# MINIMUM CLASSIFICATION ERROR TRAINING TO IMPROVE ISOLATED CHORD RECOGNITION

**J.T. Reed[1], Yushi Ueda[2], S. Siniscalchi[3],Yuki Uchiyama[2], Shigeki Sagayama[2], C.-H. Lee[1]**
[1]School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332
{jreed,chl}@ece.gatech.edu
[2]Graduate School of Information Science and Technology
The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-8656 Japan
{ueda,uchiyama,sagayama}@hil.t.u-tokyo.ac.jp
[3]Department of Electronics and Telecommunications
Norwegian University of Science and Technology, Trondheim, Norway
marco77@iet.ntnu.no

## ABSTRACT

Audio chord detection is the combination of two separate tasks: recognizing what chords are played and determining when chords are played. Most current audio chord detection algorithms use hidden Markov model (HMM) classifiers because of the task similarity with automatic speech recognition. For most speech recognition algorithms, the performance is measured by word error rate; i.e., only the identity of recognized segments is considered because word boundaries in continuous speech are often ambiguous. In contrast, audio chord detection performance is typically measured in terms of frame error rate, which considers both timing and classification. This paper treats these two tasks separately and focuses on the first problem; i.e., classifying the correct chords given boundary information. The best performing chroma/HMM chord detection algorithm, as measured in the 2008 MIREX Audio Chord Detection Contest, is used as the baseline in this paper. Further improvements are made to reduce feature correlation, account for differences in tuning, and incorporate minimum classification error (MCE) training in obtaining chord HMMs. Experiments demonstrate that classification rates can be improved with tuning compensation and MCE discriminative training.

## 1. INTRODUCTION

As online music databases continue to grow in size, more effective retrieval mechanisms are needed. In particular, recognizing certain musicological, acoustical, and cultural factors in a musical piece impact notions of similarity. One such musicological factor which has seen an increased research focus is automatic chord detection, which is a mid-level representation and a first-step in identifying the harmony of a given musical work.

Most recent approaches to identifying chords from the acoustic signal are based on using chroma features as inputs into a hidden Markov model (HMM) based system. An early approach in literature using an HMM-based system was [1], where an ergodic HMM provides the initial chord progression modeling and updated using *N*-best rescoring techniques. Sheh and Ellis [2] deal with an inadequate amount of training data by assuming that chroma vectors from the same mode (e.g., *Major*) and different pitch classes can be considered as rotated versions of one another. Bello *et al*. [3] incorporate musical knowledge into the transition probabilities and HMM parameters to improve the results. Lee and Slaney [4] increase the amount of training data available by synthesizing audio to provide accurate chord and boundary information. Improvement is made in [5] by using key-dependent ergodic HMMs and warping the chroma features into tonal centroid features [6], which gives the relation of the chroma features on the circles of fifths, minor thirds, and major thirds.

The HMM framework is inspired by automatic speech recognition (ASR), where HMMs represent words or subword units. However, the ASR community only considers the recognition rate (i.e., what was said) as important and ignores timing information (i.e., recognizing when each spoken unit begins and ends). In contrast, audio chord detection is measured in terms of frame error rate (FER), which incorporates both tasks. This paper proposes optimizing these two task separately and focuses on the problem of classification rate (i.e., recognizing what was said). To the authors' knowledge, only [2] evaluates these tasks separately. Specifically, Sheh and Ellis consider forced alignment, where the correct chord sequence is known and the timing information is extracted.

This paper implements several improvements to evaluate the limits of the chroma/HMM system for classification rates. In particular, the goal of this paper is to improve

the classification rate between highly confused chords in the feature space. For instance, one of the most common chord detection errors is between parallel modes, which share the same root, but differ in their key signature; e.g, *C Major* versus *C minor*. The cause for confusion is because the difference between a *Major* and *minor* chord is a flattening of the major third to a minor third, which may only be a few hertz.

