A computational model of harmonic chord recognition

Riana Walsh[1], Dr Donncha O' Maidin[1]

[1]Department of Computer Science & Information Systems, University of Limerick, Ireland

{riana.walsh,donncha.omaidin}@ul.ie

Abstract. *This computational model of harmonic chord recognition investigates the perception of harmonic chords by peripheral auditory processes and auditory grouping. The frequency selectivity of the auditory system is modeled using a bank of overlapping band-pass filters and a model of inner hair cell dynamics. The periodicity in each frequency channel is determined by autocorrelation. These periodicities are grouped according to pitch class. By computing the intervals between the high priority pitch classes, the model achieves considerable success in recognizing major, minor, dominant seventh, diminished and augmented chords. Its performance is evaluated on chords from instrument samples as well as from an ensemble recording.*

1. Introduction

In the auditory system, harmonics belonging to the same fundamental frequency tend to fuse and be heard as a single sound with a pitch corresponding to that fundamental frequency.

Two harmonic notes form the interval of an octave when their fundamental frequencies are in the ratio 2:1. Western music theory assumes octave equivalence: notes one or more octaves apart have the same harmonic function [Parncutt 1989]. These notes belong to the same pitch class. In Western music theory the harmonic triad is built out of three pitch classes. Both major and minor triads contain the perfect fifth interval. The diminished triad contains the diminished fifth interval (two minor 3rd intervals) and the augmented triad contains the augmented fifth (two major 3rd intervals). Along with perceiving each individual pitch in the triad we perceive the quality of the triad as a whole. Bregman [Bregman 1990] mentions that music may be thought of in terms of a horizontal and a vertical dimension. The horizontal dimension is time and the vertical dimension is pitch relations or harmonies. "The vertical organisation gives us not only the experience of chords but also other emergent qualities of simultaneous sounds, e.g. timbre, consonance and dissonance [Bregman 1990]." A chord can be thought of as a global entity that has its own emergent properties; the chord as a whole becomes the acoustic object we hear.

2. The auditory filters

The spectral analysis carried out by the basilar membrane is modeled using a parallel bank of 160 overlapping gammatone filters, each tuned to a different frequency [Patterson et al 1992]. A Matlab implementation is supplied by Malcolm Slaney [Slaney 1993]. The time domain impulse response of the gammatone filter is:

$gt(t) = a\, t^{(n-1)} \exp(-2\pi bt)\cos(2\pi ft+\theta); (t>0)$

The bandwidth of the filter is determined by b, and n is the order of the filter and so controls the slope of the skirts. The bandwidth of each filter is described by the Equivalent Rectangular Bandwidth (ERB). This is defined as [Glasberg and Moore 1990]:

$ERB = 24.7(4.37f/1000 + 1)$

With n = 4 then the filter bandwidth, b = 1.019*ERB in the current implementation [Patterson et al 1992, Slaney 1993]. The ERB increases with increasing channel centre frequency.

3. The hair cell model

Meddis [Meddis 1985, 1987] described a computational model of mechanical to neural transduction at the hair cell-auditory-nerve fibre synapse. The output excitation function in response to an acoustic stimulus is a stream of spike events precisely located in time. The model describes the production, movement and dissipation of transmitter substance in the region of the hair cell-auditory-nerve fibre synapse. Meddis [Meddis et al 1990] published a computer program of the hair cell model. This program has been implemented in Matlab by Malcolm Slaney [Slaney 1993]. The output of each filter is passed to the inner hair cell model. The model implements one hair cell per frequency channel. The model assumes that the probability of a spike occurrence is linearly related to the amount of transmitter substance in the synaptic cleft between the inner hair cell and its corresponding auditory nerve.

The following equations are computed at each time interval dt and define the operation of the model. The hair cell contains a quantity of free transmitter, q(t). A fraction of this transmitter is released between time t and time t+dt across the membrane into the synaptic cleft.

Fraction of transmitter released = $k(t)q(t)dt$

The permeability of the membrane, k(t), fluctuates as a function of the instantaneous amplitude of the acoustic stimulus, s(t).

