Reprint of a paper presented at the 1990 International Conference on Acoustics Speech and Signal Processing (p. 357-360 vol.1).

A Perceptual Pitch Detector

Malcolm Slaney
Richard F. Lyon

Apple Computer, Inc.
Cupertino, CA 95014

Abstract

We have implemented a pitch detector based on Licklider's "Duplex Theory" of pitch perception, and tested it on a variety of stimuli from human perceptual tests. We believe that this approach accurately models how humans perceive pitch. We show that it correctly identifies the pitch of complex harmonic and inharmonic stimuli, and that it is robust in the face of noise and phase changes.

This perceptual pitch detector combines a cochlear model with a bank of autocorrelators. By performing an independent autocorrelation for each channel, the pitch detector is relatively insensitive to phase changes across channels. The information in the correlogram is filtered, nonlinearly enhanced, and summed across channels. Peaks are identified and a pitch is then proposed that is consistent with the peaks.

1. Introduction

This paper describes a pitch detector that mimics the human perceptual system. Traditional approaches base a pitch decision on features of a relatively primitive representation such as the waveform or spectrum. Our pitch detector uses an auditory model. Unlike the simpler techniques, this perceptual technique works for a wide range of pitch effects, and is robust against a wide range of distortions.

The technique used was first proposed by Licklider [1] as a model of pitch perception, but it has not been taken seriously as a computational approach to pitch detection due to its high computational cost.

The representation used by the pitch detector, which corresponds to the output of Licklider's duplex theory, is the correlogram. This representation is unique in its richness, as it shows the spectral content and time structure of a sound on independent axes of an animated display. A pitch detection algorithm analyzes the information in the correlogram and chooses a single best pitch.

There are many signals, such as inharmonic tones or tones in noise, that do not have a periodic time or frequency-domain structure, yet humans can assign pitches to them. The perceptual pitch detector can handle these difficult cases and is thus more robust when dealing with the common cases. We expect that future systems will benefit by using this approach, or a cost-reduced version.

There is still considerable freedom to devise algorithms to reduce the rich correlogram representation to a pitch decision. The results we report are from a relatively simple algorithm, which does not address many of the subtle issues involved in a pitch tracker for use in a real system. Our algorithm picks a pitch for each frame of the correlogram, and does not address the decision of whether there is a valid pitch (as in the voiced/unvoiced decision in speech processing), nor does it attempt to enforce or utilize frame-to-frame continuity of pitch. Humans have complex strategies for making such decisions, depending on the task. For example in music, jumps in pitch define the melody, so continuity should not be enforced.

More work is needed to tune this model to accurately predict the pitch of inharmonic sounds. The results so far are consistent with the so-called "first effect", but not with the more subtle "second effect." Since the pitch consistent with the second effect can be seen in the correlogram, we expect that fancier algorithms can be devised to find it.

We know of only a few previous attempts to use Licklider's duplex theory as a pitch detection algorithm. The theory was originally published in 1951, but the computational cost of a cochlear model and a bank of autocorrelators remained a deterrent to its implementation until 1984, when Lyon [2] published the first image of a correlogram frame and Weintraub [3] used a cost-reduced "auto-coincidence" version as a pitch tracker for his two-voice sound separation experiments. More recently Lazzaro [4] has described a silicon VLSI implementation using auto-coincidence of action potentials, along with results of pitch experiments similar to ours. Lyon [5] has discussed issues in the VLSI implementation of this class of model.

2. The Auditory Model

As shown in Figure 1, our model of human pitch perception has three stages. The inner ear or cochlea encodes the information in the acoustic signal into a multi-channel representation that may be thought of as instantaneous nerve firing probabilities. The second stage of processing produces a correlogram, a two-dimensional image in which each row is the running short-time autocorrelation of the corresponding cochlea channel. Finally, a pitch detector combines the information in all the channels of the correlogram to decide on a single pitch. Humans can perceive multiple pitches but for the purposes of this paper we choose a "best" pitch.

