

## Perceptual constraints in systems for automatic music information retrieval

Martin F. McKinney, Michael Bruderer, Alberto Novello

Philips Research Laboratories, HTC 36, 5656 AE, Eindhoven, The Netherlands

*Abstract-* **Perceptual models and data provide useful guidelines in the development of systems for music information retrieval. We show that for music tempo extraction, similarity evaluation, and structure extraction, perceptual data are not only useful in the evaluation of automatic systems but also in their design.**

### I. INTRODUCTION

Recent advances in audio compression technology and increases in storage capacities and broadcast bandwidths allow users to access vast (and growing) amounts of music audio. With this capability comes an increased need for tools to navigate, browse and search for music. Traditionally, such tools have been based on *annotated* metadata, such as genre and artist (refs?), but recent developments in automatic music information retrieval (MIR) are making it possible to extract some types of metadata directly from the music audio signal itself (refs). These additional metadata, combined with traditional annotated metadata, provide a much richer set upon which sophisticated systems for music navigation can be built.

While many aspects of algorithms for MIR can be constrained by, or defined in terms of, musicological (refs?), sociological (refs), and/or computational (refs) factors, we have found that perceptual criteria are often invaluable guides. This is true not only for the evaluation of, but also for the development and construction of MIR systems. While the advantage of using perceptually-based constraints depends somewhat on the application context, we present here three cases in which perceptual knowledge clearly benefits or is needed for the development of algorithms. More specifically, we have examined automatic extraction of music tempo, music similarity ratings, and structural boundaries in popular music.

### II. TEMPO EXTRACTION

Music tempo is a useful statistic for music selection, comparison, mixing and playlisting. While it is common to characterize music tempo by the *notated* tempo on sheet music in beats per minute (BPM), this value often does not correspond to the *perceptual* tempo (refs). We have seen that most pieces of music, in fact, have an ambiguous perceptual tempo in that different listeners attend to different metrical

levels and thus perceive, for example, the tempo to be half or twice that perceived by others (refs). This holds for musicians as well as non-musicians.

For the applications listed above, it is in fact the perceptual tempo that is of interest and a single value is clearly not adequate. We have therefore established a method and protocol in which we report two tempo values as well as their relative perceptual salience. This method was used for the evaluation criteria in the MIREX 2005 tempo extraction contest (ref). In order to establish the “ground-truth” perceptual tempi for the contest, we performed a series of experiments in which we asked listeners to tap to the beat of musical excerpts. We then analyzed the tapped responses and computed the two most salient tempi and their relative salience. Figure 1 shows an example of a histogram of tapped tempi for a single excerpt of music. In this case the responses were almost evenly split between two tempi, each at a different metrical level of the musical excerpt.

**Fig. 1.** Histogram of subjects’ tapped tempi for a single excerpt of music. Two distinct peaks in the histogram indicate a split of perceived tempi across listeners and support the idea of annotating tempo with more than a single value.

In addition to using perceptual criteria for the *evaluation* of tempo extractors, we have also seen benefits in applying perceptually-based signal processing models to algorithms for automatic tempo extraction. Our tempo extractor conforms to a conventional two-stage structure, in which a *driving* signal is derived from the audio signal in the first stage, and is then used to drive a series of periodicity detectors in the second stage. In our case, the driving signal is a pulse train derived from an onset detection process, where each pulse represents the onset of an auditory object. The periodicity detectors are banks of comb filters, similar to those from Scheirer (refs). A block diagram of the system is shown in Fig. 2.

