Harmonic Envelope Detection and Amplitude Estimation Using Map Seeking Circuits

B. Jerry Gregoire and Robert C. Maher
Electrical and Computer Engineering, Montana State University, Bozeman, MT USA
jgregoire@ece.montana.edu ; rob.maher@montana.edu

Abstract

In this paper we introduce the use of map-seeking circuits for auditory pattern detection and classification. A map-seeking circuit is a signal processing structure used to detect a desired feature in a mixture by iteratively transforming, superposing, and comparing the composite mixture with a pattern template. The result is a mapping between the template and the position of the matched feature in the mixture. The iterative detection process is inspired by the neural connections in the human visual system. A particularly important feature of map-seeking classification is that the search operates on an additive superposition of allowable transformations of the desired feature vector, giving a linear increase in computation with increasing image complexity, rather than a brute-force feature detection that increases in computation geometrically.

1. Introduction

Human beings rely on the auditory sense for speech, musical entertainment, and event detection. The auditory sense is particularly useful for situational awareness because the ear-brain system can often determine the azimuth, elevation, and identity of a sound even when the source is not in view by the eyes. In fact, one can easily argue that humans primarily use the ears to know where to point the eyes.

Identification of sound sources, or *acoustical patterns*, has been the subject of extensive research over the last 50 years, yet no existing systems achieve identification performance anywhere near the level of a casual human listener. It is reasonable to consider that there might be another viewpoint that can give a substantial increase in efficiency and performance.

Acoustical pattern recognition is a cross-disciplinary topic. There are clear contributions from the engineering and machine intelligence disciplines, but also from neuroscience, psychology, and the artistic fields. Conversely, we can use human acoustical pattern detection as a concept proof, and then consider ways to mimic this performance computationally.

In this paper we present some preliminary work on auditory extensions to the *map-seeking circuit* pattern detection theory for images presented by Arathorn [1]. This biologically inspired procedure performs a comparison between a target pattern and systematic transformations of the observed signal. As described later in this paper, the map-seeking circuit formulation provides several key capabilities that make it well suited for the acoustical pattern detection and classification task.

The remaining sections of this paper are organized as follows. First, we review the concepts, prior work, and rationale for acoustical pattern detection. Next, we explain the biologically inspired map-seeking circuit structure and principles. We then present a preliminary set of examples, and conclude with the prospects for future work.

2. Background

One of the important cues for human situational awareness is sound, yet automated sound recognition remains a challenging and little understood problem [2, 3, 4, 5].

Acoustical signals are characterized by alternating variations in pressure as a function of time, and therefore it is common to study acoustical phenomena using a spectrogram or similar amplitude vs. frequency vs. time representations.

Many acoustical pattern recognition systems begin by deriving a set of descriptive parameters for the input signal. The parameters may include the spectral envelope, the spectral centroid, voiced/unvoiced state, time correlation among partials, inharmonicity, etc. The same parameters are derived from the target sound (or a family of target training sounds) and then a

-7803-9232-9/05/$20.00 2005 IEEE.

search is conducted to find a match between the input signal parameters and the target database [6, 7, 8, 9]. We have found this approach to be useful, but difficult to evaluate because choosing the set of analysis parameters is rather ad hoc: one must consider whether the chosen set of parameters is optimal or just arbitrary. Increasing the number of derived parameters may help, but leads to the problem of judging the importance or weighting of each feature [8, 10, 11, 12]. In any case, we have come to the conclusion that this approach has yielded only incremental improvements over the years and does not seem likely to make the strides necessary for a truly practical and useful recognition system approaching anything near the abilities of human observers. What is needed is a method with the following attributes:

- *Source-independent representation*
- *Allows frequency and duration variations*
- *Extensible to an arbitrary number of target sounds*
- *Rapid elimination of poor matches to reduce the computation explosion.*
- *Amenable to parallel processing.*
- *Low or tractable sensitivity to noise.*

2.1 Biologically Inspired Computing

One recent development has been the use of *biologically inspired* pattern recognition [1, 13]. The rationale for this is the observation that humans and other animals show astonishing skills in detecting and acting upon acoustical signals. However, adopting a biological model does not answer the algorithmic requirements, since the means and mechanisms of the biological system are still only vaguely understood. Even in the case of human observers the process involved in pattern recognition is not introspectable, and therefore difficult to capture in the form of computer software.

The cochlea is the neural transducer element of the human hearing system. The cochlea produces a spectral decomposition of sound, providing the brain with temporal and spectral information. Study of computational neurobiology may ultimately address the mechanism of human auditory pattern recognition beginning with the cochlea and leading to the higher nerve centers of the brain, but the current state of knowledge has not yet reached this level of sophistication.

