SOURCE SEGMENTATION FOR STRUCTURED AUDIO

Kathy Melih and Ruben Gonzalez

School of Information Technology,

Griffith University, Gold Coast

PMB 50 Gold Coast Mail Center QLD 9726, Australia

ABSTRACT

With increasing demand for content based manipulation of ever growing stores of audio data and the emergence MPEG-7 has come the call for structured audio representations. However, while the necessity of such a representation has been recognised and, to some extent, its essential features have been identified, its actual development and implementation have generally been relegated as problems for another time or person to solve. This paper attempts to address the shortfall by defining an audio structure that will allow content-based manipulation of audio at the level of audio objects. The paper then summarises the processes required to generate such a structure. Further, details are provided as to how the second level of this structure can be derived from a low-level perceptually based audio representation previously developed by the authors to satisfy the requirements at the lowest level of the audio structure. Finally, initial experimental results are presented.

1. INTRODUCTION

As a result of the increasing demand for content-based management and manipulation of audio data, a need for structured audio representations has been identified. The urgency of this requirement is most obvious in the light of the emerging MPEG-7 standard: "MPEG-7 shall support descriptors that can act as handles referring directly to the data, to allow manipulation of the multimedia material." [1] That is, direct access to content information must be available in or along side the encoded data stream itself. Obviously, providing such access necessarily imposes a structure on this data stream.

The authors have previously described the low level aspects of a perceptually based, structured representation that was designed specifically for content based retrieval and manipulation of audio data [2][3]. This paper details the nature of the mid-level structured description that is readily derived from this low-level representation and outlines preliminary experiments in automatically extracting the mid-level description.

2. STRUCTURED AUDIO

The general definition of structured audio representations: "description formats that are made up of semantic information about the sounds they represent and that make use of high-level or algorithmic models" [4] encompasses many representations that do not fill the requirements imposed by generic audio management. One such example is a MIDI event list. While this representation may be a useful basis for melody retrieval[5], it cannot be applied to audio of a more generic nature, neither is it a trivial matter to generate such a representation automatically from recorded music. This section briefly describes the general characteristics of a representation that would support content-based manipulation and then presents a specific structure that possesses these characteristics.

2.1 Required characteristics.

A structured representation suitable for content-based retrieval must possess several key attributes. Firstly, the data must be divided into semantically relevant units. Secondly, these units should be individually decodable and randomly accessible. Further, it is desirable that these units be extracted automatically from a raw audio stream. Indeed, in the case of separating two simultaneously occurring units (speech over background music, for example) it is impossible to do so manually. Finally, useful information about the content of the unit should be co-located with the unit itself. The current solution of including manually or semi-automatically generated text annotations is far from ideal.

2.2 Proposed structure

Figure 1 shows a structural decomposition of an audio stream useful for content-based retrieval and manipulation of audio.

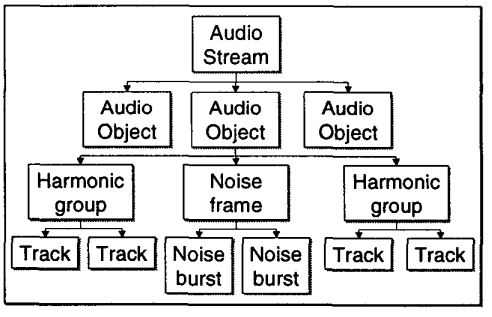

Figure 1: Structuring audio using a perceptually based representation

The lowest level contains the perceptually relevant atomic features of an audio signal. These model the features extracted by the human auditory system at the lowest levels of auditory perception. Essentially two main classes of feature are extracted: a track and a noise burst. The auditory system distinguishes two track classes: tone and sweep. We have introduced two additional classes for reasons stated in section 3.1. Track formation is detailed in [2].

Harmonic groups and noise frames appear at the next level up. The harmonic groups consist of tracks that are co-located in time and bear some resemblance to one another. In general, the frequencies of the tracks belonging to a single group will be harmonically related. Also, their frequency and amplitude contours will be scaled versions of a single 'prototype' contour. Noise frames consist of temporally adjacent noise bursts with similar characteristics (bandwidth, RMS power, spectral envelope, etc). This paper concerns itself with the formation of these mid-level groups.

The next level up in the hierarchy shown in Figure 1 contains audio objects. This level could itself be expanded, depending on

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

the context-dependent definition of an audio object. In speech, for example, the lowest level audio object might be a phoneme or single word. At the next level we might find a single spoken phrase and higher still, a long monologue. At which level the user interacts depends on what he aims to do with the objects. If automatic transcription were the aim, then the simplest level audio objects (phonemes) would suffice. On the other hand, someone wishing to interactively skim (or edit) a lengthy recorded meeting would interact with higher level objects (monologues/phrases).

