Christian Spevak, Richard Polfreman
Music Department, University of Hertfordshire
College Lane, Hatfield, Herts
AL10 9AB, UK

We present an approach to content-based sound retrieval using auditory models, self-organizing neural networks, and string matching techniques. It addresses the issues of spotting perceptually similar occurrences of a particular sound event in an audio document. After introducing the problem and the basic approach we describe the individual stages of the system and give references to additional literature. The third section of the paper summarizes the preliminary experiments involving auditory models and self-organizing maps we carried out so far, and the final discussion reflects on the overall concept and suggests further directions.

Keywords: content-based retrieval, sound classification, auditory model, self-organizing map, string matching

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 digital recordings or broadcasting archives or to aid transcription and analysis of non-notated music.

A special problem is posed by the definition of *perceptual similarity:* sound perception comprises so many different aspects (such as loudness, pitch, timbre, location, duration) 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 largely uncertain how to define a *timbre space* with respect to any underlying acoustical features (Hajda, Kendall, Carterette and Harshberger 1997). Therefore we decided to define ‘similarity‘ within the scope of our system as characterized by a similar evolution of cochleagram frames.

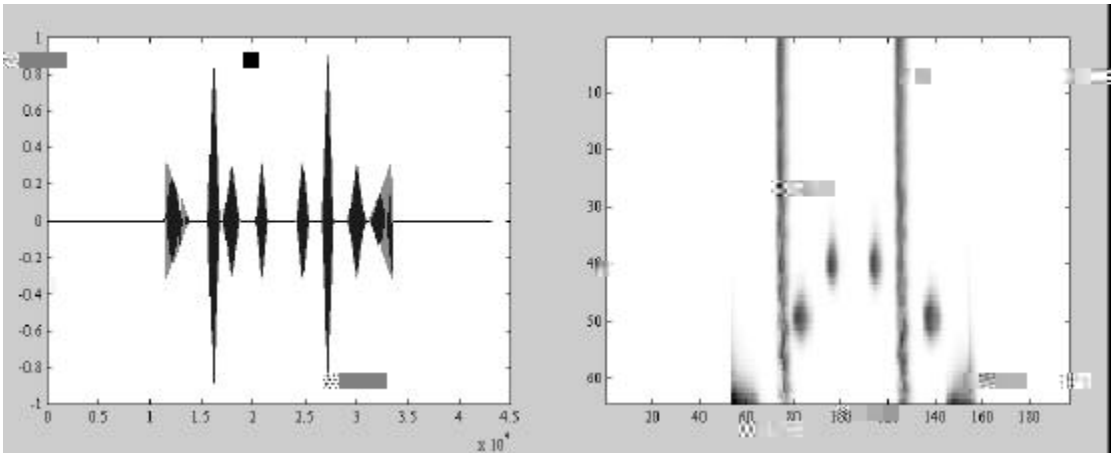

Figure 1. Waveform and cochleagram representation of a sound sample consisting of short tone and noise bursts. The cochleagram was produced by the AF/IHC model. The 44,000 samples of the waveform representation are reduced to 200 frames in the cochleagram.

We carried out a number experiments to investigate the suitability of different auditory representations within the framework of our system. The corresponding models are briefly described in the following sections.

2.1.1 Auditory filterbank and inner hair cell model (AF/IHC)

This model combines an auditory filterbank (Patterson 1992, Slaney 1993)

transform. MFCC provide a substantial data reduction, because a dozen coefficients often suffice to characterize the acoustic signal.

Self-organizing maps constitute a particular class of artificial neural networks, which is inspired by brain maps forming reduced representations of relevant facts (e.g. the tonotopic map of pitch in the auditory cortex). The SOM was developed and formalized by Kohonen (Kohonen 1982), and has meanwhile been utilized in a wide range of fields (cf. Kohonen 2000). Applications include visualization and clustering of multidimensional data as well as statistical pattern recognition.