$k(t) = g[s(t) + A]/(s(t)+A+B)$, with $[s(t)+A]>0$

$k(t) = 0$, for $[s(t)+A]<=0$, A and B are constants $B>A$.

When s(t) = 0 (no acoustic stimulation), spontaneous activity in the auditory nerve is allowed for by gA/(A+B), the transmitter leak rate. In this instance k(t) varies between 0 and g. B is the rate at which k(t) approaches its maximum value as a function of s(t) and –A defines the lowest instantaneous amplitude at which the membrane remains permeable. The synaptic cleft contains a fluctuating amount of transmitter substance c(t). The amount of transmitter lost is:

$lc(t)dt$

The amount of transmitter substance returned to the hair cell is:

$rc(t)dt$

Following the reuptake, rc(t)dt, the transmitter is subject to reprocessing delays and forms a reprocessing store, w.

dw/dt = rc(t) – xw(t)

A fraction xw(t)dt of the transmitter in this store is continuously released into the free transmitter pool. Free transmitter is replenished at a rate of y[M-q(t)]. M is the maximum amount of transmitter in the transmitter pool. The cell transmitter level q(t) is replenished through manufacture and return from the cleft but is depleted by loss into the cleft.

dq/dt = y[M-q(t)] + xw(t) - k(t)q(t)

The cleft transmitter level c(t) is replenished by release from the hair cell and depleted by loss and return to the hair cell.

dc/dt = k(t)q(t) – lc(t) – rc(t).

Spike occurrence in the auditory nerve is assumed to be linearly and probabilistically related to the residue of transmitter substance left in the cleft after the loss and reuptake. The probability of a spike occurring is:

Prob (event) = hc(t)dt, h is a constant.

The parameters of this model were optimised by Meddis [Meddis 1987] in order to simulate as closely as possible known responses at the hair cell-auditory-nerve synapse. The parameters of the model are set to simulate a high spontaneous-rate fibre. A feature of the model is its ability to simulate the decrease in phase locking for high frequency stimuli. The rate at which the transmitter is cleared from the cleft, loss and reuptake, is linked to the models ability to reflect the fine structure of the stimulus. It was noted that although this movement of transmitter causes a decrease in phase locking in the model it is not necessarily a true explanation of phase locking. Phase locking is less evident when this rate is slow relative to stimulus frequency.

4. Periodicity detection

The next stage of the model uses autocorrelation to detect periodicity in each frequency channel. The autocorrelation of a sequence x[n] is:

$$a[m] = 1/Q \sum_{n=0}^{n=Q-m-1} x[n]x[n+m]$$

Q is the length of the data sequence and m is the autocorrelation lag. For m=0, a[m] is a calculation of the total energy in the signal. The autocorrelation of a sequence defines how similar a signal is to a time-delayed version of itself. This similarity is greatest when the time delay is zero (the zero lag point). If the signal is periodic then the autocorrelation value at some non-zero value of the lag is close in value to the zero lag. The autocorrelation pattern of a periodic signal is also periodic. The periodicity value of the signal is indicated by the position of the first maximum after the zero lag maximum.

The frequency selectivity of the auditory system refers to its ability to resolve the frequency components of a complex sound. The ear has limited frequency resolution. The discrimination by the auditory system of individual frequency components in the sound depends in part on the frequency resolution of the basilar membrane. Resolution on the basilar membrane is generally described in terms of the critical band. If the action of the basilar membrane is thought of in terms of overlapping band-pass filters, then the ERB provides a practical measure of the critical band. The waveform of a frequency component that is resolved by a filter is a sine wave of that frequency. The frequency channel is resolved. When more than one harmonic, or frequency component, passes through a filter the output pulses at a rate equal to the frequency difference between the adjacent harmonics. There is a modulation present in the amplitude envelope at the output of the filter. This is an unresolved frequency channel. For a single harmonic sound the periodicity in the amplitude modulation is equal to the fundamental frequency. Autocorrelation is used to determine the main channel periodicity and the periodicity in the amplitude modulation.

