PRISM: PATTERN RECOGNITION IN SOUND AND MUSIC

ABSTRACT: PRISM is a system for visualization and pattern recognition in musical sound. Its spectral analysis from-end vastly improves upon traditional techniques. Its 2-D display, using log-frequency vs. time, directly superimposes physico-acoustic representations with musical intervals, scales, tuning systems and MIDI scores. The visual orientation of the user interface is matched by the computation: pattern-matching and polyphonic pitch detection are done by image processing operators chosen by the user. The current implementation (LightSpeed C on the MacII) supports indexing and coloring schemes that will enable many applications in sound and music: it is also "real-time ready".

Introduction. Machine perception research in sound and music offers many challenges. One issue in acoustic representation and processing that has caught our attention is how to avoid fundamental limitations of classical spectral analysis; the ear apparently gains simultaneous access to fine structure in time and in frequency. This suggests a multi-scale approach to signal analysis. For dealing with ill-defined, ill-understood, or multidisciplinary problems such as found in sound and music applications involving perception, fully automated systems are not as useful as flexible man-machine systems. We need rich visual representations of musical sound, representations that users who may be more or less naive in some of the required fields of expertise, can interact with. The ideal representation is objective, meaningful to engineers, intuitive, natural to composers and musicologists, and permits accurate analysis of polyphonic music and ethnic material. It is easy to use and require little or no special training. Traditional scoregrams, pitch tracks and conventional music notation each meet some of these criteria. The two-dimensional acoustic representation that we have found most useful, overall, relates to both scoregrams and musical scores. Physical time is displayed horizontally, and log(frequency) or semitones, vertically. The PRISM system uses this framework to display both physico-acoustic properties and musical properties.

Signal processing from-end. The PRISM spectral analysis from-end relies on the simultaneous use of multiple sample rates. Short-Time Fourier Transforms (STFT’s) are performed in quasi-parallel manner on progressively decimated versions of the original signal. Analysis performed at the original sampling rate have the highest possible time resolution. Decimated signals gradually lose time resolution and high-frequency content, but gain frequency resolution in the lower frequency range. Bandwidth-time tradeoffs are commonly interpreted in terms of a choice of resolution (or scale) for a single spectral transform, but it is better to insegregate information from multiple transforms. In PRISM, one can create and display a spectrogram-like view resulting from a specific choice of time-frequency resolution, or combine multiple views.

The first automated combination method that we designed was the Bandized-Q Frequency Transform, or BQFT. In a Constant-Q filter bank, the bandwidth $\Delta f$ of a filter is proportional to its center frequency $f$. The BQFT obtains its efficiency from the FFT, which yields a fixed $\Delta f$, but it approximates constant-Q behavior to within a factor of 2. To do so, it combines the results of the FFT’s by keeping only the upper octave of each 2FFT block at each decimation, producing the highest possible frequency resolution. Successive octaves are placed on a single display, resampled to a log scale for frequency. The efficiency of the implementation stems from the properties of recursive octave decimation, which is a repeated downsampling by a factor of two. Conceptually, signals at each sample rate are independently analyzed using STFT’s, as in Fig. 1. In actuality, the algorithm uses a halfband lowpass FIF filter followed by a decimation, as in Fig. 2.

Other combination methods actually mix the frequency and time resolutions from several views into a single one. The high frequency resolution is that of the BQFT, whereas the high time
resolution results from the STFT of the original signal. A family of variant methods have been found useful; others could be added to the system. A simple and effective way to combine already displayed views is with the point-wise gray level min operator. In practice, the resulting display does indeed appear to offer both the highest time and frequency resolution (see Fig. 3, left).

The archive stores input acquisition, filter declination banks, spectral analysis with log-f resampling, and display as four independent processes, communicating via circular buffers; it allows interrupt-driven data collection when in-line processing is required. The software's implementation will allow real-time processing with the mere addition of DSP power.

Feature extraction using image processing. We briefly review a polyphonic pitch extraction method discussed at greater length in [5]. In simple harmonic sounds, the pattern of equidistant harmonics (k+F) or (k-F) on the linear axis becomes an equally recognizable pattern on the log axis, \( \log(k+F) = \log(k) \). It is now pitch-shift-invariant. The vertical arrangement which musicians know as the harmonic series, or the collection of intervals \( \log(k) \) scaled to the display resolution, sweeps all possible fundamentals without changing shape or size.

Image convolution of the spectrum with the harmonic comb acts as a simple yet polyphonic pitch detector which operates on all timbres and frequencies in parallel. Fig. 4 shows (right) the result of a pitch extraction performed on a two voice piano tocacina displayed on the iPh in min mode. Additional preliminary processing, also defined by image convolution, improves the results obtained. A multitude of other event detection and pattern recognition problems can be addressed by visual methods on this kind of display. Space limitations do not permit their description.

Discussion. The FR1SM system is a tool for research in acoustic pattern recognition, sound source segregation and identification, and a prototype for developing applications that rely on interactive music visualization. The tool has proven effective in exploring issues and tuning analysis techniques, a step towards more automated analytic methods. The techniques described are not offered as models of the operation of the human ear. Doing the latter is also an objective of our research; yet we feel that the complexities involved in full-scale neural modeling is an effort apart from the present (more applied) focus. We believe that the approach described is a promising alternative to more traditional approaches in spectral analysis, music visualization, pitch detection and event detection. Furthermore, the integration of musical applications beneficial from the use of a visual representation readily understandable by musicians without an engineering background, yet usable by accurate automatic analysis tools. This is the key to creating environments in which modern computer technology meets effectively with the development of an increasing level of musical intelligence.

Acknowledgements. This research was supported by grants from the National Science Foundation, the System Development Foundation and the French Ministry of Foreign Affairs.

REFERENCES


ICMC GLASGOW 1990 PROCEEDINGS
Figure 1

Figure 2

Figure 3

ICMC GLASGOW 1990 PROCEEDINGS