Audio Retrieval Using Perceptually Based Structures

Kathy Melih and Ruben Gonzalez

School of Information Technology, Griffith University
PMB 50 Gold Coast Mail Centre, QLD 9726, Australia.

{K.Melih, R.Gonzalez}@gu.edu.au

Abstract

Despite growing interest in multimedia data management, audio retrieval has received little attention. In part, this can be attributed to existing unstructured audio representations that do not easily lend themselves to content based retrieval and especially browsing. This paper aims to address this oversight. It begins by reviewing existing techniques and the specific problems posed by unstructured representations. Some characteristics of audio perception that may be exploited in the solution to these problems are then presented. A new structured representation is then detailed that is designed to support content based retrieval and browsing. Finally, the suitability of this representation for its intended purpose is discussed.

1. Introduction

The sudden explosion of the Internet, in combination with the ready availability of technology to capture and store multimedia data, has resulted in large, and ever increasing stores of multimedia data. Despite a corresponding increased level of interest in multimedia data management, audio data has received very little attention. This is due, not to a lack of importance, but rather to specific difficulties posed by the medium. In particular, existing unstructured audio representations complicate the tasks required in audio data management. In particular, extracting perceptually salient index attributes and allowing non sequential access pose a challenge for conventional representations.

There are many applications that would benefit from content based audio retrieval and browsing methods. The most obvious is the extension of retrieval possibilities from the World Wide Web. Without support for random access and content based retrieval of audio, this will never be a true hypermedia system as intended. Among the many other applications that would benefit from audio retrieval techniques are content based video retrieval[1] and the management of voice mail messages.

Even a very coarse level of classification based on the type of audio (speech, music or other) can often be useful, especially during browsing. For example, locating a desired section in a concert recording is greatly simplified if it is first segmented into pieces 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 are unstructured, aiming only to reproduce the signal (or a perceptual equivalent) while providing for compact storage. In contrast, content based retrieval and especially browsing benefit from structured representations.

This paper presents a perceptually congruent structured audio representation designed specifically to support content based retrieval and browsing. The next section will review some of the general issues pertinent to audio retrieval as well as examining existing work in the area and the benefits of a structured representation. Relevant psychoacoustic principles will then be outlined. This is followed by a description of a new structured audio representation. Finally, methods of content based retrieval, based on this representation, will be outlined.

2. Audio retrieval

Two basic access methods are required in an audio retrieval system: content-based retrieval (searching) and browsing. Searching is used to recover an audio segment as the result of a specific query while browsing allows the user to navigate through the data in an orderly fashion or to find a desired passage based on loosely defined criteria.

2.1 Content based retrieval

Content based retrieval is useful when the user has a definite idea of what they wish to recover. 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 some problems mentioned later, 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).

To support content based retrieval, two things are required: segmentation and component labelling (or indexing). Segmentation involves dividing the data into cognitively significant sections. Using existing, unstructured representations, this can be a tedious or computationally intensive task in itself. The selection of index keys is the most significant issue in content based retrieval since the index keys directly influence the nature and scope of queries supported. Possibilities range from manual annotation to automatically generated statistical feature vectors.

Manual annotation suffers many drawbacks. The most obvious being that it is extremely tedious and not practical for a large database. Another drawback is the severe limitation on the scope of possible queries. These limits are imposed by the selection of index attributes which is itself limited by the fact that some features of audio are difficult, or impossible, to identify using simple textual descriptions (eg, timbre).

Recent audio retrieval systems use automatically generated feature vectors as index keys[2]. These vectors describe attributes such as the brightness, bandwidth and harmonicity of the signal and are generated by performing statistical analyses of the audio signal. An advantage of this technique is, being automatic, it is of greater practical value. Also, the non verbal description is more generic and thus more flexible. However, the scope of retrieval is still restricted by the feature analytic nature of the attributes: that is, they posses little cognitive significance. For example, bandwidth would have little semantic value to an untrained user.

