PERCEPTUAL SEGMENT CLUSTERING FOR MUSIC DESCRIPTION AND TIME-AXIS REDUNDANCY CANCELLATION

Tristan Jehan
Massachusetts Institute of Technology
Media Laboratory

ABSTRACT

Repeating sounds and patterns are widely exploited throughout music. However, although analysis and music information retrieval applications are often concerned with processing speed and music description, they typically discard the benefits of sound redundancy cancellation. We propose a perceptually grounded model for describing music as a sequence of labeled sound segments, for reducing data complexity, and for compressing audio.

1. INTRODUCTION

Typical music retrieval applications deal with large databases of audio data. One of the major concerns of these programs is the meaningfulness of the *music description*, given solely the audio signal. Another concern is the efficiency of *searching* through a large space of information. With those considerations, some recent techniques for annotating audio include psychoacoustic preprocessing models [1], and/or a collection of frame-based (i.e., 10-20 ms) perceptual audio descriptors [2] [3]. The data is highly reduced, and the description hopefully relevant. However, although the annotation is appropriate for sound and timbre, it remains complex and inadequate for describing *music*, a higher-level cognitive mechanism. We propose a meaningful, yet more compact description of music, rooted on the segmentation of audio events.

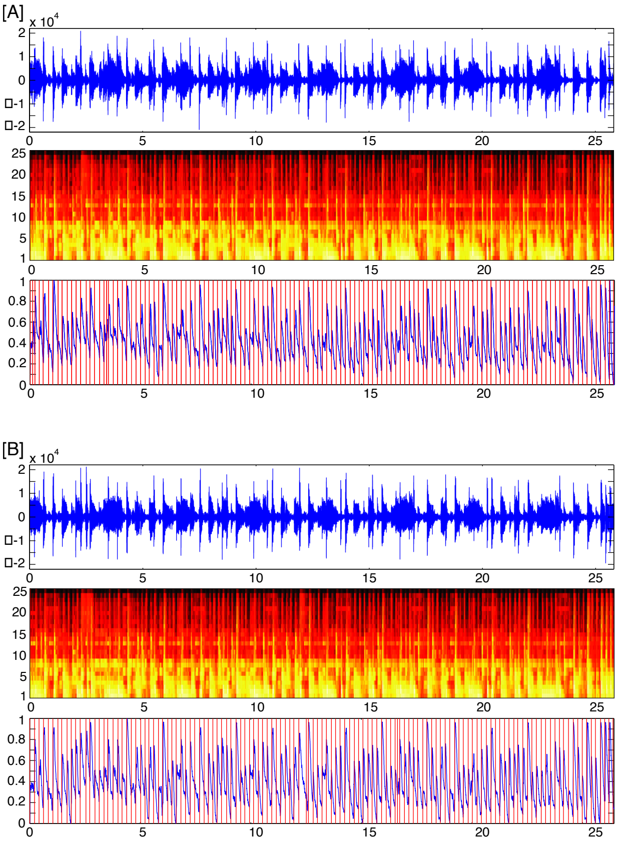

Figure 1. [A] 26-second audio excerpt of “Watermelon Man” by the *Headhunters* (1974). From top to bottom: waveform, auditory spectrogram, and loudness curve with segmentation markers (129 segments of about 200 ms). [B] resynthesis of the piece with only 30% of the segments (less than 8 seconds of audio). From top to bottom: new waveform, auditory spectrogram, loudness curve, and segmentation. Note that there are few noticeable differences, both in the time and frequency domains.

In [4], Jonathan Foote and Matthew Cooper introduced a novel approach to musical structure visualization. They used self similarity of Mel-frequency cepstral-coefficient feature vectors as a signature for a given audio piece. From the resulting matrix could be derived a representation of the rhythmic structures, which they called *beat spectrum*. In [5], they proposed a statistically-based framework for segmenting and clustering large audio segments via *Singular Value Decomposition*. The analysis could for instance return the structural summarization of a piece, by recognizing its “most representative” chorus and verse patterns. Our approach, on the other hand, starts with a perceptual technique for describing the audio spectral content *first*, derives a meaningful segmentation of the musical content *then*, and only *later* computes a matrix of similarities. Our goals are both description and resynthesis. We assume no prior knowledge about the music being analyzed. For instance, the segment sizes are automatically derived from the music itself.

