Focus

Soft Computing 6 (2002) 357 – 364 © Springer-Verlag 2002
10.1007/s00500-002-0189-3

Content-based audio classification and retrieval using a fuzzy logic system: towards multimedia search engines

M. Liu, C. Wan, L. Wang

Abstract In recent years, available audio corpora are rapidly increasing from fast growing Internet and digital libraries. How to classify and retrieve sound files relevant to the user's interest from large databases is crucial for building multimedia web search engines. In this paper, content-based technology has been applied to classify and retrieve audio clips using a fuzzy logic system, which is intuitive due to the fuzzy nature of human perception of audio, especially audio clips with mixed types. Two features selected from various extracted features are used as input to a constructed fuzzy inference system (FIS). The outputs of the FIS are two types of hierarchical audio classes. The membership functions and rules are derived from the distributions of extracted audio features. Speech and music can thus be discriminated by the FIS. Furthermore, female and male speech can be separated by another FIS, whereas percussion can be distinguished from other music instruments. In addition, we can use multiple FISs to form a "fuzzy tree" for retrieval of more types of audio clips. With this approach, we can classify and retrieve generic audios more accurately, using fewer features and less computation time, compared to other existing approaches.

Keywords Fuzzy logic, Audio classification, Audio retrieval, Feature extraction, Query-by-example

1 Introduction

Web surfers frequently use text search engines such as Yahoo and Lycos to find their wanted web pages; however, commercial search engines for multimedia database, especially for audios, are lacking. Users can benefit from the ability to directly search these media, which contain rich information but could not be precisely described by text. Hence, content-based indexing and retrieval technologies are the first crucial step towards building such multimedia search engines.

In recent years, much research has been conducted on content-based audio classification and retrieval, as well as in other relevant fields, such as audio segmentation, indexing, browsing and annotation. Generally, audio can be categorized into three major classes: speech, music, and sound. Different techniques have been employed to process these three types of audios individually. Speech signals are the best studied. With automatic speech recognition systems becoming mature, speech and spoken document retrievals are often carried out by transforming the speeches into texts. Traditional text retrieval strategies are then used [1–3].

Music retrieval is sometimes treated as a string matching problem. In [4], a new approximate string matching algorithm is proposed to match feature strings, such as melody strings, rhythm strings, and chord strings, of music objects in a music database. Kosugi et al. [5] described a retrieval system that enables a user to obtain the name of a desired song from an audio database by humming a part of a melody as a query. A music information retrieval system dealing with MIDI files using complex-valued recurrent neural networks is proposed in [6].

Besides speech and music, general sounds are the third major type of audios. Some research has been devoted to classification of this kind of audios, and others focus on even more specific domains, such as classification of piano sounds [7] and ringing sounds [8].

In spite of different techniques applied in the audio classification and retrieval process, the underlying procedure is the same, which can be divided into three major steps, audio feature extraction, classifier mapping, and distance ranking.

The first step towards these content-based audio database systems is to extract features from sound signals. Features can be extracted from time, frequency, and coefficient domains. Time domain features include root mean square (RMS), silence ratio, zero-crossing ratio, and so on. Frequency domain features include spectral centroid, bandwidth, pitch, etc. The mel-frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC), which are widely used in speech recognition, are also adopted for classification of general sounds. Recently, some researchers have paid much attention to wavelet coefficients [9, 10]. They argued that the multi-resolution property of wavelet coefficients and their better time-frequency resolution are suitable for indexing and searching.

Based on the features extracted, various classifiers can then be used for sound classification. In [11], a multidimensional Gaussian maximum *a posteriori* (MAP) estimator, a Gaussian mixture model (GMM) classifier, a spatial partitioning scheme based on a *k–d* tree, and a nearest neighbor classifier were examined in depth to discriminate speech and music. In [12], a threshold-based heuristic rule procedure was developed to classify generic

M. Liu, C. Wan (✉), L. Wang
School of Electrical and Electronic Engineering,
Nanyang Technological University,
Block S2, 50 Nanyang Avenue, Singapore 639798
e-mail: {p147508078, ecrwan, elpwang}@ntu.edu.sg

audio signals, which was model-free. The Hidden Markov Model was used in [13] to classify TV programs into commercial, basketball, football, news, and weather based on audio information. A novel nearest feature line classifier for audio classification was proposed in [14]. Wold et al., used perceptual and acoustical features for content-based audio classification and retrieval [15] and are now developing SoundFisher™ software in JAVA for commercial usage.