2.2 Browsing

Browsing is required when a user can’t specify a query exactly, or to review search results where several possibilities have been retrieved. The first instance requires that broad queries (‘find all sections containing speech/modulation/transitions’) be supported while the second requires non-sequential access and a logical, hierarchical structure. This structure may be inherent in the data itself, as a result of some grammatical property, or may be the result of a classification based on attributes of the data. The former can be applied to musical and speech data whilst the latter is applicable to instances of discrete sounds, such as general environmental sounds.

Music has a structure that may be viewed in a number of ways[3][4]. The basic hierarchy divides pieces into phrases that in turn are composed of notes. Speech may be organised according to speaker transitions[5][2] then into individual phrases or words by silence detection. Discrete environmental sounds have no grammar upon which a structure can be based. In this case, a hierarchical structure may be developed by performing a categorisation on qualities such as loudness, pitch or harmonicity of the signals[2].

In the few existing systems where browsing is supported, the structure is generally temporal rather than content based. Time-line representations[5] or time compressed play-back[6] provide only a very low level of browsing support. In such systems, the user still needs to be reasonably familiar with the content of each section for there to be any benefit. True content based browsing will often be of much greater value.

To support content based browsing, not only does the data require a structure, but also perceptually significant attributes need to be readily accessible at a fairly high level of temporal resolution. For example, a user may wish to recover all sections of a recording that contain a particular rhythm pattern. Existing generic audio retrieval systems that perform segmentation of the data based only on the type of sound do not have the necessary resolution to support such browsing. This shortcoming can be attributed to the reliance on unstructured audio representations and segmentation and classification based on attributes that have limited cognitive significance.

2.3 Existing work

The existing work in audio retrieval tends to divide the domain into two distinct sections: speech and non-speech, with the non-speech category further subdivided into music and general environmental sounds. Of the existing systems, most focus on only a single category. Speech data has received by far the most attention and very little existing work encompasses all categories. As a result, many methods exist to segregate speech from other forms of audio[7][8] These may be useful to provide a coarse index by type but do not fulfil the requirement of content based retrieval.

For speech, transcriptions derived using automatic speech recognition would seem to be an ideal method of generating an index. However, this is not yet possible in unconstrained environments[9]. Thus manual intervention is often required to create an accurate index. Also, speech signals contain much semantic content that would be lost in a simple transcription (eg prosodics).

Indexing musical data requires some means of accessing melodic information. MELDEX, a score based retrieval system, takes queries by example, transcribes them into musical notation and uses this description to search through a database of musical scores[10]. Ghias *et al*[11] propose a system for melody retrieval that relies on converting queries into strings of relative pitch transitions. Searches are performed in an index containing similar strings as keys. This index is generated directly from MIDI files, which contain score information. Both systems are akin to searching a speech database using textual transcriptions and thus suffer similar drawbacks, including the loss of semantically significant information. Also, these methods are highly constrained.

Systems for the retrieval of general environmental sounds involve the calculation of feature vectors for use as index keys at query time[2][12]. In [12], speech recognition techniques are used to create an index. The biggest disadvantage of this method is that such an index could not support queries of the nature “find recordings that sound like…” This problem is solved by calculating statistical feature vectors based on generic acoustical properties of the audio signal[2]. However, these vectors only describe discrete sounds in a holistic sense so extracting higher level information, such as pitch contours, requires further processing. Thus, while some content based browsing is supported, it is of a very low level. For example, it may be possible to gain ready access to all musical sections in the collection but it would be difficult to find changes of key within a piece.

All these systems can only handle a single 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. This results in the introduction of processing overheads. Additionally, these systems rely on separate index or metafiles, thus, increasing storage requirements for data that is already by nature voluminous. Finally, very little, if any, support for browsing exists within these systems.

2.4 Benefits of a structured representation

Many of the problems suffered by existing audio retrieval systems can be attributed to the reliance on unstructured audio representations. Existing audio representations give very little indication of the actual content of the data encoded. With such representations, processing and storage overheads are an inevitable consequence of the desire to provide support for retrieval and browsing. Also, the underlying structure of the data is not directly apparent in these unstructured representations, which is counter to the requirements for browsing.