2. PSYCOACOUSTICALLY INFORMED SEGMENTATION

Segmenting is the process of dividing the musical signal into smaller units of sounds [6]. A sound is considered *perceptually meaningful* if it is timbrally consistent, i.e., it does not contain any noticeable abrupt changes. We base our segmentation on an *auditory model*. Its goal is to remove the information that is the least critical to our hearing sensation, while retaining the important parts.

We first apply a running STFT, and warp the spectrum to a 25-critical-band Bark scale. We then model the non-linear frequency response of the outer and middle ear [7], and apply frequency and temporal masking [8], turning the outcome into a "what-you-see-is-what-you-hear" type of spectrogram [9] (see figure 1-[A]-2). A *loudness* function is easily derived by summing energy across frequency channels (see figure 1-[A]-3).

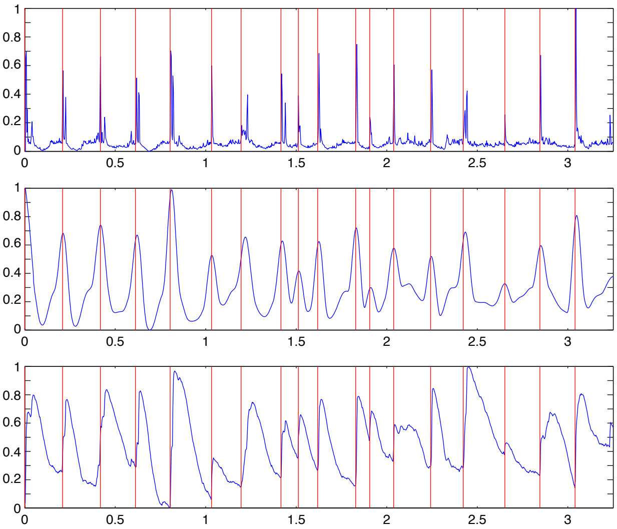

Figure 2. Short 3.2-second audio excerpt extracted from figure 1. From top to bottom: the unfiltered event detection function, the event detection function convolved with a 150-ms Hanning window, the loudness curve. Vertical red lines represent onset markers.

We convert the spectrogram into an *event detection function* by first calculating the first-order difference function for each spectral bands, and by summing across channels (see figure 2-1). Transients are localized by *peaks*, which we smooth slightly by convolving the function with a 150-ms Hanning window to combine those perceptually fused together (i.e., two events separated in time by less than 50 ms [10]). The required onsets can finally be found by extracting every local maxima within that function (see figure 2-2). Since our concern is resynthesis by concatenating audio segments, we refine the onset location by searching the corresponding previous local minimum in the loudness function, and the closest zero-crossing in the waveform (see figure 2-3).

3. LABELING AND SIMILARITIES

Music could be described as an *event-synchronous path* within a perceptual multidimensional space of audio segments. Musical patterns can be recognized as loops within that path. A perceptual multidimensional scaling (MDS) of sound is a geometric model which provides us with the determination of the Euclidean space that describes the distances separating timbres as they correspond to listeners' judgments of relative dissimilarities. It was first exploited by Grey [11] who found that traditional monophonic pitched instruments could be represented in a three-dimensional timbre space with axes corresponding roughly to attack quality (temporal envelope), spectral flux (evolution of the spectral distribution over time), and brightness (spectral centroid). However, little work has previously been done on the similarity of rich polyphonic arbitrary sound segments. We seek to label these segments in a way that the perceptually similar ones fall in the same region of the space. Redundant segments get naturally clustered, and shall be *coded* only once.

