Review of State of The Art of Chord Recognition

Abhishek Kulkarni[1], Utkarsh Gulumkar[2], Onkar Karale[3], Sohan Joshi[4]
Zeal College of Engineering and Research, Pune

Abstract- *In this paper we present a comprehensive review of the two of the recent audio chord recognition systems. This review paper focuses on automatic chord estimation rather than future chord prediction. This paper attempts to provide systematic analysis of two popular and validated techniques. We are integrating theoretical analysis of Audio chord with Deep Chroma extractor development using artificial neural network, and a more traditional estimation approach using Temporal Correlation Support Vector Machine.*

Index Terms- *Chord Estimation, Chroma Features, Deep Chroma Extractor, Pitch Class Profile, Recurrent Neural Network, SVM, Temporal Correlation*

I. INTRODUCTION

Chord recognition has become an important research field after receiving attention from the MIR (music information research) community. It shows promise to become an essential tool of understanding, playing and creating Music for learners, artists and musical instrument makers.

There are three standard processing steps used in chord recognition systems i.e. Feature extraction, pattern matching and filtering. In current state-of-the-art automatic chord recognition systems, machine learning techniques are commonly used to perform feature extraction and pattern matching.

A chromagram/Variants of the pitch class profile (PCP) is a representation of harmonic content by time dependent Chroma vectors, first described by Fujishima (1999). This technique categorizes chords well for pure tones, but has lower accuracy for real world sounds that include spectral noise and frequency band widening, percussion, mistuning, and harmonics or timbre dependency.

Temporal Correlation Support Vector Machine (TCSVM) proposes two step minor major chord recognition. In first step, robust principal component analysis is used to distinguish vocals and chord features in music and enhanced logarithmic pitch class profile (ELPCP) is used to extract low rank component of spectrogram matrix. In second step, improved support vector machine algorithm (TCSVM) is proposed to further derive the temporal correlations among extracted LPCP chord features.

Deep Chroma Extractor method argues that the main issue of noisy features in standard Chroma extractors such as Chromagram can be solved by feeding artificial neural network, contextual audio spectrum instead of single frame input and train the network to selectively compensate noise to remove harmonic ambiguities.

- Abhishek Kulkarni is currently pursuing Bachlor degree program in Compuer engineering in Savitribai Phule Pune University, India, PH-8446985390. E-mail: kulkarnis.abhishek@gmail.com

- Utkarsh Gulumkar is currently pursuing Bachlor degree program in Compuer engineering in Savitribai Phule Pune University, India, PH-8412084615. E-mail: utkarsh.gulumkar@gmail.com

- Onkar Karale is currently pursuing Bachlor degree program in Compuer engineering in Savitribai Phule Pune University, India, PH-9021935966. E-mail: karaleonkar19@gmail.com

- Sohan Joshi is currently pursuing Bachlor degree program in Compuer engineering in Savitribai Phule Pune University, India, PH-9167858583. E-mail: sohanjoshi844@gmail.com

A number of works used neural networks in the context of chord recognition. Humphrey and Bello implemented Convolutional Neural Networks to classify major and minor chords end-to-end. Boulanger-Lewandowski and Sigtia proposed Recurrent Neural Networks as a post-filtering method, where the former used a deep belief net, the latter a deep neural network as underlying feature extractor

II. RELATED WORK

2.1 Temporal Correlation Support Vector Machine:

The system is two-step process: Feature extraction and chord classification as shown in figure 1. In first step, vocal signals from music are separated and beat intervals are calculated followed by extracting a set of vectors for the PCP.

In next step combination of SVM classification and Viterbi Algorithm is used to employ temporal correlation of chords leading to TCSVM.

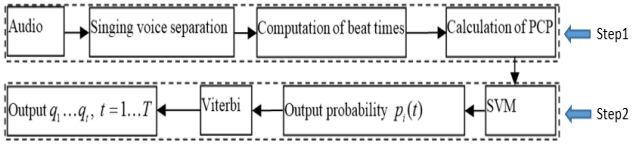


Figure 1. Chord estimation system.

2.1.1 Step 1: Enhanced PCP Feature:

The system extracts feature vector from raw audio as a PCP vector using following steps: (1) using the constant Q transform to calculate the 36-bin chromagram; (2) mapping the spectral chromagram to a particular semitone; (3) median filtering; (4) segmenting the audio signal with a beat-tracking algorithm; (5) reducing the 36-bin chromagram to a 12-bin chromagram based on beat-synchronous beat-synchronous segmentation(using beat tracking algorithm proposed by Ellis); (6) normalizing the 12-bit chromagram.