There are many standard distance metrics that can be used for classification. A new metric for measuring the similarity between two PDF's (probability density functions) of mixture type was proposed and used with GMMs [16].

Once an audio has its label, it can be indexed and annotated for browsing and retrieval. Contrary to using keywords in queries for text retrieval, examples are used in queries for sounds. Usually, the similarities between the audio samples in the database and the query example are calculated, a distance-ranking list is given as the retrieval result. Query-by-humming is another kind of query-by-example, where the example is generated by a user on site to search for approximations. For long signals, segmentation is firstly conducted to segment the audio into homogeneous clips before classification [17].

In this paper, we focus on classification and retrieval of two major audios, speech and sounds of music instruments, using fuzzy logic. The fuzzy nature of audio searching lies in the facts that (1) both the query and target are approximation of the user's memory and desire and (2) exact matching is sometimes impossible or impractical. Therefore, fuzzy logic system is a natural choice in audio classification and retrieval.

In the literature, there exists some research in the audio domain using fuzzy logic. In [18], a new method for multilevel speech classification based on fuzzy logic has been proposed. Through simple fuzzy rules, their fuzzy voicing detector system achieves a sophisticated speech classification, returning a range of continuous values between extreme classes of voiced/unvoiced. In classification of audio events in broadcast news [19], when fuzzy membership functions associated with the features are introduced, the overall accuracy of hard threshold classifier can be improved by 4.5% to achieve 94.9%. All these related work has demonstrated the ability of fuzzy logic to enhance classification performance and thus given more or less hints for us to conduct our research of audio classification and retrieval with fuzzy inference systems.

Normally, there are a large number of features adopted as inputs to the existing systems, while in our system, the fuzzy classifier based on only two extracted features can discriminate audio from speech and music, female speech from male speech, sounds of music instruments from percussion and others. We can also further use multiple fuzzy inference systems, called "fuzzy trees", for hierarchical audio classification and retrieval.

The rest of the paper is organized as follows. Audio feature extraction and normalization, which are necessary steps in a content-based system, are discussed in Sect. 2. The proposed fuzzy inference system for audio classification is described in Sect. 3. Fuzzy-tree search and retrieval are explained in Sect. 4. The experimental results are shown in Sect. 5. Finally, conclusion is given in Sect. 6.

2
Audio feature extraction and normalization

In order to classify audios automatically, features are to be extracted from raw audio data source at the beginning. The audio database being classified in this paper is described in Table 1. The lengths of these files range from about half a second to less than 10 s. They are sorted into two major categories: speech and music. Speech includes female and male speech, and percussion is distinguished from the rest of music instrument because of its inharmonic nature. This database is a subset of the database used in [15]. The original database has 16 classes, two from speech (female and male speech), seven from music (percussion, trombone, cello, oboe, tubular-bell, violin-bowed, violin-pizzicato), and seven from other sounds. Here in this paper, the classes of female speech, male speech and percussion remain the same as the original ones. The rest of six music instrumental sounds are combined to form one larger class "Other Music". The seven classes from other sounds are removed from the original data set. Fuzzy logic will be applied to hierarchically classify the audio into their corresponding classes. The inputs to the fuzzy inference system (FIS) are some of the extracted features.

We extract features from the time, frequency, and coefficient domain. They are obtained by calculating the mean and standard deviation of frame-level characteristics. These characteristics are computed from 256 samples per frame, with 50% overlap between two adjacent frames from hamming-windowed original sound. Silence frames, in which the frame energy is below a small threshold, are ignored during the rest of processing.

2.1
Time domain features

Time domain features include RMS (root mean square), ZCR (zero-crossing ratio), VDR (volume dynamic ratio) and silence ratio:

RMS: It is a measure of loudness of the frame.

ZCR: A zero-crossing is said to occur if successive samples have different signs. The zero-crossing ratio is the number of the time-domain zero-crossings and total number of samples in a frame.

VDR: It is the difference of maximum and minimum RMS normalized by the maximum RMS of the frame audio signal. The magnitude of VDR is dependent on the type of the sound source.

Table 1. Structure of the audio database used in this paper

Class name	Number of files
1. Speech	53
1.1 Female	36
1.2. Male	17
2. Music	299
2.1 Percussion	102
2.2 Others	197
Total	352

Silence ratio: It is the ratio of silent frames (determined by preset threshold) and the entire frames. It is noted that the silence frame ignored before feature extraction is termed absolute silence. Here, it is relative silence. The threshold is set to 10% of RMS.

