, music and sound discrimination), and different further treatment depended on the classification result. For example, the speech recognition and word spotting (identification of keywords) can be conducted for speech, while music classification such as instrument recognition can be carried out for music, and sound spotting (identification of predefined sound classes) can be performed for sound with necessary user annotation when applicable. Based on all these processings, an audio item is constructed including (1) raw audio file and extracted features as its resource; (2) proper ID (including URL and Title), the start and end time obtained from segmentation, descriptions obtained from the corresponding classification procedure, manually added copyright information as its metadata; and (3) the audio ontology as its structure. Note that the steps in dashline blocks need user interactions. During the audio item retrieval process, the search engine can utilized both power of text search ability based on its structure and metadata and content-based query-by-example search ability based on its resource. For example, the search engine can go through text search based on ontology to see whether there is a match.

Then, it can perform content-based similarity distance calculation in that category instead of the whole database to retrieve audio files to meet with user's requirements.

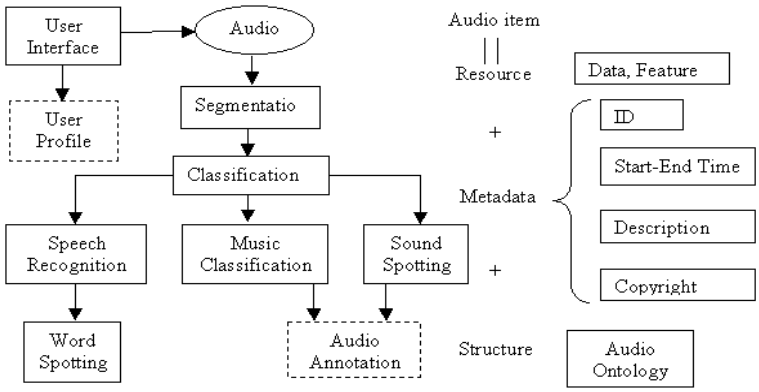


Fig. 13.9 The framework for audio retrieval using "audio item" concept with ontology

13.7 Conclusions and Outlook

Motivated by the ambition to build an audio search engine, which allow users to retrieval their interested audio files just as convenient as retrieval text documents from today's Internet, we discuss techniques in automatic content-based audio classification and retrieval. The content-based audio signal processing is conducted from speech, music to general audio signals. Based on extract features, the statistical and neural networks approaches for audio classifications are carried out. For audio retrieval, a hierarchical retrieval scheme and relevance feedback techniques are proposed to improve performance. Then, the utilization of audio ontology and audio item in audio retrieval is demonstrated.

Future content-based audio retrieval systems should be adaptable to provide access to any content or semantic concept such as 'happy music', not just limited concepts predefined by the system. This way, more intelligence can be added towards building a smart online audio retrieval systems with adaption to fast growing Internet. The ultimate goal of content-based audio processing is to make the audio can be managed as similar as text document, where user can not only hear it at ease but also "read" it inside out.

Bibliography

Bai Bo-Ren and Chien Lee-Feng and Lee Lin-Shan. (1996). Very-large-vocabulary Mandarin voice message file retrieval using speech queries, *Fourth International Conference on Spoken Language.* **3**, pp. 1950–1953.

Chou Ta-Chun and Chen, A.L.P. and Liu Chih-Chin. (1996). Music databases: indexing techniques and implementation, *Proceedings of International Workshop on Multimedia Database Management Systems.* pp. 46–53.

Eronen, A. and Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features, *IEEE International Conference on Acoustics, Speech, and Signal Processing.* **2**, pp. 63–71.

Foote Jonathan T.. (1997). Content-Based Retrieval of Music and Audio, *Proc. of SPIE.* **3229**, pp. 138–147.

Grossman David A. and Frieder Ophir, (1998). Information Retrieval: Algorithms and Heuristics, *Kluwer Academic Publishers.*

Gruber T.R. (1993). A translation approach to portable ontology specifications, *Knowledge Acquisition.* **5**, 2, pp. 199–220.

Johnson, S.E. and Jourlin, P. and Moore, G.L. and Jones, K.S. and Woodland, P.C (1999). The Cambridge University spoken document retrieval system, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* **1**, pp. 49–52

Johnson, S.E. and Woodland, P.C. (2000). A method for direct audio search with applications to indexing and retrieval, *IEEE International Conference on Acoustics, Speech and Signal Processing.* **3**, pp. 1427–1430.

Kaminsky, I. and Materka, A. (1995). Automatic source identification of monophonic musical instrument sounds, *IEEE International Conference on Neural Networks.* **1**, pp. 189–194.

Kashino, K. and Smith, G. and Murase, H. (1999). Time-series active search for quick retrieval of audio and video, *IEEE International Conference on Acoustics, Speech, and Signal Processing.* **6**, pp. 2993–2996.

Kashino, K. and Kurozumi, T. and Murase, H. (2000). Feature fluctuation absorption for a quick audio retrieval from long recordings, *15th International Conference on Pattern Recognition.* **3**, pp. 98–101.

Keislar, D. and Blum, T. and Wheaton, J. and Wold, E. (1995). Audio Databases

with Content-Based Retrieval, *the International Computer Music Conference.* pp. 199–202.

Khan, L. and McLeod, D. (2000). Audio structuring and personalized retrieval using ontologies, *Proceeding of IEEE Advances in Digital Libraries.* pp. 116–126.

Kurimo Mikko. (2002). Thematic indexing of spoken documents by using self-organizing maps, *Speech Communication.* **38**, 1-2, pp. 29–44.

Lambrou T. and Kudumakis P. and Sandler M. and Speller R. and Linney A.. (1998). Classification of Audio Signals using Statistical Features on Time and Wavelet Transform Domains, *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Lap Yip Chi and Kao B. (2000). A study on n-gram indexing of musical features, *Proceedings of IEEE International Conference on Multimedia and Expo.* **2**, pp. 869–872.