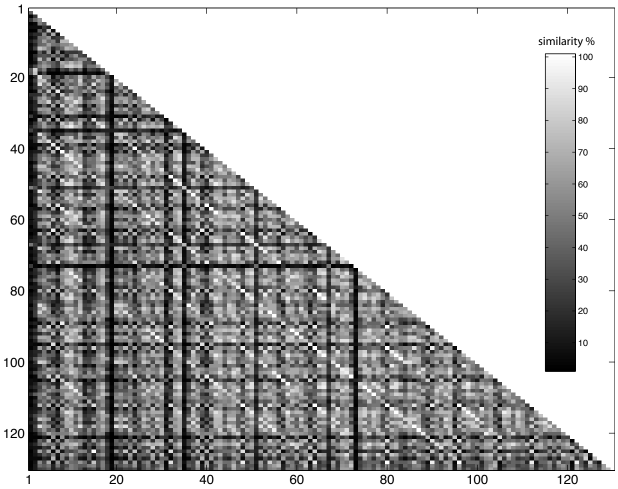

Figure 3. Matrix of perceptual self similarities for the 129 segments of the "Watermelon Man" excerpt of figure 1-[A]. White means very similar, and black very dissimilar. Note the black lines, which represent very unique segments, and the white diagonal stripes, which illustrate pattern redundancies in the music, although the music was fully performed and not loop-based, i.e., no digital copies of the same material.

Our current representation describes sound segments with 30 normalized dimensions. Because segments are small and consistent, 25 dimensions are derived from the average amplitude of critical bands of the auditory spectrogram over time, and 5 are derived from the temporal loudness function (normalized loudness at onset and at offset, maximum loudness, length of the segment, and relative location of the maximum loudness). A more accurate representation taking into account the complete dynamic variations of the spectral envelope, and a dynamic programming approach is currently under development (a collaboration with J.J. Aucouturier from Sony CSL).

However, our preliminary results have been satisfactory.

A very compact, yet perceptually meaningful vector description of the time structure of *musical events* (much like an "audio DNA" symbolic sequence) is now established. We can finally compute the self similarity matrix between segments with, for example, a simple mean squared distance measure (see figure 3). Other distance measures could very well be considered.

4. CLUSTERING AND COMPRESSION

Since the space is Euclidean, a simple k-means algorithm can be used for clustering. An arbitrary small number of clusters may be chosen depending on the targeted accuracy and compactness. The process is comparable to vector quantization: the smaller the number of clusters, the smaller the lexicon and the stronger the quantization. Figure 4 depicts the segment distribution for a short audio excerpt at various *segment ratios* (defined as the number of segments retained divided by the number of original segments). Audio examples corresponding to the resynthesis of this excerpt at various segment ratio settings (see description below), as well as many other examples are available at: `www.media.mit.edu/~tristan/ISMIR04/`

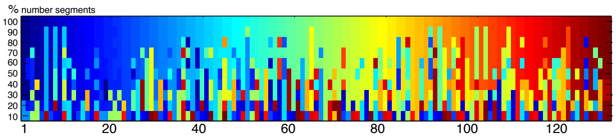

Figure 4. Color-coded segment distribution for the 129 segments of the "Watermelon Man" piece of figure 1-[A] at various segment ratios. 100% means that all segments are represented, while 10% means that only 13 different segments are retained. Note the time-independence of the segment distribution, e.g., here is an example of the distribution for the 13 calculated most perceptually relevant segments out of 129:

33 33 66 66 23 122 23 15 8 112 42 8 23 42 23 15 112 33 33 66 66 66 108 23 8 42 15 8 128 122 23 15 112 33 66 115 66 122 23 15 8 128 42 66 128 42 23 15 112 33 66 115 8 108 23 15 8 42 15 8 128 122 23 115 112 33 66 115 86 128 23 33 115 112 42 8 128 42 23 115 112 8 66 8 66 108 86 15 23 42 15 8 128 122 23 115 112 8 66 115 86 128 23 122 8 112 42 8 108 42 23 115 112 8 66 115 66 108 86 122 23 42 122 23 128 122 23 128 128

Compression is the process by which data is reduced into a form that minimizes the space required to store or transmit it. While modern lossy audio coders efficiently exploit the limited perception capacities of human hearing in the frequency domain [12], they do not take into account the perceptual redundancy of sounds in the time domain. We believe that by canceling such redundancy, we can reach further compression rates. The segment ratio indeed highly correlates with the compression rate that is gained over traditional audio coders.