The baseline system adopted in this study is the current state-of-the-art and placed first in the 2008 MIREX Audio Chord Detection task (Task 2: no pre-training) [7]. To attenuate percussive sounds, the harmonic-percussive source separation (HPSS) algorithm [8] is used in the baseline system to isolate the harmonic part of the spectrum prior to chroma extraction and maximum likelihood (ML) estimation. This paper incorporates the improvements of automatic tuning compensation and minimum classification error (MCE) training [9].

The automatic tuning algorithm is a simplification of the one proposed in [3]. Small, uniform databases, such as the Beatles Chord Database [10], experience improved performance with tuning normalization because slight differences in tuning cause confusion between highly competing chords. This can lead to confusion among parallel modes since they differ by a single note, for example. Due to the trade-off between spectral and time-based resolution in frame-based music processing, a slight difference in tuning could allow for energy to bleed into neighboring energy bands, leading to confusion.

MCE, a highly successful discriminative training approach, enhances ASR performance by overcoming two assumptions made by parametric approaches. Like speech, the assumption that the true distribution of chroma vectors is an HMM is an approximation to yield a parametric fit. ML techniques estimate parameters corresponding to the mode of the likelihood function. However, if the true distribution differs from the assumed model, the ML technique is not guaranteed to yield an optimal performance. In addition, the strength of the parametric fit relies on accurate parameter estimates. However, current acoustic chord databases are quite small in size and contain around 100 songs from one to five artists. With such small artist diversity, it is unlikely that current databases are a good representation of the entire acoustic space. MCE integrates a discriminative training approach into the parameter estimation problem by directly optimizing the performance classification; i.e., classification error. Specifically, a logistic transform incorporates the classification error rate into the objective function so that gradient probabilistic descent will yield an improved set of parameters.

The baseline algorithm, is described in Section 2 and improvements are detailed in Section 3. Experimental results in Section 4 compare modifications to the ML baseline. Finally, Section 5 gives concluding remarks.

## 2. BASELINE ALGORITHM: MIREX 2008 SUBMISSION

### 2.1 Harmonic/Percussion Source Separation

As noted in [11], transients and noise decrease the chord recognition accuracy in chroma-based approaches. This is largely due to percussive sources, which spread energy across the entire frequency spectrum. While the authors in [11] use a median filter to smooth percussive effects, this paper uses the HPSS algorithm [8], which integrates the harmonic and percussive separation into the objective function

$$J(\mathbf{H},\mathbf{P}) = \frac{1}{2\sigma_H^2}\sum_{k,n}(H_{k,n-1}-H_{k,n})^2 + \frac{1}{2\sigma_P^2}\sum_{k,n}(P_{k-1,n}-P_{k,n})^2 \quad (1)$$

where $H_{k,n}$ and $P_{k,n}$ are the values of the power spectrum at frequency index $k$ and time index $n$ for the harmonic spectrum, $\mathbf{H}$, and the percussive spectrum, $\mathbf{P}$, respectively. The parameters $\sigma_P^2$ and $\sigma_H^2$ need to be set experimentally. To ensure that each time-frequency component of the harmonic and percussive spectrum components sum to a value equal to the original spectrum, $W_{k,n}$, and to ensure that power spectrums remain positive, the following constraints are added to the minimization of (1)

$$H_{k,n} + P_{k,n} = W_{k,n} \quad (2)$$

$$H_{k,n} \geq 0 \quad (3)$$

$$P_{k,n} \geq 0 \quad (4)$$

Note that minimizing (1) is equivalent to maximum likelihood estimation under the assumption that $(H_{k,n-1} - H_{k,n})$ and $(P_{k-1,n} - P_{k,n})$ are independent Gaussian distributed variables. This simplification leads to a set of iterative update equations for the harmonic and percussive spectrums. At the output of HPSS are two waveforms; one of these contains a percussive-dominated spectrum and the other a harmonic-dominated spectrum. Further details can be found in [8].