2.2 Frequency domain features

The features used in this domain include frequency centroid, bandwidth, four sub-band energy ratios, pitch, salience of pitch, spectrogram, first two formant frequencies, and formant amplitudes:

Frequency centroid (brightness): It represents the balancing point of the spectrum.

Bandwidth: It is the magnitude-weighted average of the difference between the spectral components and the frequency centroid. It can quantitatively express the range of frequencies over which the power or energy density spectrum is concentrated.

Sub-band energy ratio: The frequency spectrum is divided into 4 sub-bands with intervals $\left[0, \frac{\omega_0}{8}\right]$, $\left[\frac{\omega_0}{8}, \frac{\omega_0}{4}\right]$, $\left[\frac{\omega_0}{4}, \frac{\omega_0}{2}\right]$, $\left[\frac{\omega_0}{2}, \omega_0\right]$, where ω_0 is the half sampling frequency. The sub-band energy ratio is measured by the energy in the sub-band divided by the total energy. Sub-band energy ratios, when used together, reveal the distribution of spectral energy over the entire frame.

Pitch: Pitch refers to the fundamental period of a human speech waveform. We compute the pitch by finding the time lag with the largest autocorrelation energy.

Salience of pitch: It is the ratio of the first peak (pitch) value and the zerolag value of the autocorrelation function.

Spectrogram: It describes the overall energy distribution at different time and frequency.

First two formant frequencies and amplitudes: Formant is caused by resonant cavities in the vocal tract of a speaker. The first and second formants are most important.

2.3 Coefficient domain features

We use two kinds of coefficients, i.e., MFCCs and LPCs:

Mel-Frequency Cepstral Coefficients: MFCC calculation includes discrete Fourier transform of audio frames, followed by filtering by a triangular bandpass filter bank, logarithmic scaling and discrete cosine transform. Here, 13 MFCCs are used.

Linear prediction coefficients: The LPC coefficients are a short-time measure of the speech signal, which describe the signal as the output of an all-pole filter. The first 13 orders of LPC parameters are chosen, which are calculated every 20 ms in this implementation.

2.4 Feature computation and normalization

Among these features, computing the time features needs the shortest period of time, followed by frequency features. The computing of coefficient features is most complex and it takes the longest period of time, especially for MFCCs. As our purpose is for online web application, time and frequency features are preferred.

Normalization can ensure that contributions of all audio feature elements are adequately represented. Each audio feature is normalized over all files in the database by subtracting its mean and dividing by its standard deviation. The magnitudes of the normalized features are more uniform, which keeps one feature from dominating the whole feature vector.

3 Fuzzy inference system

There are several important issues in building a Fuzzy Inference System (FIS), such as selecting the right features as inputs of the system, constructing proper membership functions and rules, and tuning parameters to achieve a better performance.

3.1 Selecting Features as Inputs

In order to select appropriate features as inputs to the FIS from the extracted ones [20], we use a simple nearest neighbor (NN) classifier and a sequential forward selection (SFS) method to choose the appropriate features. The entire data set is divided into two equal parts for training and testing the NN classifier.

Firstly, the best single feature is selected based on classification accuracy it can provide. Next, a new feature, in combination with the already selected feature, is added in from the rest of features to minimize the classification error rate, in order to find the combination of two features that leads to the highest classification accuracy. Our objective is to use as few features as possible to achieve a reasonable performance. As experiments show that even the best single feature is not sufficient to do the classification alone, we select the first two best single features as inputs. These two features are thus chosen as inputs to the FIS as well. Through experiments, we find the spectrogram and pitch salience ratio are the first two features for discriminating speech from music; pitch and pitch salience for distinguishing female and male speech; pitch salience and first MFCC coefficient for separating percussion from the rest of music instruments.

3.2 Membership function and rule construction

We use one fuzzy classifier in the first hierarchy for discriminating speech and music as an example to show the design of our system. The two normalized feature (spectragram and pitch salience ratio) histogram of the two classes are shown in Figs. 1 and 2, respectively. Each histogram is normalized by its peak value. After determining the inputs, the key to constructing the fuzzy classifier is to design the membership function and extract rules. In fact, the membership functions of each input and output, as well as the rules, can be derived from simulating the feature distributions. We chose Gaussian membership functions, which is fully parameterized by the mean and the standard deviation. We calculate these parameters directly from the statistics of the features among the whole data source. We use "small" and "large" to denote their membership according to their class distribution. The resulting simplified Gaussian membership functions simu-