The onset detector features two processing components, modeled after perceptual phenomena, that help boost performance (ref: Schrader masters thesis). The psychoacoustic phenomena of forward masking (refs, Durrant, Schrader) and “loudness sustain” (refs, Zwicker, Schrader) suggest

**Fig. 2.** Block diagram of the tempo extractor. The two stage system detects auditory-object onsets in the first stage and looks for periodicities in those onsets in the second stage. Non-linear processing of the signal envelope, modeled after perceptual auditory processing, improves performance of the onset detector.

that listeners are more responsive to the onsets (attacks) of sounds than to their offsets (releases). To model this behavior we have included a non-linear lowpass filter in the first stages of the onset detector, with a faster time constant (5 msec) for rising portions of the signal than for decaying portions (100 msec). In addition, to model loudness sustain, we employed an unconventional signal differentiator in which the difference is taken between the current sample and the minimum in window (33 msec). These two components improve onset detection by 22% as measured by the error function:

$$Err = \frac{N_{fp} + N_m}{N_{ann}} \cdot 100\%, \quad (1)$$

where  $N_{fp}$  is the number of false positives,  $N_m$  is the number of misses, and  $N_{ann}$  is the number of annotated onsets, i.e., ground truth labels. This more accurate onset detector led to a 22% improvement in tempo extraction.

### III. MUSIC SIMILARITY

Music similarity is an important characteristic for music browsing and playlisting. There has been much recent work on methods for the automatic evaluation of similarity between two musical pieces based on signal analysis (refs, including MIREX). While some studies have used listening tests as a method for measuring the performance of such systems, very little work has been done on actually characterizing perceptual similarity. Data on, or a model of, music similarity perception would be a valuable tool in guiding systems for music similarity evaluation.

We recently developed a methodology for characterizing the perceptual similarity between pieces of music using triadic comparisons (Novello, ISMIR). We then performed multidimensional scaling (MDS) on the data to yield a 4-dimensional “perceptual similarity” space, in which all eighteen musical excerpts were placed. Initial results show that music genre is a strong factor that correlates highly with similarity ratings and some genres are perceptually closer to each other than others. Figure 3 shows the results of the MDS plotted on two dimensions.

**Fig. 3.** MDS plot of musical excerpt (dis)similarity ratings (shown in two dimensions). Distances between points reflect perceptual similarity distances between excerpts.

A logical next step is to now broaden our database with a larger-scale rating experiment and then use the data on perceptual similarity to calculate a mapping between the music audio feature space and the perceptual similarity space. With this mapping one could better predict the perceptual similarity between two songs based on the extracted features.

### IV. MUSIC SECTIONING

The ability to automatically extract structural elements from music audio is necessary for automatic music summarization and can lead to more sophisticated playlisting and browsing tools. Previous work on music structure analysis has shown that evidence of musical structure exists in plots made from correlation analyses of various signal features (cite Foote and others), however, no method has been found that reliably extracts structural boundaries across a broad range of music. A confounding factor is that structural boundaries can be defined in terms of musicological, computational or perceptual criteria and it is not clear whether these definitions are always in agreement. In many applications, it is perceptual boundary criteria that are most relevant and so, as with tempo and music similarity, we have established a methodology for quantifying and modeling perceptual structural boundaries (Bruderer ICMPC and ISMIR).

We conducted an experiment in which we focused on Western popular music and asked subjects to indicate sectional or phrase boundaries while listening through a piece of music. We found high agreement across subjects for some boundaries and low agreement for others. In a second phase of the experiment, we asked subjects to rate the salience of particular boundaries and found a high correlation between salience ratings and the number of subjects who indicated a particular boundary. This finding confirms an assumption held by others (refs) that boundary salience is reflected in the global response across subjects.

We are using the data collected in this experiment to construct a model for music sectioning based on music theoretical cues and extracted features. Current results show that while some cues are more dominant than others in indicating sectional boundaries, there is a large variance across songs. Future models will likely have to include adaptive elements in order to properly select among the potential cues for structural boundaries.

### V. CONCLUSION

We have seen performance benefits from using perceptual criteria and models in systems for automatic music information retrieval. Our most comprehensive example comes from our music tempo extractor where it was not only the evaluation criteria that were perceptually-based, but also acoustic processing components as well. Our work in music sim-

# Paper ID

ilarity and music structure analysis suggest that careful collection and treatment of perceptual data can provide robust design and evaluation guidelines in these areas as well.

## VI. REFERENCES