2.2 Map-Seeking Circuits

The map-seeking circuit theory includes several novel features reflecting the architecture of the human visual system, including a rapid means for locating translated, scaled, and rotated versions of a target object within a scene. The possibility that similar transformations might be applicable to modeling the auditory system is the basis of our investigation. The map-seeking scheme is particularly appealing because it uses the ordering principle of superposition and a *nonlinear feedback topology* to converge rapidly without spending effort on unpromising candidate matches.

Although counterintuitive at first, the notion of superposing multiple transformations of the input image actually mimics the neural structure of the visual system, and unlike the more well-known computational detection algorithms using correlation-like structures, the superposition increases in complexity *additively*, rather than *multiplicatively*. The resulting reduction in computation is substantial, and perhaps helps explain the ability of the human auditory system to obtain rapid and reliable pattern detection without double-precision computation elements.

The map-seeking circuit (MSC) concept is shown in Figure 1.

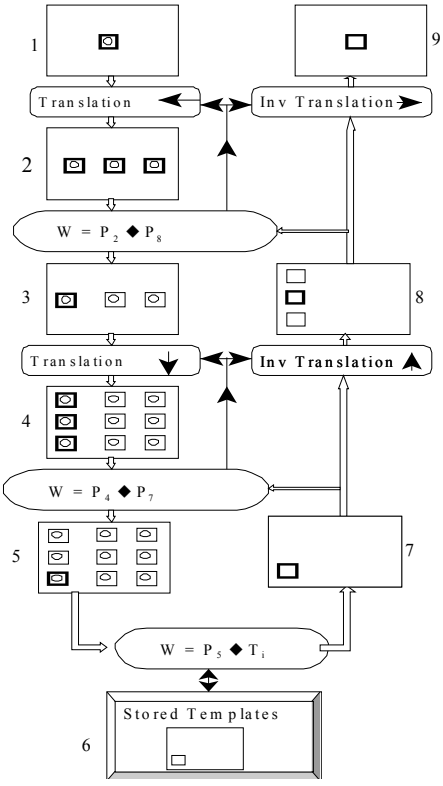

Figure 1: Map-seeking circuit topology

The MSC has two paths, a forward path and a backward path. The input is assumed to be a mixture of objects, possibly including a target object that is a transformed (e.g., rotated, scaled, translated) version of

-7803-9232-9/05/$20.00 2005 IEEE.

a reference object. As the input mixture progresses through the forward path, it is transformed according to chosen parameters such as translation and rotation in the case of images, or musical pitch and spectral envelope in the case of audio signals. The reverse path attempts to find the best match between a set of transformations and the target object, or template. The process is iterative, so the processed mixture is passed sequentially through the forward and reverse paths. If a set of transformations is found that describe the mapping of the template to the target in the input mixture, the target is "found" and the output contains the target object without the other components of the mixture. If the process does not converge to a match after some selected number of iterations, the process is terminated with a null output.

To illustrate, consider the visual example shown in Figure 1. The MSC is configured with a rectangle template, and the allowed transformations in this deliberately simple example are limited to horizontal and vertical translations only. The input mixture in this example contains both a rectangle (the target) and a circle. The figure shows nine panels divided into two paths, with the forward path shown descending on the left and the backward path shown ascending on the right.

The input mixture consisting of the rectangle and circle (see Panel 1 of Figure 1) is presented to the MSC. Beginning on the upper left, the entire mixture is transformed by translating it successively to the left. All the translations are then added together as shown in Panel 2. The second transform is a vertical translation, replicating the Panel 2 image as shown in Panel 4. A similarity calculation is made between the Panel 5 superposition and the stored template rectangle, such as performing a simple dot product. The similarity calculation favors the rectangle indicated in Panel 7 and propagates up the backward path on the right hand side. Panel 7 is then compared to Panel 4 to produce another similarity measure that weights the best transform. In this case the translation that shifted the input mixture "down one row" in the vertical direction is preferred over others. This is shown in Panel 5. On the backward path the inverse transform chosen in layer 2—"shift up one row" in this case—is applied, as depicted in Panel 8. The backward path intermediate result of Panel 8 is then compared to Panel 2 to choose the favored layer 1 transform ("shift left one column"), as shown in Panel 3. One final inverse transform, "shift right one column," produces Panel 9, and the process is continued for a second pass through the forward and backward paths. As the incorrect transforms in the forward path are iteratively attenuated, the match with Panel 6 and the set of favored transforms are strengthened. After several iterations the favored transformations are uniquely identified with the best target. At this point the target representation is now mapped via the transformation set to the rectangle in the original image.