Perceptual clustering allowed us to reduce the audio material to the most perceptually relevant segments. These segments can be stored along with a list of indexes and locations. Resynthesis of the audio consists of juxtaposing the audio segments from the list at their corresponding locations (see figure 1-[B]). Note that no cross-fading between segments or interpolations were used in our examples.

Currently, our implementation allows us to define a segment ratio, regardless of the music content. However, too few clusters may result in *musical distortions* at resynthesis, i.e., the sound quality is fully maintained, but the musical "syntax" may audibly shift from its original form. A more useful system would in fact adapt its segment ratio to the music being compressed (i.e., the more redundant, the more compressed), and would prefer a *perceptual accuracy* control parameter to our static segment ratio setting. This is currently under development.

5. DISCUSSION AND FUTURE WORK

Reducing audio information beyond current state-of-the-art perceptual codecs by structure analysis of its *musical* content is arguably a bad idea. Purists would certainly disagree with the benefit of cutting some of the original material altogether, especially if the music was entirely performed. There are obviously great risks for music distortion currently and the method applies naturally better to certain genres, including electronic music, pop, or rock, where repetition is an inherent part of its qualities. Formal experiments could certainly be done on measuring the *entropy* of a given piece and the *compressibility* across subcategories.

We believe that, with a real adaptive strategy and an appropriate perceptually grounded error estimation, the principle has great potential, primarily in devices such as cell phones, and PDAs, where bit rate and memory space matter more than sound quality. At the moment, segments are compared and concatenated as raw material. There is no attempt to transform the audio itself. However, a much more refined system would estimate similarities independently of certain perceptual dimensions, such as loudness, duration, aspects of equalization or filtering, and possibly pitch. Resynthesis would consist of transforming *parametrically* the retained segment (e.g., amplifying, equalizing, time-stretching, pitch-shifting, etc.) in order to match its target more closely. This could greatly improve the musical quality, increase the compression rate, and refine the description, consequently enabling additional analysis tasks.

Perceptual coders have already provided us with a valuable strategy for estimating the perceptually relevant audio surface (by discarding what we cannot hear). Describing musical structures at the core of the codec is an attractive concept that may have great significance for many higher-level information retrieval applications, including song similarity, genre classification, rhythm analysis, transcription tasks, etc.

6. CONCLUSION

We propose a *low-rate* perceptual description of music signals based on a psychoacoustic approach to segmentation. The description can be *quantized* meaningfully by clustering segments, and the audio *compressed* by retaining only one segment per cluster. Although the technique is not fully developed yet, promising results were obtained with early test examples. We believe that such approach has potential both in the music information retrieval, and the perceptual audio coding domains.

7. REFERENCES

[1] Elias Pampalk, Simon Dixon, and Gerhard Widmer, "Exploring music collections by browsing different views," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, MD, October 2003.

[2] Perfecto Herrera, Xavier Serra, and Geoffroy Peeters, "Audio descriptors and descriptor schemes in the context of MPEG-7," *International Computer Music Conference*, 1999.

[3] Martin McKinney and Jeoren Breebaart, "Features for audio and music classification," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, MD, October 2003.

[4] Jonathan Foote and Matthew Cooper, "Visualizing musical structure and rhythm via self-similarity," in *Proceedings International Computer Music Conference*, La Habana, Cuba, 2001.

[5] Matthew Cooper and Jonathan Foote, "Summarizing popular music via structural similarity analysis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, October 2003.

[6] George Tzanetakis and Perry Cook, "Multifeature audio segmentation for browsing and annotation," in *Proceedings IEEE Workshop on applications of Signal Processing to Audio and Acoustics*, October 1999.

[7] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, pp. 155–182, 1979.

[8] T. Painter and A. Spanias, "A review of algorithms for perceptual audio coding of digital audio signals," 1997, Available from `www.eas.asu.edu/ speech/ndtc/dsp97.ps`.

[9] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer Verlag, Berlin, 2nd edition, 1999.

[10] Bob Snyder, *Music and Memory: an Introduction*, MIT Press, Cambridge, MA, 2000.

[11] J. Grey, "Timbre discrimination in musical patterns," *Journal of the Acoustical Society of America*, vol. 64, pp. 467–472, 1978.

[12] Marina Bosi and Richard E. Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publishers, Boston, December 2002.