A COMPUTATIONAL MODEL OF MUSIC LISTENING

M. Piszczalski & B. A. Galler

Computer & Comm. Sci. Dept.

University of Michigan

ABSTRACT

We studied musical sounds to learn how human music recognition is accomplished. To do this, we sought key perceptual features by computer analysis of the sounds. The results of our analyses can be displayed in motion graphics; the color, pseudo three-dimensional graphs can be animated to change in synchrony with the sounds they represent. On our unique system sounds are analyzable into several different perceptual states. These states can be viewed graphically and/or aurally using computer sound synthesis techniques.

Keywords: music recognition, motion graphics, graphics-with-sound output

BACKGROUND

Our research has centered on how human sound perceptions are derived from sounds. More specifically, we have investigated musical sounds to learn how pitch, musical notes, and other musical concepts are carried in the sound waves (Fig. 1) and are comprehended from the sound medium by the typical human listener.

Because musical sounds from natural sources do not produce simple, consistent sound wave patterns (Seashore, 1938), we hypothesized that musical perceptions must result from the listener's selective attention to certain parts of the sound signal. The primary thrust of our work therefore has been (1) speculating on what those significant parts of the sounds may be, (2) isolating them in a "pure" state, (3) evaluating the isolated information to see how adequately it carries the sought-after music percept, and finally, (4) postulating how the extracted "key" data can be characterized simply. The success of this approach relied heavily on aids that help us make good hypotheses about sound structure and allow us to evaluate newly isolated parts of sounds rapidly. Bear in mind that despite the definite, clear-cut impressions musical sounds generate in listening, the sound itself is poorly understood, in the scientific sense of knowing how sound patterns map into perceptual patterns. Another problem occurs when sounds are in digital form; they then become represented by far too much data for us to understand the underlying musical significance by directly inspecting the thousands of numbers that represent each second of music.

An early step in the process is to reduce the amount of digital musical information to more manageable proportions. The data can be reduced in a number of ways, although we wanted to find procedures that gave a reduced set of data that still retained the key musical elements such as the musical pitch. To evaluate if the reduced set still had sufficient data to convey a musical attribute, we generated new sounds electroncially by using only the reduced set to describe synthesized sound output. (See Fig. 2.) Conceiving new techniques for reducing the original sound data, on

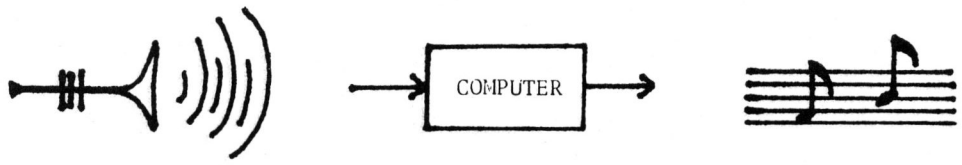


Figure 1. Sound-to-music notation was studied using computational methods.

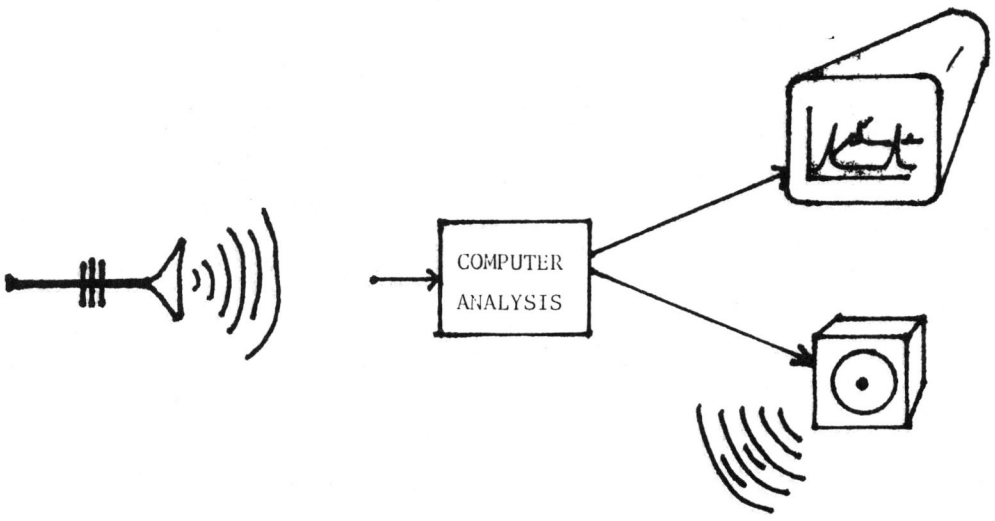


Figure 2. The computer analyzes the sounds and may display the intermediate data on a CRT screen or generate synthesized sounds based on the detected sound features.

the other hand, was a most difficult problem; fortunately, computer graphics was extremely helpful here. Because we wished to abstract general patterns from the highly detailed original sound information, graphics displays aided us by revealing the breath of information any procedure would have to handle to accomplish this reduction task. The procedure's output could likewise be evaluated and compared to its input, using only graphical forms, so we could to better understand how the procedure transformed the original data.

