EE391 Special Report (Spring 2005)
Automatic Chord Recognition Using A Summary Autocorrelation Function

Advisor: Professor Julius Smith

Kyogu Lee
Center for Computer Research in Music and Acoustics (CCRMA)
Music Department, Stanford University
kglee@ccrma.stanford.edu

Abstract

In this paper, a novel approach based on human perception for automatic chord recognition from the raw audio data is proposed. To this end, templates of summary autocorrelation function (SACF) for 24 major/minor chords are first generated from the synthetic tones. The SACF of the input audio is then obtained over time and correlated with the 24 template SACFs to give the ratings for each chord. A chord with the maximum correlation is recognized as the chord of input audio. The results show the proposed method outperforms the traditional methods which use as feature set Chromagram or Pitch Class Profile (PCP).

1 Introduction

A musical chord can be defined as a set of simultaneous tones, and succession of chords over time, or chord progression, forms a core of harmony in a piece of music. Hence analyzing the overall harmonic structure of a musical piece often starts with labeling every chord in it. This is a difficult and tedious task even for experienced listeners with the scores at hand. Automation of chord labeling thus can be very useful for those who want to do harmonic analysis of music. Once the harmonic content of a piece is known, it can be further used for higher-level structural analysis. It also can be a good mid-level representation of musical signals for such applications as music segmentation, music similarity identification, and audio thumbnailing. For these reasons and others, automatic chord recognition has recently attracted a number of researchers in a Music Information Retrieval society.

A chromagram or a Pitch Class Profile has been the choice of the feature set in automatic chord recognition or key extraction since Fujishima introduced it (Fujishima 1999). Perception of musical pitch has two dimensions - *height* and *chroma*. Pitch height moves vertically in octaves telling which octave a note belongs to. On the other hand, chroma tells where it stands in relation to others within an octave. A chromagram or a pitch class profile is a 12-dimensional vector representation of a chroma, which represents the relative intensity in each of twelve semitones in a chromatic scale. Since a chord is composed of a set of tones, and its label is only determined by the position of those tones in a chroma, regardless of their heights, chromagram seems to be an ideal feature to represent a musical chord.

There are some variations to obtain a 12-bin chromagram, but it usually follows the following steps. First, the DFT of the input signal $X(k)$ is computed, and the constant Q transform X_{CQ} is calculated from $X(k)$, which uses a logarithmically spaced frequencies to reflect the frequency resolution of the human ear (Brown 1990). The frequency resolution of the constant Q transform follows that of the equal-tempered scale, and thus the kth spectral component is defined as

$$f_k = (2^{1/B})^k f_{min},$$

where f_k varies from f_{min} to an upper frequency, both of which are set by the user, and B is the number of bins in the constant Q transform. Once $X_{CQ}(k)$ is computed, a chromagram vector CH can be easily obtained as:

$$CH(b) = \sum_{m=0}^{M-1} \left| X_{CQ}(b + mB) \right|,$$

where $b = 1, 2, \cdots, B$ is the chromagram bin index, and M is the number of octaves spanned in the constant Q spectrum. For chord recognition, only $B = 12$ is needed, but

$B = 24$ or $B = 36$ is also used for pre-processing like fine tuning.

The remainder of this paper is organized as follows: Section 2 reviews related work on this field; Section 3 starts by stating the problems in the previous work caused by choosing the chromagram as the feature set, and provides a solution by suggesting a different feature vector called Summary Autocorrelation Function (SACF). In Section 4, the comparison of the two methods with real recording examples as well as a test signal is presented, followed by discussions. In section 5, conclusions are made and directions for future work are suggested.

2 Related Work

A chromagram or a pitch class profile (PCP) based features have been almost exclusively used as a front end to the chord recognition or key extraction systems from the audio recordings. Fujishima developed a realtime chord recognition system, where he derived a 12-dimensional pitch class profile from the DFT of the audio signal, and performed pattern matching using the binary chord type templates (Fujishima 1999). Gomez and Herrera proposed a system that automatically extracts from audio recordings tonal metadata such as chord, key, scale and cadence information (Gomez and Herrera 2004). They used as the feature vector, a Harmonic Pitch Class Profile (HPCP), which is based on Fujishima's PCP, and correlated it with a chord or key model adapted from Krumhansl's cognitive study. Similarly, Pauws used the maximum-key profile correlation algorithm to extract key from the raw audio data, where he averaged the chromagram features over variable-length fragments at various locations, and correlate them with the 24 major/minor key profile vectors derived by Krumhansl and Kessler (Pauws 2004). Harte and Sandler used a 36-bin chromagram to find the tuning value of the input audio using the distribution of peak positions, and then derived a 12-bin, semitone-quantized chromagram to be correlated with the binary chord templates (Harte and Sandler 2005).

