A Neural Network Based Classifier and Biofeedback Device for Improving Clarinet Tone-Quality

Ian R. Fasel, Kurt D. Bollacker, Joydeep Ghosh
ianfasel@mail.utexas.edu, kdb@ece.utexas.edu, ghosh@ece.utexas.edu
Department of Electrical and Computer Engineering, University of Texas,
Austin, TX 78712

Abstract

This paper describes an automated tool for classifying tone quality (a quality related to timbre). This tool provides real-time visual feedback to players of clarinet to help improve tone production technique. A neural network architecture is employed to build a graphical biofeedback device that allows the user to immediately "see" what changes in technique lead to better tone-quality. The tone is also classified and probability estimates are shown in a bar graph, giving quantitative feedback to the user.

Introduction

In musical sounds, timbre is that quality which is left after pitch, duration and intensity are accounted for; it is the quality that distinguishes the instruments in an orchestra. However, once an instrument and its pitch have been identified, there still remains to be evaluated the tone-quality of the instrument relative to other instruments of its type. Tone-quality is a primary way of distinguishing skilled players from novice ones, and a refined tone is a common goal of all instrumentalists. Because tone-quality perception is a refined skill that beginning students often have difficulty with, an automated tool that can provide continuous multi-modal feedback is desirable as a pedagogical tool.

Tone quality is a feature that is difficult to analyze in physical and mathematical terms because it depends on a great number of parameters. The fact that the human auditory system can make distinctions and decisions concerning tone-quality without difficulty, in spite of this high dimensionality, has led researchers to study the mechanisms by which biological auditory systems can achieve such a reduction in dimensionality. Connectionist approaches, such as Kohonen's self organizing map algorithm, seem a logical approach to this problem, particularly in light of the fact that several kinds of ordered feature maps are known to appear in human structures for sensory perception, e.g., in somatosensory maps connected with the sense of touch, tonotopic organization in the primary auditory cortex, retinotectal mapping in the primary visual cortex, etc.

Laden [1] has studied neural network based methods for identifying pitch from complex signals, and Cosi *et al* [2] and Toivainen [3] have investigated neural network based methods for identifying which instrument has produced a signal (timbre). Self-organizing maps have been used by a number of researchers for artificially modeling timbre classification (Cosi [2]; Feiten & Gunzel [4]), while others have employed similarity scaling techniques to determine acoustic parameters contributing to perception of timbre in psychological subjects, using Multidimensional Scaling (MDS) to map similarity ratings of numerous samples into a low-dimensional space (Wedin [5], Grey [6]). A widely accepted result of several of these experiments is Grey's Timbre space, a three dimensional space for clustering timbre using MDS. Grey's timbre space is successful because its three dimensions can be roughly labeled as describing a) power spectrum b) synchronicity in attack, and c) presence of high frequency inharmonic noise in the attack stage.

While the present study is not a study of timbre but of tone-quality classification, tone-quality is a similar attribute, in some respects simply a more refined kind of timbre classification that further describes different instruments and players once their instrument has been identified. Thus it is reasonable to assume that tone-quality perception is processed by mechanisms quite similar to

timbre perception, and that similar methods for investigation can be useful. In particular, is it possible to embed tone-quality metrics into a very low-dimensional space while retaining significant similarity information? Our work investigates this question for a two-dimensional topological space that can be readily visualized, and answers it in the affirmative. This result forms the core for a powerful feedback and tutorial system for improving clarinet tone quality.

Like previous studies, this study uses Fourier analysis to preprocess samples, producing an auditory image which is then presented to a self-organizing map. As previous researchers have pointed out, using Fourier analysis to produce an auditory image effectively eliminates useful time-domain information, which Grey's timbre space has shown to be significant in timbre perception. However, while time varying effects such as high frequency inharmonic noise in the attack stage are important in identifying particular instruments, this study was primarily concerned with steady state tone-colors for a single instrument, making the attack parameters of Grey' space less significant. However, this study would ideally include better consideration of time-varying effects as well as provide better auditory modeling (such as cochlear modeling) for simulating the physical mechanisms of sound perception. Current research indeed involves investigation the use of various methods such as wavelet analysis to extract useful time-domain information while still considering the frequency information that is known to be significant in perception.

The overall system architecture is depicted in Fig. 1.

The following sections describe the different components in more detail.