A number of key attributes appear to be extracted from an audio signal during the process of human audio perception. Basing a representation on these attributes is an obvious means of aiding content based retrieval since a mid-level index is effectively ‘built in’ to the representation. As these attributes are perceptually based, they will most likely support higher level queries. For example, minimal processing is required to determine the type of a sound (eg speech) given access to such attributes. Although each sound type might eventually be treated differently, segmentation does not introduce significant processing overheads

In order to support browsing, identifying a structure in the data is vital. Having isolated the key cognitive features of an audio signal, psychoacoustic principles can be applied to identify any inherent structure in the data. An additional benefit is that since this structure is perceptually congruent, it is better able to support semantically useful content based browsing.

3. Perceptual considerations

The representation presented in this paper exploits aspects of human audio perception to achieve three aims. The first, common in audio coding[13], 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 known as a time-frequency distribution (TFD).

This process displays several interesting phenomena relevant to the representation presented in this paper. The first is that the TFD 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, humans are able to distinguish 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[14]. Three audio classes exist: frequency sweeps, tone bursts and noise bursts. The representation of these classes is shown in Figure 1.

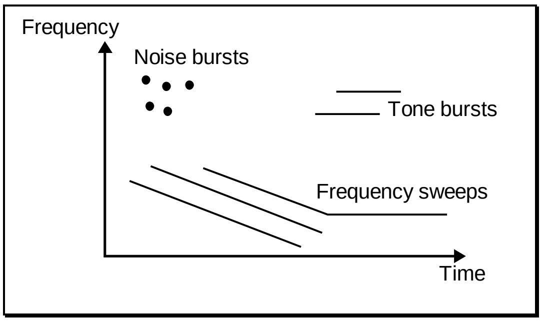

Figure 1: Mental representation of audio

There appears to be a set of general principles that are applied in achieving the task of stream segregation. These principles include[15]:

- 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. Structured audio

4.1 Overview

The representation developed is essentially a parametric description of the three sound classes identified in section 4.2. Incoming audio data is analysed to produce a TFD. The peaks in amplitude of this distribution are found and tracked through time and frequency. Each resultant track is then classified according to type (tone, noise or sweep) and parametrically recorded. Finally, the tracks are grouped according to psychoacoustic principles and encoded in these groups. The process is illustrated in Figure 2.

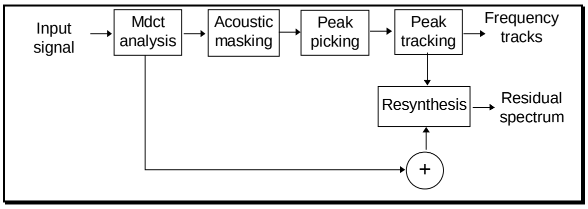

Figure 2: Coding algorithm

4.2 TFD generation

Sinusoidal transform coding[16] represents audio signals as a sum of sine waves with time varying amplitude and frequency as follows:

$$s(n) = a_i(n)\exp(j\phi_i) \qquad (1)$$

where $s(n)$ is the sampled audio signal and a_i and ϕ_i are the time varying amplitude and phase respectively of the *i*th sine wave. Thus, the signal is described as a series of amplitude trajectories through time-frequency space. This technique would seem ideal for the purpose, however, it is not completely suitable and has been adapted in two ways.

Conventionally, the parameters are estimated using a short time Fourier transform (STFT) and the TFD is sampled uniformly in time and frequency. This leads both to redundancy and poor noise performance. To eliminate redundancy and to avoid undesirable blocking effects, a modified discrete cosine transform (MDCT) is used instead of the customary STFT. Further, the TFD is perceptually tuned, mimicking the time-frequency resolution of the ear. The tiling in the time-frequency plane is shown in Figure 3.