### 2.2 Chromagram

Chroma vectors are the most common features in audio chord detection algorithms and describe the energy distribution among the 12 chromas; i.e., pitch classes. To derive chroma vectors, the harmonic-emphasized music signal is first downsampled to 11025 Hz. Next, the signal is broken into frames of 2048 samples with a 50% overlap. The constant Q transform [12] provides spectral analysis using a logarithmic spacing of the frequency domain, whereas the traditional discrete Fourier transform (DFT) uses a linear spacing of the frequency domain. The resulting spectrum, $S$, of the audio signal $s(t)$ is given by

$$S(k) = \sum_{t=0}^{T(k)-1} w(t,k)s(t)e^{-j2\pi f_k t} \quad (5)$$

where the analysis window, $w(t,k)$, and the window size, $T(k)$, are functions of the frequency bin index, $k$. The center frequency of the $k$-th bin is designed to match the equal-temperament scale [13]. For example, if it is desired to have one bin per note on an 88-piano keyboard, then the bin center frequencies are

$$f_l = 2^{l/\beta} f_{\text{ref}} \tag{6}$$

with the number of bins per octave, $\beta$, set to 12, the minimum reference frequency, $f_{\text{ref}}$, set to the frequency of *A0* (i.e., 27.5 Hz), and $l = \{1, 2, ..., 88\}$. The resulting chroma vector for frame $n$ is

$$c_n(b) = \sum_{r=0}^{R} |S(b + r\beta)| \tag{7}$$

where $b = \{1, 2, ..., \beta\}$ is the chroma bin number and $R$ is the number of octaves considered.

### 2.3 HMM classifier

The optimal chord sequence, $W^*$ is decoded such that [14]

$$\begin{aligned} W^* &= \arg\max_W P(W|C) \\ &= \arg\max_W \frac{P(C|W)P(W)}{P(C)} \\ &\propto \arg\max_W P(C|W)P(W) \end{aligned} \tag{8}$$

where $C = \{c_1, c_2, \ldots, c_N\}$ is the sequence of chroma vectors. The probabilities of the acoustic model and tonality model are $P(C|W)$ and $P(W)$, respectively. Note that in speech, $P(W)$ is the language model; i.e., the prior probability for a sequence of words, $W$. For this paper, the tonality model assumes that every chord is equally likely. The reason for the proportionality in (8) is that $P(C)$ is the same for all chord sequences. The acoustic model is the probability of producing the observed chroma vectors for chord $W$ and is modeled with a HMM; i.e.,

$$P(C|W) = \pi_{q_0} \prod_{n=1}^{N} a_{q_{n-1}q_n} b_{q_n}(c_n) \tag{9}$$

where $\pi_{q_0}$ is the initial state probability, $a_{q_{n-1}q_n}$ is the transition probability from state $q_{n-1}$ to $q_n$, and $b_{q_n}(c_n)$ is the output likelihood, which is modeled by a Gaussian mixture model (GMM)

$$b_{q_n}(c_n) = \sum_{d=1}^{D} \omega_d N(\mu_d, \Sigma_d) \tag{10}$$

where $D$ is the number of mixtures, $\omega_d$ is the mixture weight for component, $d$, and $N(\mu_d, \Sigma_d)$ is a Gaussian density with mean $\mu_d$ and covariance $\Sigma_d$. If the features are uncorrelated, then the covariance matrix is diagonal, $\Sigma_d = diag(\sigma_1, \sigma_2, ..., \sigma_{12})$. Note that a single state HMM is equivalent to a GMM.