Sheh and Ellis proposed a statistical learning method for chord segmentation and recognition, where they used the hidden Markov models (HMMs) trained by the Expectation-Maximization (EM) algorithm, and treated the chord labels as hidden values within the EM framework (Sheh and Ellis 2003). Bello and Pickens also used the HMMs with the EM algorithm, but they incorporated musical knowledge into the models by defining a state transition matrix based on the key distance in a circle of fifths, and by avoiding random initialization of a mean vector and a covariance matrix of observation distribution, which was modeled by a single Gaussian (Bello and Pickens 2005). In addition, in training the model for parameter estimation, they selectively update the parameters of interest on the assumption that a chord template or distribution is almost universal, thus disallowing adjustment of distribution parameters.

In the following section, we state the problems with the chromagram-based approaches in chord/key estimation application, and propose a novel method which uses the autocorrelation function in place of the chromagram.

3 Chord Recognition Using Autocorrelation

All of the aforementioned work on chord recognition or key extraction, while the details of the algorithms may vary, have one thing in common - they all use a chromagram as the feature vector. To identify a chord, some use a template matching algorithm (Fujishima 1999; Gomez and Herrera 2004; Pauws 2004; Harte and Sandler 2005), whereas others use a probabilistic model such as HMMs (Sheh and Ellis 2003; Bello and Pickens 2005), but the front end to the recognition systems is always the 12-dimensional chromagram. This may cause serious problems, particularly when used with the template matching algorithm.

3.1 Problems with Chroma-based Approach

In chord recognition systems with a template matching algorithm, templates of chord profiles are first generated. For example, since a C major triad comprises of three notes C (root), E (third), and G (fifth), the template for a C major triad is [1,0,0,0,1,0,0,1,0,0,0,0], and for a G major triad, it will be [0,0,1,0,1,0,0,0,1,0,0,1], where the template labeling is [C,C#,D,D#,E,F,F#,G,G#,A,A#,B]. As can be seen in these examples, every template in 12 major triads will be just a shifted version of the other, and for the minor triad, it will be the same as the major triad with its third shifted by one to the left; *e.g.*, [1,0,0,1,0,0,0,1,0,0,0,0] is a C minor triad template. Templates for augmented, diminished, or 7th chords can be defined in a similar way. This kind of template matching may cause a lot of confusion to the recognition systems especially for noisy input signals. Furthermore, when a real acoustic instrument play a note, it not only produces a tone at its fundamental frequency, but also produces many partials, some of which have frequencies that are harmonics of another note. Therefore, when three notes are played simultaneously, as in a triad, there will be a number of partials at pitch classes other than those of the notes, and some of them will overlap as well. Hence templates of all-or-none type described above are not suitable to represent chords played by the real world

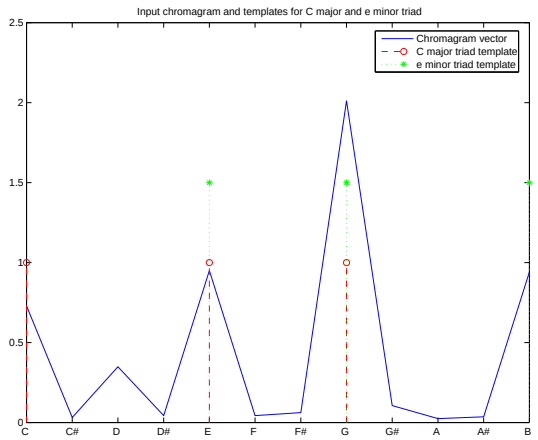

Figure 1: Chromagram of a test C major triad with 2nd inversion and binary templates for C major and E minor triads