As the iterations progress, the *Ordering Principle of Superposition* ensures that if the target exists in the input mixture the transforms that best map the template to the target will have greater weightings than the other allowable transforms. This allows the transforms to be added together (superposed), which greatly reduces the computational complexity. If the number of transforms required at each layer is N and there are M layers in the circuit, this would require N^M comparisons if done by an exhaustive search method. However, by applying superposition, this number is only $N \cdot M$.

3. Method

As an elementary but useful starting point, consider a task in which we would like to compare the frequency spectrum of a sustained musical instrument tone with a predetermined template. The term *spectral envelope* is used to describe the overall energy distribution of the signal's magnitude spectrum as a function of frequency. For example, a sustained musical sound with a periodic waveform has a spectrum consisting of harmonic spectral peaks with magnitudes that vary with frequency. The spectral envelope of such a signal could be defined rather crudely as a curve that smoothly interpolates each of the harmonic spectral peaks. More sophisticated methods for defining the spectral envelope involve separating the harmonic features attributable to the instrument's excitation mechanism (e.g., a vibrating reed) from the broader spectral resonances due to the instrument's passive body and mechanical couplings (e.g., an air column with finger holes).

To demonstrate the map-seeking procedure for audio spectral envelope matching we use a one layer MSC with a single template, as shown in Figure 2. The task in this case is to detect a particular spectral envelope rather than searching among many different representations. In a single detection MSC the downward link from Panel 3 to the stored representation is not needed. The purpose of that link is to choose which stored representation best matches the target, but in this case we have only one target.

The flow for a single layer MSC is otherwise identical to the general description given in Section 2.2 above. The target in Figure 2 is represented in Panel 1. Panel 2 shows the sum of the transforms that represent an amplitude scaling. The function of Panels 3, 4, 5 and 6 are respectively the same as 3, 6, 7, and 9 shown in Figure 1.

-7803-9232-9/05/$20.00 2005 IEEE.

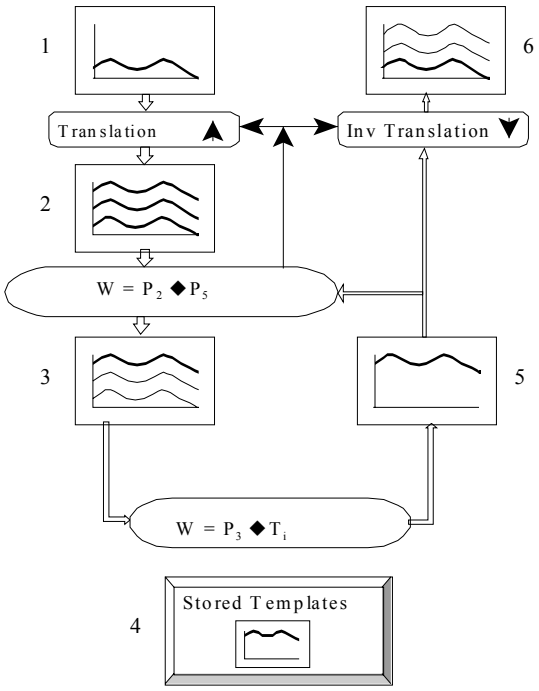

Figure 2: Single-layer map-seeking example for spectral envelopes

To utilize the MSC algorithm, we represent the spectral envelope as a two-dimensional frequency vs. time image. By shifting the 2-D image in the vertical direction, a match is then made with the spectral envelope of the template. This shift is the transform of the MSC layer.

3.1 Sparse Data for Robust Matching

To ensure that the target transforms in the neighborhood of the correct template will provide a match even if the precise details differ, the algorithm can benefit from "blurring" (deliberate fuzziness) in the target and template. This is discussed below for both the amplitude and frequency dimensions.

To create a spectral representation suitable for the MSC, we use a gammatone filter bank [14]. This has several advantages. The constant-Q filter bandwidths significantly reduce the amount of data compared to the uniform resolution of a raw FFT, thereby increasing the computational efficiency. Since the individual bandpass filter outputs overlap, the spectral energy is smeared across several filter outputs, producing the desired fuzziness on the spectral axis. Finally, using a gammatone filter bank is a practical model for the spectral sensitivity of the human hearing system.