Static (non-motion) displays of the time-varying music information were the first we developed, for example, Fig. 3. In studying output from this type of display music information, we first would find a time reference in the graph. Once this point was established we could then go on to recognize other graphical features and associate them to the sounds. For example, if the graph contained several musical notes played in sequence, we would look for the beginning of the first note, and then identify the other notes proceeding from this known reference point. Problems arise, however, when passages of hundreds of consecutive musical notes are to be studied. A single graph cannot contain all the details. Several more detailed graphs taken from middle sections of the time sequence, however, were hard to interpret because of the difficulty in finding an obvious landmark in the graph that could serve as a reference point for interpreting the rest of the graph. This graphics problem led us to develop motion graphics-with-sound output, as is described below.

TECHNICAL FACILITIES

An integrated hardware and software system was developed, enabling us to explore sound-based data using a broad variety of techniques and analysis procedures.

Sound input aid - SMPTE reader: Because we worked with sound that had been recorded on a four-track audio tape recorder, we wished to know precisely where the sound segment began on the audio tape. We therefore put a SMPTE time code (Mallon, 1979) on an unused track of the tape prior to computer processing. This track could then be continuously monitored by the computer to get the current position of the playing tape. When we wished to enter sounds into the computer we could do so by specifying the SMPTE time when we wished to begin digitizing. A computer program would then look at the SMPTE time-code track for the specified target time and begin digitizing the analog sound signal when the time was found (i.e., when the target position of the tape was at the read head of the recorder). This freed us from storing vasts amounts of digitized data for archival purposes, since we could always reenter the same segment of the original analog sounds with less than a 10-msec restart error. This SMPTE reading capability was accomplished without any new hardware; we used the system's A/D converter and software instead of purchasing a hardware SMPTE reader costing several thousand dollars. The synchronization between the audio tape recorder and our computer also enabled us to have graphical output appear in synchronization with the recorded sound as we will describe shortly.

Sound output: A 16-bit digital-analog converter (D/A) was custom-built for the system. (See Piszczalski, et. al. (1981) for additional technical details.) It was used extensively, both for (1) reconstructing sounds from data automatically extracted from original sounds ("analysis-resynthesis"), and (2) evaluating simpler, totally-synthesized sounds for psychoacoustic tests. The D/A capability was a tremendous aid in both of these modes. We could audibly evaluate the efficacy of our processes in the analysis-resynthesis mode and also pinpoint significant perceptual-acoustical relationships in the tightly controlled second mode. Several software programs and subroutines were developed for the generation of synthesized sounds.