A self-organizing map can be imagined as a latticed array of neurons, each of which is associated with a multidimensional weight vector. The weight vectors must have the same number of components as the input vectors to enable a mapping of the input data onto the lattice. Self-organization takes place during the training phase, where the preprocessed data is repeatedly presented to the network. For each input vector, a *best-matching unit* is determined and its weight vector adjusted towards the input vector. By adapting not only the best-matching unit, but also its neighbours, the network ‘learns’ the global topology of the input data and forms a set of *ordered discrete reference vectors*. These reference vectors 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 vectors by their index numbers only and disregard their mutual relations except for the binary distinction between ‘equal’ and

procedures and results can be found in previous publications (Spevak and Polfreman 2000, Spevak, Polfreman and Loomes 2001).

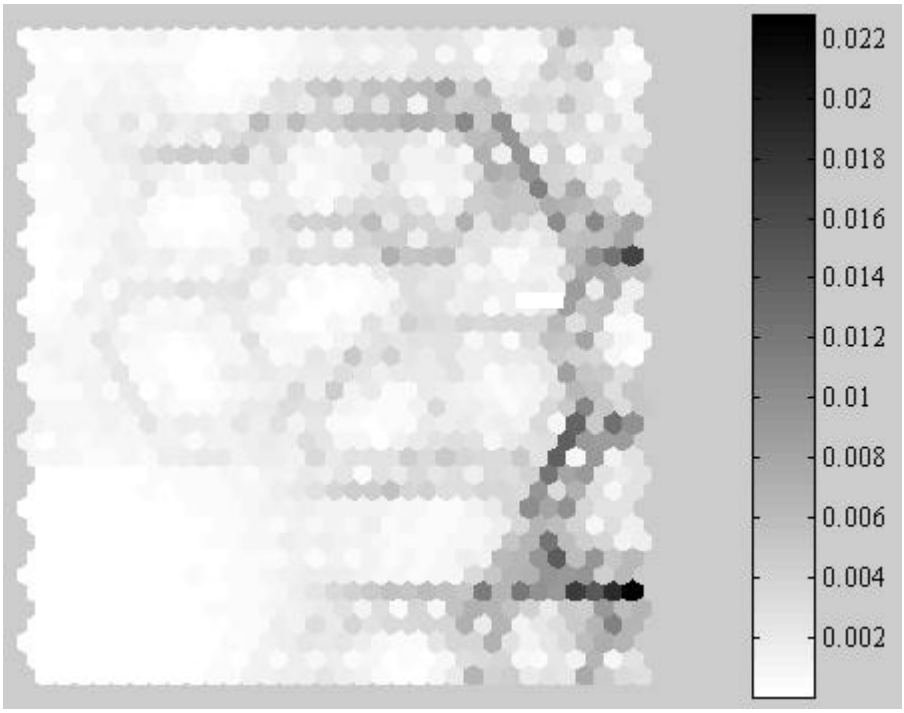

Figure 2. U-matrix of a SOM comprising twenty units by seventeen after training it with the test sounds preprocessed by the AF/IHC model. Different shades of grey represent the weight space distances between adjacent units on the lattice; cluster borders are indicated by darker colours.

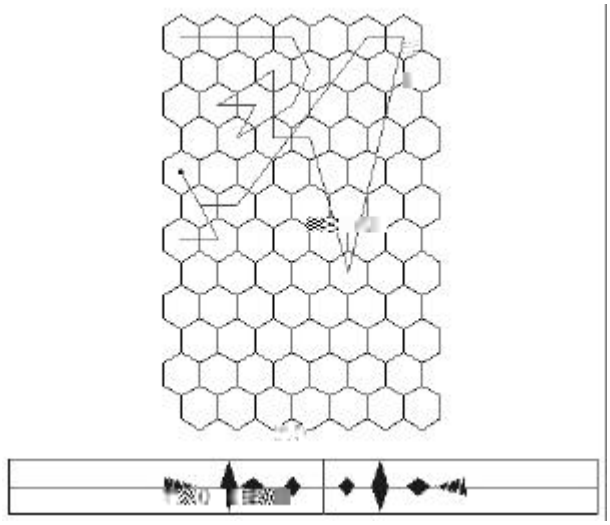

Figure 3. Still frame from a film visualizing the trajectory produced by a sequence of quickly alternating tone and noise bursts, preprocessed with Lyon's cochlear model, on a seven by twelve SOM.