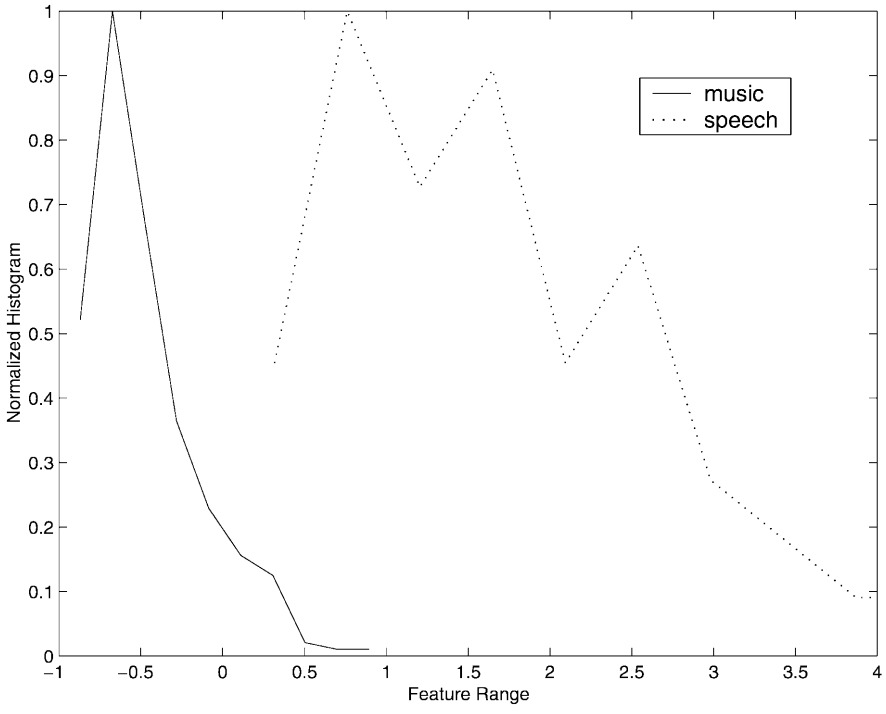

Fig. 1. The feature distribution of spectrogram for speech and music

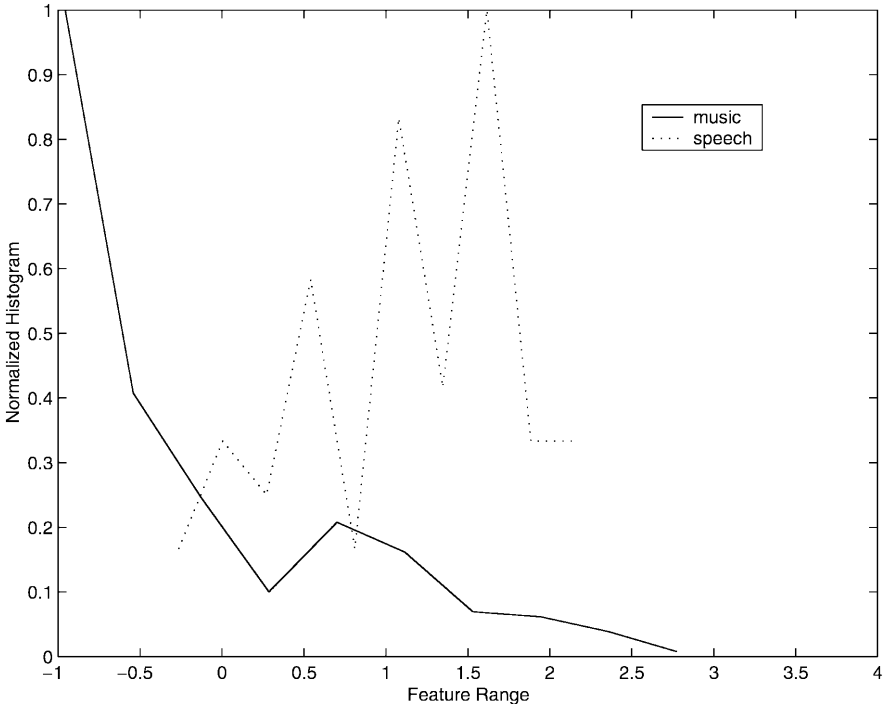

Fig. 2. The feature distribution of pitch salience ratio for speech and music

lating the feature distributions are shown in Figs. 3 and 4. Another two Gaussian membership functions are chosen for output, shown in Fig. 5. One mean is zero and another is one, with same standard deviation that makes their equal probability at center of distribution. An overview of the fuzzy classifier for discriminating speech and music are given in Fig. 6. The four rules in the FIS, i.e., each input has two rules, are listed below.

