

# Chord Estimation from Audio using Hidden Markov Models

**Roy Ben Haroosh**

Department of Computer Science  
Your Institution

`roy.benharoosh@post.runi.ac.il`

**Gal Davidi**

Department of Computer Science  
Your Institution

`gal.davididi@post.runi.ac.il`

March 20, 2025

## Abstract

Automatic chord recognition is a significant task within Music Information Retrieval (MIR). Hidden Markov Models (HMMs) using beat-synchronous chroma features have proven effective for chord estimation. This report explores the use of Hidden Markov Models (HMM) and investigates pre-trained audio embeddings, such as L3-Net, to enhance acoustic modeling for chord recognition.

## 1 Introduction

Automatic chord recognition aims to identify and segment chords within audio recordings automatically. Since 2008, it has been a task in the Music Information Retrieval Evaluation eXchange (MIREX), achieving notable success. Hidden Markov Models (HMM) coupled with chroma vectors have established themselves as effective methods. This work evaluates extensions to classical HMM methods by integrating neural network-based acoustic models and leveraging pre-trained embeddings model.

## 2 Related Works

Previous studies have shown HMMs combined with Expectation-Maximization (EM) as robust techniques for chord recognition. Notably, studies like *Chord Segmentation and Recognition using EM-Trained HMM* have highlighted the efficacy of such models. Recent advancements leverage deep neural networks and

deep embeddings, such as Deep Belief Networks (DBNs), OpenL3 embeddings, and VGGish features, significantly improving acoustic modeling quality.

### 3 Architecture

Our proposed architecture leverages an NN/HMM hybrid approach:

1. **Chroma Extraction:** Extract beat-synchronous chroma vectors using Short-Time Fourier Transform (STFT).
2. **OpenL3 Embedding:** Extract OpenL3 embeddings using L3-Net model
3. **Temporal Model:** Hidden Markov Model integrates temporal dependencies between chord progressions.

The probability of an observed sequence  $O$  given a state sequence  $Q$  in an HMM is defined by:

$$P(O|Q, \lambda) = \prod_{t=1}^T P(o_t|q_t, \lambda) \quad (1)$$

The forward algorithm computes the probability of the partial observation sequence up to time  $t$  as:

$$\alpha_t(i) = \left[ \sum_{j=1}^N \alpha_{t-1}(j) a_{ji} \right] b_i(o_t) \quad (2)$$

where  $a_{ji}$  represents transition probabilities and  $b_i(o_t)$  emission probabilities.

### 4 Experiments

We evaluated various configurations of acoustic models:

- Baseline: Classical HMM with chroma features.
- Data improvement: removing percussive sounds from the audio signal.
- Inference improvement: polling (majority vote).
- Pre-trained embeddings from VGGish and L3net as input features.

#### 4.1 Baseline

The training process involves several key steps. Initially, labeled chord annotation files and corresponding audio files are loaded, from which beat-synchronous chromagrams are extracted. These chroma vectors represent pitch class intensities aligned with the 12-tone Western dominant scale. Subsequently, we calculate the mean chroma vector and covariance matrix for each chord state, defining

the emission probabilities of the Hidden Markov Model (HMM). Formally, the emission probability for an observation  $o_t$  given a state  $q_t$  is modeled as:

$$b_{q_t}(o_t) = \mathcal{N}(o_t; \mu_{q_t}, \Sigma_{q_t}) \quad (3)$$

where  $\mu_{q_t}$  is the mean vector and  $\Sigma_{q_t}$  the covariance matrix for the chord state  $q_t$ . The transition probabilities between chords,  $a_{ij}$ , indicating the probability of transitioning from chord state  $i$  to state  $j$ , are computed empirically from the training set and arranged in the transition matrix  $A$ :

$$a_{ij} = P(q_{t+1} = j | q_t = i) \quad (4)$$

The likelihood of an observed chroma sequence  $O = \{o_1, o_2, \dots, o_T\}$  given a state sequence  $Q = \{q_1, q_2, \dots, q_T\}$  is calculated by:

$$P(O|Q, \lambda) = \prod_{t=1}^T b_{q_t}(o_t) \quad (5)$$

Finally, we optimize the model parameters using the Expectation-Maximization (EM) algorithm, iteratively maximizing the likelihood of the observed sequences. This involves computing forward probabilities defined recursively as:

$$\alpha_t(i) = \left[ \sum_{j=1}^N \alpha_{t-1}(j) a_{ji} \right] b_i(o_t) \quad (6)$$

These steps collectively refine the HMM to achieve accurate chord recognition.