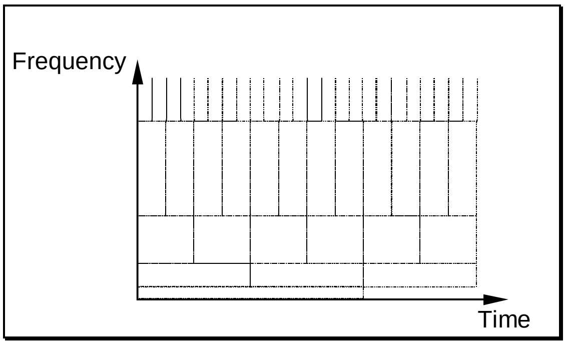

Figure 3: Time-frequency resolution of the TFD

The TFD is generated using the filtering operation described in Figure 4.

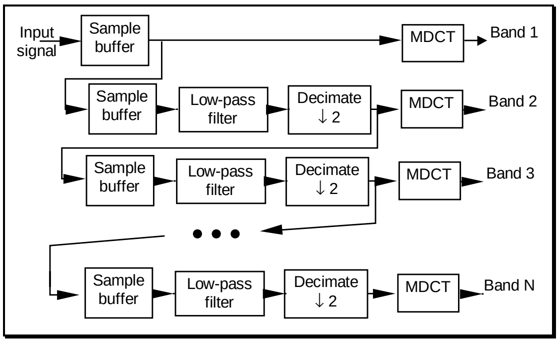

Figure 4: DCT-based TFD generation

The result of the operation shown in Figure 4 is five sets of MDCT coefficients. The details of the time-frequency resolution of each band are given in Table 1. Note that the temporal resolution is half the window length since the MDCT is implemented with 50% overlap.

Band Number	Maximum Frequency (kHz)	Frequency Resolution (Hz)	Temporal Resolution (ms)
1	16	31.25	16
2	8	15.63	32
3	4	7.81	64
4	2	3.91	128
5	1	1.95	256

Table 1: Frequency band data

To facilitate the down sampling operation, which removes redundant data in the low frequency bands, the input signal is recursively filtered. A FIR filter is used which has the impulse response described by (2). To reduce pass band ripple, a hamming window is applied to the filter coefficients. The frequency response of the filter is shown in Figure 5.

$$h(n) = \frac{\sin\left(\omega_c\left(n - \frac{N-1}{2}\right)\right)}{\pi\left(n - \frac{N-1}{2}\right)}, \qquad (2)$$

where $\omega_c = \frac{\pi}{2}$ and $N = 81$.

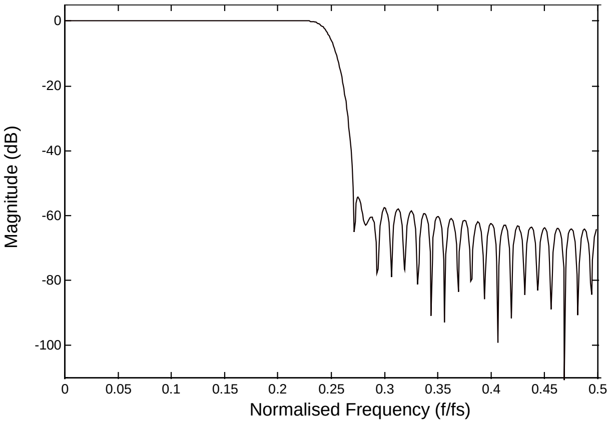

Figure 5: Filter frequency response

Let $x_m(n)$ be a window of the signal at time m. The MDCT coefficients are given by[17]:

$$X_m(k) = \sum_{n=0}^{N-1} h(N-1-n)\, x_m(n) \cos\left(\frac{2\pi\left(k+\frac{1}{2}\right)\left(n+n_0\right)}{N}\right), \qquad (3)$$

$$0 \le k \le N-1$$

where $h(n) = \sin\left(\frac{2\pi(2n+1)}{4N}\right)$ is a window function and $n_0 = \frac{N}{4} + \frac{1}{2}$ is a phase term. Since the MDCT returns only N/2 unique coefficients for a length *N* window, 50% overlapped windows can be used without increasing the data rate. The use of this overlap is desirable because it helps to eliminate blocking effects.