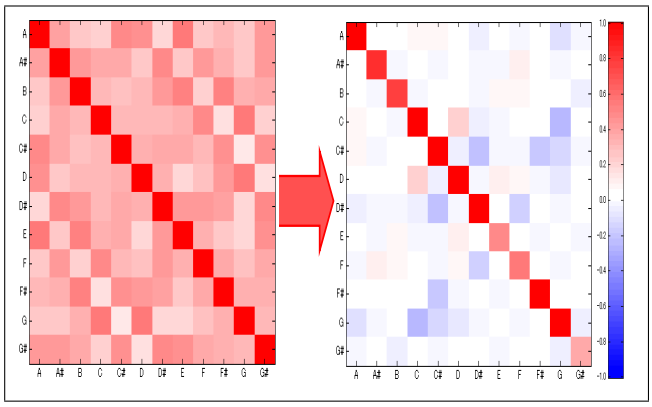

**Figure 1**. Cross correlations of chroma features. Right: original chroma features. Left: DFT chroma features. Dark shades (or red if in color) indicate higher correlation (light shades (or blue in color) indicate low correlation.

## 3. IMPROVED ALGORITHM

Since the baseline algorithm is in a format compatible with the automatic speech recognition paradigm, it provides a good framework to test more advanced speech processing techniques, such as MCE, as an alternative model estimation step. In addition, parameter reduction and tuning compensation are implemented and compared to the baseline.

### 3.1 Fourier Transform Chroma Features

As noted in [3], chroma features are highly correlated because harmonics of different pitch classes overlap and is demonstrated in the left part of Figure 1. For instance, the third harmonic of *C4* (261.63 Hz fundamental, 784.89 Hz third harmonic) is highly confusable with *G5* (783.99 Hz fundamental). However, as shown on the right of Figure 1, the resulting feature dimensions have less cross-correlation after applying a DFT on the chroma features.

### 3.2 Tuning Compensation

A second enhancement is tuning compensation. Standard tuning is such that the *A* note above middle *C* on a piano keyboard (i.e., *A4*) is approximately 440 Hz. However, artists may intentionally or unintentionally have a reference tuning different from the standard. This can lead to confusion in algorithms which assume that all music is tuned to the standard reference, as shown in Figure 2. In the upper part of Figure 2, a 12-dimensional chroma is applied to a piece of music whose energy distribution is higher in frequency than standard tuning ($A4 \simeq 440Hz$). Therefore, the signal energy is distributed between the intended note (e.g., *C3*) and the neighboring note (e.g., *C#3*). Considering that *Major* and *minor* chords differ by only one semi-tone in a single note, this can lead to large confusion between the *Major* and *minor* modes of a given chord.

To account for differences in tuning, a simplified version of [3] is implemented. The original tuning compensation algorithm uses 36 bins per octave ($\beta = 36$ in (7)) in the calculation of chroma vectors. A peak picking algorithm produces a histogram, which gives information about

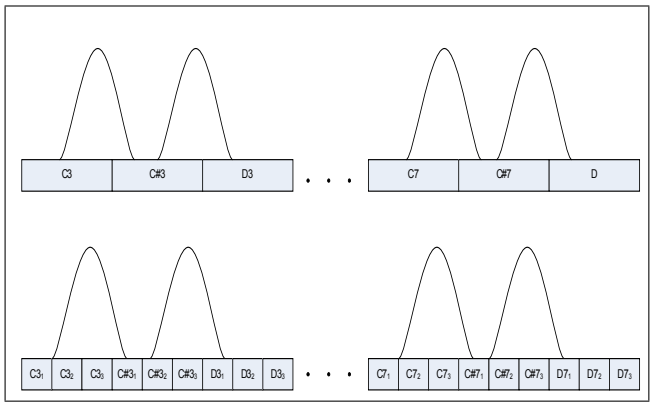


**Figure 2**. Upper: Hypothetical mistuned energy distribution. Bottom: Find tuning alignment giving maximum energy distribution at sampled points.

the tuning of the piece. A circular shift is then applied to the chroma vector as a corrective factor. The reason for peak picking is that noise sources (e.g., percussion and transients) corrupt the chroma vectors with non-harmonic sources.