Graphical displays: We interfaced a

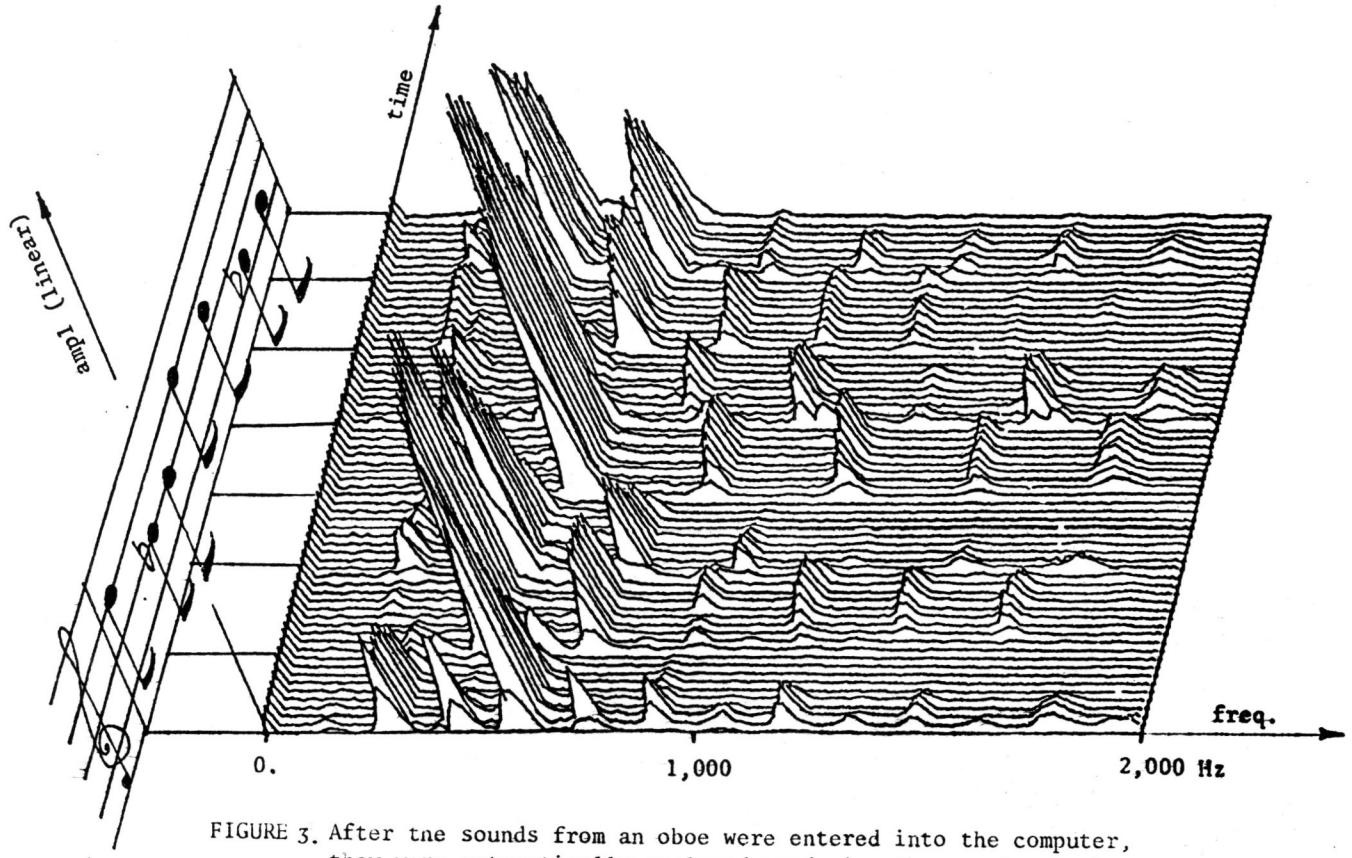


FIGURE 3. After the sounds from an oboe were entered into the computer, they were automatically analyzed producing the spectrograph-like data seen here. The musical notes on the left show the melody that was played.

raster-scan, color display susbsystem manufactured by Ramtek Corporation to our computer system. Data could be transferred at high speeds (i.e., under DMA control) between our host computer and this unit. Several programs were written, the most useful of which took two-dimensionally arrayed data and produced a fast, pseudo-three-dimensional image using a simple color-bar type format. We first transmitted color vectors to the background areas of the screen and then successively wrote the vectors to the foreground areas; hence, the ordering of the "write" calls produced an effective hidden-line removal effect with no computations needed for actually removing hidden lines - they are simply overwritten.

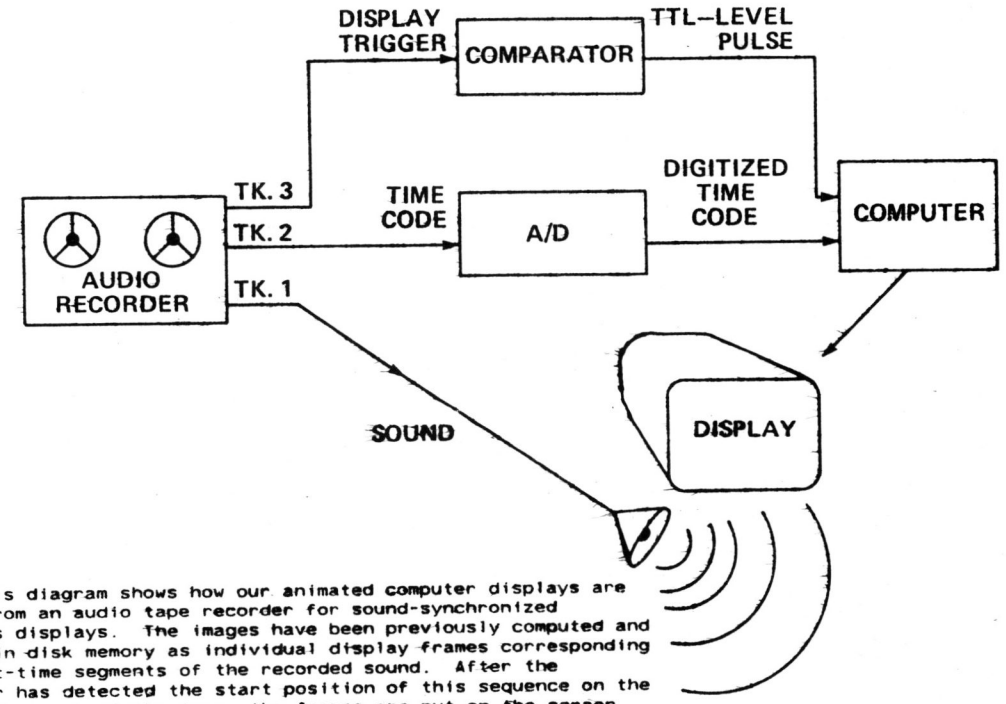
Sound-synchronized displays. Because our graphics presented data from sounds, we wanted a system that would show a continuously changing image, changing in synchrony with the sounds it represented. At first we put together the necessary software and hardware for an automated computer animation system for producing 16-mm films of the changing graphics. The sounds associated with the visual images were then added later on the standard optical track on the film.