Collecting and Pre-Processing the Data

A limited set of sound samples, representative of several distinct tone qualities, were utilized to train the initial network. Samples were taken with a Sun amplified microphone using the recording software built into the Sun OS on a ROSS Hyperstation. These samples were taken with 16 bits at a sampling frequency of 44.1 kHz and were then edited and prepared for use on an Apple Macintosh 8600/300 using Macromedia's Sound Edit 16 version 2.

To train the system, a clarinet player from the UT School of Music provided 300 samples of clarinet tones of varying quality, where each sample could be classified in one of four classes of tone quality. In addition to the examples of good tone-quality, three examples of poor tone quality were sampled, each identified by what fault was introduced into the tone production method. These were 1) closed throat, 2) low tongue position, and 3) poor air support. Samples were taken for each all of these four tone qualities across three pitches. The pitches were chosen to represent the three registers of the clarinet (middle C for the chalameau register, F2 for the clarion register, and D3 for the altissimo register) in order to account for the perceived change in timbre in each of these registers. Indeed, the harmonic content of each register is slightly different (essentially, lower harmonics are removed as the register number is increased), while harmonic content of each tone-quality is relatively uniform within each

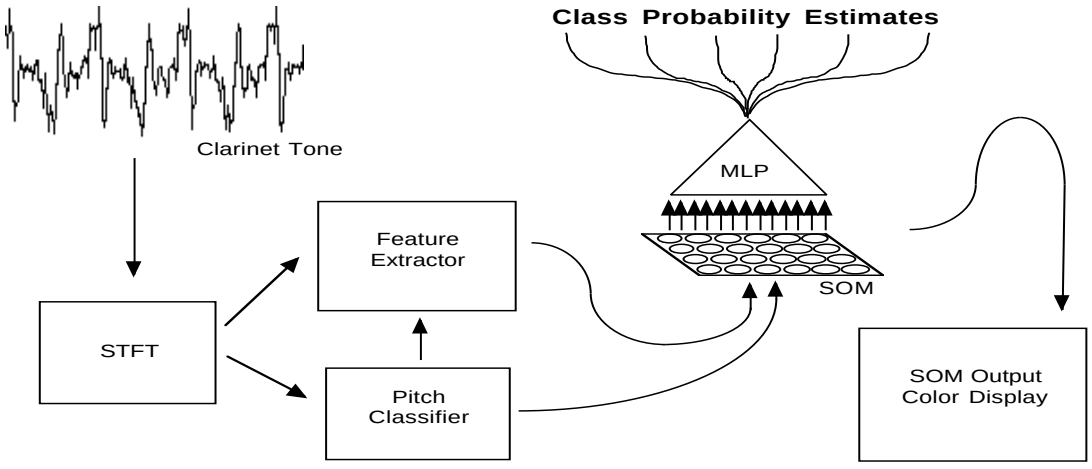

Figure 1. The clarinet tone is first preprocessed by an FFT, and the power spectrum is normalized. The frequency vector is then used to classify the pitch, which is used to decide which harmonics to extract to form the harmonic peak vector. This vector is fed into the appropriate SOM, whose response an MLP uses to classify the tones, while an intensity map visually displays the most intense SOM output node responses.

register.

Once samples were collected, they were preprocessed with a windowed Fourier transform using Welch's averaged periodogram method to produce an auditory image suitable for input into a SOM. Each sample was divided into overlapping sections, each of which was detrended, then hanning windowed, then zero-padded to length 8196. The magnitude squared of the DFTs of the sections were then averaged to form a power spectral density vector. This vector was normalized to make the samples intensity invariant by forcing the maximum peak of each sample to be equal to 10 dB. Finally, because the power spectrum density vector was several thousand elements wide, the bands corresponding to the harmonic peaks (typically appearing in integral multiples of the fundamental frequency) were extracted to make a reduced-dimensional vector of 32 elements, corresponding to the first 32 harmonics present in the tone (up to 12 kHz for F2). This dimension reduction was performed primarily to reduce computational effort and speed up response, but there is also evidence that a similar dimension reduction occurs in the human auditory system. Auditory fibers tend to fire in a "phase-locked" way in response to low-frequency stimuli, resulting in a more prominent response in those frequency bands that are integral multiples of the principal stimulus period, thus emphasizing the natural harmonics of musical instruments (Cosi 1994). Thus, this method of feature extraction has some biological plausibility.