4.1 Masking thresholds

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

The first step in calculating acoustic masking thresholds is to transform the MDCT spectrum from the physical domain, *f*, into the critical band (Bark) domain, *z*. This is achieved using the expression[18]:

$$z = \frac{26.81f}{(1960 + f)} - 0.53 \qquad (4)$$

The critical band spectrum is then convolved with the basilar membrane spreading function to generate the masking thresholds. The basilar membrane spreading function is illustrated in Figure 6 and can be obtained using[19]:

$$\Lambda(z)_{\mathrm{dB}} = 15.8114 + 7.5(z + 0.474) - 17.5\left(1 + (z + 0.474)^2\right)^{1/2} \qquad (5)$$

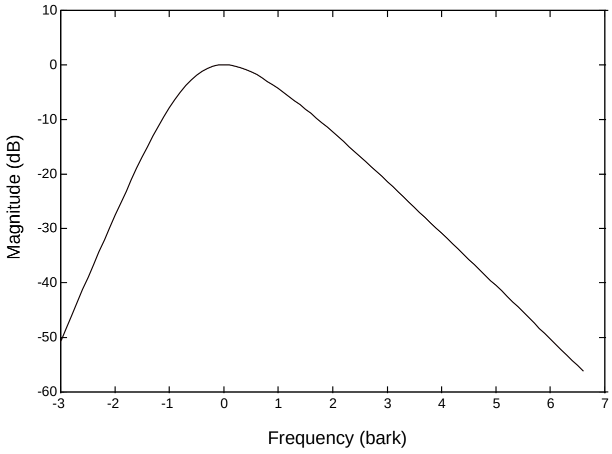

Figure 6: Basilar membrane spreading function

In addition to acoustic masking thresholds, quiet thresholds are also applied. These thresholds mask out any components that are too low in intensity to be audible. Thresholds are those provided in the MPEG1 audio standard [20]. Linear interpolation is used to derive intermediate values not reported in the standard. The resultant masking curve is shown in Figure 7.

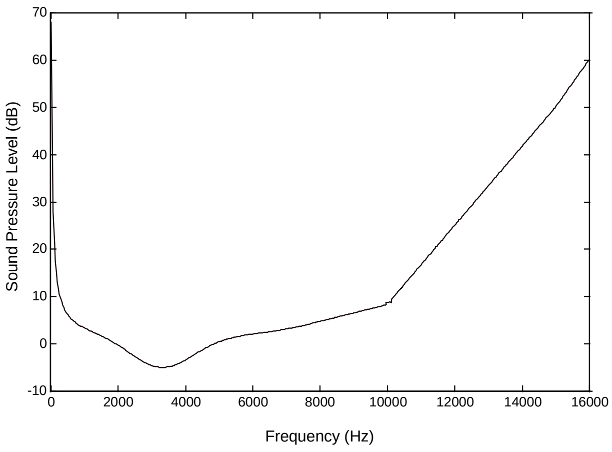

Figure 7: Absolute threshold of hearing

Having obtained the thresholds, their application is relatively straightforward. Firstly, the acoustic masking thresholds are transformed back into the physical frequency domain. These thresholds are then compared with the quiet thresholds. Whenever the masking threshold falls below the quiet threshold, it is replaced with the quiet threshold. The signal spectrum is then compared with the threshold. All coefficients that fall below the threshold are set to zero.

4.2 Peak picking and tracking

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:

$$\begin{gathered} X_t(f_i) > 0 \text{ AND } \left(X_t(f_{i-1}) < X_t(f_i) > X_t(f_{i+1})\right), \\ \text{OR} \\ X_t(f_i) < 0 \text{ AND } \left(X_t(f_{i-1}) > X_t(f_i) < X_t(f_{i+1})\right) \end{gathered} \qquad (6)$$

where $X_t(f_i)$ is the amplitude at the ith frequency, f_i, in the current time frame, t. Expression (6) is used instead of a more straight forward test for the maximum of absolute values since problems were encountered with coefficients of similar magnitudes but opposite signs. Reconstruction quality suffers when the ‘negative peaks’ are ignored. This is a characteristic of the MDCT, which is a result of its basis functions.