PCP values are normalized using p-norm and logarithm. After applying the logarithm and normalization, the chromagram is called the LPCP.

$$QPCP_{\log}(p) = \log_{10}[C \cdot QPCP_{12}(p) + 1] \quad (1)$$

$$QPCP_{norm}(p) = QPCP_{\log}(p) / \| QPCP_{\log} \| \quad (2)$$

2.1.2 Enhanced PCP with Vocal signals removal:

The framework of vocal signals separated is shown in Figure 2 by enhanced PCP (EPCP) in two steps. In first step spectrogram of signals in matrix D calculated from STFT. In second step, the inexact augmented Lagrange multiplier (ALM) algorithm is used to solve A+E=|D| where A is low rank matrix of music accompaniment and E is sparse matrix of Vocal signals

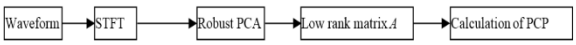


Figure 3. Calculation of enhanced PCP (EPCP).

Input: matrix D, parameter λ
$Y_0^* = D/J(D); E_0 = 0; \mu_0 > 0; \rho > 1; k = 0.$
while not converged **do**
// lines 4–5 solve $A_{k+1} = \underset{A}{\operatorname{argmin}} L(A, E_k, Y_k, \mu_k)$.
$(U, S, V) = svd(D - E_k + \mu_k^{-1} Y_k);$
$A_{k+1} = US_{\mu_k^{-1}}[S]V^T.$
// line 7 solves $E_{k+1} = \underset{E}{\operatorname{argmin}} L(A_{k+1}, E, Y_k, \mu_k)$.
$E_{k+1} = S_{\lambda\mu_k^{-1}}[D - A_{k+1} + \mu_k^{-1} Y_k].$
$Y_{k+1} = Y_k + \mu_k[D - A_{k+1} - E_{k+1}].$
$\mu_{k+1} = \rho\mu_k.$
$k = k + 1.$
end while
Output: (A_k, E_k).

2.1.3 Automatic chord recognition using Support vector machine classification:

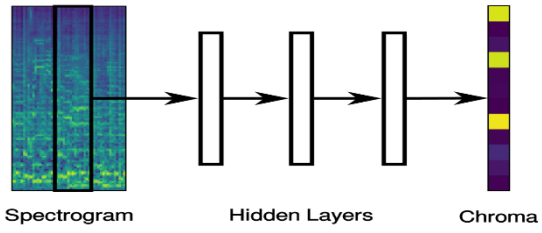
SVM is commonly used machine learning technique for classification, regression and other learning tasks. The Viterbi algorithm uses the probability estimates and trained state transition probability to estimate the chord of the music. Because of the temporal correlation of chords, we combine the SVM classification with the Viterbi algorithm and call the system TCSVM (Temporal Correlation Support Vector Machine).

Algorithm 2: Viterbi algorithm

1 Initialization:
$\delta_t(S_i) = \Pi_i P(Q_1|S_i),\ \psi_t(i) = 0,\ 1 \leqslant i \leqslant K;$
2 Recursion:
$\delta_t(S_i) = \max_{1 \leqslant j \leqslant N} (\delta_{t-1}(S_j) \cdot A(S_j, S_i)) \cdot P(Q_t|S_i), 2 \leqslant t \leqslant T\ \psi_t(i) = \arg[\max_{1 \leqslant j \leqslant N} (\delta_{t-1}(S_j) \cdot A(S_j, S_i))];$
3 Termination:
$q_T^* = \arg\max_{1 \leqslant i \leqslant N} [\delta_t(S_i)],\ P^* = \max_i [\delta_t(S_i)];$
4 Path backtracking: $q_t^* = \psi_{t+1}(q_{t+1}^*)\ t = T-1, T-2 \ldots 1.$

2.2 Deep Chroma Extractor:

The Deep Chroma Extractor replaces the PCP algorithm with a recurrent neural network (RNN). It first decomposes the audio signal into a short-time Fourier transform (STFT), instead of the constant Q-transform used in the PCP algorithm, as STFT has deeper information and the RNN is not prone to the issues that PCP has. The STFT is then fed into a 3 layer, 512 units per layer, RNN, that outputs to 12 output units corresponding to the 12 chroma bins. The RNN is trained on 7 out of 8 Beatles albums in the MIREX database, and compared with other algorithms.