In compliance with the requirements stated earlier, the elements of the structure are randomly accessible and individually decodable. From a user's perspective, the relevance of this increases significantly at each step up the hierarchy. While the elements at lower levels form perceptually significant units, they contain incomplete cognitive information. Decoding a single track at the lowest level results only in a single sinusoid that may or may not be frequency modulated. This obviously has very little audition value. Inverting a single harmonic group may well result in an intelligible signal. However very few naturally occurring sounds are purely tonal or noise-like. Hence, it is much more likely that a combination of at least two mid-level elements is required to create an inversion that is both intelligible and of acceptable quality. It follows that inverting an individual audio object should result in a signal that is both intelligible and of reasonable perceptual quality.

The utility of inverting individual low-level elements may appear somewhat limited; nevertheless, the ability to gain access to such components is highly desirable. Firstly, much useful information about a sound can be inferred from the characteristics of its constituent tracks and noise bursts. Further, manipulation of these elements may well be desired to edit the sound in a manner that would be impossible at the level of raw audio stream. Finally, it is obviously impossible to form the higher level elements without first decomposing the signal into these fundamental units.

2.3 Generating and using the structure

Figure 2 gives an overview of the audio structuring process.

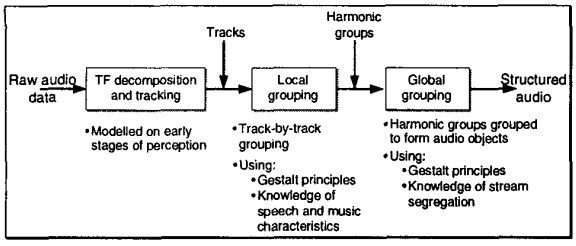

Figure 2: Overview of structured audio generation

Strictly speaking, the representation can already be considered as structured according to the definition appearing in the introduction after only the first stage of processing. The algorithmic model used is the same as that used in the lower levels of human audition. Also, much useful semantic information (pitch contours leading to melody description and others [2]) is easily derived from this track-based description. However, insufficient organisation exists at this level to enable interaction with and manipulation of the data beyond the most trivial cases. The situation is akin to attempting to edit a text document one character at a time. While it may be necessary to interact with the data at this level on small isolated regions (to fix spelling errors, for example) it is tedious to do so when reviewing the overall structure of the document. Further, it is impossible to determine the semantic content of a document from the characteristics of an individual character. Indeed, the derivation of semantically useful information mentioned earlier presupposes that the tracks under analysis belong to a single audio object (i.e., that the higher level structuring has already taken place).

At the other extreme, we find the most structured and highly organised view of the data. At this level the user interacts with a single semantically complete unit. They may, for example, wish to change the order of pieces in a concert recording or to remove the background music from a recorded speech. In the text analogy, this is equivalent to interacting with a document at the paragraph or page level. The former example corresponds to "cutting and pasting" blocks of text while the latter most closely resembles the overlaying of text and images. However, while interaction at this level is the most powerful and useful, automatic generation of such descriptions is far from trivial. Manual generation, on the other hand, is at best tedious and at worst, especially in the case of separating simultaneously occurring objects, virtually impossible.

The mid-level harmonic groups and noise frames serve as a bridge between the two extremes. Implicit in the formation of mid-level objects, is source separation. Tracks derived from a single source tend to exhibit similar time-frequency-amplitude characteristics. It follows that in grouping tracks accordingly, one separates tracks originating from individual sources. Thus, harmonic groups are an ideal basis for audio objects. In the text analogy, mid-level objects would most closely resemble individual words or picture elements (for overlaid graphics). It is possible that an individual group will form an object by itself, as in phoneme extraction for speech recognition. However, it is much more likely that even the lowest level audio object will consist of more than one temporally adjacent and/or simultaneous harmonic group and/or noise frame.

3. FORMING HARMONIC GROUPS

3.1 Track shape determination

The first stage in creating harmonic groups is to determine the shape of the individual tracks. This is relatively straightforward given appropriately encoded tracks. In [3] chain code was suggested as a useful description from both coding and shape description perspectives. A more efficient parametric representation describes the frequency and amplitude contours with cubic splines. In either case, the 'shape' of a track is determined by matching its frequency contour against a small set of generalised track shapes. This set is shown in Figure 3.