The gammatone filter bank is utilized as follows. We rectify and accumulate each filter output to obtain the short-time amplitude representation for each center frequency in the filter bank, expressed in decibels. The final step is to find and plot the spectral peaks with the assumption that the spectral envelope is the same for small changes in pitch and amplitude. The spectral peaks are then plotted with the amplitude and frequency as the ordinate and abscissa, respectively. The amplitudes are blurred vertically ±0.75 dB with a function that decreases as the square of the distance from the amplitude position. The primary purpose of the blurring function is to emphasize the measured amplitude while allowing a match between the target and template with a reasonable tolerance.

4. Results

We report two experiments. The first demonstrates the MSC algorithm's ability to identify and map an input to a stored template through amplitude scaling, and the second demonstrates the rejection of a signal for which no allowable mapping exists.

4.1 Experimental Setup

The target samples for both experiments were obtained from the musical instrument recordings compiled by the University of Iowa Experimental Music Studios [15]. Both experiments used the steady state portion of a signal for the test input. The target signals were not normalized to a common power level, so the power levels were found to range from -1 dB to -6 dB referenced to the template power. The template was synthesized by finding the spectral envelope of four notes played on the oboe: B3, C4, D4 and G4. These four notes have similar spectral envelopes between 200 Hz and 2,000 Hz. As such, this range was used to synthesize the template. The raw steady-state spectra were mathematically altered to have the same average rms power. The resulting partials were plotted to create a representative spectral envelope. Linear interpolation was used to estimate the envelope between the partials. Each note was played fortissimo, that is, loudly, so that any spectral dependence on intensity would be minimized for the experiment.

4.2 Identification and Mapping Example

The first task of the MSC was to report a match if the target sound was an oboe and find the amplitude of the input with respect to the template. Each transformation represented 0.1dB with respect to the template. The target sounds tested covered three octaves, ranging from B-flat 3 to A-flat 6. With the exception of D4, the mapping process converged to a match when the input oboe sound was between B-flat

-7803-9232-9/05/$20.00 2005 IEEE.

3 and G4. If the input was higher than G4, the amplitude image was found to have an insufficient match with the template, giving a null condition in the MSC output. The MSC's rejection of the D4 match requires further investigation. The MSC has an empirically determined threshold to determine if a mapping is allowed. It is possible that further refinement of this threshold would have yielded a positive mapping for D4. The lowest amplitude error was with E4, -0.1dB. The mapping error was progressively greater the further away the musical pitch was from E4. The worst error was -2.4 dB for G4 and the average amplitude error was 0.77 dB. The amplitude discrepancies may or may not be significant depending on the nature of the matching task.

The figures below illustrate the MSC detecting the harmonic envelope of one of the notes, D-flat 4, during the iterative matching cycle for experiment one. Although convergence is typically between 50 and 100 iterations, in this particular case, the MSC did not resolve the mapping after 500 iterations, which is the arbitrary convergence limit in our test software. Such a result is not without merit however, as it has been our experience that if a mapping cannot be found, the maximum transform weight will generally decay relatively quickly, e.g., within 10 to 20 iterations. In this case we can be reasonably certain that a mapping exists and it corresponds to the maximum weight of 1.0. In this case, the best mapping was 2.5 dB. This represents an error from the expected 3.0 dB of -0.5 dB. This is within the +/- 0.75 dB 'fuzziness' of the algorithm, and is therefore a useful outcome.

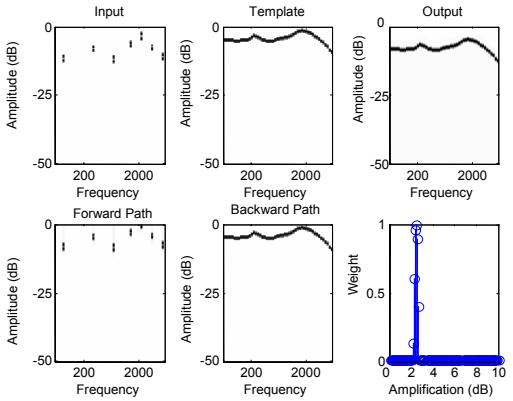

Figure 3: MSC matching between a test envelope and a target template

At the particular step in the iteration depicted in Figure 3, the MSC has mapped the template to the input with the mapping shown in the lower right panel. The panels labeled *Forward Path* and *Backward Path* are the summations of the transforms. At the end of the iteration cycle they are very nearly identical in the vertical position as one would expect with a probable mapping.

