Sound Spotting – a Frame-Based Approach

Christian Spevak
Music Department
University of Hertfordshire
College Lane, Hatfield, AL10 9AB, UK
+44 1707 28 4442
c.spevak@herts.ac.uk

Richard Polfreman
Music Department
University of Hertfordshire
College Lane, Hatfield, AL10 9AB, UK
+44 1707 28 6473
r.p.polfreman@herts.ac.uk

ABSTRACT
We present a system for content-based retrieval of perceptually similar sound events in audio documents ('sound spotting', using a query by example. The system consists of three discrete stages: a front-end for feature extraction, a self-organizing map, and a pattern matching unit. Our paper introduces the approach, describes the separate modules and discusses some preliminary results and future research.

1. PROBLEM
The possibility of storing large quantities of sound or video data on digital media has resulted in a growing demand for content-based retrieval techniques to search multimedia data for particular events without using annotations or other meta-data. This paper presents an approach to a task that can be described as *sound spotting:* the detection of perceptually *similar* sounds in a given document, using a *query by example*, i.e. selecting a particular sound event and searching for 'similar' occurrences. The proposed system could be applied to content-based retrieval of sound events from broadcasting archives or to aid transcription and analysis of non-notated music.

A particular problem is posed by the definition of *perceptual similarity:* sound perception comprises so many different aspects that it is very hard to define a general perceptual distance measure for a pair of sounds. Even if the variability is restricted to timbre alone, it is still uncertain how to construct a *timbre space* with respect to any underlying acoustical features. Within the scope of our system we decided to focus on the spectral evolution of sounds by calculating a time-frequency distribution and splitting the signal into a series of short-time frames. Similarity can then be assessed by comparing sequences of frames.

2. APPROACH
Our concept builds on various connectionist approaches to modelling the perception of timbre that have been investigated over the last ten years [1, 2, 3]. These systems typically consist of some kind of auditory model to preprocess the sounds, and a self-organizing map to classify the resulting feature vectors. The reported experiments involved the classificaton of a small number of test sounds equalized in pitch, duration and loudness. To extend these models towards evolutions of timbre, pitch and loudness we have pursued a dynamic, frame-based approach involving three stages.

First the raw audio data is preprocessed by an auditory model performing a *feature extraction*. The signal is divided into short-time frames and represented by a series of feature vectors. In the current system we use a parametric representation adopted from automatic speech recognition, mel-frequency cepstral coefficients (MFCC).

Second a *self-organizing map* (SOM) is employed to perform a vector quantization while mapping the feature vectors onto a two-dimensional array of units. The SOM assigns a best-matching unit to each input vector, so that a sound signal corresponds to a sequence of best-matching units.

Finally a pattern matching algorithm is applied to search the entire source for sequences 'similar' to a selected pattern. Currently we refer to the SOM units simply by discrete symbols (disregarding the associated weight vectors and topological relations) and perform an *approximate matching* on the resulting sequences.

3. SYSTEM COMPONENTS

3.1 Feature Extraction
Besides their application in speech recognition mel-frequency cepstral coefficients have been successfully utilized for timbre analysis [4] and music modeling [5]. MFCC calculation involves the following steps: the signal is divided into short frames (10-20 ms), a discrete Fourier transform is taken of each frame and its amplitude spectrum converted to a logarithmic scale to approximately model perceived loudness. The spectrum is smoothed by combining Fourier bins into outputs of a 40 channel mel-spaced filterbank (*mel* being a psychological measure of pitch magnitude). Finally a discrete cosine transform is applied to extract principal components and reduce the data to typically 13 components per frame.

3.2 Self-Organizing Map
Self-organizing maps constitute a particular class of artificial neural networks, which is inspired by brain maps such as the tonotopic map in the auditory cortex [6]. A self-organizing map can be imagined as a lattice of neurons, each of which possesses a multidimensional weight vector. Feature vectors are mapped onto the lattice by assigning *a best-matching unit* to each vector.

Self-organization of the map takes place during a training phase, where the entire data is repeatedly presented to the network. The SOM 'learns' the topology of the input data and forms a set of ordered discrete reference vectors, which can be regarded as a reduced representation of the original data.

To enable an efficient pattern matching process in the third stage of the system we represent the best-matching units by their index number only and disregard their mutual relations. A sound sample then corresponds to a string of symbols, which can be further processed by means of string searching algorithms.