Training

Once the training set was collected, a 10x10 hexagonally connected self organizing map was constructed using the SOM toolbox for Matlab. Due to the differences between registers, three SOMs were trained, one for each register. Once the SOMs were trained, the response of each SOM to each sample was recorded, resulting in 270 response vectors, where each of the one hundred elements of the vector represented the response of one node of the SOM. In order to help visually distinguish different groupings of samples, the twenty five strongest responding nodes were isolated by zeroing out the lower valued responses, resulting in a one-hundred element vector with seventy-five zeroes and twenty-five non-zero values for each sample. This forced Matlab to make all lower responding notes uniformly colored when the SOM response was displayed.

The response vectors of the SOM to the elements of the training set were then used as the training set for a multilayer perceptron (MLP), which used the high-response isolated vectors as inputs and the labels of each vector as the outputs. A schematic of the MLP is shown in Figure 2. Because there were three SOMs, one for each register, there were three corresponding MLPs, one for each SOM.

Output

Once the SOM and MLP had been trained, a production system was built which evaluates new samples in real time. First a sample is taken using a simple amplified microphone attached to the computer. This sample is then preprocessed in the same way as the original training samples: The tone is analyzed using a STFT, and the resulting vector is used to classify the pitch of the sample. The pitch information is used to identify the sample as being in register one, two or three, which is then used to

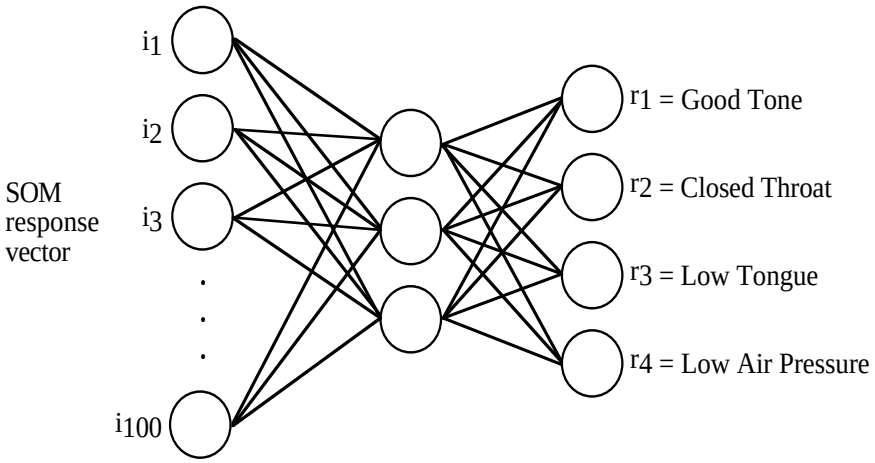

Figure 2: The multilayer perceptron (MLP) is trained on the response vectors from the SOM. Each output node represents one possible tone quality.

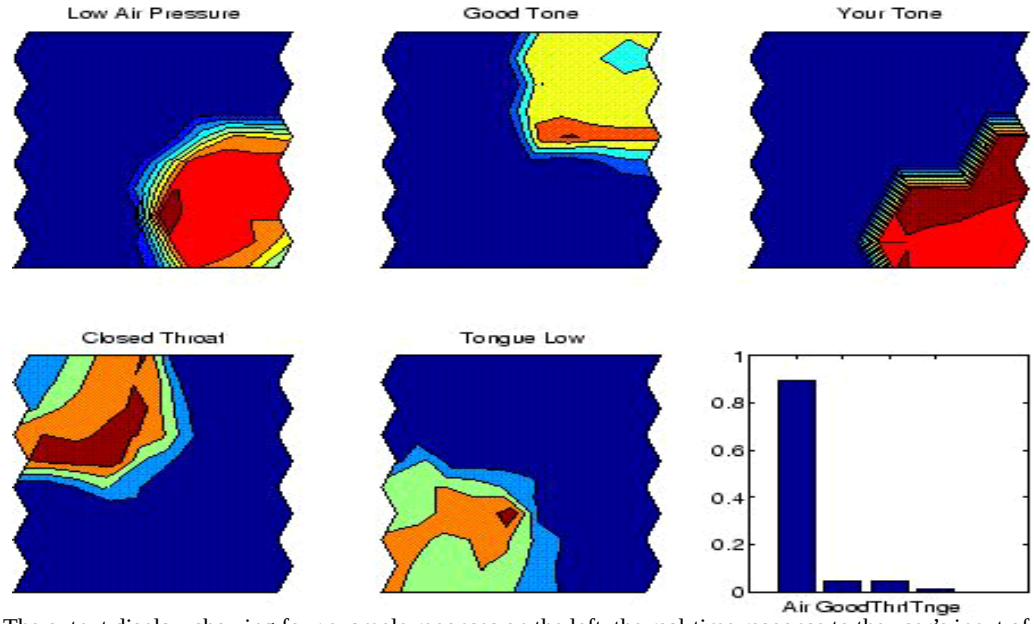

Figure 3: The output display, showing four example reponses on the left, the real-time response to the user's input of the upper right, and the MLP psterior class probablity estimates on the bottom right.