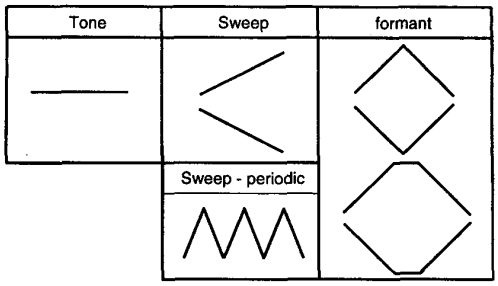

Figure 3: Track shapes and their classification

Figure 3 shows the only allowed track shapes. Any tracks that do not fall into one of the categories shown are sub-divided until each individual section can be classified according to Figure 3. This ensures that each track describes some portion of a single acoustical event. For example, long tracks that are generated for music signals are broken up at individual note boundaries. In order to properly represent noise bursts, a second description is necessary. The information required to adequately characterise a

noise burst is the overall RMS power for the burst and the general shape of the magnitude spectrum of the noise. This set of primitive elements is an extension of the three primitives hypothesised to exist in human audio cognition: sweep, tone and noise burst. The formant and periodic sweep classes have been introduced to deal with speech and vibrato respectively without breaking tracks into unnecessarily short (sub-elemental) lengths.

3.2 Group formation rules

Having determined the shape of each track, the next stage involves the actual grouping. The group formation procedure applies rules derived from the following Gestalt principles[6]:

- Belongingness: All partials[1] must be assigned to a group.
- Exclusive allocation: A partial may only be assigned once.
- Similarity: Similar partials tend to form a single stream.
- Common fate: Partials that begin and/or end simultaneously and vary similarly tend to be fused into a single stream.
- Continuity: Partials are grouped to favour gradual changes rather than abrupt changes along the stream.
- Closure: Gaps in partials tend to be filled when there is evidence that they have been masked.
- Proximity: Partials in close temporal or spectral proximity tend to be fused into the same stream.

The first four principles in the list are most useful for harmonic group formation while the latter three items in the list are most useful for audio object formation. In addition to these Gestalt principles, one key characteristic of speech and music signals influences group formation: harmonicity. The tracks making up most speech and music sounds are harmonically related, hence the tracks comprising a harmonic group are constrained to be similarly related. The rules governing harmonic group formation are summarised in Figure 4 below.

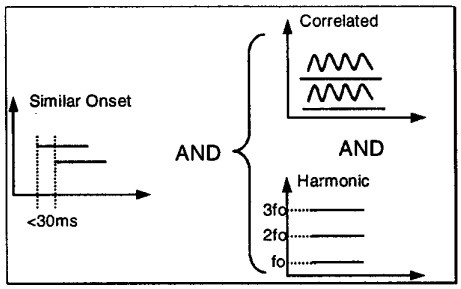

Figure 4: Harmonic group formation rules

3.3 Algorithm

A relatively straightforward algorithm achieves harmonic group formation. The list of tracks is searched for the lowest, ungrouped track. The length of this track is constrained to be greater than a threshold value to prevent the inclusion of very short noise tracks. (These short tracks will eventually be subsumed into noise frames). The lowest track found is then used as the basis for a model track shape. This model track shape is basically a spline description of the track's frequency and amplitude contours. The frequency contour of the model is then multiplied by successive harmonic numbers and the list of tracks is searched for any tracks that coincide, within limits, with the model over the extent to which they overlap in time. This test is slightly less rigorous than implied by the rules illustrated in Figure 4 since the common onset time restriction is ignored. This modification is necessary to cope with tracks describing speech or rapid musical passages where onset times are often uncertain due to noise sections at these onsets. An improvement yet to be implemented is to relax the common onset rule only for short tracks that succeed a noise frame. Figure 5 details the algorithm.

```
Set group number to 1
WHILE ungrouped tracks remain
    Find lowest track of length greater than min
    Extract frequency contour
    Set harmonic number to 1
    WHILE max frequency in contour < max allowed
        find all tracks that match frequency contour
        assign to group
        increment harmonic number
        multiply contour freq^s by harmonic number
    END WHILE
```

Figure 5: Harmonic group formation algorithm

4. PRELIMINARY RESULTS

At this early stage of algorithm development, three simple test files have been used to verify the validity of the technique. However, even this simple case poses some non-trivial problems that will be discussed in the succeeding sections. The files contained speech, music and a combination of the two respectively. The speech file consisted of a female speaker counting from one to four. The music file contained the first four notes of the G major scaled played in the middle register of a solo flute. The combined file was an artificial mix of the other two signals such that each spoken digit occurred simultaneously with the approximate onset of each note. All files were windows PCM WAV files sampled at 32kHz with 16 bit resolution.

Figure 6 shows the tracks generated for the mixed source data.

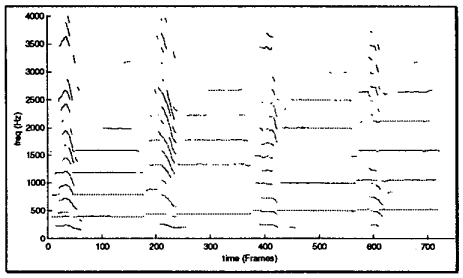

Figure 6: Tracks generated for mixed source data

The tracks obviously display characteristics of both speech and music. Of particular interest is the proximity in frequency of the formants in the spoken sections to some of the harmonics in the music sections. This poses something of a challenge for the grouping, namely: when a single track is an equally good match (in harmonic frequency terms) for two groups, to which should it be assigned? This question is resolved by matching track shapes. In the case illustrated, the tracks (which have been broken by the track classification procedure) are assigned to the group corresponding to the speech tracks. Figure 7 illustrates the result.