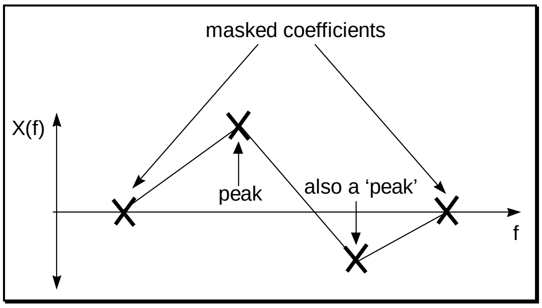

Figure 8: Peak picking

The result of the peak picking stage is a list of peak amplitudes and frequencies for each time frame. Tracking is performed according to the algorithm in [16] which involves matching peaks from adjacent time frames that are closest in frequency, within set limits. The process can be summarised as follows:

1. Assume that a set of tracks is currently in existence (initially, the first frame of peaks are taken as the existing tracks). Denote their frequencies as ω_{mk}, where k is the track number and m is the frame number.
2. For each track, search the next frame of peaks, $\omega_{(m+1)j}$, (j is the peak number) for all peaks whose frequency falls within Δ of ω_{mk} of the last frequency in the track. Mark these as suspect peaks.

3. For each suspect peak, check to see if there is a better match in frame m. If a better match cannot be found for one or more suspects, the closest in frequency to ω_{mk} is appended to track k. Otherwise, track k is said to ‘die’.
4. Any remaining unmatched peaks in frame m+1 are added to the list of tracks.

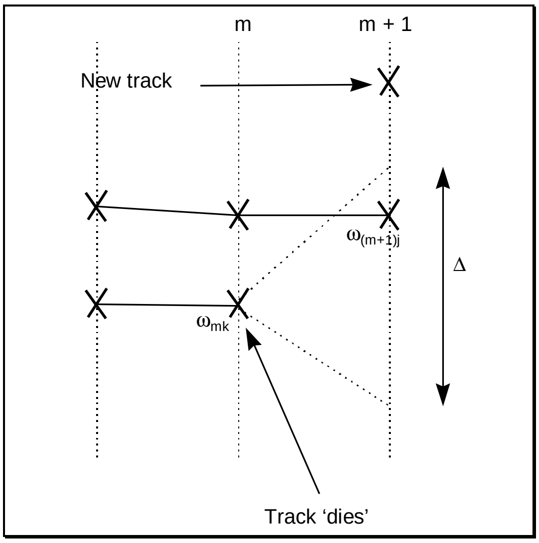

Figure 9: Peak tracking

This algorithm has been modified to overcome specific some problems encountered. Firstly, the varying resolution of the TFD means that adjacent frames at low frequencies are much further apart in time than those at higher frequencies. This means that two frames that are adjacent at low frequencies will not be adjacent in higher bands. This is resolved by fixing the time index, m, relative to the shortest analysis window (m = 1 corresponds to t = 16msec) and recording the current time index each time a peak is assigned to a track.

Secondly, ambiguous cases often arise where two peaks are equally likely candidate matches for a single track. This is particularly relevant to the case of frequency sweeps in close frequency proximity. To resolve such cases, amplitude matching is performed in addition to frequency matching. In particular, this has helped to remove some of the horizontal bias that the algorithm appears to display. As another solution to this problem, Hough transforms, is being considered as an alternative to this algorithm.

4.3 Track description

The peak picking and tracking stage results in a set of tracks which describe the audio. The following information is 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.