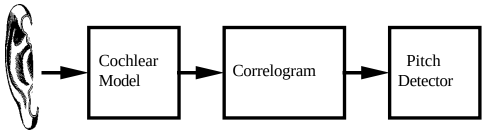

Figure 1. Three stages of neural processing are used in our model of pitch perception.

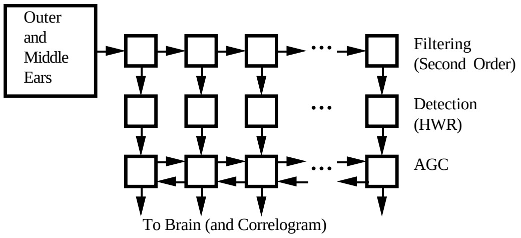

Figure 2. Our cochlear model combines filtering, detection and automatic gain control.

We use a cochlear model designed by Lyon and described by Slaney [6] to convert a sound waveform into a vector of numbers that represent the information sent to the brain. This system is diagrammed in Figure 2. It is important to remember that the cochlear model used here does not try to accurately model the internal structure of the ear but only to approximate the information contained in the auditory nerve. Other, more accurate models can be substituted to get better results.

A cascade of second order filters is used to model the propagation of sound along the Basilar Membrane (BM.) At each point along the cochlea the BM responds best to a broad range of frequencies and it is this movement that is sensed by the Inner Hair Cells. The "best" frequency of the cochlea varies smoothly from high frequencies at the base to low frequencies at the apex.

Inner Hair Cells only respond to movement of the BM in one direction. This is simulated in the cochlear model with an array of Half Wave Rectifiers (HWRs) that detect the output of each second order filter. The HWR nonlinearity serves to convert the motion of the BM at each point along the cochlea into a signal that represents both the envelope and fine time structure.

Finally, four stages of Automatic Gain Control (AGC) allow the cochlear model to compress the dynamic range of the input to a level that can be carried on the auditory nerve. The multi-channel coupled AGC used here simulates the ear's adaptation to spectral tilt as well as to loudness.

3. The Correlogram

The correlogram is an animated picture of the sound that shows frequency content along the vertical axis and time structure along the horizontal axis. If a sound is periodic, the autocorrelation functions for all cochlear channels will show a peak at the horizontal position that corresponds to a correlation delay equal to the period of repetition. This is generally equal to the perceived pitch period. Since the peaks in all channels, or rows of the image, occur at the same delay, or horizontal position, they form a vertical line in the image. The duplex theory says that sounds with a perceived pitch, even if they are not periodic, will produce a vertical structure in the correlogram image at the delay equal to the perceptual pitch. On the other hand, formants, or narrow resonances in the frequency domain, are displayed as horizontal bands in the correlogram.

The correlogram is computed by finding the (short-time, windowed) autocorrelation of the output of each cochlear frequency channel. The autocorrelator can be implemented with an FFT, as was done in this study for efficiency reasons. But it is more likely that the brain computes it using a neural delay line. This structure is very similar to the cross-correlator structures that have been found in the brains of owls [7] and cats [8] for spatial localization, but the structures that could compute the correlogram for pitch have yet to be found. Perhaps the hardest feature of the correlogram to neurophysiologically justify is the long time delay needed (on the order of 10ms.) Other schemes, for example, based on mechanical delays in the cochlea, have been proposed to implement a correlation [9].

An autocorrelation function requires twice as much dynamic range as the input signal that it represents. The required dynamic range is reduced in our model by partially normalizing each autocorrelation by the energy at that frequency. If each autocorrelation was normalized by its energy, the correlogram would not show any variation as a function of frequency and formants would not be seen. Instead we normalize each correlation with the energy raised to the 3/4 power. This serves to reduce the dynamic range of the correlogram to half the dynamic range of the input, but does not hide the important features, such as formants.

The correlogram is similar to another scheme that has been proposed to model auditory perception. Patterson's Pulse Ribbon Model [10] delays the outputs of individual neurons and then searches for common firings across cochlear channels. The correlogram is more robust since it sums up the firings from multiple channels and across time. But more importantly, an autocorrelation is only concerned with time differences so it effectively zeros out the phase between channels.

