# AN AUDIO REPRESENTATION FOR CONTENT BASED RETRIEVAL

*Kathy Melih, Ruben Gonzalez and Philip Ogunbona*
School of Information Technology, Griffith University, QLD, Australia.
{K.Melih, R.Gonzalez}@gu.edu.au

## ABSTRACT

Despite the increasing interest in multimedia data retrieval audio data has received little attention. This is due, not to a lack of interest but rather to unique difficulties posed by the medium. In particular existing unstructured audio representations do not easily lend themselves to content based retrieval and especially browsing. This paper aims to address this oversight by developing an audio representation that provides direct support for browsing and content based retrieval. This support is the result of a structured representation based on psychoacoustic principles in which salient attributes of audio are directly accessible. In addition, the representation is compact thus addressing the requirement for minimisation of storage.

## 1. INTRODUCTION

There are many applications in which content based retrieval of audio is desirable. The frustrating situation of knowing what a piece of music sounds like but not its name or composer is a common example. So too, is attempting to locate a desired short section of audio in a long recording. Content based access to video can also benefit from audio retrieval methods[1]. Even the most basic level of classification based on the type of audio (speech, music or other) can often be useful. For example, locating a desired section in a concert recording is greatly simplified if it is first segmented into pieces (and perhaps movements) by the location of non-music sections (applause, silence or speech).

However, traditional audio coding techniques result in representations that make extracting even this low level information difficult. This is because traditional representations aim only to reproduce the signal (or a perceptual equivalent) while providing for compact storage, resulting in unstructured representations. In contrast, content based retrieval and especially browsing benefit from structured representations. This paper presents a structured audio representation designed specifically to support content based retrieval and browsing.

## 2. ISSUES IN AUDIO RETRIEVAL

### 2.1. General Issues

The main issue in multimedia data management is the selection of the index keys. This is influenced by the nature of the underlying data representation. The nature of index keys influences the effectiveness of searching and browsing.

#### *2.1.1. Data representation*

Counter to human cognition of audio signals (section 3.2), traditional audio representations are unstructured. Such representations make it difficult to extract indexing information. As a result, support for retrieval relies on separate index files. This situation is undesirable for several reasons. The existence of a separate index file introduces storage overheads: an unwelcome addition to storage hungry data. Further, the lack of structure in the data makes browsing difficult, if not impossible. Finally, in the case of manually generated indexes or supervised automatic systems, the task of generating index files is tedious.

A solution to these problems is a structured audio representation that provides direct access to salient attributes of the data. The need to manage separate index files is eliminated since indexing information can be extracted directly from the representation. If the structured representation is also made compact, a further reduction in storage requirements can be achieved. Also, such a structured representation can be generated automatically thus reducing workload.

#### *2.1.2. Index Keys*

Selection of index keys restricts the nature and flexibility of searches and is the defining factor in any content based retrieval system. In the case of audio, existing methods have used indexes ranging from manually generated text based labels to automatically generated statistical feature vectors[2].

Text based labelling suffers many obvious drawbacks. The first is that the range of possible queries is severely limited. The limits are imposed by the selection of index attributes and the fact that some features of audio are difficult, or impossible, to describe verbally (eg, timbre). In the case of speech, transcriptions derived from automatic speech recognition would seem ideal. However, this is not yet possible in unconstrained environments[3][4]. Also, speech contains much semantic content that would be lost in a simple transcription (eg prosodics).

Recent audio retrieval systems use automatically generated feature vectors as index keys[5]. These vectors are generated by performing statistical analyses of the audio signal and describe attributes such as the loudness and

harmonicity of the signal. This non verbal description is more generic making it more flexible. However, the scope of retrieval is restricted by the low level, feature analytic nature of the attributes. Also, these methods do not directly reveal the underlying structure of the audio and therefore provide little support for browsing.

In human audio perception there appear to be a number of key attributes extracted from the audio signal. Using these attributes as a basis for index keys is an obvious means of creating a flexible content based retrieval system. Since these attributes are perceptually based, they will most likely provide better support for higher level queries. Additionally, psychoacoustic principles can be applied to these attributes to identify structure in the data, providing better support for browsing.