The encoding of this representation is yet to be implemented. However, at this stage, two possibilities are being considered for the amplitude and frequency contours: 3 dimensional chain codes and polynomial description. These need to be evaluated in terms of coding gain, complexity and the ease with which contour information can be extracted.

4.4 Track classification and segregation

Once the tracks have been generated, each track is classified according to type: noise burst, tone burst or frequency sweep. Very short tracks are classified as noise bursts. Examining the frequency contour of the track makes the decision as to whether a long track belongs to the class of tone burst or sweep.

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 is applied. All tracks with similar onset times are considered as possible members of a single group. If there is a further relationship between these tracks, a group is formed. This relationship may be harmonicity, or correlated frequency or amplitude modulation. These principles are illustrated in Figure 10. Finally, the tracks are encoded in groups.

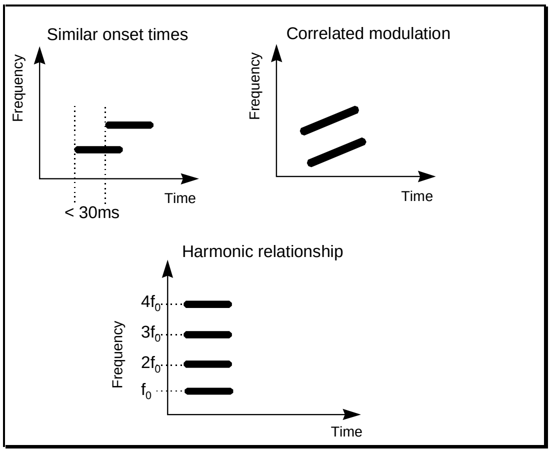

Figure 10: Psychoacoustic grouping rules

5. Suitability for 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 exhibits unique characteristics in the time-frequency domain that are directly visible in the track representation.

Identifying silence is a trivial matter as it is characterised simply 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. Another characteristic of speech that is directly visible in this representation is its almost unique formant structure[7]. 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 depend on the type of sound.

For music data, the melody line can be determined by following the pitch along tracks. Similarly, rhythm can be extracted by analysing the amplitude contour. 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 or rhythm information extracted and used as a basis for comparison.

In the case of speech, change of speaker or gender may be determined by examining the pitch. 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 an instant of time (noise or tone). The suitability of this representation for speech recognition is yet to be investigated. However, the work of [21] suggests that query by example should be supported since it is proposed that, although the same utterance spoken at different times will posses slightly different track structures, there will be a simple transformation between the two structures which is constant across tracks.

There are at least two methods by which pitch information can be extracted from this representation. The first is spectral compression and the second involves calculating two pitch measures (known as spectral pitch and virtual pitch)[22]. Both methods require that the perceptually significant tonal components of the signal first be isolated. This is precisely what the frequency tracks of the representation presented in this paper describe. Thus, the extraction of pitch information is simplified over existing representations.

It should be noted that all the indexing attributes currently used in audio retrieval systems can also be derived from this representation. Indeed, the extraction of this information will generally be simplified. Pitch information is one example discussed earlier. Other examples include loudness and harmonicity. Loudness is easily determined from the amplitude contours. Harmonicity follows directly from the grouping of tracks. In addition, the variation of these qualities can be determined over time, instead of being confined to an average value over the entire audio segment. Thus, the structured audio presented here is capable of supporting all existing query methods as well as cognitively significant structure based queries.

With index attributes directly accessible, the representation clearly supports content based retrieval. Support for browsing may be less obvious but is nonetheless accounted for in a number of ways. Firstly, random access is furnished by the track structure since individual semantically congruent passages can be easily recovered and decoded. In the case of signals with inherent structure (ie music and speech) the track representation reveals the perceptually relevant aspects of this structure, for example phrases are indicated by short periods of silence which in turn correspond to sections containing no tracks. Also, broad classification according to type, as discussed earlier, can also be useful for browsing. These access methods are basic extensions of time-based browsing methods. The most significant advantage of structured audio is its ability to support true content based browsing.