3.2.1 Auditory models

The functional similarity of the two auditory models – AF/IHC and Lyon's cochlea model – as opposed to the MFCC representation was clearly reflected in the the organization of the SOMs and the course of the trajectories. The trajectories produced by the auditory models were generally smoother than those obtained with MFCC, which was mainly caused by the

lowpass filtering in the data reduction stage. The MFCC trajectories reacted immediatedly to changes in the sound signal and tended to oscillate between two or more units even for perceptually steady sounds.

quantization tool and regard the reference vectors as an abstract symbolic representation of the sound data, which can then be subjected to efficient string searching techniques.

The question whether such a system will be able to retrieve perceptually valid matches remains unanswered yet. After implementing the string matching stage we will address that issue by comparing the system's performance with similarity ratings from human listeners. The vague definition of 'sound similarity' clearly introduces an element of uncertainty, because different listeners will presumably pay attention to different kinds of similarity. A possible way out of this dilemma would be a more analytic approach, in which the preprocessing extracts a set of well-defined sound features (such as sound level, spectral centroid and periodicity) that can be related to particular perceptual dimensions (loudness, brightness and pitch). Important contributions in that direction have been made e.g. by Wold, Blum, Keislar and Wheaton (1999) and McAdams and colleagues (McAdams, Winsberg, Donnadieu, De Soete and Krimphoff 1995, Peeters, McAdams and Herrera 2000).

Cosi, P., De Poli, G. and Lauzzana, G. 1994. Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research* 23: 71-98.

Cosi, P., De Poli, G. and Prandoni, P. 1994. Timbre characterization with mel-cepstrum and neural nets. *Proceedings of the International Computer Music Conference*, pp. 42-5.

Crawford, T., Iliopoulos, C. S. and Raman, R. 1998. String-matching techniques for musical similarity and melodic recognition. In W. B. Hewlett and E. Selfridge-Field (eds.) *Melodic Similarity: Concepts, Procedures, and Applications*, pp. 73-100. MIT Press.

De Poli, G. and Prandoni, P. 1997. Sonological models for timbre characterization.

Hawkins, H. L., McMullen, T. A., Popper, A. N. and Fay, R. R. (eds.) 1996. *Auditory Computation*. Springer.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43: 59-69.

Kohonen, T. 2000. *Self-Organizing Maps*. Third edition. Springer.

Lemström, K. 2000. String Matching Techniques for Music Retrieval. Report A-2000-4. University of Helsinki.

Logan, B. 2000.

Spevak, C. and Polfreman, R. 2000. Analyzing auditory representations for sound classification with self-organizing neural networks. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-00)*, pp. 119-24.

Spevak, C., Polfreman, R. and Loomes, M. 2001. Towards detection of perceptually similar sounds: investigating self-organizing maps*. Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*, pp. 45-50.

Stephen, G. A. 1994. *String Searching Algorithms*. Singapore: World Scientific.

Toiviainen, P. 1996. Optimizing auditory images and distance metrics for self-organizing timbre maps. *Journal of New Music Research* 25(1): 1-30.

Toiviainen, P. 1997. Optimizing self-organizing timbre maps: two approaches. In M. Leman (ed.) *Music, Gestalt, and Computing. Studies in Cognitive and Systematic Musicology*, pp. 337-50. Springer.

Toiviainen, P. 2000. Symbolic AI versus connectionism in music research. In E. R. Miranda (ed.) *Readings in Music and Artificial Intelligence*, pp. 47-67. Amsterdam: Harwood Academic Publishers.

Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M. and Näätänen, R. 1998. Timbre similarity: convergence of neural, behavioral, and computational approaches. *Music Perception* 16: 223-42.

Ukkonen, E. 1985. Algorithms for approximate string matching. *Information and Control* 64: 100-118.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. 2000. SOM Toolbox for Matlab 5. *Technical Report* A57. Helsinki University of Technology.

Warren, R. M. 1999. *Auditory Perception: A New Analysis and Synthesis.* Cambridge University Press.

Wold, E., Blum, T., Keislar, D. and Wheaton, J. 1999. Classification, search, and retrieval of audio. In B. Furht (ed.) *Handbook of Multimedia Computing*. CRC Press.