#### *2.1.2. Searching and Browsing*

Searching is useful when the user has a definite idea of what they wish to retrieve. Searches may be based on broad queries to find data of a single 'type' (eg, 'retrieve all occurrences of speech') or on specific queries based on the semantic content of an audio record (eg, 'find the song that contains the melody...'). There are a number of methods by which queries can be posed. The lowest level involves specifying the numerical values of the index keys directly. This is obviously of little practical use. Text based queries, while suffering the problems mentioned earlier, may be useful when specifying broad search categories. The most natural, and useful, form of query from an audio database is by example (eg, the desired melody is hummed into a suitable interface to form the query).

Browsing is required when a user can't specify a query exactly or to review search results where several possibilities have been retrieved. Browsing requires that the data be provided with a logical, hierarchical structure. This structure may be inherent in the data or may result from classification of the data based on its attributes. The former can be applied to musical data and speech whilst the latter is applicable to instances of discrete sounds, such as general environmental sounds. Music has a structure which may be viewed in a number of ways [6][7]. Speech may be organised according to speaker transitions [3][5] then into individual phrases or words by silence detection.

### 2.2 Existing Work

Of the existing audio retrieval systems, most focus on a single type of audio (speech, music or other). Many methods exist to segregate speech from other forms[6][7] and may be useful to provide a coarse index by type but do not fulfil the requirement of content based retrieval.

MELDEX, a score based retrieval system, takes queries by example, transcribes them into musical notation and uses this to search through a database of musical scores[8]. Ghias et al[9] propose a system for melody retrieval that relies on converting queries into strings of relative pitch transitions and performing searches in a similar index created from MIDI files. Both systems are akin to searching a speech database using textual transcriptions and are highly constrained.

Systems for the retrieval of general environmental sounds involve the calculation of feature vectors for use as indexes at query time[5][10]. A separate feature vector is required for each individual sound.

All these systems, can only handle one type of audio at a time. A mixed collection of general sounds must first be segmented before creating the index, usually in a completely separate process. Little processing is required to determine the type of sound from a representation with direct access to salient attributes. Although each sound type might eventually be treated differently, segmentation does not introduce significant processing overheads.

## 3. PERCEPTION OF AUDIO

The representation presented exploits aspects of human audio perception to achieve two aims. The first, common in audio coding[11], is to reduce redundancy. Less common applications of psychoacoustics are to provide the data with structure and to isolate key cognitive features.

### 3.1 Peripheral processing

Audio signals undergo a frequency transformation effected by the basilar membrane in the inner ear. The result is a representation of the input audio in a three dimensional (time, frequency, intensity) space.

This process displays several interesting phenomena. The first is that the frequency transformation consists of axes that are non-uniformly sampled. Frequency resolution is coarse and temporal resolution is fine at high frequencies while temporal resolution is coarse and frequency resolution is fine at low frequencies. Also, the amplitude axis displays a frequency dependent non-linearity.

Another interesting phenomenon is masking. If two signals in close frequency proximity are presented simultaneously, the less intense sound may be rendered inaudible. The two signals may be tones or noise. Masking can also occur when two signals are presented in close temporal proximity.

### 3.2 Mental Representation of Audio

The signal reaching the ear is a mix of signals from many different sources. However, we are capable of distinguishing individual sounds. The process responsible is stream segregation. Stream segregation involves decomposing the signal into its constituent parts (partials) then grouping them into streams: one for each sound.

At a basic level, one can model audio representation in the human mind as a series of peak amplitude tracks in a time-frequency-intensity space[12]. Three audio classes exist: frequency sweeps, tone bursts and noise bursts. The representation of these classes is shown in Figure 1.

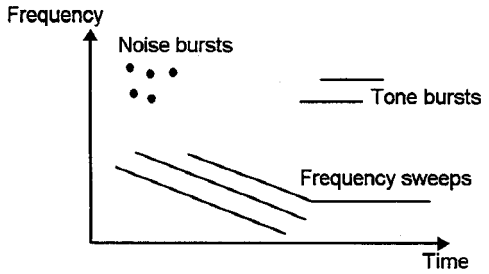


**Figure 1.** Mental representation of audio

There appears to be a set of general principles that are applied in achieving the task of stream segregation. These principles include[13]:

- Similarity: tones similar in frequency tend to be fused.
- Continuation: partials representing smooth transitions tend to fusion. Rapid transitions tend separation.
- Common fate: partials with similar onset times and/or correlated modulation tend to be fused.
- Disjoint allocation: in general, each partial can belong to only one stream.

