Neural Network Pitch Tracking over the Pitch Continuum
I. J. Tayle and M. Greenough
Department of Physics and Astronomy, University of Wales College of Cardiff
PO Box 913, Cardiff, CF2 3YB, Wales, UK
Telephone: (+44) 1222 874040 Ext 6954 Fax: (+44) 1222 874056
email: im@astro.cf.ac.uk

ABSTRACT: This paper describes an ANN pitch-determining system which can accurately determine pitch from a wide variety of musical instruments across the pitch continuum. The system incorporates an ART MAP ANN initially trained on "semitone-bin distributions" (created by a neural-like input mapping from an FFT to semitone bins) for notes in concert pitch. At this stage it is able to classify pitch within a semitone. Further spectral analysis then allows an estimate of the pitch with a typical accuracy of about 10 cents.

1. Introduction: Pitch-tracking systems are much sought after in the musical world, for automated score-printing of annotated music, in ethnomusicological studies, and computer interaction with human musicians in live performances. Ideally, a pitch-tracking system which accurately simulates the way humans perceive pitch is needed. To this end, most researchers have incorporated, in their systems, ideas relating to certain pitch-perception theories and pitch-determination algorithms which are based (e.g. summing the subharmonics [TSS92] and [DWS82]) or matching the harmonic-ideal template against the input spectrum to find the closest fit [Go78]). Although such methods work well in general, there are instruments which produce a much depleted or inharmonic set of spectral components. Such spectral patterns may well confuse systems which use simple comparisons to determine pitch. Although such algorithms may be extended to cater for a greater variety of instruments, the process involves further pitch analysis and re-coding of computer implementations. Our method involves training an ART MAP ANN [CR91a] with a wide variety of spectrally different patterns, so that the information relevant to the pitch-determining process can be extracted. This has been shown to produce a more robust pitch-determining system capable of handling many more spectrally diverse cases. e.g. [TG94a] have shown that the method of subharmonic summation made 1.3 times more mistakes for absolute pitch and 4.5 times more for pitch chroma compared to an ART MAP network consisting of an "S-Art" network as the ART1 module and an ART 1 network as the ART2 module.

This paper describes an ANN pitch-determining system called SPeCS (SMARTMap Pitch Classifying System) which can accurately determine pitch from a wide variety of musical instruments across the pitch continuum. SPeCS self-organises an ANN consisting of two "S-Art" networks. The network within the supervised ARTMAP infrastructure to produce SMARTMAP (S-Art MAP Modified ARTMAP)

2. SPeCS Outline: SPeCS is outlined in figure 1. Briefly, the CD-quality signal is Low-Pass filtered, re-sampled, adapted, last Fourier transformed, interpolyed and transformed into a spectrual distribution before it is presented to the SMARTMAP network. After learning, the neural network can classify pitch within an accuracy of a semitone and then further examination of the spectrum reveals the pitch more precisely. The next six sections show the role of each of these operations in detail.

(i) Sampling, Low-Fass Filtering and Re-Sampling: The sound/s were sampled at CD-quality (44.1 kHz) and stored in sound files on a SUN SPARC & Workstation. Although, such a high sampling rate is not needed for determining pitch (i.e. only frequencies up to 4-5 kHz are needed), we chose this rate for three reasons. Firstly, our long-term intention is to make this system as general as possible so that it can work on a variety of computer platforms and ADC interfaces, some of which may have limited sampling-speed choices. Secondly, choosing a lower sampling rate would produce aliasing if there was no external low-pass filtering. Lastly, we are intending to build up a large database of musical sounds for analysis. High fidelity, therefore, could be important in some cases e.g. the analysis of timber.

The eventual goal is for this system to perform real-time pitch classification and at least 20 classifications per second are needed to handle e.g. fast trills. Since a recursive time-domain low-pass filter
was used it was convenient to process a continuous time series and so 2048 samples (nearest power of 2 to 44100/20) were processed at a time. This allows about 21.5 classifications per second. A 7th order time-domain Butterworth filter with a cut-off frequency of 4 kHz was used (designed by Tony Fisher's interactive Filter Design Program at York, UK, available through the World Wide Web). This gave a sufficiently good roll-off so that after re-sampling to 11025 Hz by taking every 4th sample a reasonably clean signal of 512 samples consisting of frequencies up to 5.5 kHz remained.

(ii) Adaptive Line Enhancing (ALE): [SD88]. This technique uses an LMS algorithm to update a set of coefficients (or weights) which are connected to previous points in the time-domain signal. The effect is to reduce the noise in the signal.

For example, figure 2a shows the Fourier Spectrum of a sine wave saturated by Gaussian noise. The frequency of 2713.184 Hz was chosen to be exactly divisible by the frequency resolution of the spectrum (i.e. 21.533 Hz). This minimises the effects of the frequency spilling over on to adjacent frequency points. Figure 2b shows the same signal after processing by the ALE algorithm. ALE is an optional part of SPeCS which is only used when there is a poor signal-to-noise ratio.