In the next stage the envelope of the autocorrelation pattern for each frequency channel is extracted. The envelope of the autocorrelation pattern is the contour formed by the maxima in the pattern. For a resolved frequency channel the envelope pattern forms a straight line. If the frequency channel is unresolved there is a modulation present in the envelope pattern. The least squares method is used to detect envelope modulation in the autocorrelation pattern. A straight line is fit to the envelope of the autocorrelation maxima in each frequency channel. For a channel that is resolved, the straight line will fit quite well to the envelope maxima. For a channel that is unresolved the contour of the envelope maxima will deviate from a straight line. The correlation coefficient is used as a measure of how far the envelope contour of the maxima deviates from the straight line; it indicates the amount of envelope modulation in each frequency channel. A correlation coefficient value of 0.997 was chosen as a threshold value to indicate the presence or absence of envelope modulation. A value greater than this threshold value indicates that no envelope modulation is present. The closer the envelope of the maxima is to a straight line the less modulation there is in the channel. Frequency channels are divided into those with and without envelope modulation.

If the envelope modulation is periodic, its periodicity value is calculated. The criterion for periodicity used is success in a heuristic search for the first maximum which has corresponding maxima at integer multiples of its lag value in the autocorrelation pattern.

5. Interval and chord identification

This algorithm identifies the major, minor, diminished, augmented triads and the dominant 7th chord. A harmonic triad is made up of three pitch classes the root, third and fifth of the triad. More than three notes in the triad chord indicates an octave doubling. For major and minor triads the interval between the root and the fifth is a perfect fifth. For the major triad the interval between the root and the third is a major third and for the minor triad the corresponding interval is a minor third. The diminished triad can be considered as consisting of a minor third interval and a diminished 5th interval from the root, or two adjacent minor third intervals. The augmented triad is made up of a major third interval and an augmented 5th interval from the root or two adjacent major 3rd intervals. The dominant 7th has four distinct pitch classes and consists of a major triad with an added minor 7th interval above the root.

In general if the ratio between the two fundamentals of a musical interval is m:n, with m and n integers, every nth harmonic of the first fundamental will correspond in frequency with every mth harmonic of the second. When the ratio is small integers, such as 2:1 in the octave interval, there will be many exact correspondences of frequencies. However in equal tempered tuning, which was formed out a requirement for equally spaced intervals (in terms of frequency ratio) regardless of tonality, all intervals (apart from the unison and octave) match approximately to the corresponding interval found in the harmonic series. The difference between the nearly coinciding harmonics is small, within 5% of the critical bandwidth [Howard and Angus 2001].

The frequency channels are grouped according to pitch class. The envelope modulation periodicity and the main channel periodicities are collated. Each pitch class is assigned a priority based on the number of channels with that periodicity. The highest priority pitch class is the one that occupies the maximum number of channels. For a single harmonic sound, or an octave interval, the highest priority pitch class is the pitch class to which the fundamental frequency belongs. This is because the fundamental frequency appears as a beat periodicity in addition to possibly being present as a resolved frequency component.

The periodicities corresponding to the root, third and fifth generally appear as high priority pitch classes. The interval size in semitones is computed between all high priority pitch classes, searching for the fifth, and then for the major or minor third.

6. Results

This model was tested on chords created from instrument samples taken from the McGill University master sample (MUMS) CD set and on a recording of the second movement of Bach's Brandenburg Concerto No. 2 in F major. The instruments in the recording are recorder, oboe, violin, cello and harpsichord. From both sets of results it can be seen that the model performs well in identifying chords created from different harmonic instruments.

Table 1 holds the results of the instrument sample chords. The individual instrument samples were not balanced for equal loudness when creating the chords but kept at the original sample level. The durations were measured from the start of the instrument sample. Table 2 shows results for chords taken from the second movement of J.S. Bach's Brandenburg Concerto. The analysis was run on single time frames with starting time points corresponding as close as possible to one fifth, two fifths, three fifths and four fifths of the duration of the sound. The model performed well on major, minor and augmented chords. The performance of the model was less reliable on the dominant 7th and diminished chords. It may be noted that the diminished chord forms part of a dominant 7th chord (i.e. C, E, G, Bb is a dominant 7th and E, G, Bb is a diminished chord).