decide which harmonics should be extracted to form the 32 channel acoustical image. The register information is also important in deciding which of the three SOMs (and their corresponding MLP) to use. Once the 32 channel image is created, the vector is presented to the SOM, and, as in the training set, the 25 highest responding nodes are isolated by setting the values from the 75 lower response nodes to zero.

This response vector is then used in two ways. First, the vector is presented to the trained MLP, which gives the new sample a particular classification. The output of the MLP is shown in a bar graph, so that the user can visually see how closely their tone corresponds to one of the predefined classes. The accuracy of this MLP was tested in a leave-one-out cross-validation, and an accuracy of 93 percent was observed.

Second, the response vector is plotted directly in a two dimensional intensity map, where high responses are represented by colors early in the spectrum (such as red and orange), while low intensity responses are represented by darker, bluer colors. Figure 3 shows the output display seen by the user.

The visual output is enhanced in two further ways. First, in addition to the two displays described above, the user is also shown several example SOM responses to tone samples in the training set. This way the user can make visual comparisons between his tone and other tones. Second, the entire process of matching and displaying a tone sample takes less than a second, so the display of the user's tone quality can be updated in near real time. The sampling, classifying and displaying algorithms are placed in a continuous loop, so that the display is updated slightly faster than once per second. The resulting effect is that a changing tone-quality appears to result in a moving color region on the two-dimensional map.

Conclusions

The SOM and MLP make for novel and useful tools for musical analysis and pedagogy. This is an excellent example of how useful neural networks are for working with extremely noisy and mathematically poorly characterized data like musical signals.

There are several ways in which this system can be improved. First, it would be useful to allow many different clarinetists to test and provide feedback about the system to find more useful ways for representing tone-quality and possible solutions. Current results are based entirely on the opinions of one expert player. Second, because preprocessing is based on Fourier analysis, this machine is currently only capable of providing feedback about steady-state "long tones". It is not capable of providing feedback about the attack or decay stage of notes, nor can it respond effectively to time-varying changes in tone color that the more advanced clarinet players produce. Furthermore, while the clarinet players at UT do

not use vibrato, some clarinetists do, as do many other instruments. To this end, current research is investigating the use of wavelet analysis to replace the Fourier analysis module of the system in order to identify invariant features in the time-domain samples. This makes the case of feature reduction much more difficult, as there is no method as straightforward as selecting integral harmonics available here. One possible direction is the use of Kohonen's adaptive subspace SOM, which has been shown to spontaneously generate a wavelet-like solution through training. Until these possible measures are implemented, however, we must be satisfied with long tone results.

References

[1] B. Laden, "A parallel learning model of musical pitch perception," *Journal of New Music Research*, vol 23, pp. 133-144, 1994.

[2] P. Toivianen, M. Kaipainen, J. Louhivuori, "Musical Timbre: Similarity ratings correlate with computational feature space differences," *Journal of New Music Research*, vol 24, pp. 283-297, 1995.

[3] P. Cosi, G. De Poli, and G. Lauzzana, "Auditory modeling and self organizing neural networks for timbre classification," *Journal of New Music Research*, vol 23, pp. 71-98, 1994.

[4] B. Feiten and S. Günzel, "Automatic indexing of a sound database using self-organizing neural nets," *Proc. Of X Colloquium on Musical* Informatic, pp. 102-108, 1993.

[5] L. Wedin and G. Goude, "Dimension analysis of the perception of instrumental timbre," *Scandinavian Journal of Psychology*, vol 13, pp. 228-240, 1972.

[6] J.M. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol 63, pp. 1493-1500, 1977.

[7] T. Kohonen, "Self-organization and associative memory," Berlin: Springer-Verlag, 1984

[8] T. Kohonen, "The self-organizing map," *Proc of the IEEE*, vol 78, no 9, pp. 1464-1480, 1990.