MIREX 2013: LARGE VOCABULARY CHORD RECOGNITION SYSTEM USING MULTI-BAND FEATURES AND A MULTI-STREAM HMM

Taemin Cho, Juan P. Bello
Music and Audio Research Lab (MARL)
Music Technology, New York University
New York, USA

ABSTRACT

This paper describes the submitted systems to the MIREX 2013: Audio Chord Estimation task.

1. INTRODUCTION

The submitted systems are:

- CB1 - Pre-trained : This model is trained with 495 songs consisting of 100 RWC pop songs, 195 us-pop songs, 200 songs from The Beatles and Queen songs.
- CB2 - for train-test evaluation

These systems can recognize 157 chord types, one of the largest chord vocabularies among state-of-the-art systems. This paper includes a new feature extraction and a new modeling technique used in the submitted systems. Since these approaches are based on a simple architecture, they have the advantage of a higher processing speed than existing systems.

2. FEATURE VECTOR MODIFICATION

Although 12-dimensional chroma features are yet the most powerful representation of music for chord recognition, particularly for detecting major and minor triads, they may be insufficient for discriminating a large number of chords including more complex chord types. Since choma features contain information about only the average energies of 12 pitch classes, the sharing of notes causes significant overlapping of chords in chroma space. This overlapping causes confusion between chords, and thus recognition performance decreases as the number of chords in a lexicon increases. In this paper, a novel variation of chroma is proposed, which can alleviate the overlap problem by providing more information about pitch content in a signal.

The basic idea of this variation is to split the whole frequency band into multiple sub-bands and then to generate individual chromagrams for each. To do this, a pitch spectrum $P(p)$ is obtained using constant-Q transform, where p is the MIDI note number. The proposed approach, then, starts by splitting $P(p)$ into K number of sub-bands using a set of Gaussian filters defined as:

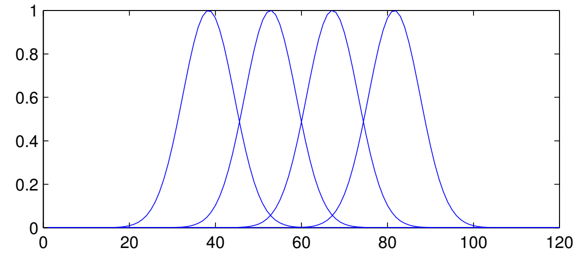

Figure 1. Filter bank where $K = 4$

$$W_k(p) = \exp\left(-\frac{(p-\mu_k)^2}{2\cdot\sigma_K^2}\right) \tag{1}$$

where $k \in [1, K]$ is the sub-band index, μ_k is the center of the k^{th} sub-band, and σ_K is the standard deviation that controls the width of the filter. μ_k and σ_K are defined as follows:

$$\mu_k = 72\cdot\left(\frac{k}{K+1} - 0.5\right) + 60 \tag{2}$$

$$\sigma_K = \frac{15}{(K+1)/2} \tag{3}$$

Such that, a six octave pitch range (72 semitones) centered around C4 ($p = 60$) is equally divided into K sub-bands with half-overlap. The resulting filter-bank, for $K = 4$, is shown in Figure 1.

The k^{th} sub-band chroma vector $C^{(k)}(c)$ with indexes $c \in [0, 11]$ is calculated as:

$$C^{(k)}(c) = \sum_{p\,:\,p\equiv c \pmod{12}} W_k(p)\cdot P(p) \tag{4}$$

All sub-band chroma vectors are normalized to have unit norm.

3. COMBINATION OF SUB-BAND CHORD MODELS

The general idea of the method presented in this section is to build independent chord models for each sub-band chromagram (i.e. $C^{(k)}$), and then to combine them to make a global recognition decision. In this system, the streams of sub-bands are assumed to be conditionally independent of each other and thus modeled by separate HMMs. These HMMs are then combined into a single larger HMM, a multi-stream HMM.