## 4.2 Input data improvement

In our experiments, we investigated the impact of removing percussive sounds from the audio signal. To accomplish this, we utilized the Librosa library, specifically applying the Harmonic-Percussive Source Separation (HPSS) algorithm. The HPSS method decomposes the audio signal into harmonic and percussive components through the following process: Short-Time Fourier Transform (STFT)  $\rightarrow$  HPSS  $\rightarrow$  Inverse Short-Time Fourier Transform (ISTFT). Formally, we applied the HPSS function provided by Librosa as follows:

```
y_harmonic, y_percussive = librosa.effects.hpss(y, sr=sr)
```

where  $y$  represents the original audio signal and  $sr$  is the sampling rate. After separation, we utilized only the harmonic component to extract chroma features and subsequently train our model. Experimental results demonstrated a slight improvement in chord recognition accuracy using this preprocessing step.

## 4.3 Inference improvement

We also explored the impact of incorporating polling into the HMM inference process. Polling involves aggregating multiple predictions to enhance the final

chord recognition decision. we used the same model to predict the chord for each beat in the song, and then we used the majority vote to determine the final chord for that beat.

## 4.4 Pre-trained embeddings

we tried to use the OpenL3 (L3-Net) embeddings as input features to the HMM.

# 5 Results

Preliminary results indicate:

- Classical HMM with chroma achieved accuracy around 25% F1-score around 29%. Precision around 38% and recall around 25%.
- Data improvement slightly improved accuracy, reaching 27% F1-score around 30%. Precision around 39% and recall around 26%.
- Polling slightly improved accuracy, reaching 36% F1-score around 37%.
- Incorporation of pre-trained embeddings (OpenL3) further increased accuracy to approximately Z%.

Detailed results are available in the provided notebook and will be continuously updated.

# 6 Training Data

The quality and composition of training data play a crucial role in the performance of our chord recognition system.

## 6.1 Dataset Overview

Our model was trained on the Beatles dataset, a widely used benchmark in Music Information Retrieval (MIR) research. This dataset consists of 50 songs by The Beatles with expert-annotated chord labels aligned with the audio recordings. The annotations provide precise timing information for each chord change throughout the songs.

## 6.2 Data Characteristics

The training data exhibits the following key characteristics:

- **Size:** 50 songs with approximately
- **Features:** Beat-synchronous chromagrams (12-dimensional vectors representing the intensity of each pitch class)

- **Chord vocabulary:** 24 chord types (major and minor triads in all 12 keys)
- **Musical diversity:** Spans the Beatles' career from 1962-1970, covering various musical styles and harmonic complexities

### 6.3 Preprocessing Steps

Before training, we applied several preprocessing techniques:

- **Beat tracking:** Audio was segmented according to beat positions to create beat-synchronous features
- **Chroma extraction:** 12-dimensional chroma vectors were computed for each beat segment using librosa
- **Normalization:** Chroma features were normalized to ensure consistent scaling across different audio segments
- **Embedding extraction:** For the pre-trained embeddings, we used the L3-Net model.

### 6.4 Data Splitting

The dataset was divided using a standard 80%-20% split for training and testing. This approach ensures that our model evaluation reflects its performance on unseen data while maintaining the statistical properties of the original dataset. We ensured that songs from the same album were kept in the same split to avoid data leakage.

## 7 Conclusion

Our findings demonstrate the effectiveness of integrating neural network acoustic modeling and pre-trained embeddings with traditional HMMs for chord recognition tasks. Further experimentation and hyperparameter tuning may yield additional improvements.

## Code and Media

The code for this project is available at: [https://github.com/caiomiyashiro/music\\_and\\_science/tree/master/Chord%20Recognition](https://github.com/caiomiyashiro/music_and_science/tree/master/Chord%20Recognition)

Media and additional results can be accessed here: [https://drive.google.com/drive/folders/1YmfEPtX\\_QLlpo0sR0kQwFV4-Lz6Sd01W](https://drive.google.com/drive/folders/1YmfEPtX_QLlpo0sR0kQwFV4-Lz6Sd01W)