Lee Berners T. and Hendler J. and Lassila O., (2001). The Semantic Web, *Scientific American.* pp. 34–43.

Li S. Z.. (2000). Content-based audio classification and retrieval using the nearest feature line method, *IEEE Transactions on Speech and Audio Processing.* **8**, 5, pp. 619–625.

Liu Chih-Chin and Hsu Jia-Lien and Chen A.L.P. (1999). An approximate string matching algorithm for content-based music data retrieval, *IEEE International Conference on Multimedia Computing and Systems.* **1**, pp. 451–456.

Liu Mingchun and Wan Chunru. (2001). A study on content-based classification and retrieval of audio database, *International Database Engineering and Application Symposium.* pp. 339–345.

Liu Mingchun and Wan Chunru. (2001). Feature Selection for Automatic Classification of Musical Instrument Sounds, *Proceeding of ACM/IEEE Joint Conference on Digital Library'01(JCDL'01).* pp. 247–248.

Liu Mingchun and Wan Chunru. (2002). Boolean Search for Content-based Audio Retrieval Using Fuzzy Logic, *Proceeding of 1st International Conference on Fuzzy Systems and Knowledge Discovery(FSKD'02).*

Liu Mingchun and Wan Chunru and Wang Lipo. (2002). Content-Based Audio Classification and Retrieval Using A Fuzzy Logic System: Towards Multimedia Search Engines, *Journal of Soft Computing.* **6**, 5, pp. 357–364.

Liu Mingchun and Wan Chunru, (2003). Weight Updating for Relevance Feedback in Audio Retrieval, *Proceeding of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*

Liu Zhu and Huang Qian. (2000). Content-based indexing and retrieval-by-example in audio, *IEEE International Conference on Multimedia and Expo.* **2**, pp. 877–880.

Makhoul, J. and Kubala, F. and Leek, T. and Liu Daben and Long Nguyen and Schwartz, R. and Srivastava, A. (2002). Speech and language technologies for audio indexing and retrieval, *Proceedings of the IEEE.* **88**, 8, pp. 1338–1353.

Martin, K. D. and Kim, Y. E. (1998). Musical instrument identification: a pattern-recognition approach, *the 136th Meeting of the Acoustical Society*

of America.

Melih. K and Gonzalez, R, (1998). Audio retrieval using perceptually based structures, *IEEE International Conference on Multimedia Computing and Systems.* pp. 338–347.

Melih Kathy and Gonzalez Ruben, (1998). Identifying Perceptually Congruent Structures for Audio Retrieval, *5th International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services.* pp. 125–136.

Miiva, T. and Tadokoro, Y. (1999). Musical pitch estimation and discrimination of musical instruments using comb filters for transcription, *2nd Midwest Symposium on Circuits and Systems.* **1**, pp. 105–108.

Piamsa-Nga, P and Alexandridis, N.A and Srakaew, S and Blankenship, G.C. and Jr, Subramanya, S.R. (1999). In-clip search algorithm for content-based audio retrieval, *Third International Conference on Computational Intelligence and Multimedia Applications.* pp. 263–267.

Pye, D. (2000). Content-based methods for the management of digital music, *IEEE International Conference on Acoustics, Speech, and Signal Processing.* **4**, 4, pp. 2437–2440.

Qi, Y. and Hunt, B.R. (1999). Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier, *IEEE Transactions on Speech and Audio Processing.* **1**, 2, pp. 250–255.

Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **21**, 4, pp. 360–370.

Sawhney Nitin. (1997). Situational Awareness from Environmental Sounds, *MIT Media Lab.*

Scheirer, E.D. (1999). Towards music understanding without separation: segmenting music with correlogram comodulation, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.* pp. 99–102.

Smith, G and Murase, H and Kashino, K. (1998). Quick audio retrieval using active search, *IEEE International Conference on Acoustics, Speech and Signal Processing.* **6**, pp. 3777–3780.

Tseng Yuen-Hsien. (1999). Content-Based Retrieval for Music Collections, *SIGIR: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pp. 176–182.

Viswanathan, M. and Beigi, H.S.M. and Dharanipragada, S. and Tritschler, A (1999). Retrieval from spoken documents using content and speaker information, *Proceedings of the Fifth International Conference on Document Analysis and Recognition.* pp. 567–572

Vlassis N and Likas A. (1999). Kurtosis-based dynamic approach to Gaussian mixture modeling, *IEEE Transactions on Systems, Man and Cybernetics, Part A.* **29**, pp. 393–399.

Wold, E. and Blum, T. and Keislar, D. and et al. (1996). Content-based classification, search, and retrieval of audio, *IEEE Multimedia.* pp. 27–36.

Zhang Tong and Kuo C.-C. Jay. (1998). Hierarchical system for content-based audio classification and retrieval, *SPIE's Conference on Multimedia Storage*

and Archiving Systems III. **3527**, 2, pp. 398–409.

Zhang Tong and Kuo C.-C. Jay. (1999). Classification and retrieval of sound effects in audiovisual data management, *the 33rd Annual Asilomar Conference on Signals, Systems, and Computers.* **3527**, 2, pp. 398–409.

Zhang Tong and Kuo C.-C. Jay. (1999). Hierarchical classification of audio data for archiving and retrieving, *IEEE International Conference On Acoustics, Speech, and Signal Processing.* **6**, pp. 3001–3004.

ISO/IEC JTC1/SC29/WG11 (2001). Vision, Technologies and Strategy, MPEG Document: ISO/IEC JTC1/SC29/WG11, *ISO/IEC TR 21000-1 Part 1.* **N3939**.

http://www.cselt.it/mpeg.