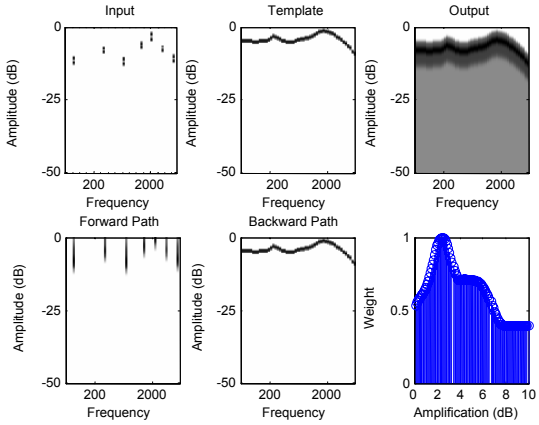

Figure 4: MSC matching at an intermediate step in the iterative process

To demonstrate further the iterative nature of the MSC, Figure 4 shows an intermediate step as the MSC is attempting to find a match. The darker portions in panels *Forward Path* and *Output* represent the favored transformations on the 11th iteration. If we had more than one template, the *Backward Path* panel would likewise have a summation. It this case the MSC would be in the process of culling the undesirable templates.

4.3 Null Case

The second experiment was to demonstrate that the MSC could reject a signal that does not have an allowable mapping to the template. To demonstrate this, a violin playing E-flat 4 was used as the input with the oboe template. The MSC correctly returned a null condition by forcing all transform weights to zero. The null condition is triggered when the matches on the backward path of the template and transforms fails to be greater than an empirically determined threshold.

5. Future work

There are three main areas for further study: robustness, system complexity and optimum parameters. Further study needs to be done to determine the robustness of MSC in the presence of noise and or other potential targets. To be useful the system must, of course, be able to work in a real world environment and to identify sound targets in a mixture of competing signals.

The examples shown here assume that the target's spectral envelope is roughly constant over a range of

-7803-9232-9/05/$20.00 2005 IEEE.

musical pitches and intensities, which is not generally true. Additional layers and transforms need to be identified to account for these natural variations. These natural and predictable transformations may require a new search domain, such as a corellogram.

Finally, properly rejecting a sound that is a mismatch for the stored template is dependent on a threshold of similarity in the backward path of the MSC. Determination of this threshold is not yet well understood analytically, and is performed empirically. There needs to be a more formal determination of the optimum level so that the detection is entirely automatic.

6. References

[1] Arathorn, D. W. (2002). *Map-seeking circuits in visual cognition: A computational mechanism for biological and machine vision*. Stanford University Press, Stanford, CA, USA.

[2] Duda, R. and Hart, P. (1973). *Pattern classification and scene analysis*. John Wiley & Sons.

[3] Scheirer, E. and Slaney, M. (1997). "Construction and evaluation of a robust multifeature speech/music discriminator." In Proc. 1997 IEEE ICASSP, Munich, pp 1331-1334.

[4] Rossignol, S., Rodet, X., Soumagne, J., Collette, J-L, and Depalle, P. (1998). "Feature extraction and temporal segmentation of acoustic signals." Proceedings of the ICMC, pp. 199-202.

[5] Foote, J.T. (1999). "An overview of audio information retrieval." ACM Multimedia Systems, 7:2-10.

[6] Bregman, A. (1990). *Auditory Scene Analysis*. MIT Press, Cambridge, MA, USA.

[7] Ellis, D. (1996). *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT Dept. of Electrical Engineering and Computer Science.

[8] Wold, E., Blum, T., Keislar, D., and Wheaton, J. (1996). *Content-based classification, search and retrieval of audio*. IEEE Multimedia, 3(2):27-36.

[9] Tzanetakis, G. and Cook, P. (1999). *Multifeature audio segmentation for browsing and annotation*. In Proc.1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA99, New Paltz, NY.

[10] Foote, J.T. (1997). Content-based retrieval of music and audio. In Multimedia Storage and Archiving Systems II, Proceedings of SPIE, pp. 138-147.

[11] Kimber, D. and Wilcox, L. (1996). Acoustic segmentation for audio browsers. Proc. Interface Conference (Sydney, Australia).

[12] Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. J.Acoust.Soc.Am, 103(1):588,601.

[13] Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. Journal of Neuroscience, 13:4700-4719.

[14] Slaney, M. (1993). *An efficient impementation of the Patterson-Holdsworth auditory filter bank*. Technical Report # 35, Apple Computer.

[15] Fritts, L. (1997). *University of Iowa Musical Instrument Samples*, URL: http://theremin.music.uiowa.edu/

-7803-9232-9/05/$20.00 2005 IEEE.