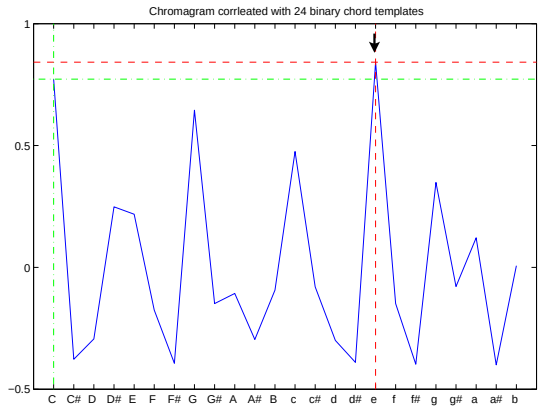

Figure 2: Correlation with 24 major/minor triads using chromagram

instruments. This can be even more problematic particularly when the chord is inverted; *i.e*, when the third note (1st inversion) or the fifth note (2nd inversion) in a triad forms a bass note instead of the root note. Figure 1 shows a snapshot of the chromagram of a C major triad with 2nd inversion, played by a cello (G4), a flute (C5), and a trumpet (e5), which are excerpted from the McGill sample CDs. As can be seen in the figure, even if the instruments are just playing notes in a C major triad, all 12 bins in the chromagram have some energy because of the harmonics they produce. Overlaid are two binary templates for a C major triad (dashed line with circle) and for an E minor triad (dotted line with asterisk). Since 12th bin for pitch class B contains quite a lot of energy due to harmonics, an E minor triad template will have higher correlation with the input chromagram vector than a C major triad template. Figure 2 shows the correlation result with all 24 binary templates.

3.2 Summary Autocorrelation Function

Autocorrelation analysis has long been used to explore the theory of pitch perception in that a periodic signal is found to have peaks at multiples of its period in an autocorrelation function. Licklider proposed a duplex theory of pitch perception where the frequency analysis is done in the cochlea by means of auditory filter banks, and the autocorrelation analysis is performed by the neural part of the system (Licklider 1951). Based on Licklider's model, Meddis and Hewitt proposed another time-domain pitch perception model, where they introduced a Summary Autocorrelation Function (SACF) to explain a number of phenomena seen in pitch perception experiments such as the missing fundamental, ambiguous pitch, pitch shift of equally spaced, inharmonic components, and so on (Meddis and Hewitt 1991).

One of advantages of time-domain pitch perception models such as Meddis and Hewitt's is they can easily explain a periodicity pitch or a missing fundamental. This is a phenomenon that human can perceive a pitch of the fundamental from the upper harmonics even if the fundamental is missing. For example, if there are tones at $400, 600, 800, 1000$ Hz, the perceived pitch is 200 Hz. While no spectral component exists at the fundamental frequency in the Fourier transform, a strong peak is found in the autocorrelation function, the inverse of which corresponds to the fundamental frequency. This perception of a missing fundamental coded by human auditory mechanism plays a very important role in chord perception. This helps us recognizing chords with inversions as the same chords as the one in root position. It is shown in the same example describe above - C major triad with 2nd inversion. Figure 3 illustrates the SACF of the same test chord (solid line).

In Figure 3, the SACF has three peaks at time lags whose inverses correspond to G4 (392 Hz), C5 (523.2 Hz), and E5 (659.2 Hz) as indicated. However, it has strongest peak at 130 Hz corresponding to the root note of the chord (C3).

3.3 Implementation

First, the SACF templates of 24 major/minor triads were generated from the synthetic tones. Following the technique that Krumhansl used in her experiments for quantifying harmonic hierarchies, each triad consisted of 15 sine waves over the entire five-octave range, with 5 components for each of the three chord tones. From these synthetic audio files, the

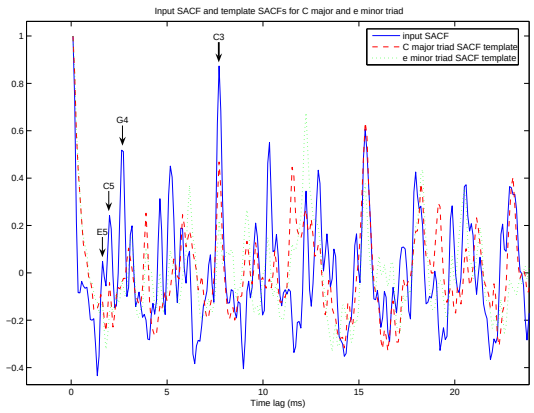

Figure 3: SACF of a test C major triad with 2nd inversion and SACF templates for C major and E minor triads

SACF templates of 24 major/minor triads were computed once and for all. In Figure 3, the SACF template of the C major triad is overlaid in dashed line, and that of the E minor triad in dotted line. For each frame of the input audio, the SACF of the same length and properties as those of the templates was generated, and it is correlated with all of the 24 templates resulting in ratings for 24 triads. The chord with maximum correlation is recognized as the chord in the audio file. Figure 4 shows the correlation result for the same C major triad example with 2nd inversion used above.