However, because of HPSS, a simplified procedure filters percussive noise sources and leaves energy due to harmonic sources. The new algorithm takes a 36-dimension chroma vector for each frame in a song, so that each note considered is divided into a three bins

$$\tilde{c}_n^{(\alpha)}(b) = \sum_{r=0}^{R} |S(b + \alpha + r\beta)| \tag{11}$$

where $\alpha = \{1, 2, 3\}$ and $b = \{1, 2, \ldots, 12\}$. The algorithm then retains the set the $\alpha$ which produces the chroma vector with the greatest Euclidean length

$$c_n = \arg\max_{\tilde{c}_n^{(\alpha)}} \left( \tilde{c}_n^{(\alpha)} \cdot \tilde{c}_n^{(\alpha)} \right) \tag{12}$$

### 3.3 Minimum Classification Error Learning

As mentioned in the Introduction, MCE is a highly successful discriminative training approach to improving automatic speech recognizers over ML and MAP estimation. The optimization criterion in MCE is to minimize the estimated classification loss

$$L(\Lambda) = \frac{1}{J} \sum_{j=1}^{J} \sum_{m=1}^{M} l_m(X_j; \Lambda) 1(X_j \in \Omega_m) \tag{13}$$

where $\Lambda$ are the model parameters, $J$ is the number of training examples, $\{X_1, X_2, \ldots, X_J\}$, $M$ is the number of categories (i.e., chords), $l_m(\cdot)$ is a loss function, and $1(X_j \in \Omega_m)$ is one if $X_j$ is in category $\Omega_m$ and zero otherwise. Typically, a 0-1 loss is used for $l_m(\cdot)$, which makes the objective function discrete and difficult to optimize. However, a common approximation for the loss function is to replace the 0-1 loss with a logistic function [9],

$$l_m(X_j; \Lambda) = \frac{1}{1 + \exp(-\gamma d_m(X_j; \Lambda) + \theta)} \tag{14}$$

where $\gamma$ and $\theta$ are experimental constants and $d_m(X_j; \Lambda)$ is a misclassification measure, which is negative with a correct classification and positive when a classification error is made.

A good indication of misclassification is the distance between the correct class and competing classes; therefore, the chosen misclassification measure is based on the generalized log-likelihood ratio [9]:

$$d_m(X; \Lambda) = -\log g_m(X; \Lambda) + \log \left[ G_m(X; \Lambda) \right]^{1/\eta} \tag{15}$$

where

$$g_m(X; \Lambda) = \max_q \pi_{q_0}^{(m)} \prod_{n=1}^{N} a_{q_{n-1} q_n}^{(m)} b_{q_n}^{(m)}(c_n) \tag{16}$$

$$G_m(X; \Lambda) = \frac{1}{M - 1} \sum_{p, p \neq m} \exp[g_p(X; \Lambda)\eta] \tag{17}$$

where $\eta$ is an experimental positive constant and the superscript $^{(m)}$ refers to the $m$-th HMM. Note the misclassification measure in (15) compares the probability of the target class against a geometric average of the competing classes. The parameter $\eta$ determines the importance of the competing classes by the degree of competition with the target class. In particular, as $\eta \to \infty$, (17) returns only the most competitive class. A gradient probabilistic descent procedure [9] produces a set of parameters that yields a local optimum of (13) through the update equations

$$\Lambda_{\tau+1} = \Lambda_\tau - \epsilon \frac{\partial l_m(X_j; \Lambda)}{\partial \Lambda} \bigg|_{\Lambda = \Lambda_\tau} \tag{18}$$

In order to keep the necessary constraints for an HMM density, the following transformations are used [9]:

$$\tilde{\mu}_d^{(m)}(b) = \frac{\mu_d^{(m)}(b)}{\sigma_d^{(m)}(b)} \tag{19}$$

$$\tilde{\sigma}_d^{(m)}(b) = \log \sigma_d^{(m)}(b) \tag{20}$$

## 4. RESULTS

### 4.1 Experimental Setup

The evaluation database is the set of studio albums by *The Beatles*, which were transcribed at the chord level by Chris Harte *et. al* [10]. As in the 2008 MIREX contest, only the *Major* and *minor* chords are used for the evaluation. All extended chords, *Augmented*, and *diminished* chords are mapped to the the base root, *Major*, and *minor* chords, respectively. A two-fold evaluation is implemented, where half the albums are used as a training set and the remaining half are used as a test set in the first fold. In the second fold, the roles of the training and test set are swapped. Note all songs from a particular album occur in either the test or training set for each individual fold. This is the same setup as MIREX 2008, but the third fold in MIREX 2008 was removed because it was observed that the test cases where already covered in the first two folds. Prior to HPSS, audio is downsampled to 11025 Hz. In addition, chord boundary

