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

This paper describes our submission to the 2009 MIREX Audio Chord Detection Contest. Improvements over the submission from last year [1] reduce feature correlation, account for differences in tuning, and incorporate minimum classification error (MCE) training in obtaining chord-level HMMs. Experiments demonstrate that classification rates improve with tuning compensation and MCE discriminative training. The task is evaluated under the same setup as the 2008 MIREX Audio Chord Detection Contest must be used.

1. IMPLEMENTATION OVERVIEW

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) [3]. To attenuate percussive sounds, the harmonic-percussive source separation (HPSS) algorithm [4] 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, decorrelation through a DFT of the chroma vector, and minimum classification error (MCE) training [5].

2. FEATURE EXTRACTION

2.1 Harmonic/Percussion Source Separation

As noted in [6], 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. This paper uses the HPSS algorithm [4], 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 σ_P^2 and σ_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 containing a percussive-dominated spectrum and the other containing a harmonic-dominated spectrum. The harmonic-dominated spectrum is retained for further processing and the percussive-dominated spectrum is discarded. Further details can be found in [4].

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 [7] 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 center frequency of each bin is designed to match the equal-temperament scale [8]. Next, chromagram features are calculated for frame n as

$$c_n(b) = \sum_{r=0}^{R} |S(b + r\beta)| \tag{5}$$

where $b = \{1, 2, ..., \beta\}$ is the chroma bin number, β is the dimensionality of the chroma vector, and R is the number of octaves considered. In this submission, β is set to 60 and reduced to 12 during tuning compensation.

2.3 Tuning Compensation

Because HPSS is very effective in seperating percussive sources and other transients from the harmonic spectrum, a simplified form of the tuning compensation in [2] is implemented. Specifically,a 60-dimension chroma vector for each frame in a song is extracted, so that each note considered is divided into five bins

$$\tilde{c}_n^{(\alpha)}(b) = \sum_{r=0}^{R} |S(b + \alpha + r\beta)| \tag{6}$$

where $\alpha = \{1, 2, 3\}$ and $b = \{1, 2, \ldots, 12\}$. The algorithm then retains the set the member of α that produces the chroma vector with the greatest Euclidean length; i.e., maximum energy:

$$c_n = \arg\max_{\tilde{c}_n^{(\alpha)}} \left(\tilde{c}_n^{(\alpha)} \cdot \tilde{c}_n^{(\alpha)} \right) \tag{7}$$

2.4 Dynamic Features

In speech recognition, using dynamic features such as delta cepstrums are often used with static features and known to increase recognition rates. Similar to delta cepstrums, dynamic features of chroma vectors (delta chroma vectors) can be used in chord detection. The motivation behind the use of delta chroma vectors is to obtain higher accuracy on chord boundaries since delta chroma vectors have large values on sound changes.

To reduce noise in the derivative calculation, delta chroma vectors can be obtained from a weighted regression analysis of chroma vector sequences. The delta chroma vector at time τ can be calculated using the δ previous and future samples:

$$\Delta c(l, \tau) = \frac{\sum_{k=-\delta}^{\delta} k \cdot w(k) c(l, \tau + k)}{\sum_{k=-\delta}^{\delta} k^2 w(k)} \tag{8}$$

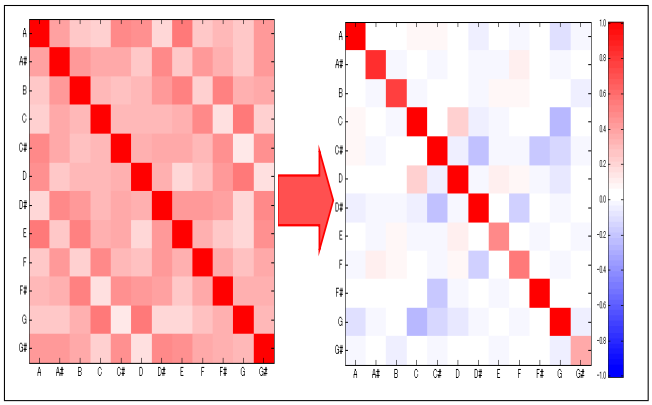

Figure 1. Cross correlations of chroma features. Right: original chroma features. Left: DFT chroma features. Dark shades (or red if in color) indicate higher correlation (light shades (or blue in color) indicate low correlation.

where the weights, $w(k)$, are an even function where $c(l, \tau)$ is chroma feature l at time τ.