IF *(Spectrogram is small)*, THEN *(Type is music)*
IF *(Spectrogram is large)*, THEN *(Type is speech)*
IF *(Salience ratio is small)*, THEN *(Type is music)*
IF *(Salience ratio is large)*, THEN *(Type is speech)*

Similarly, the rules for the second FIS to classify female and male speech are:

IF *(Pitch is small)*, THEN *(Type is male speech)*
IF *(Pitch is large)*, THEN *(Type is female speech)*
IF *(Pitch salience is small)*, THEN *(Type is male speech)*
IF *(Pitch salience is large)*, THEN *(Type is female speech)*

The rules for the third FIS to identify percussion from non-percussion musical instrumental sounds are:

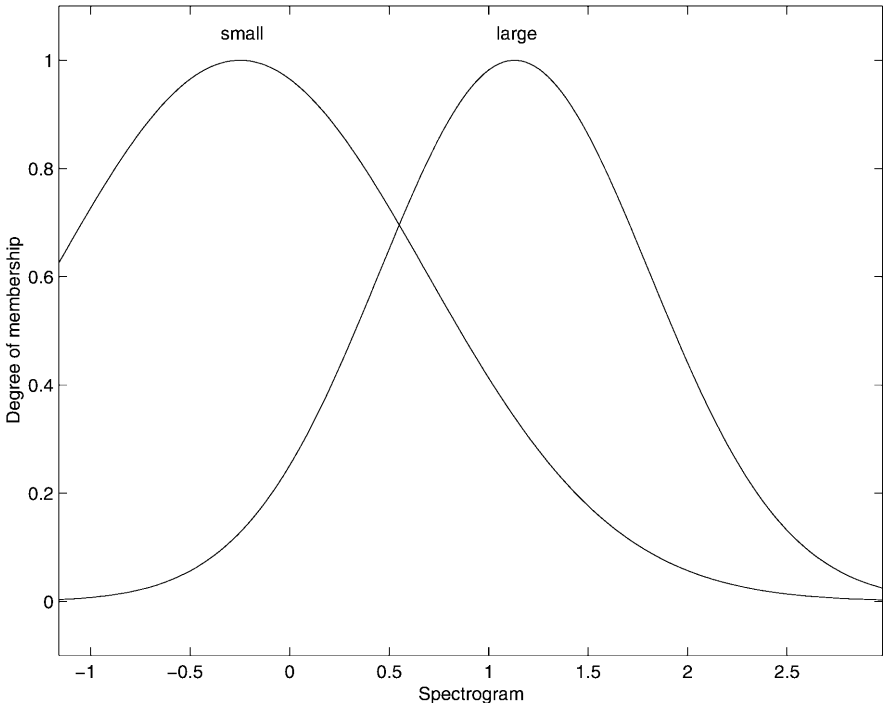

Fig. 3. The Gaussian membership function simulating the feature distribution of spectrogram for speech and music

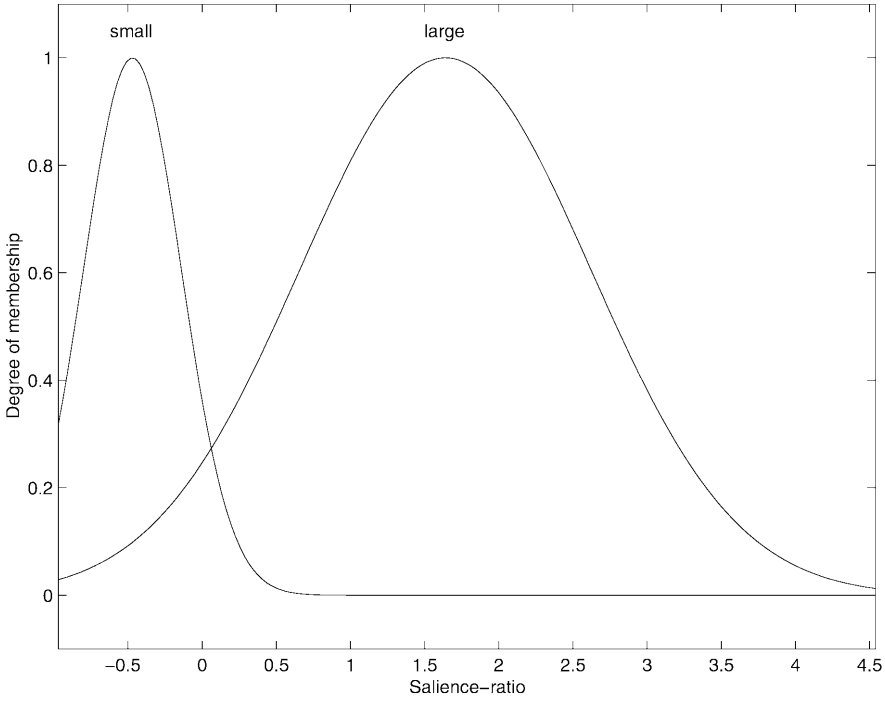

Fig. 4. The Gaussian membership function simulating the feature distribution of pitch salience ratio for speech and music