| Fold | BL | FT | FT+TC | FT+TC+MCE |
|---|---|---|---|---|
| 1 | 57.84 | 57.84 | 61.91 | 73.59 |
| 2 | 61.74 | 61.74 | 64.97 | 71.35 |
| Total | 59.95 | 59.95 | 63.57 | 72.37 |

**Table 1**. Classification accuracies for Fourier transform features and tuning compensation (BL = baseline, FT = Fourier Transform, TC = tuning compensation, MCE = minimum classification error).

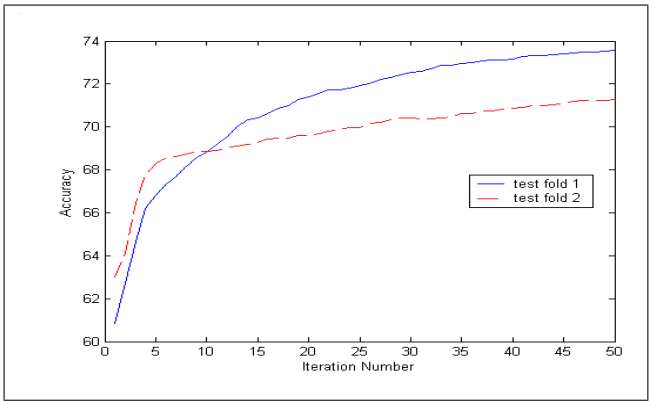


**Figure 3**. Classification accuracy versus iteration number.

information is assumed to be known and results are given in percentage of correctly recognized isolated chord segments, except in Section 4.3, where results are given in frame accuracy.

### 4.2 Isolated Chord Recognition Results

Table 1 details the improvement over the ML approach, where chords are modeled with a single Gaussian distribution with a full covariance matrix for the baseline, Fourier transform, and Fourier transform with tuning compensation cases. The MCE results listed used 50 iterations of the gradient probabilistic descent algorithm. The parameters $\gamma$, $\theta$, and $\eta$ were found experimentally by using a set of five songs from the training set as a cross-validation set. After cross-validation, the entire training set is used to retrain the system.

Tuning compensation provides a modest, but consistent gain in performance. Applying a Fourier transform does not change the performance from the baseline. However, the main advantage of applying a discrete Fourier transform is to attenuate the correlation in chroma features, and is equivalent to a discrete cosine transform for real, symmetric data [14]. In addition speech processing algorithms, such as MCE, assume diagonal covariance matrices in the GMM observation probability. Therefore, applying MCE is straightforward and results in a drastic increase in performance over ML estimation. In particular Figure 3 demonstrates, generally, each iteration of the gradient probabilistic descent algorithm improves the classification rate.

To understand the types of errors that remain, the confusion matrix is presented in Table 2. It is observed that *Major* chords are classified more accurately than *minor* chords. Specifically, many errors are due to recognizing *minor* chords with the correct root, but wrong mode; i.e., the parallel *Major* chord. For example, 82% of *c minor* chords are recognized as *C Major*. The second most common type of error is in mistaking a *Major* chord for its *minor*, which are chords that share the same key signature, but differ in the root note. For example, *e minor* is confused with *G Major* 12% of the time. Note, that no language model is used in this current paper since the goal of this paper is to study the confusions that arise due acoustic confusability in the chroma/HMM framework.