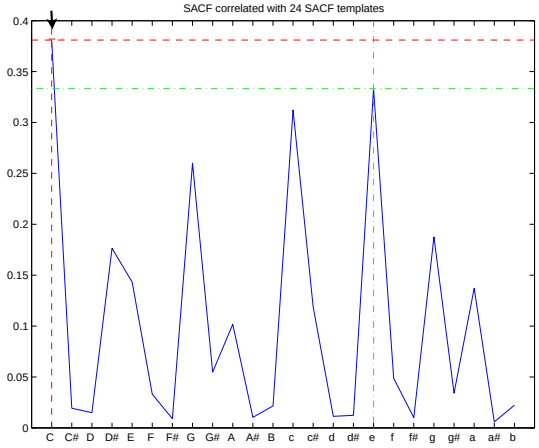

Figure 4: Correlation with 24 major/minor triads using SACF

For evaluation with the real audio recording examples, they were downsampled to 11025Hz, and a Hamming window of length 8192 samples and a hopsize of 1024 samples were used for both the chromagram and the SACF calculation. For the SACF calculation, 16-channel ERB filterbank was used, and the maximum time lag was 1024 samples, which corresponds to the size of the SACF feature vector.

4 Experimental Results and Discussions

Figure 5 compares the results of the frame-level chord recognition from the real recording *Another Crossroads* by Michael Chapman. Dashed line with circles represents identified chords using a pattern matching algorithm using the chromagram as the feature set and the binary chord templates. Without any pre-processing or post-processing, the results look very noisy, and have many recognition errors. On the other hand, chord recognition using the SACF in solid line with x's shows great improvement in performance. Not only there is much less error in chord labeling, but also it detects the chord changes very accurately. Vertical lines represent the ground truth chord boundaries with chord names.

A simple lowpass filtering across a number of frames in the chromagram or in the SACF can improve results since transient and noisy signals such as percussion sounds can obscure harmonic contents of the signal. Figure 6 displays a smoothed version of chord recognition task. This reduces errors quite a lot, but an SACF-based method still outperforms a chromagram-based method.

As illustrated in the above example, the SACF features showed a better performance than the conventional chromagram in a chord recognition task using the template matching algorithm. This improvement in performance can be explained from the perspective of neurobiological study. Tramo *et al.* found, using physiological, psychoacoustic, and neurological methods, that harmonic perception is governed by properties of the auditory system, which include the capacity of peripheral auditory neurons to encode temporal regularities in acoustic fine structure of signals (**?**). This is essentially the same information that an autocorrelation function contains .

5 Conclusions

The Summary Autocorrelation Function was proposed as a feature vector to be used as an front end to the chord recognition system with a pattern matching algorithm. The ACF has long been used for pitch identification or pitch perception, and the Summary ACF has been proven by (Meddis and Hewitt 1991) to explain many phenomena observed in pitch perception experiments. Experimental results show that the SACF outperforms the chromagram in real recording examples. This can be explained partly by

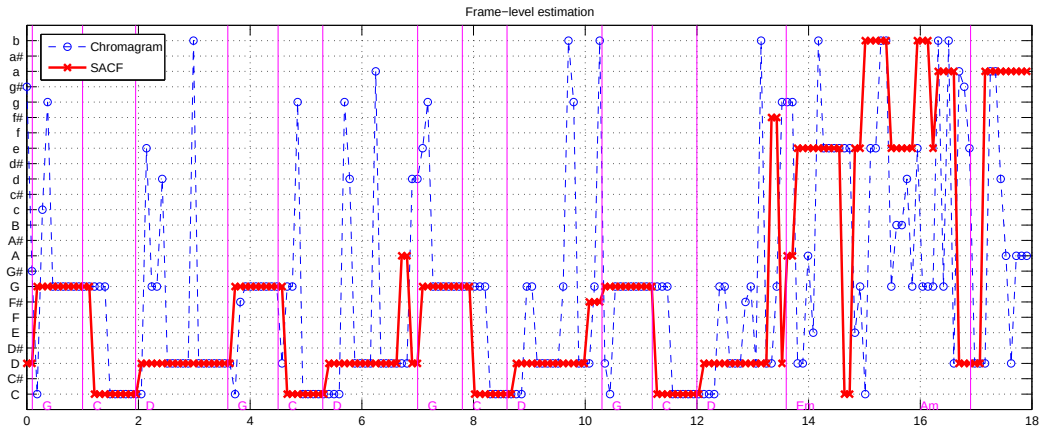

Figure 5: Frame-level recognition with real recording example (*Another Crossroads* by Michael Chapman)