In table 1 four out of the seven chords were recognized correctly in all the time frames tested. Every chord was identified correctly in the first time frame. In the second and third time frame four out of the seven were correct. In the final time frame six out of the seven were correctly recognized. The G major dominant 7th chord was incorrectly identified in the second and third time frame. In the second time frame it was identified as a minor triad with root B. The F# is the third and sixth harmonic of the oboe B and is strong in this time frame. This same error occurred with the diminished triad in the second and third time frames. For the A major dominant 7th chord the identification as a C# minor triad (C#, E, G#) arises because of the presence of a strong G#, a third and sixth harmonic of the C# on the oboe, and the E is already present to complete the triad.

In the second table three out of the eleven chords were correctly recognised in all time frames tested. In the first and second time frames five and six out of eleven were correctly recognised. In the third and fourth time frame ten out of eleven chords were correctly identified. The second D minor chord in table 2 was identified as an F major chord in the first two time frames. This is due to the C, which is present as the third and sixth harmonic of F and occurs in more channels than D, for these two time frames. The third D minor chord in table 2 was identified as a major chord with root note Bb. The Bb arises as a difference periodicity in the unresolved channels. Both D and F are strong periodicities in this time frame. A and Bb occur with equal strength. In the G major dominant 7th chord analysis the F# is in a significant number of channels. The F# is a third and sixth harmonic of B. Some harmonics present may be stronger than the fundamental frequency.

Table 1. Results of chords built from recorded instrument samples

		Starting point (of duration) for analysis			
Chord type and duration (s) of chord	**Instruments**	**1/5**	**2/5**	**3/5**	**4/5**
F major (0.6s)	cello C3, violin F5, oboe A5	Major, root F	Major, root F	Major, root F	Major, root F
G minor (0.5s)	cello G3, violin Bb5, flute D6	Minor, root G	Minor, root G	Minor, root G	Minor, root G
G major dominant 7th (0.5s)	cello G2, violin F5, oboe B4, flute D6	Major, root G dominant 7th	*minor, root B	*undefined	Major, root G dominant 7th
A major dominant 7th (0.9s)	Cello A3, oboe C#4, piano G4, flute E6	Major, root A dominant 7th	Major, root A	Major, root A	*Minor, root C#
Diminished B, D, F (0.9s)	oboe B4, flute D6, violin F5	Diminished B, D, F	*Minor triad, root B	*Minor triad, root B	Diminished B, D, F
Augmented D, F#, Bb (0.9s)	oboe D4, piano F#3, violin Bb5	Augmented D, F#, Bb	Augmented D, F#, Bb	Augmented D, F#, Bb	Augmented D, F#, Bb
Augmented C#, A, F (0.9s)	oboe C#4, piano F4, oboe A5	Augmented C#, F, A	Augmented C#, F, A	Augmented C#, F, A	Augmented C#, F, A

Table 2. Results of chords from the 2nd movement of Bach's Brandenburg Concerto No. 2 in F major