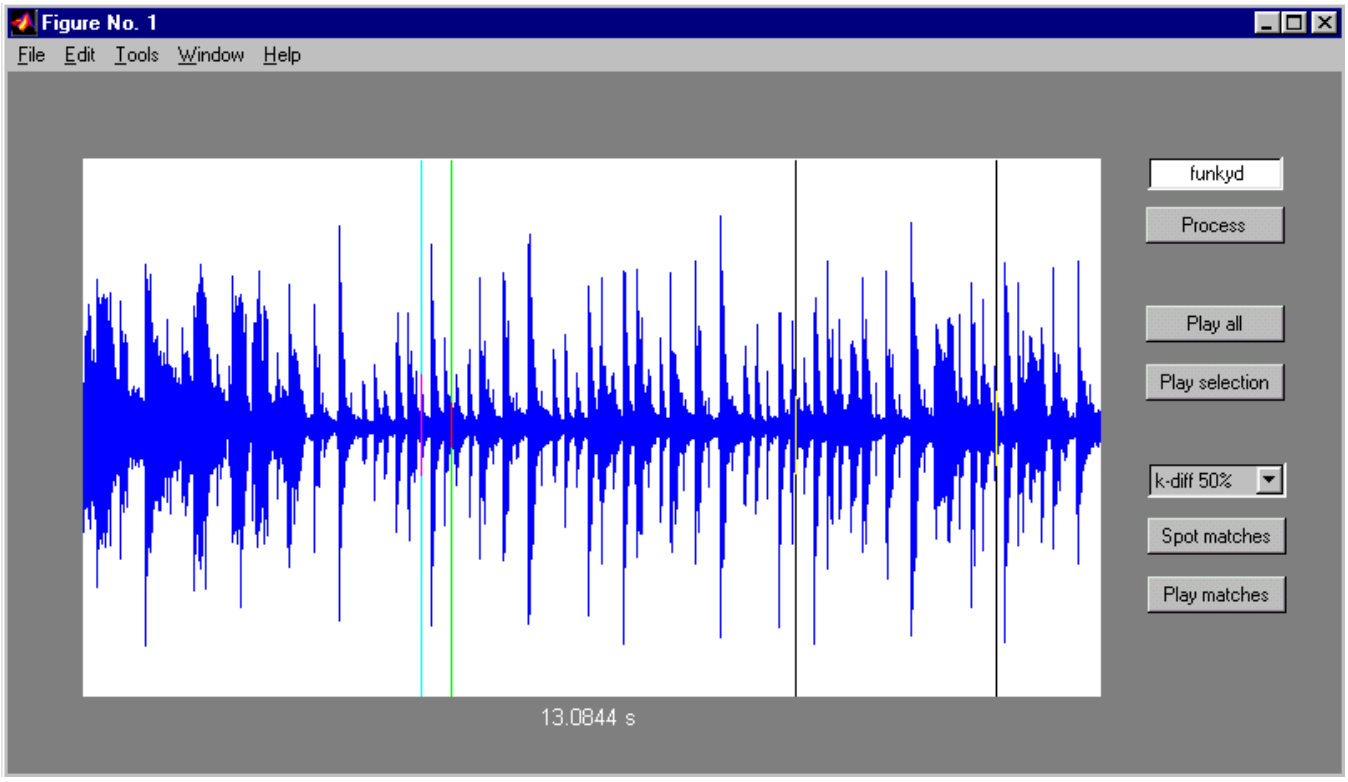

Figure 1. Graphical user interface of the prototype implementation.

3.3 Pattern Matching

We use a *k-difference inexact matching* algorithm to retrieve approximate matches of a selected *pattern* from the entire *text* [7]. The algorithm retrieves matches differing by an *edit distance* of at most *k*, where edit distance denotes the minimum number of operations needed to transform one string into another, permitting insertion, deletion and substitution of symbols. *k* can be specified with respect to the pattern length (e.g. 40%).

4. DISCUSSION

Initial experiments conducted with a MATLAB prototype implementation (see Figure 1) have demonstrated varying degrees of success in retrieving perceptually similar sounds. Encourageing results have been obtained for instance with drum loops, where similar sounds could easily be detected. Difficulties arise when an event has to be detected in a mixture of different sounds. The reduction of the multidimensional feature vectors to index numbers and the use of a simple string matching algorithm clearly entails a significant loss of potentially important information, which could be avoided by a more sophicated distance measure in conjunction with a suitable pattern matching algorithm. These issues will be addressed in future research.

5. REFERENCES

[1] Piero Cosi, Giovanni De Poli and Giampaolo Lauzzana (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research* 23(1): 71-98.

[2] Bernhard Feiten and Stefan Günzel (1994). Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal* 18(3): 53-65.

[3] Petri Toiviainen (2000). Symbolic AI versus connectionism in music research. In: Eduardo Reck Miranda (ed.) *Readings in Music and Artificial Intelligence*, 47-67. Harwood Academic Publishers.

[4] Giovanni De Poli and Paolo Prandoni (1997). Sonological models for timbre characterization. *Journal of New Music Research* 26: 170-197.

[5] Beth Logan (2000). Mel frequency cepstral coefficients for music modeling. *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*.

[6] Teuvo Kohonen (2000). *Self-Organizing Maps.* Third edition. Springer.

[7] Graham A. Stephen (1994). *String Searching Algorithms.* World Scientific Publishing, Singapore.