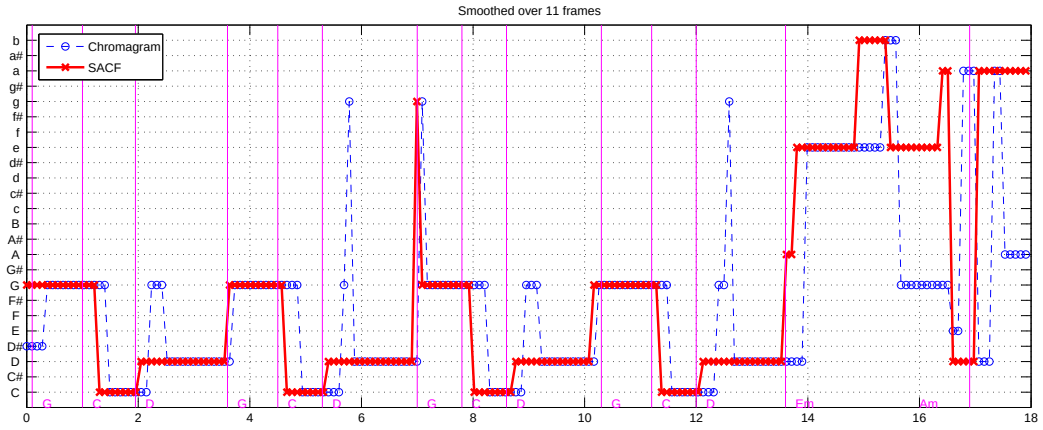

Figure 6: Smoothed over 11 frames with the same example (*Another Crossroads* by Michael Chapman)

a strong peak seen in the SACF at the missing fundamental, which plays an important role in chord perception, especially in major chords. The disadvantage of the SACF is its high dimensionality. In the experiments, the vector size of the SACF was 1024 compared with 12 in the chroma vector. This can be a serious problem if we are to use the SACF in machine learning models such as the HMMs or the SVMs, whose performance can be degraded by far if the feature size is too big. There are techniques to reduce the dimension of vector like the PCA or the SVD, and we may use them for our application. Another drawback with the SACF is that it is computationally expensive. The SACF proposed by (Meddis and Hewitt 1991) used 128 auditory channels. In the present paper, it was reduced down to 16 channels without a big loss in performance. It can be further reduced using a 2-channel model proposed by Tolonen and Karjalainen (Tolonen and Karjalainen 2000) where they used just two channels (below and above 1 kHz) for multipitch estimation. Possible future work may involve improving these two drawbacks.

References

Bello, J. P. and J. Pickens (2005). A robust mid-level representation for harmonic content in music signals. In *Proceedings of the International Symposium on Music Information Retrieval*, London, UK.

Brown, J. C. (1990). Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America 89*(1), 425–434.

Fujishima, T. (1999). Realtime chord recognition of musical sound: A system using Common Lisp Music. In *Proceedings of the International Computer Music Conference*, Beijing. International Computer Music Association.

Gomez, E. and P. Herrera (2004). Automatic extraction of tonal metadata from polyphonic audio recordings. In *Proceedings of the Audio Engineering Society*, London. Audio Engineering Society.

Harte, C. A. and M. B. Sandler (2005). Automatic chord identification using a quantised chromagram. In *Proceedings of the Audio Engineering Society*, Spain. Audio Engineering Society.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia 7*(4), 128–134.

Meddis, R. and M. J. Hewitt (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i:pitch identification. *Journal of the Acoustical Society of America 89*(6), 2866–2882.

Pauws, S. (2004). Musical key extraction from audio. In *Proceedings of the International Symposium on Music Information Retrieval*, Barcelona, Spain.

Sheh, A. and D. P. Ellis (2003). Chord segmentation and recognition using em-trained hidden Markov models. In *Proceedings of the International Symposium on Music Information Retrieval*, Baltimore, MD.

Tolonen, T. and M. Karjalainen (2000). A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing 8*(6), 708–716.