The next approach we developed avoided the lengthy film developing steps of the previous method by producing a simulated real-time sound-synchronized display. This procedure began by first noting the SMPTE tape time when the sounds were entered into the computer. While the sounds were being entered into the computer, a trigger track would be simultaneously recorded on an unused track of the audio recorder a pulse for every 512 samples entered into the computer, i.e., the "trigger" pulses. Next the sound data was analyzed and the display images were calculated off-line. At this stage everything had been processed and the sounds were now represented by vector-display data, grouped into display frames and stored in the host computer's disk memory. We then rewound and started the tape. A computer program waited for the beginning SMPTE time to be detected on the time-code track. Once this happened, the trigger track of the audio tape would externally pace the rate at which the host computer sent display frames to the screen. Sound-visual synchronization was thus insured by locking the display rate to the tape speed of the original sounds as they were being replayed on the audio tape recorder. (See Figure 4.)

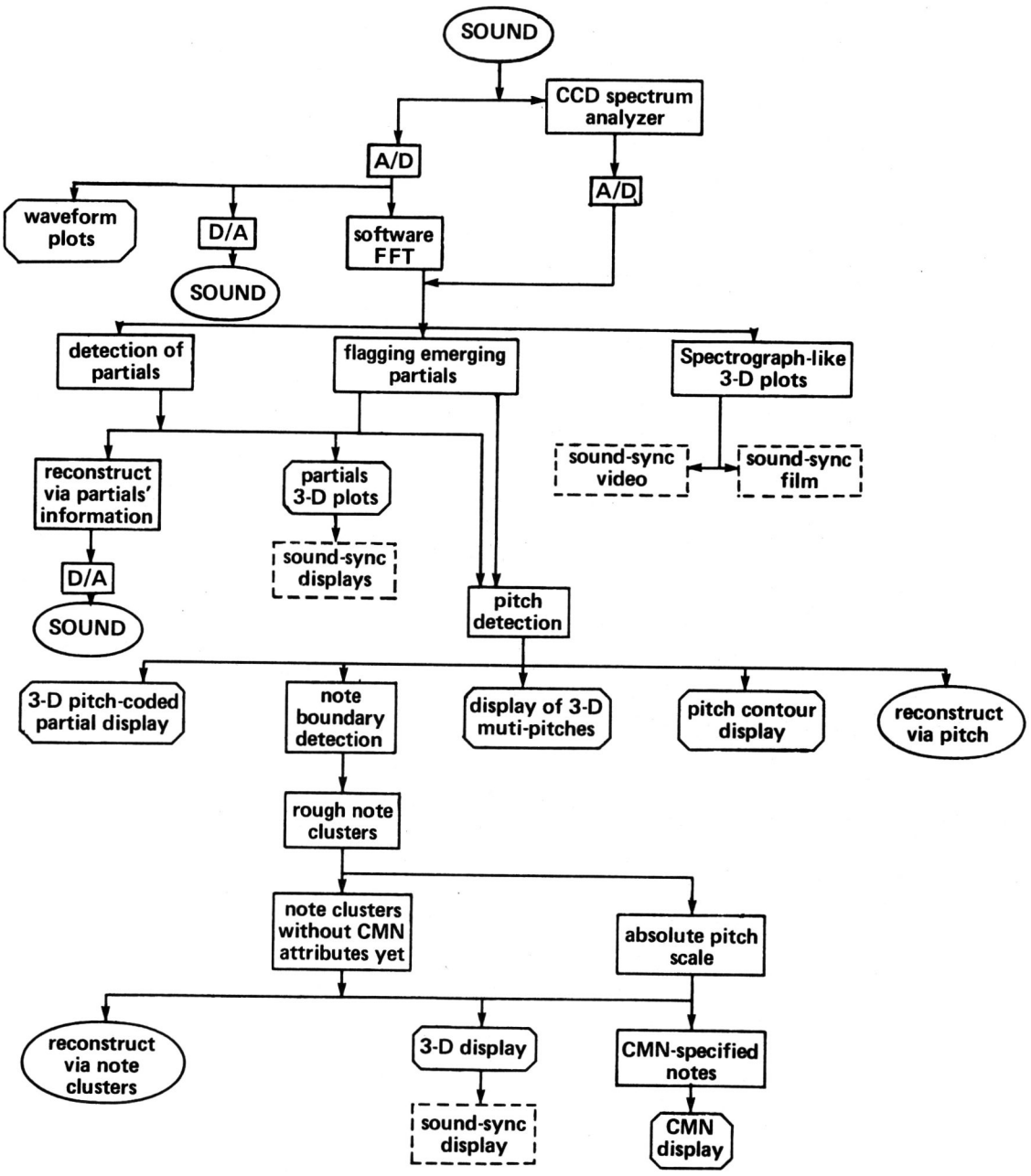
The strength of our overall system lies in the ability to monitor at several points what the analysis routines are doing to the representation of the sounds. Graphics or reconstructed sounds give results that may be quickly evaluated; sound-synchronized displays are especially effective. Figure 5 gives a broad overview of the analysis system showing where sounds and graphics can be generated from the data available at several stages in the processing; unfortunately, describing the behavior and rationale for each of these components in the system is beyond the scope of this report. (See Piszczalski and Galler, 1978.)