(iii) Spectral Processing: A 512-point fast Fourier Transform was then applied to the signal. Each bin in the Fourier spectrum therefore had a frequency resolution of 21.533 Hz. A process of interpolation then attempts to pin-point the peak frequency more accurately. Several interpolation tables involving amplitude ratios and relative shifts were set up, catering for the whole frequency range. The ratios are calculated by considering the relative amplitudes of the spectral points neighbouring the peak e.g.

\[ r = (S_j - S_{j-1})/(S_j - S_{j+1}) \]  

where \( S_j \) is a peak amplitude. Peaks are identified as those whose amplitude is significantly greater than the mean of 11 randomly chosen spectral points. Testing the method with 4000 sine waves of random frequency and phase indicated a typical accuracy of 0.4 Hz.
(iv) Fourier-to-Semitone bin Mapping Scheme: Groups of frequency components lying within bandwidths of a semitone are mapped to individual 'semitone bins' of a representative intensity. Thus the Fourier spectrum is transformed into a distribution of intensity over the semitones of the chromatic scale. The mapping scheme here uses a 'cheb-hat' function which identifies the frequency with the largest amplitude in the area within a bandwidth of a semitone around the semitone's centre frequency. This is given in the equation below.

\[ A_s = \begin{cases} 1 & \text{if } \frac{1}{2}(F_s + F_{s-1}) \leq f \leq \frac{1}{2}(F_s + F_{s+1}) \\ 0 & \text{otherwise} \end{cases} \]

where \( F_s \) is the frequency of semitone \( s \), \( f \) is the actual frequency under consideration and \( A_s \) is the amplitude at semitone-bin \( s \). Figure 3 shows how a spectrum of a C5 note played on a Cello is transformed from an amplitude spectrum to an interpolated spectrum, and a semitone distribution.

Figure 3: (a) 256-point spectrum with frequency resolution of 21.533 Hz; (b) with peak positions more accurately determined by interpolation; and (c) the resulting semitone distribution.

(v) SMARTMAP: ART 2-A [CGR95] is an algorithm which simulates the main properties of the ART 2 ANN. In fast-learn mode it accurately duplicates the behaviour of ART 2 and the authors also indicate schemes for simulating ART 2 in the slow-learn mode (called intermediate learning in ART 2-A). ART 2-A runs two to three orders of magnitude faster than ART 2. sART [TG94b] is a modification of the ART 2-A algorithm which speeds up ART 2-A in the intermediate learning case. By attaching, to each output node, a different learning rate which is decreased according to the amount of learning the particular node has had, the training-presentation time is significantly reduced, and thus a speed-up of up to two orders of magnitude can be achieved. The sART system has essentially the same architectural structure as ART 2-A but differs in its finer architecture levels and function. sART consists of a layer of \( M \) \( F_N \) nodes, a layer of \( N \) \( F_M \) nodes, fully connected by a set of bottom-up adaptive weights, and an orienting subsystem which incorporates the reset mechanism (see figure 4a).

Recently, a parallel S.ART algorithm has been implemented on a Tr Anastasian distributed-memory parallel computer. The Paramid consists of 48 i860-XP nodes, each with a processing capability of 100 MFlops. Two parallel S.ART networks were then connected by a dynamic map-field network to produce parallel SMARTMAP. The whole network’s architecture can be seen in figure 4b. These networks were implemented, essentially, in the same way as our parallel ART 2-A and ARTMAP networks [Txy95].

(vi) Estimating the Exact Pitch: The amplitudes of the significant spectral peaks are held in an array, along with their frequencies (now known more precisely through interpolation). The ANN identifies the pitch within a semitone. Reference is then made to the array to find a significant peak (the fundamental) within a quarter-tone of this pitch. Subsequent peak-hunts are conducted, roughly around the 2nd, 3rd etc harmonic frequencies, but guided by a running mean of subharmonics reflecting the best estimate so far of the fundamental frequency. This procedure can cope with any missing or mistuned harmonics and determines the pitch with a typical accuracy of 10 cents.

3. Conclusions: The system was trained with 49 chromatic notes (C5 to C6) taken from 11 instruments, chosen for their spectral variety, including piano, tubular bells, trumpet and banjo. The 539
training patterns were learned in 20 seconds by the parallel SMARTMAP network using 3 processors, and
indicators are that the system will comfortably classify pitch in real-time. On a SPARC 5 workstation
it cas cope with about 30 pitches a second.

References

[CGR91a] Gall A. Carpenter, S. Grossberg, and J.H. Reynolds. ARTMAP: Supervised real-time learning
and classification of nonstationary data by a self-organizing neural network. Neural Networks,

[CGR91b] Gall A. Carpenter, S. Grossberg, and D. B. Rosen. ART 2-A: An Adaptive Resonance algo-

tation of Goldstein's theory of pitch perception. Journal of the Acoustical Society of America,


[SD88] Samuel D. Stearns and Ruth A. David. Signal processing algorithms, Chapter 12. Prentice-

[Tay95] Ian Taylor. Parallel supervised atmap learning of analog multi-dimensional maps using self


[TGG94a] Ian Taylor and Mike Greenough. Evaluation of artificial-neural-network types for the deter-

[TGG94b] Ian Taylor and Mike Greenough. S.ART: A modified ART-2-A algorithm with rapid inter-
mediate learning capabilities. Proceedings of the IEEE International Conference on Compu-

[TSS82] E. Terhardt, G. Stoll, and M. Seewald. Algorithm for extraction of pitch and pitch salience