| Fold | BL | FT | FT+TC | FT+TC+MCE |
|---|---|---|---|---|
| 1 | 74.96 | 74.96 | 77.04 | 77.90 |
| 2 | 72.95 | 72.95 | 73.54 | 74.51 |
| Total | 73.46 | 73.46 | 75.20 | 76.10 |

**Table 3**. Frame accuracy for continuous chord recognition.

| # Frames | BL | FT+TC | FT+TC+MCE |
|---|---|---|---|
| 0 | 73.91 | 75.20 | 76.12 |
| 1 | 76.59 | 77.92 | 78.93 |
| 2 | 78.61 | 79.94 | 80.99 |

**Table 4**. Frame accuracy versus number of frames removed at chord boundary.

### 4.3 Continuous Chord Recognition

While this paper is mainly concerned with isolated chord classification, an additional experiment demonstrates the performance of continuous chord recognition. In this case, the frame accuracy is used as the performance metric. The results are presented in Table 3. As expected, improvement is less pronounced with adequate boundary information. Further analysis shows that one reason for the performance drop is due to identifying chord boundaries. As shown in Table 4, allowing a tolerance region of two frames on either side of a true chord transition point increases the frame accuracy. Specifically, 20% of the error occurred within two frames of a chord transition point when at least one chord to either side of the transition point was detected correctly. In [2], it was demonstrated that chroma/HMM setup performed well during forced alignment (i.e., when the chord sequence is given), but poorly when no information on the chord sequence was given. These results indicate that the chord detection problem might benefit from treating the two tasks separately and optimizing each task individually.

## 5. CONCLUSIONS

This paper considers audio chord detection as two separate tasks: (1) classifying what chords are played and (2) determining when chords begin and end. Several advanced pre-processing techniques are implemented such as HPSS, which attempts to separate transients and percussive sources from the harmonic spectrum. Further, eliminating

| | C | C# | D | D# | E | F | F# | G | G# | A | A# | B | c | c# | d | d# | e | f | f# | g | g# | a | a# | b |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 91 | | | | | 2 | | 2 | | | 2 | | | | | | | | | | | 2 | | |
| C# | | 71 | 2 | | 2 | | 12 | | | | | 5 | | 7 | | | | | | | | | | |
| D | 1 | | 89 | | 1 | | | 2 | | 2 | | | | | | | | | | | | 1 | | 3 |
| D# | | | | 81 | | 1 | | | | | 9 | | 1 | | | 6 | | | | | 1 | | | |
| E | | | | | 94 | | | | | 1 | | | | | | | 2 | | | | | | | 1 |
| F | 3 | | | | | 91 | | 2 | | | 1 | | 1 | | 1 | | | | | | | 2 | | |
| F# | 1 | | | 1 | | | 91 | | 1 | 1 | | | | 3 | | | | | 2 | | | | 1 | |
| G | 1 | | 1 | | | | | 94 | | | | | | | | | 1 | | | | | | | |
| G# | | | | | | | 3 | | 83 | | | | | | | | | 3 | | | 10 | | | |
| A | | | 1 | | 1 | | | | | 94 | | | | | | | | | 1 | | | 1 | | |
| A# | | | | 1 | 1 | 1 | | | | | 91 | 1 | | | 1 | | | | | 3 | | 1 | 1 | |
| B | | | | 1 | 1 | | 1 | | 1 | 3 | 1 | 85 | | | | | | | 1 | | 2 | | | 3 |
| c | 82 | | | | | | | 9 | | | | | 9 | | | | | | | | | | | |
| c# | | 1 | | | 3 | | | | 1 | 5 | 1 | 2 | | 87 | | | | | | | | | | |
| d | 6 | | 12 | | | 13 | | 3 | | 3 | 1 | | | | 57 | | | | | | | 4 | | 1 |
| d# | | | | | | | | | | | | 33 | | | | 67 | | | | | | | | |
| e | 5 | | 1 | | 19 | | | 12 | | 5 | | | | | | | 50 | | | | | 8 | | 1 |
| f | 1 | 1 | | | 1 | 11 | | | 16 | | 6 | | | | | | | 63 | | | | | 1 | |
| f# | | | 2 | 1 | 2 | | 7 | | | 11 | | | | | | | | | 76 | | | | | 1 |
| g | 3 | | | | | | | 14 | | | 9 | | | | 1 | | 1 | | | 71 | | | | |
| g# | | | | | 4 | | | | | | | 12 | | | | | | | | | 85 | | | |
| a | 5 | | | | 2 | 2 | | | | 13 | | | | | | | | | | | | 78 | | |
| a# | | 2 | | 2 | | | | | | | 40 | | | | | | | | | | | | 56 | |
| b | | | 6 | | | | | 1 | | | | 20 | | | | | | | | | | | | 71 |