Our research facility apparently differs from that of any other in the world concerned with hearing research in the following respects: (a) sounds can be reconstructed at several stages in the analysis, corresponding roughly to different perceptual levels; (b) the data can be graphically displayed in several different formats corresponding to these different levels of sound representation; and (c) either the original or any of the reconstructed sounds can be presented in synchrony with any of the graphical images derived from the same, original sound segment.

In computer-vision and image-processing research, reconstruction of an original photograph is often done to check what each process in a complex system is doing to the image representation. We believe our facility is unique in hearing research in our extensive sound reconstruction cabilities that give comparable auditory feedback throughout the processing. As for displays, we know of no other music research facility that presents sound information by showing conceptual groupings, particularly among the more primitive acoustical elements. For example, in



4. This diagram shows how our animated computer displays are paced from an audio tape recorder for sound-synchronized graphics displays. The images have been previously computed and stored in disk memory as individual display frames corresponding to short-time segments of the recorded sound. After the computer has detected the start position of this sequence on the time code track of the tape, the frames are put on the screen synchronized with the sounds playing on the tape recorder. The trigger pulses on another audio track paces the rate of display, locking the frame rate to the audio-tape speed and therefore to the sounds themselves.



KEY: SOUND display sound-sync display analysis

5. Hundreds of routines performing a variety of input, analysis, display and sound synthesis functions were integrated into our system. The intermediate data in several places in the overall processing could be monitored via graphics or reconstructed sounds. Also shown are places where sounds and graphics can be simultaneously presented in simulated real time, a tool we found to be especially helpful.

one spectrograph-like display, all the partials contributing to the same pitch are shown in the same color. Regarding sound-synchronized, animated displays, we know of no other which can display higher-level information above the primitive grey level representation of the raw spectrogram. Furthermore, our various windows into the processing can be used on longer sound segments, including hundreds of connected musical notes played on traditional instruments. This is in contrast to most other facilities that can digitally process only a few seconds of sounds from the outside world - not the connected, longer sound patterns we hear in everyday listening.

ACKNOWLEDGEMENTS

We wish to thank Ramtek Corporation, Santa Clara, California, for supporting this research. The work was done on the facilities of the Bioelectrical Sciences Laboratory, University of Michigan. We likewise acknowledge the critical past support of the National Science Foundation (Grant No. MCS-780905).

REFERENCES

Mallon, B. "SMPTE time code comes to audio," db pp. 39-41, November, 1978.

Piszczalski, M., and B. A. Galler, "Analysis and transcription of musical sound," Proceedings of the 1978 Int'l Computer Music Conference, Vol. 2, pp. 585-618, Northwestern University, Evanston, IL.

Piszczalski, M., B. A. Galler, R. Bossemeyer, M. Hatamian, and F. Looft, "Performed music: analysis, synthesis and display by computer," J. Audio Engineering Society, Vol. 29, No. 1, pp. 38-46. Jan.-Feb., 1981.

Seashore, C. E., The Psychology of Music, New York: McGraw-Hill,