## 4. PROPOSED REPRESENTATION

The representation proposed is essentially a parametric representation of the three sound classes identified in section 3.2. Sinusoidal transform coding[14] allows audio signals to be described as a series of amplitude trajectories through time-frequency space and would seem ideal for the purpose. However, it is not completely suitable and has been adapted in two ways. To eliminate redundancy and to avoid undesirable blocking effects, a modified discrete cosine transform (MDCT) is used instead of the customary short time Fourier transform. Also, time-frequency-distribution (TFD) is perceptually tuned to mimic the time-frequency resolution of the ear.

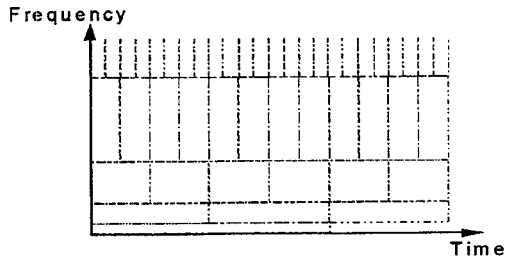


**Figure 2:** Time-Frequency resolution of the TFD

The arrangement shown in Figure 2 better models the frequency transformation effected by the ear well as eliminating redundant data and providing better noise performance. This variable resolution is achieved by dividing the TFD into a series of resolution bands separated by octaves. The bands are generated using constant length, variable resolution analysis windows obtained from the filtering operation shown in Figure 3.

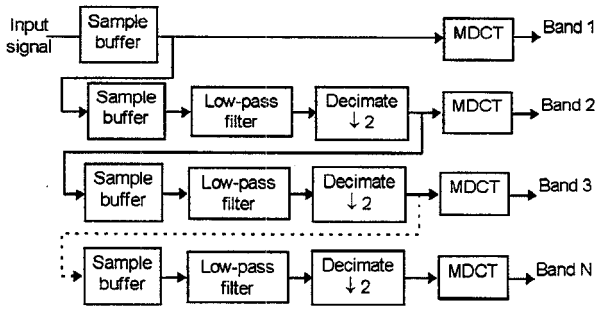


**Figure 3:** MDCT-based TFD generation

Having generated a variable resolution TFD, acoustic masking and quiet thresholds are applied to eliminate perceptually redundant data. This helps to compact the final representation as well as simplifying the following stages of processing.

The next stage of processing involves peak picking and tracking. Peaks are found by searching for all points in the TFD that satisfy the condition:

$$|X_t(f_{i-1})| < |X_t(f_i)| > |X_t(f_{i+1})|$$

where $X_t(f_i)$ is the amplitude at the $i$th frequency, $f_i$, in the current time frame, $t$. Tracking is performed according to the algorithm in [14] which involves matching peaks in adjacent time frames that are closest in frequency within set limits.

The peak picking and tracking stage results in a set of tracks which describe the audio. Tracks are described parametrically with the following information available for each track:

- Start time;
- Finish Time;
- Frame Numbers;
- Amplitude Contour; and
- Frequency Contour

The unit of time corresponds to the shortest analysis frame length. The list of frame numbers is required since the variable resolution means that amplitude and frequency values may not be available at all times along a track: this is of particular relevance to frequency sweeps.

Given this description, each track is then classified according to type: noise, tone or sweep. Very short tracks are classified as noise. The decision as to whether a track is

a tone or a sweep is made by examining the frequency data along the track.

Having classified the tracks, psychoacoustic principles can be applied to segregate them into streams. At this stage, the aim of segregation is simply to remove correlation in the data so only a very basic set of principles are applied. Finally, the tracks are encoded in groups.

## 5. CONTENT BASED RETRIEVAL

Determining the nature of a given segment of audio data follows directly from this representation. Audio data can be classified into one of four categories: speech, music, silence and noise. Each of these categories exhibit unique characteristics in the time-frequency domain that are directly visible in the track representation. Silence is identified by the absence of any tracks. Music consists of long harmonically related tracks with few periods of silence. Speech is identified by the presence of relatively short noise bursts, tone bursts and frequency sweeps interspersed with frequent short periods of silence. Noise consists entirely of noise bursts.