**Table 2**. Chord confusion matrix (%). Rows are true chords, columns are hypothesized chords. Capital letters represent *Major* chords and lowercase letters represent *minor* chords.

the correlation between chroma features allows for the use of many speech processing tools because these tools are built using the assumption of diagonal matrices in the observation probability densities.

In this paper, tuning compensation and MCE enhance the chord recognition task over traditional maximum likelihood by reducing the confusion due to noise in the feature extraction stage. In the future, the authors hope to incorporate other advanced speech processing techniques, such as *N*-best re-scoring, to combat other areas of confusion such as the confusion between *minor* chords and their relative and parallel *Major* equivalents. Finally, it was observed that even when chords are detected correctly, 20% of the error occurred at chord transition points. Therefore, the authors are investigating chord transition detection algorithms to optimize the second task of chord detection.

## 6. REFERENCES


[1] T. Kawakami, M. Nakai, H. Shimodaira, S. Sagayama: "Hidden Markov Model Applied to Automatic Harmonization of Given Melodies," *IPSJ Technical Report*, 99-MUS-34, pp. 59-66, Feb., 2000. (*in Japanese*)

[2] A. Sheh and D.P.W. Ellis: "Chord Segmentation and Recognition Using EM-trained Hidden Markov Models," *Proc. ISMIR*, pp. 183–189, 2003.

[3] J.P. Bello and J. Pickens: "A Robust Mid-level Representation for Harmonic Content in Musical Signals," *Proc. ISMIR*, pp. 304-311, 2005.

[4] K. Lee and M. Slaney: "Automatic Chord Recognition from Audio Using an HMM with Supervised Learning," *Proc. ISMIR*, pp. 133-137, 2006.

[5] K. Lee and M. Slaney: "Acoustic Chord Transcription and Key Extraction from Audio Using Key-dependent HMMs Trained Synthesized Audio," *IEEE TASLP*, Vol. 16, No. 2, pp. 291-301.

[6] C.A. Harte, M.B. Sandler, and M. Gasser: "Detecting Harmonic Change in Musical Audio," *Proc. Audio Music Comput. Multimedia Workshop*, pp. 21-26, 2006.

[7] J. Downie: "Music Information Retrieval Evaluation eXchange (MIREX)," [Online]. Available: http://www.musicir.org/mirex/2008/index.php/

[8] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, S. Sagayama "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," *Proc. EUSIPCO*, 2008.

[9] B.-H. Juang, W. Chou, C.-H. Lee: "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE TSAP*, Vol. 5, No. 3, pp. 257-265, 1997.

[10] C. Harte, M. Sandler, S. Abdallah, and E. Gómez "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations," *Proc. ISMIR*, pp. 66-71, 2005.

[11] H. Papadopoulos and G. Peeters: "Large-scale Study of Chord Estimation Algorithms Based on Chroma Representation and HMM," *Intern. Wkshp. Content-Based Multimedia Indexing*, pp. 53-60, 2007.

[12] J. Brown: "Calculation of a constant Q spectral transform," *J. Acoust. Society America*, Vol. 89, No. 1, pp. 425-434, 1991.

[13] S. Kostka and D. Payne: *Tonal Harmony*, McGraw Hill, 2004.

[14] L. Rabiner and B.-H. Juang: *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.