[1] Note: A "partial" roughly corresponds to a "track" and a "stream" roughly corresponds to either a "group" or an "audio object".

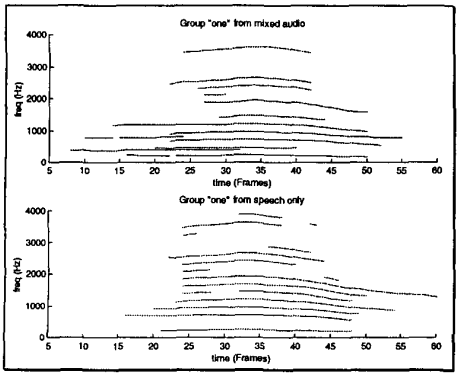

Figure 7: Track groups automatically generated for the number "one" in both the single source speech and mixed data files

This solution is both pragmatic and perceptually motivated as it fulfils the requirements of the first three Gestalt principles listed in section 3.2. However, as a result of this decision, the simultaneously occurring music group is left incomplete with one of two possibilities. The first of these, occurring if the speech begins sometime after the beginning of the note, is that the music group will most likely form two separate groups. While this may appear to be a sub-optimal grouping, application of the last three Gestalt principles listed in section 3.2 resolves this apparent problem. If, on the other hand, the speech utterance begins more or less simultaneously with the note, the note onset is completely lost. The result of this is illustrated in Figure 8.

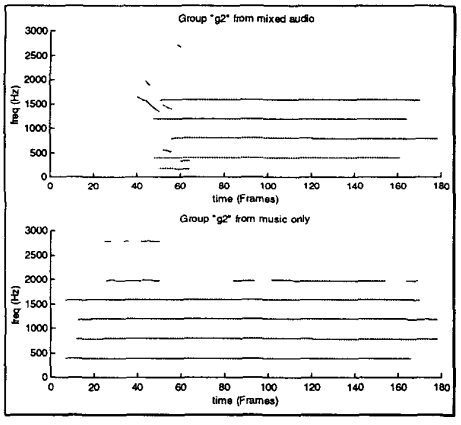

Figure 8: Harmonic group extracted for the note G2 from both single source and mixed source files. Note that the onset information is lost due to interaction between the harmonics of the music signal and the formants of the speech signal.

This case does pose a significant challenge for deciding on the actual onset characteristics of the tracks. However, this same problem exists for the human perceptual system with ambiguities often resulting in errors that cause auditory illusions[6]. As it stands, the grouping will be sufficient for the purposes of identifying "what" sources occur approximately "when" and thus structuring the audio according to Figure 1. However, the loss of onset information is unacceptable from an inversion perspective since the characteristics of the onset heavily influence the perceived timbre. The strategy that will be adopted to deal with such instances is yet to be determined.

5. DISCUSSION AND CONCLUSIONS

A method for source segregation using a perceptually motivated, structured audio representation has been presented. The source segregation forms the basis of a fully structured representation aimed at providing access to individual audio objects within a generic audio stream. In short, the representation essentially hopes to solve the one issue that is at the heart of the MPEG-7 requirements but is yet to be resolved.

This paper has described the essential features of the structure and given a brief overview of the processes required to generate such a structure. Further, the utility and essential features of the elements at each level of the structure has been detailed. Next an algorithm for the generation of the mid-level elements has been presented. Finally, the results of initial experiments in implementing this procedure have been reported.

6. REFERENCES

[1] ISO, "MPEG-7 Requirements Document V.9", N2859, Vancouver Canada, July 1999.

[2] Melih, K. and Gonzalez, R., "Audio Retrieval Using Perceptually Based Structures", *Proc. IEEE International Conf. On Multimedia Computing and Systems '98*, Austin, Texas, 28 June - 1 July 1998, 338-347.

[3] Melih, K. and Gonzalez, R., "Structured Coding for Content Based Interactive Audio", *Proc. IEEE International Conf. On Multimedia Computing and Systems '99*, Florence, Italy

[4] Vercoe, B. L., Gardner, W. G., Scheirer, E. D., "Structured Audio: Creation, Transmission, and Rendering of Parametric Sound Representations", *Proc. of the IEEE*, vol. 86, no. 5, May 1998, pp. 922-940.

[5] Ghias, A., Logan, J., Chamberlin, D., Smith, B.C., "Query by Humming", *Proc. ACM Multimedia 95*, Nov 1995, 231-236

[6] Bregman, A.S., Auditory Scene Analysis: the Perceptual Organisation of Sound, MIT Press, 1990