Once this coarse level of classification has been performed, an individual segment of audio can be further analysed based on its type. Given that the tracks are parametrically represented, analysis basically involves comparing the parameter values of individual tracks. The types of higher level information that can be inferred from the tracks depends on the type of sound.

In the case of speech, change of speaker or gender may be determined by examining the lowest frequency track. Speaker emphasis is visible in the variation of relative amplitude of the tracks. Voicing information is directly visible by the nature of the tracks at the instant of time (noise or tone). The suitability of this representation for speech recognition is yet to be investigated.

For music data, the melody line can be determined by following the pitch along tracks. The representation should also permit query by example. Queries input via an appropriate audio interface can be analysed into the track representation and then the melody information extracted can be used as a basis for comparison.

## 6. CONCLUSIONS

Existing unstructured audio representations make content based retrieval difficult. As a result, current systems rely on the use of separate index files. This paper has discussed the disadvantages of this situation. A solution to the problem, a structured audio representation, has been proposed. This representation is based on psychoacoustic principles and has been designed to provided direct access to salient attributes of audio signals.

## 7.0 REFERENCES


[1] R. Kazman, R. Al-Halimi, W. Hunt and M. Mantei, "Four Paradigms for Indexing Video Conferences", *IEEE Multimedia*, Spring 1996, pp. 63-73.

[2] M. Abdel-Mottaleb, H-L. Wu and N. Dimitrova, "Aspects of Multimedia Retrieval", *Philips J. Res.*, **50** (1996), pp. 227-251.

[3] D. Hindus, C. Schmandt and C. Horner, "Capturing, Structuring and Representing Ubiquitous Audio", *ACM Trans. On Information Systems*, v. 11, n. 4, Oct 1993, pp 376-400.

[4] A. G. Hauptmann, M. J. Witbrock, A. I. Rudnicky and S. Reed, Speech for Multimedia Information Retrieval, UIST '95, pp. 79-80.

[5] E. Wold, T. Blum, D. Keislar and J. Wheaton, "Content-Based Classification, Search and Retrieval of Audio, *IEEE Multimedia*, Fall 1996, pp. 27-36.

[6] A. S. Tanguine, "A Principle of Correlativity of Perception and its Application to Music Recognition", *Music Perception*, Summer 1994, 11 (4), pp. 465-502.

[7] P. Aigrain, P.J.V. Longueville, P. Lepain, "Representation-based user interfaces for the audiovisual library of year 2000", Proc. SPIE Multimedia and Computing and Networks 1995, vol. 2417, Feb 1995, pp. 35-45.

[5] L. Wilcox, D. Kimber, and F. Chan, "Audio Indexing Using Speaker Identification", *SPIE* v. 2277, pp. 149-157.

[6] J. Hoyt, H. Wechsler, "Detection of Human Speech in Structured Noise", IEEE ICASSP, vol 2. 1994, pp 237-240.

[7] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music", IEEE ICASSP, 1996, pp 993-996.

[8] R. J. McNab, L. A. Smith, D. Bainbridge and I. H. Witten, "The New Zealand Digital Library MELody inDEX", *D-Lib Magazine, May 1997*, http://www.dlib.org/dlib/may97/meldex/05witten.htm.

[9] A. Ghias, J. Logan, D. Chamberlin and B. C. Smith, "Query By Humming: Musical Information Retrieval in An Audio Database", Proc. ACM Multimedia '95, San Fransisco, pp 231-236.

[10] R. S. Goldhor, "Recognition of Environmental Sounds", IEEE ICASSP, vol 1, 1993, pp 149-152.

[11] N. Jayant, J. Johnston and T. Safranek, "Signal Compression Based on Models of Human Perception", *Proc of the IEEE*, vol. 81, no. 10, Oct 1993, pp1383-1421.

[12] D. P. W. Ellis, B. L. Vercoe, "A Perceptual Representation of Audio for Auditory Signal Separation", presented at the 23rd meeting of the Acoustical Society of America, Salt Lake City, May 1992.

[13] B. C. J. Moore, An Introduction to the Psychology of Hearing, fourth edition, Academic Press, 1997.

[14] T. F. Quatieri, R. J. McAulay, "Speech Transformations Based on a Sinusoidal Representation", *IEEE Trans. ASSP*, vol. ASSP-34, no. 6, Dec 1986, pp. 1449-1463.