4. A Pitch Detector

The correlogram clearly shows many aspects of auditory perception. A voiced sound will excite several parts of the cochlea by different amounts. Each frequency in the voice is modulated by the vocal cords; the common periodicity across cochlear channels, which is seen as a vertical structure in the correlogram, is an indication of pitch.

Humans can easily perceive multiple pitches. An excellent example of this is described in McAdams' thesis [11]. McAdams separated the harmonics of an oboe into even and odd components, and then created a new sound by adding independent vibrato to the two sets of harmonics. The resulting sound clearly has two independent voices, separated in pitch by an octave. An animated correlogram of this "oboe" clearly shows two independently moving "objects"; but separating multiple sounds and calculating their pitches is not the subject of this paper. Instead we will show an algorithm that can choose a single "best" pitch to describe a sound.

Our pitch detector consists of four steps. A preprocessing step modifies the correlogram to enhance the peaks. The value at each time lag in the enhanced correlogram is then summed across all frequencies. Peak locations at this stage give estimates of all the possible periodicities in the correlogram. The third step is to combine evidence at the subharmonics of each pitch to make the pitch estimate more robust. Finally, the largest peak is picked, being careful to avoid octave errors, and a numerical value of the pitch is determined based on the location of the peak. This sequence of steps is shown in Figure 3 for the vowel /u/.

We use two stages of preprocessing. First, the correlogram is convolved with an operator that enhances vertical lines in the

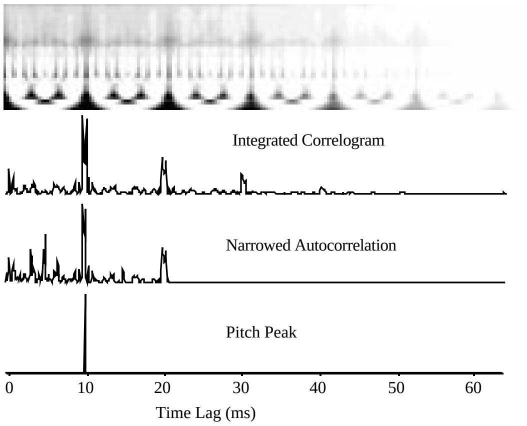

Figure 3. The correlogram and perceived pitch for a male speaker saying the vowel /u/.

correlogram. One such operator is shown in Figure 4. The second stage enhances the correlogram by passing it through an expansive non-linearity (half-wave rectification and squaring). This serves to enhance the peaks which are indicative of the periodicities in the sound. These preprocessing steps have been used in the examples that follow.

The correlogram, C(τ,f), is integrated across channels to calculate a one dimensional estimate of the pitches present in the sound,

$$P(\tau) = \sum_i C(\tau,f_i)\ df.$$

The pitch function, P(τ), is a function of autocorrelation lag, τ, and represents the likelihood that a pitch is present with a frequency of 1/τ.

Brown [12] defines a Narrowed Autocorrelation (NAC) function which is computed from the pitch function by

$$P_N(\tau) = \sum_i^N (N-i)\ P(i\tau).$$

In the absence of our nonlinear enhancement stage, the NAC function would be equivalent to a modified version of the autocorrelation given by

$$C_N(f,\tau)=\int[R(f,t\ )+R(f,t+\tau)+\ldots+R(f,t+N\ \tau)]^2 dt$$

where R(f,t) is the instantaneous firing rate of auditory nerves with a center frequency of f at time t. For example, a pitch of 100 Hz will show a peak in the integrated correlogram at 10 ms and at the subharmonics of 20 ms, 30 ms and so on. The NAC allows these subharmonics to be considered when the pitch is determined.

We use a technique described by Nishihara [13] to judge the location of the pitch peaks. In general the peaks in the pitch function are symmetric and an accurate estimate of their center is made by fitting a polynomial to the points near the peak. Using multiple points to determine the location of the peaks allows the pitch period to be determined with a resolution finer than the sampling interval (in low noise situations), and a more robust estimate to be made when noise is present.

$$\begin{matrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{matrix}$$

Figure 4. This convolutional operator is used to emphasize the vertical structure in the correlogram.