2.2.1 Step 1: Quarter Tone Spectrogram:

The input audio file is converted into a quarter spectrogram, defined as logarithmic frequency ranges binned into 24 bins per octave. The paper takes a STFT of the audio file with sample rate 44100 Hz. This is then fed into a triangular wave filter to convert the linear frequency scale into a logarithmic one. This is then binned into bins corresponding to frequencies between 30Hz and 5500 Hz and 24 bins per octave, resulting in 178 bins total. The bins are then logarithmically compressed once more and averaged over 0.7s intervals, which are finally outputted to the RNN.

2.2.2 Step2: Recurrent Neural Network:

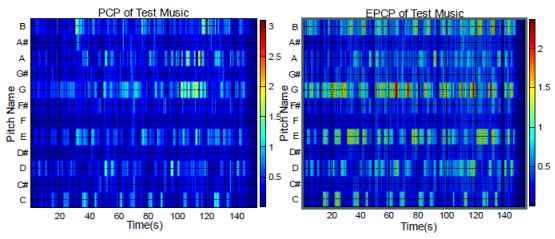
The quarter tone spectrogram is fed into a 3 layer, 512 units each, with 12 unit output layer, Recurrent Neural Network. The network is trained on 7 Beatles albums. It is trained using the cross-entropy loss function between the predicted chroma vector and the target chroma vector, and back propagates the error using the ADAM update rule. The training applies unit dropout with probability 0.5.

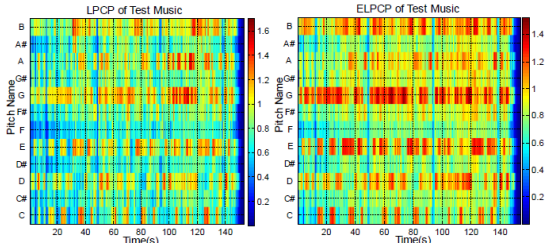
III. OBSERVATIONS

3.1 In this section we review observations. Figure 4 shows the PCP and EPCP of an audio music piece (‘Baby It’s You’ song by Beatles). Figure 5 shows the LPCP and ELPCP of same musical piece. Figure 4 shows that EPCP recognition is more improved than PCP furthermore Figure 5 shows that ELPCP most clear representation of audio music.

3.2 The Deep Chroma Extractor:

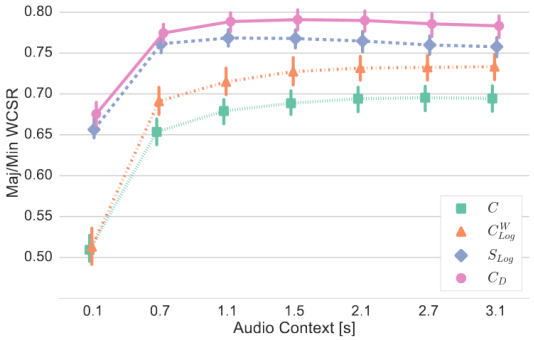
The Deep Chroma Extractor (CD) is compared against three common algorithms: the standard PCP chromogram (C), a chromogram with frequency weights and logarithmic compression of the constant-q (CW Log), and the purely quarter tone spectrogram (SLog). The algorithms are tested using MIREX datasets. The results are shown below:





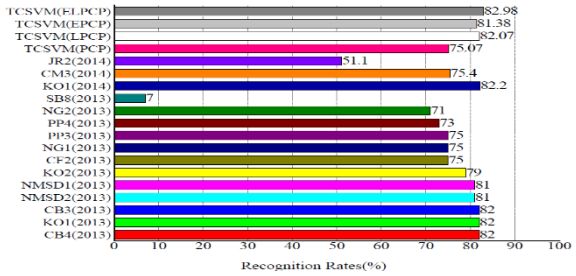
3.2 The Deep Chroma Extractor:

The Deep Chroma Extractor (CD) is compared against three common algorithms: the standard PCP chromogram (C), a chromogram with frequency weights and logarithmic compression of the constant-q (CW Log), and the purely quarter tone spectrogram (SLog). The algorithms are tested using MIREX datasets. The results are shown below:

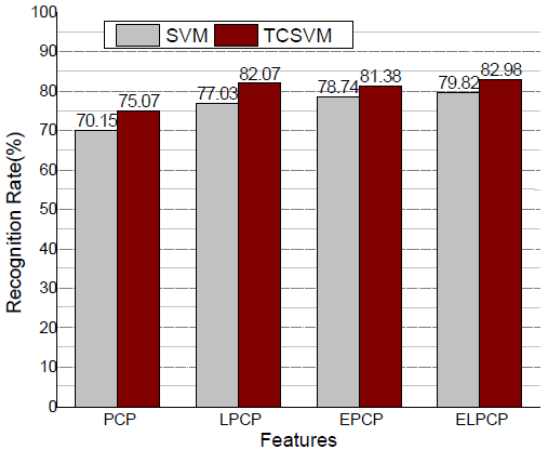


IV. EXPERIMENTAL ANALYSIS AND RESULTS

The results of the chord recognition by TCSVM are evaluated based on MIREX’09, MIREX’13 and MIREX’14 databases.



The recognition results of the TCSVM method with ELPCP are superior, as shown in Figure 8. Because EPCP and ELPCP consider the temporal correlation of music, the rates show few differences between the SVM and TCSVM



	Btls	Iso	RWC	RW	Total
C	71.0±0.1	69.5 ±0.1	67.4±0.2	71.1±0.1	69.2±0.1
C^W_{Log}	76.0±0.1	74.2 ±0.1	70.3±0.3	74.4±0.2	73.0±0.1
S_{Log}	78.0±0.2	76.5 ±0.2	74.4±0.4	77.8±0.4	76.1±0.2
C_D	**80.2**±0.1	**79.3**±0.1	**77.3**±0.1	**80.1**±0.1	**78.8**±0.1

V. CONCLUSION

The original PCP algorithm proposed by Fujishima is still competitive with modern techniques, and modified PCP algorithms like ELPCP have some of the best accuracies of any algorithm. Recurrent Neural Networks have only recently been used in chord recognition algorithms and have become almost as good as modified PCP algorithms, showig 80.2% accuracy on Bealtes data sets compared to 82% for ELPCP.

REFERENCES

[1] Zhongyang Rao, Xin Guan and Jianfu Teng Chord Recognition Based on Temporal Correlation Support Vector Machine, Journal of Applied Sciences, May 2016

[2] F. Korzeniowski, G. Widmer. Feature Learning for Chord Recognition: The Deep Chroma Extractor. 15th International Society for Music Information Retrieval Conference (ISMIR), New York, USA, 2016.

[3] Fujishima, T. Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music. In Proceedings of the International Computer Music Conference, Beijing, China, 22-27 October 1999; pp. 464–467.

[4] Ueda, Y.; Uchiyama, Y.; Nishimoto, T.; Ono, N.; Sagayama, S. HMM-based approach for automatic chord detection using refined acoustic features. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010), Dallas, TX, USA, 14–19 March 2010; pp. 5518–5521.

[5] Harte, C.; Sandler, M. Automatic Chord Identifcation Using a Quantised Chromagram. In Proceedings of the Audio Engineering Society Convention 118, Barcelona, Spain, 28–31 May 2005.

[6] Degani, A.; Dalai, M.; Leonardi, R.; Migliorati, P. Real-time Performance Comparison of Tuning Frequency Estimation Algorithms. In Proceedings of the 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA), Trieste, Italy, 4–6 September 2013; pp. 393–398.

[7] Morman, J.; Rabiner, L. A system for the automatic segmentation and classification of chord sequences. In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 1–10.

[8] Lee, K. Automatic Chord Recognition from Audio Using Enhanced Pitch Class Profile. In Proceedings of the International Computer Music Conference, New Orleans, LA, USA, 6–11 November 2006.

[9] Varewyck, M.; Pauwels, J.; Martens, J.-P. A novel chroma representation of polyphonic music based on multiple pitch tracking techniques. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, BC, Canada, 26–31 October 2008; pp. 667–670.

[10] Müller, M.; Ewert, S. Towards timbre-invariant audio features for harmony-based music. IEEE Trans. Audio Speech Lang. Process. 2010, 18, 649–662.

[11] Nwe, T.L.; Shenoy, A.; Wang, Y. Singing voice detection in popular music. In Proceedings of the 12th Annual ACM International Conference on Multimedia, New York, NY, USA, 10–16 October 2004; pp. 324–327.

[12] Oudre, L.; Grenier, Y.; Févotte, C. Template-based Chord Recognition: Influence of the Chord Types. In Proceedings of the Internation