IF *(Pitch salience is small),* THEN *(Type is percussion)*
IF *(Pitch salience is large),* THEN *(Type is non-percussion)*
IF *(First MFCC coefficient is small),* THEN *(Type is percussion)*
IF *(First MFCC coefficient is large),* THEN *(Type is non-percussion)*

3.3 Tuning the FIS

Although the fuzzy inference systems are thus constructed completely, there are ways to improving the performance, for example, by tuning parameters of those membership functions, choosing other types of membership function corresponding to the feature distribution, or using neural networks to train the membership functions for a closer approximation. Since those features selected by the sequential forward selection method are sub-optimum inputs, we may also try other combinations of features as input to improve accuracy.

4 Fuzzy-tree search and retrieval

Content-based audio search and retrieval can be conducted as follows. When a user inputs a query audio file

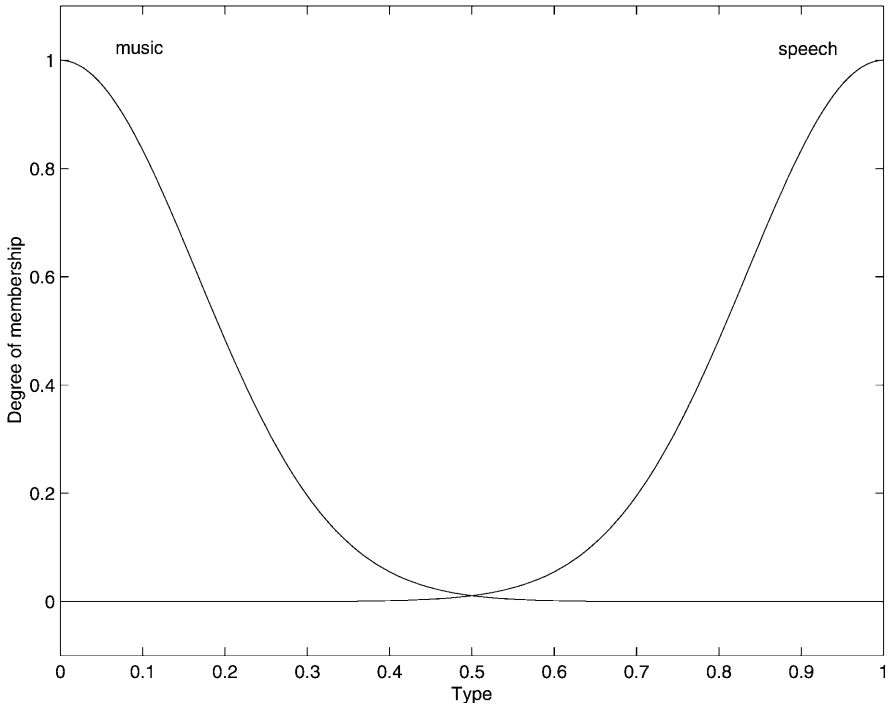

Fig. 5. The Gaussian membership function simulating the output for speech and music