As noted in [2], 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. This is similar to the use of the discrete cosine transform to reduce correlation in image processing and in the calculation of Mel-frequency cepstral coefficients

3. CHORD MODELING

3.1 HMM classifier

The optimal chord sequence, W^* is decoded such that [9]

$$\begin{aligned} W^* &= \arg\max_W P(W|C) \\ &= \arg\max_W \frac{P(C|W)P(W)}{P(C)} \\ &\propto \arg\max_W P(C|W)P(W) \end{aligned} \tag{9}$$

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. For this paper, a bigram tonality model is estimated from the training data. The acoustic model is the probability of producing the observed chroma vectors for a chord W and is modeled with a HMM with a GMM observation probability. Note that each chord is modeled with an individual HMM versus the ergodic model seen in most chord detection algorithms, where each chord is assigned a single state. Further, an HMM with a single state reduces to a GMM.

3.2 Minimum Classification Error Learning

MCE is a highly successful discriminative training approach which improves automatic speech recognizers over ML and maximum a posteriori estimation [5]. The optimization criterion in MCE is to minimize the estimated classification loss

$$L(\Lambda) = \frac{1}{J}\sum_{j=1}^{J}\sum_{m=1}^{M} l_m(X_j;\Lambda)1(X_j \in \Omega_m) \qquad (10)$$

where Λ 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 the indicator function, which is one if X_j is in category Ω_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 [5],

$$l_m(X_j;\Lambda) = \frac{1}{1+\exp(-\gamma d_m(X_j;\Lambda)+\theta)} \qquad (11)$$

where γ and θ are experimental constants and $d_m(X_j;\Lambda)$ is a misclassification measure.

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 [5]:

$$d_m(X;\Lambda) = -\log g_m(X;\Lambda) + \log\left[G_m(X;\Lambda)\right]^{1/\eta} \qquad (12)$$

where

$$g_m(X;\Lambda) = \max_q \pi^{(m)}_{q_0}\prod_{n=1}^{N} a^{(m)}_{q_{n-1}q_n} b^{(m)}_{q_n}(c_n) \qquad (13)$$

$$G_m(X;\Lambda) = \frac{1}{M-1}\sum_{p,p\neq m}\exp[g_p(X;\Lambda)\eta] \qquad (14)$$

where η is an experimental positive constant and the superscript $^{(m)}$ refers to the m-th HMM. Note the misclassification measure in (12) compares the probability of the target class against a geometric average of the competing classes. The parameter η determines the importance of the competing classes by the degree of competition with the target class. In particular, as $\eta\rightarrow\infty$, (14) returns only the most competitive class. A gradient probabilistic descent procedure [5] produces a set of parameters that yields a local optimum of (10) through the update equations

$$\Lambda_{\tau+1} = \Lambda_\tau - \epsilon\frac{\partial l_m(X_j;\Lambda)}{\partial\Lambda}\bigg|_{\Lambda=\Lambda_\tau} \qquad (15)$$

In order to keep the necessary constraints for an HMM density, the following transformations are used [5]:

$$\tilde{\mu}^{(m)}_d(b) = \frac{\mu^{(m)}_d(b)}{\sigma^{(m)}_d(b)} \qquad (16)$$

$$\tilde{\sigma}^{(m)}_d(b) = \log\sigma^{(m)}_d(b) \qquad (17)$$

4. REFERENCES

[1] Yuki Uchiyama, Kenichi Miyamoto, Nabutaka Ono, Shigeki Sagayama: "Automatic chord detection using harmonic sound emphasized chroma from musical acoustic signal," *MIREX 2008*, http://www.music-ir.org/mirex/2008/abs/uchiyamamirex2008.pdf

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

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

[4] 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.

[5] 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.

[6] 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.

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

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

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