		Starting point (of duration) for analysis			
Chord type and duration (s) of chord	**Bar number**	**1/5**	**2/5**	**3/5**	**4/5**
D minor 0.7s (cello D3, violin A4, oboe F5)	29 (3rd beat)	Minor, root D	Minor, root D	Minor, root D	Minor, root D
D minor 0.5s (cello D3, oboe F4, recorder A5)	7 (3rd beat)	*Major, root F	*Major, root F	Minor, root D	Minor, root D
F major 0.6s (cello C3, violin F5, oboe A5)	35 (1st beat)	*Minor, root C	Major, root F	Major, root F	Major, root F
G minor 0.4s (cello D3, oboe G5, recorder Bb5)	37 (1st beat)	Minor, root G	Minor, root G	Minor, root G	Minor, root G
G minor 0.5s (cello G3, violin Bb5, recorder D6)	37 (3rd beat)	*Major, root C	*Undefined	Minor, root G	Minor, root G
C major dominant 7th 0.3s (cello C3, violin E5, oboe Bb4, recorder G5)	18 (3rd beat)	*Undefined	Major, root C, dominant 7th	Major, root C, dominant 7th	Major, root C, dominant 7th
G major dominant 7th 0.5s (cello G2, violin F5, oboe B4, recorder D6)	22 (3rd beat)	*Minor, root B	*diminished B, D, F	*Minor, root B	Major, root G
A major dominant 7th 0.3s (cello A3, violin C#6, oboe G4, recorder E6)	28 (3rd beat)	Major, root A, dominant 7th	Major, root A, dominant 7th	Major, root A, dominant 7th	Major, root A, dominant 7th
D minor 0.6s (cello D3, violin A4, recorder F5)	57 (3rd beat)	Minor, root D	*Major, root Bb	Minor, root D	Minor, root D
Diminished G#, F, B 0.6s (cello G#2, oboe F4, recorder B5)	8 (1st beat)	Diminished G#, F, B	*undefined	Diminished G#, F, B	Diminished G#, F, B
Augmented A, 0.2s, C#, F (cello A2, violin C#6, oboe A4, recorder F6)	28 (3rd beat)	Major, root A	Augmented A, C#, F	Augmented A, C#, F	Augmented A, C#, F

7. Conclusions

The performance of a computational model of harmonic chord recognition has been demonstrated. The model was evaluated on a set of chords created from instrument samples and on chords from a recording of the second movement of J.S. Bach's Brandenburg Concerto No. 2. The results demonstrate that the model works well on harmonic chords created from different instruments. Results are good for the major, minor and augmented chords but the performance is less reliable in the case of the dominant 7th and diminished chord. The results for the chords from the Bach recording tend to improve towards the middle and end of the sound (the last two time frames). Errors at the start of the chord may be due to transients or some reverberation of the previous harmony in the recording. In the case of chords created from the instrument samples errors did not occur at the start of the chords.

References

Bregman A. S. (1990) 'Auditory Scene analysis: The perceptual organisation of sound.' Cambridge, MA: MIT Press.

Hartmann W. M. (1998) 'Signals, Sound, and Sensation.' Copyright © 1998 Springer-Verlag New York, Inc.

Howard D.M and Angus J. (2001) 'Acoustics and Psychoacoustics.' Second Edition, Music Technology Series, Focal Press.

Meddis R. (1986) 'Simulation of mechanical to neural transduction in the auditory receptor.' J. Acoust. Soc. Am. 79 (3), March 1986

Meddis R. (1988) 'Simulation of auditory-neural transduction: Further studies.' J. Acoust. Soc. Am. 83 (3), March 1988.

Meddis R., Hewitt, M. J. and Shackleton T. M. (1990) 'Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse.' J. Acoust. Soc. Am. 87 (4), April 1990.

Meddis R. and Hewitt M. J. (1991) 'Virtual pitch and phase sensitivity of a computer model of the auditory periphery. 1:Pitch identification.' J. Acoust. Soc. Am. 89(6), June 1991.

Moore B.C.J. (2003) 'An introduction to the Psychology of Hearing.' Fifth Edition, Academic Press Copyright © 2003, Elsevier Science (USA).

Parncutt R. (1989) 'Harmony: A Psychoacoustical Approach.' Copyright © Springer-Verlag 1989

Patterson R. D., Robinson, K., Holdsworth, J., Mckeown, D., Zhang, C. and Allerhand, M. (1992) 'Complex Sounds and Auditory Images.' Auditory Physiology and Perception, editors Y. Cazals, L. Demany, K. Horner. 1992

Roederer J. G. (1995) 'The Physics and Psychophysics of Music, An Introduction.' Third Edition, © 1995 by Springer-Verlag New York, Inc.

Slaney, M. (1993) 'An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank.' Apple Computer Technical Report #35. Perception Group-Advanced Technology Group © 1993, Apple Computer, Inc.

Slaney, M. (1993) ‘Auditory Toolbox: A Matlab toolbox for auditory modeling work version 2’ Technical Report #1998-010. © Apple Computer, Inc. 1993-1994, © Interval Research Corporation 1994-1998.