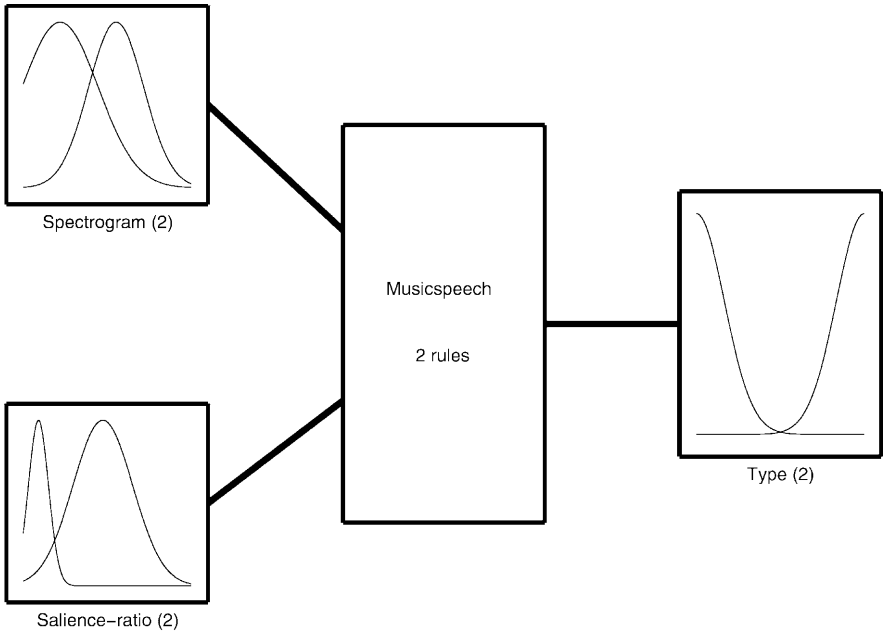

Fig. 6. The FIS input–output diagram

and requests to find relevant files, both the query and each audio file in the database are represented as feature vectors. A measure of the similarity between the query feature vector and a stored feature vector is evaluated and a list of files based on the similarity are fed back to the user for listening and browsing. The user may refine the query to get audios more relevant to his or her interest by feedbacks. The purpose of classification is to give the unknown audio a label from a set of existing labels, whereas in retrieval, we need to find samples relevant or similar to the query example. The performance of classification is measured by accuracy, while the performance of retrieval is measured by precision and recall defined as follows:

$$\text{Precision} = \frac{\text{Number of relevant files retrieved}}{\text{Total number of files retrieved}} \quad (1)$$

$$\text{Recall} = \frac{\text{Number of relevant files retrieved}}{\text{Total number of files relevant}} \quad (2)$$

Precision indicates the quality of the answer set, while recall indicates the completeness of the answer set. In an ideal situation, precision is always 1 at any recall point. In classification, we concentrate on precision, whereas in retrieval, we wish to retrieve as many relevant files as possible (high recall).

A fuzzy-tree architecture as shown in Fig. 7 can be constructed for retrieval. One of two classes can be dis-

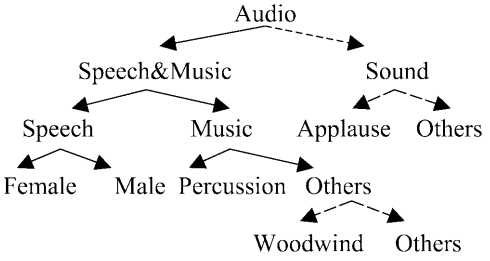

Fig. 7. The hierarchical fuzzy tree for retrieval for retrieva

tinguished at a time. The dashed lines are under development. Firstly, we use the fuzzy classifier to classify the query example into a particular class as a node in the tree. Then, a Euclidean distance method is adopted to select the most similar samples in the database in that class. We assume that the files in a same class are relevant; otherwise, they are irrelevant. When the database grows, new classes can be added to the tree. Only links between that class and its immediate upper level are updated, with the rest of the tree unchanged.

5
Experimental results

In an integrated "fuzzy-tree" classification system, all classifications can be done automatically. For example, when an audio is submitted to the system for classification, we use the first FIS with inputs of spectrogram and pitch salience ratio to distinguish speech from music. If the result is speech, we then use the second FIS with inputs of pitch and pitch salience for discriminating female and male speech; if the decision is music, we can use the third FIS with inputs of pitch salience and first MFCC coefficient for separating percussion from the rest of music instruments. With more domain knowledge collected, we may discover new features and new rules which are fit for identifying environmental sounds such as thunder, laughter, etc, from music and speech at the very beginning, and then recognize some sounds with semantic meanings like applause among the generic sounds, or recognize more individual music instrument from the rest of instrument family by studying their vibration characteristics.

As for retrieval, when a query input is presented, the direct search may result in mixture types of audio clips being retrieved. If we firstly classify the query into a particular node of the fuzzy tree, we can then search relevant files only in that subspace instead of the whole database. For example, both speech and music clips can be in the search results of a speech query. If we classify the query before search, the results will be in one same type. Thus, the precision will increase and the searching time will decrease. If the classification is wrong, we can go to search in another direction with user's feedback, since there are only two leaves for each node in the tree.

Three experiments have been done hierarchically to get the performance of all these fuzzy classifiers. At the first level of the fuzzy tree in Fig. 7, each audio file is used as input to the fuzzy music-speech classifier. It can distinguish music and speech with an accuracy of 92%.

