UNSUPERVISED AUDIO KEY AND CHORD RECOGNITION

Yun-Sheng Wang
Department of Computer Science,
George Mason University
Fairfax, USA
ywange@gmu.edu

Harry Wechsler
Department of Computer Science,
George Mason University
Fairfax, USA
wechsler@gmu.edu

ABSTRACT

This paper presents a new methodology for determining chords of a music piece without using training data. Specifically, we introduce: 1) a wavelet-based audio denoising component to enhance a chroma-based feature extraction framework, 2) an unsupervised key recognition component to extract a bag of local keys, 3) a chord recognizer using estimated local keys to adjust the chromagram based on a set of well-known tonal profiles to recognize chords on a frame-by-frame basis. We aim to recognize 5 classes of chords (major, minor, diminished, augmented, suspended) and 1 N (no chord or silence). We demonstrate the performance of the proposed approach using 175 Beatles' songs which we achieved 75% in F-measure for estimating a bag of local keys and at least 68.2% accuracy on chords without discarding any audio segments or the use of other musical elements. The experimental results also show that the wavelet-based denoiser improves the chord recognition rate by approximately 4% over that of other chroma features.

1. INTRODUCTION

The ability to extract local keys and chords from audio signals is an important step toward music transcription and segmentation using machines. Transcription of music typically requires the understanding of scale degrees used in a music piece as well as the analysis of harmony which correspond nicely to local keys and chords, respectively. On the other hand, music segmentation is the process of partitioning the target music signals into multiple sections so that each section is homogeneous within its boundary but distinct from its neighboring sections; it usually serves as an intermediate step to solve a larger problem such as content-based information retrieval. In [1], six types of segmentation cues - cadence patterns, key schemes, text, instrumentation, rhythm, and harmony - were discussed; using extracted local keys and chords, a multi-dimensional harmonic rhythm can be constructed for segmenting rock or popular music. However, the interpretation of keys and chords are often subjective [2]. For keys, the presence and exact locations of key modulations are often interpreted differently by musicians. For chords, when power chords are played, are they major or minor triads? Should a chord be extended to the 7^{th}? Due to such uncertainties, in this paper, we present a probabilistic approach to estimate, from audio signals, a "bag of local" (BOL) keys and use the extracted keys to recognize chords. Our previous work [3] adopted similar unsupervised approach in estimating keys and chords of symbolic music (MIDI); in this paper, we extend our previous approach to wave audio signals.

2. RELATED WORK

In this section, we review recent work that extracts keys and chords simultaneously from wave audio signals with concentration on those that utilized the unsupervised approach. To analyze keys or chords from audio signals, the most common front end is to transform sound waves into the frequency domain which is subsequently mapped into a chromagram to represent the energy level of the 12 pitch classes, pioneered by [4].

Most recent unsupervised estimation of local keys and chords uses a probabilistic framework [5, 6, 7] by modeling the acoustic likelihood p(X|K,C) to find the best K and C using dynamic programming search technique in 24 keys × 48 chords space. Specifically in [7], a key-chord model and state transition probabilities comprising three sub models (duration, key, and chord) was proposed using the same search space in [5]. Cosine similarity was computed between key template, proposed by [8] enhanced from the pioneering profiles [9], and observed data. The chord model determines the likelihood of observation given a chord being played. The best key-chord sequence is determined by search using the Viterbi algorithm as proposed in [6]. In [10], a probabilistic framework was also used, where the overall chord probabilities were estimated directly from the music piece using the EM algorithm. The likelihood of each chroma frame given chord templates was modeled as a mixture where the estimated overall chord probabilities are the mixing proportion. The likelihood function is treated as the similarity measure between the chroma vector and chord templates. They achieved 71% overlap accuracy for 3 types of chords (maj, min, and 7).

The majority of recent supervised approach involves Hidden Markov Models (HMM) which requires labeled training data and is capable of incorporating other facet of musical elements such as beats or bass line information. In [11], constant tempo, 4/4 time beat pattern, and one global key were assumed; bass pitches from melody lines were incorporated into a probabilistic-based key/chord recognition system. Chroma features were modeled using Gaussian mixture model (GMM) whose parameters were estimated using the EM algorithm and number of Gaussian components were preset to 1, 2 4, 8, and 16. For 150 Beatles' songs, they achieved 73.7% recognition rate for two classes of chords (maj & min). In [12], a six-layer dynamic Bayesian network was used to simultaneously estimate chord sequence, bass notes, metric positions of chords and keys in four layers while the other two observed layers model low-level bass and treble audio features. They achieved 71% accuracy on 176 audio tracks from the MIREX 2008 Chord Detection Task.