5. Results

The results of three experiments are described here. Two of the sound examples are from a wonderful compact disc of auditory examples produced by the Acoustical Society of America [14].

A pitch can be perceived for sounds that do not have any energy at the fundamental. This is called residue pitch and there are many excellent examples of this phenomena on the ASA CD. Demonstration 22 includes an example of residue pitch with low frequency noise (approximately –20 dB signal-to-noise ratio or SNR.) The pitch computed from one frame of the correlogram of this sound is shown in Figure 5. While in this particular case the lowpass noise can be easily removed with a filter, the perceptual pitch detector described here does equally well with the opposite case, a low frequency pitch with highpass noise.

The example above can be explained by the periodicities of the filtered waveforms, but the same can not be said of inharmonic sounds. Consider a three tone complex produced by AM modulating a 2kHz tone at 200 Hz. This will be perceived with a pitch of 200Hz since the three tones represent the 9th, 10th and 11th harmonics of a 200 Hz fundamental. Now, if the carrier frequency is moved up 40Hz to 2040 Hz the tones no longer form a harmonic complex. To a first approximation the sound is perceived to have a pitch of 204 Hz, as if the sound is still harmonic and is approximately the 9th, 10th and 11th harmonics of a 204 Hz fundamental. This result is shown in Figure 6.

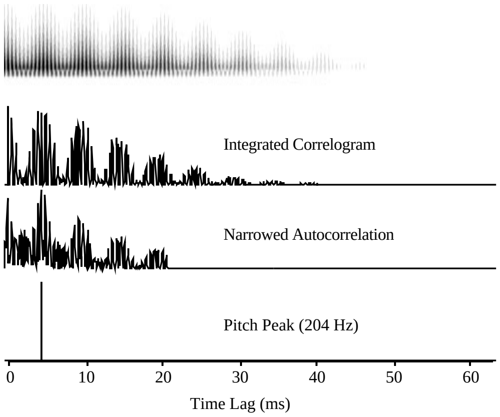

Figure 5. The correlogram and perceived pitch of an inharmonic tone is shown here.

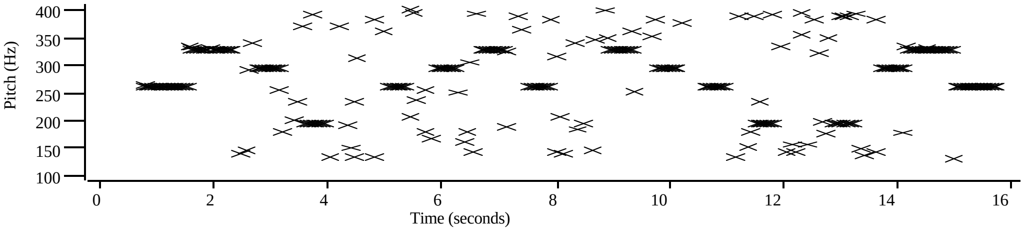

Figure 5. The perceived pitch of example #22 from the ASA CD. In this example, the Westminster chimes melody is played by alternating a low tone (fundamental only) and a residue tone at the same pitch. During the middle 12 seconds of this example low pass noise is added to give an approximate SNR or -20 dB. Our pitch detector does not make a decision about whether valid pitch is present so during the times between residue pitches (indicated by solid black horizontal lines) the pitch is determined by the noise.

As the carrier frequency of this example is raised, people will hear the pitch fall slowly and then jump to a higher pitch. The slow fall of pitch as the carrier frequency is raised is known as the second effect [15]. While the correlogram shows this effect at the low end of the almost-vertical line, we do not know how to work this result into a robust algorithm.

Finally, Figure 7 shows the pitch perceived by our algorithm due to the Shepard tones. Demonstration 27 of the ASA CD contains a continuous example of the Shepard tones (done by Jean-Claude Risset) that is always decreasing in pitch. This pitch of this sound is ambiguous, and typically listeners follow one pitch for a while until it becomes unreasonably low and then a new pitch is chosen. Our algorithm also shows an ambiguity by jumping between possible choices.