The information that is readily accessible in the track description gives powerful support to content based browsing. This is because it is relatively straightforward to recover all sections of a recording that contain some semantically significant attribute. Examples of such attributes include vibrato, specific chord sequences and changes of key. This level of information is impossible to extract from holistic statistical descriptions. In some cases, such as vibrato or timbre, even complete transcriptions will fail. Thus by using perceptually congruent structures, an increased level of support for browsing has been achieved.

6. Conclusions and future work

Having developed the audio representation, there are two directions for future work. The first is to resolve the encoding issues and the second is to verify the suitability of the representation for its intended purpose of audio retrieval. In addressing the first issue, the desire to provide a compact representation must be carefully approached so as not to compromise access to the salient features of the structure.

The second task will receive the most attention and will involve the development of methodologies and algorithms, based on the newly developed structure, to perform classification, content based retrieval and browsing of audio. As has been indicated in Section 5, the cognitive congruence of the attributes promises to simplify many retrieval tasks as well as introducing many new query possibilities.

Of all the media types, audio retrieval has received the least attention. This paper has attempted to address this oversight by reviewing the relevant issues and proposing a solution. Specific problems encountered by the few existing audio retrieval systems have also been reviewed.

Existing unstructured audio representations have been shown to make content based retrieval difficult and browsing virtually impossible or of little value. A structured audio representation, based on psychoacoustic principles has been suggested as a solution to the problem and the benefits of such a representation stated. The relevant perceptual attributes of audio have also been discussed.

The new structured representation has been described in detail. This representation is based on psychoacoustic principles and has been designed to provide direct access to perceptually salient attributes of audio signals. In addition, the structure has been greatly influenced by cognitive principles. Finally, the suitability of this representation for content based retrieval and browsing has been discussed.

The key feature of the audio structure presented here is its cognitive congruence. The information that can be readily extracted from the track description has true semantic significance. This is in contrast to the feature sets of existing systems that rely on statistical attributes that have only incidental significance.

7. References

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

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

[3] S. Tanguine, “A Principle of Correlativity of Perception and its Application to Music Recognition”, *Music Perception*, summer 1994, 11 (4), pp. 465-502.

[4] P.J.V. Aigrain, P. Longueville, 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] D. Hindus, C. Schmandt, C. Horner, “Capturing, Structuring and Representing Ubiquitous Audio”, *ACM Transactions on Information Systems*, vol 11, no 4, Oct 1993, pp 376-400.

[6] B. Arons, “SpeechSkimmer: Interactively Skimming Recorded Speech”, Proc. USIT 1993: ACM Symposium on User Interface Software and Technology, Nov 1993.

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

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

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

[10] 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.

[11] 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.

[12] S. Goldhor, “Recognition of Environmental Sounds”, IEEE ICASSP, vol 1, 1993, pp 149-152.

[13] 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.

[14] 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.

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

[16] 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.

[17] P. Duhamel, Y. Mahieux, J. P. Petit, “A Fast Algorithm for the Implementation of Filter Banks Based on ‘Time Domain Aliasing Cancellation’”, IEEE ICASSP, pp. 2209-2112, 1991.

[18] M. Paraskevas, J. Mourjopoulos, “A Differential Perceptual Audio Coding Method with Reduced Bitrate Requirements”, *IEEE Trans ASSP*, v. 3, n. 6, Nov 1995.

[19] M.R. Schroeder, B. S. Atal, J. L. Hall, “Opimizing digital speech coders by exploiting masking properties of the human ear”, *J. Acoust. Soc. Amer*., **66** (6), Dec 1979, pp 1647-1651.

[20] ISO/IEC 11 172-3.

[21] A. B. Fineberg, R. J. Mammone, “Detection and Classification of Multicomponent Signals”, Proc. 25th Asilomar Conference on Computer, Signals and Systems, Nov 4-6, 1991.

[22] E. Terhardt, G. Stoll, M. Seewann, “Algorithm for extraction of pitch and pitch salience from complex tonal signals”, *J. Acoust. Soc. Am*., **71** (3), March 1982.