Table 2. Classification performance

FIS classifier	Music–Speech	Female–Male	Percussion–Others
Classification accuracy	92%	89%	81%

At the second level of the fuzzy tree in Fig. 7, 53 speech files are submitted to the female-male classifier and 299 music files are fed into the percussion-others classifier. They can separate female and male speech with 89% accuracy and divide percussion and others with 81% accuracy respectively. All classification results are summarized in Table 2.

6
Conclusions

In this paper, we propose the fuzzy inference system for audio classification and retrieval, as a first step towards a multimedia search engine for the Internet. The benefits of the fuzzy classifier lie in the facts that no further training is needed once the fuzzy inference system is designed. Thus, classification can be performed very quickly. In addition, when the database grows, new classes can be added to the fuzzy tree. Only links between that class and its immediate upper level are updated, with the rest of the tree unchanged. With this architecture, fast online web applications can be built. Future work along this direction is to use neural networks to train the parameters to obtain better membership functions, and to explore new features and rules to classify various audios with the so-called "fuzzy tree" for hierarchical retrieval.

References

1. **Makhoul J, Kubala F, Leek T, Daben Liu, Long Nguyen, Schwartz R, Srivastava A** (2000) Speech and language technologies for audio indexing and retrieval, Proc IEEE **88**(8): 1338–1353
2. **Viswanathan M, Beigi HSM, Dharanipragada S, Tritschler A** (1999) Retrieval from spoken documents using content and speaker information, ICDAR '99 pp. 567–572
3. **Gauvain J-L, Lamel L** (2000) Large-vocabulary continuous speech recognition: advances and applications, Proceedings of the IEEE, **88**(8): 1181–1200
4. **Chih-Chin Liu, Jia-Lien Hsu, Chen ALP** (1999) An approximate string matching algorithm for content-based music data retrieval, IEEE Int Conf Multimedia Comp Syst **1**: 451–456
5. **Kosugi N, Nishihara Y, Kon'ya S, Yamamuro M, Kushima K** (1999) Music retrieval by humming-using similarity retrieval over high dimensional feature vector space, IEEE Pacific Rim Conf Commun, Comp Signal Processing 404–407
6. **Kataoka M, Kinouchi M, Hagiwara M** (1998) Music information retrieval system using complex-valued recurrent neural networks, IEEE Int Conf Syst, Man, and Cybernetics, **5:** 4290–4295
7. **Delfs C, Jondral F** (1997) Classification of piano sounds using time-frequency signal analysis, ICASSP-97 **3:** 2093–2096
8. **Paradie MJ, Nawab SH** (1990) The classification of ringing sounds, ICASSP-90, 2435–2438
9. **Guohui Li, Khokhar AA** (2000) Content-based indexing and retrieval of audio data using wavelets, ICME 2000 **2:** 885–888

10. **Subramanya SR, Youssef A** (1998) Wavelet-based indexing of audio data in audio/multimedia databases, Proc Int Workshop on Multi-Media Database Management Sys 46–53
11. **Scheirer E, Slaney M** (1997) Construction and evaluation of a robust multifeature speech/music discriminator, ICASSP-97 **2**: 1331–1334
12. **Tong Zhang, Jay Kuo, C-C** (1999) Heuristic approach for generic audio data segmentation and annotation, ACM Multimedia'99, pp. 67–76
13. **Liu Z, Huang J, Wang Y** (1998) Classification TV programs based on audio information using hidden Markov model, IEEE Second Workshop on Multimedia Signal Processing, pp. 27–32
14. **Li SZ** (2000) Content-based audio classification and retrieval using the nearest feature line method, IEEE Trans Speech and Audio Processing **8**(5): 619–625
15. **Wold E, Blum T, Keislar D, Wheaten J** (1996) Content-based classification, search, and retrieval of audio, IEEE Multimedia **3**(3): 27–36
16. **Zhu Liu, Qian Huang** (2000) Content-based indexing and retrieval-by-example in audio, ICME **2**: 877–880
17. **Kemp T, Schmidt M, Westphal M, Waibel A** (2000) Strategies for automatic segmentation of audio data, ICASSP **3**: 1423–1426
18. **Beritelli F, Casale S, Russo M** (1995) Multilevel speech classification based on fuzzy logic, Proc IEEE Workshop on Speech Coding for Telecommunications, pp. 97–98
19. **Zhu Liu, Qian Huang** (1998) Classification of audio events in broadcast news, IEEE Second Workshop on Multimedia Signal Processing, pp. 364–369
20. **Jang J-SR** (1993) ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans Syst, Man and Cybernetics **23**(3): 665–685