6. Acknowledgements

We are grateful for assistance from our colleagues, Richard Duda (San Jose State), Bill Stafford (MIT) and Steve Greenberg (Wisconsin). We also appreciate the support of the participants of the Auditory Perception reading group at Stanford's CCRMA.

7. References

[1] J. C. R. Licklider, "A duplex theory of pitch perception," in *Psychological Acoustics*, E. D. Schubert (ed.), Dowden, Hutchinson and Ross, Inc., Stroudsburg, PA, 1979.

[2] Richard F. Lyon, "A computational model of binaural localization and separation," *Proceedings of IEEE ICASSP*, Boston, MA, 1983.

[3] Mitchel Weintraub, "A computational model for separating two simultaneous talks," *Proceedings of IEEE ICASSP*, Tokyo, 1986.

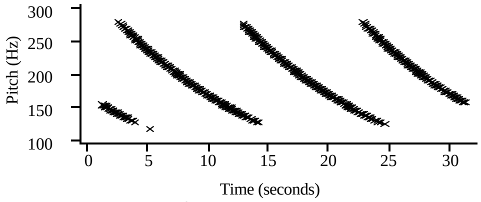

Figure 7. The pitch of a continuous Shepard tone is shown here. Note, that the pitch is ambiguous and this pitch detector can show different pitches in adjacent frames.

[4] John Lazzaro and Carver Mead, "Silicon models of auditory localization," *Neural Computation*, vol. 1, pp. 41-70, 1989.

[5] Richard F. Lyon, "Analog VLSI Hearing Systems," in *VLSI Signal Processing, III*, Robert W. Brodersen and Howard S. Moscovitz (eds.), IEEE Press, 1988.

[6] Malcolm Slaney, "Lyon's Cochlear Model," Apple Technical Report #13 (available from the Apple Corporate Library, Cupertino, CA 95014), 1988.

[7] C. E. Carr and M. Konishi, "Axonal delay lines for time measurement in owl's brainstem," *Proc. Natl. Acad. Sci.*, vol. 85, pp 8311-8315, 1988.

[8] T. C. T. Yin and S. Kuwada, "Neuronal mechanisms of binarual interaction," in *Dynamic Aspects of Neocoritical Function*, G. M. Edelman, W. E. Gall and W. M. Cowan (eds),Wiley, New York, 1984.

[9] Shihab A. Shamma, Naiming Shen and Preetham Gopalaswamy, "Stereausis: Binaural processing without neural delays," *Journal of the Acoustical Society of America*, vol. 86 (3), pp. 989-1006, September 1989.

[10] Roy D. Patterson, "A pulse ribbon model of monaural phase perception," *Journal of the Acoustical Society of America*, vol. 82 (5) pp. 1560-1586, November 1987.

[11] Steve McAdams, "Spectral fusion, spectral parsing and the formation of auditory images," Ph.D. Dissertation, Stanford University, May, 1984.

[12] Judith Brown and Miller S. Puckette, "Calculation of a 'narrowed' autocorrelation function," *Journal of the Acoustical Society of America*, vol. 85 (4) pp. 1595-1601, April 1989.

[13] H. Keith Nishihara and P. A. Crossley, "Measuring photolithographic overlay accuracy and critical dimensions by correlating binarized lapalacian of gaussian convolutions." *IEEE Transactions on Pattern Analysis and Machine Intellicence*, vol. 10 (3), pp 17-30, January 1988.

[14] A. J. Houtsma, T. D. Rossing, W. M. Wagenaars, *Auditory Demonstrations* (Compact Disc), Acoustical Society of America, (500 Sunnyside Boulevard, Woodbury, NY, 10797), 1987.

[15] Arnold W. Small, "Periodicity Pitch," in *Foundations of Modern Auditory Theory*, Jerry V. Tobias (ed), Academic Press, New York, 1970.