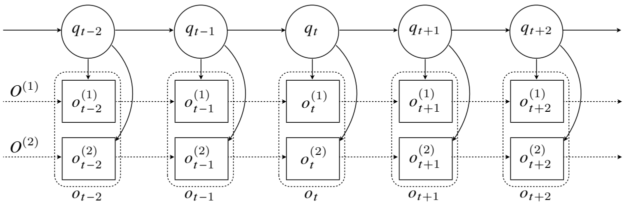

Figure 2. A multi-stream HMM with 2 streams.

An example of a multi-stream HMM is shown in Figure 2. In this figure, the two observation streams, denoted as $O^{(1)}$ and $O^{(2)}$, are the outputs of a single state sequence. More generally, in a K-stream HMM (i.e., a multi-stream HMM with K streams), the i^{th} state consists of K independent emission probability distributions each of which corresponds to the k^{th} observation stream. Let $b_i^{(k)}$ be the pdfs of the i^{th} state for the k^{th} stream, and $o_t = \left\{o_t^{(1)}, \ldots, o_t^{(K)}\right\}$ is an observation at time t, where $o_t^{(k)}$ be the observation sub-vector associated with the k^{th} stream. Under the assumption of synchronicity and independence, the observation probability of the i^{th} state for o_t can be defined as:

$$b_i(o_t) = \prod_{k=1}^{K} \left[b_i^{(k)} \left(o_t^{(k)} \right) \right]^{w_{ik}} \tag{5}$$

where w_{ik} is a weight factor reflecting the reliability of the respective streams, and is restricted to $\sum_{k=1}^{K} w_{ik} = 1$ and $0 \leq w_{ik} \leq 1$. Although (5) is not a true probability distribution, it is preferred in many implementations due to its simplicity, including the well-known HTK[1] [2]. A more sophisticated definition of $b_i(o_t)$ is introduced in [1]. However, in this definition, the complexity of $b_i(o_t)$ grows exponentially with both the number of Gaussian components and the number of streams. For these reasons, this chapter uses $b_i(o_t)$ defined in (5) where the computational time increases linearly with K. w_{ik} is set to $1/K$, making all streams have the same importance. A K-stream HMM is then decoded using the standard Viterbi algorithm.

4. SYSTEM SPECIFICATIONS

The submitted systems detect 157 chord types consisting of (maj, min, maj7, min7, 7, maj6, min6, dim, aug, sus4, sus2, hdim7 and dim7) × 12 keys plus no-chord. This lexicon includes all chord types which account for more than 0.1 % of all chords in our data collection. Extremely rare chords, such as the minor-major-7th chord (e.g., CmM7), are thus omitted. This lexicon does not cover all possible chords, omitting extended (tension) chords (e.g., E7$^{\sharp 9}$) and chord inversions (or slash chords, e.g., Cmaj7/E). In these cases, an extended chord is interpreted as its basis triad or tetrad without any tension notes (e.g., E7$^{\sharp 9}$ is interpreted as E7), and a chord inversion (slash chord) is mapped to its root position chord by simply removing the bass indicated after the slash (e.g., Cmaj7/E is considered as Cmaj7).

The submitted systems use $K = 4$, and multivariate GMMs (5 Gaussians with full covariance matrices) are used for chord models. For training, in order to compensate for the limited amount of training data, training samples are transposed so that their root note is always C. In other words, each k^{th} band chromagram (i.e., $C^{(k)}$) is transposed to the C-root for training. Then, 14 C-based chord models are trained using the transposed features, and the trained models are re-transposed to other roots to complete a set of full chord models.

[1] http://htk.eng.cam.ac.uk

5. REFERENCES

[1] MJF Gales and SS Airey. Product of gaussians for speech recognition. *Computer Speech & Language*, 20(1):22–40, 2006.

[2] S.J. Young, G. Evermann, MJF Gales, D. Kershaw, G. Moore, JJ Odell, DG Ollason, D. Povey, V. Valtchev, and PC Woodland. *The HTK book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.