

LEARNING HARMONIC RELATIONSHIPS IN DIGITAL AUDIO WITH DIRICHLET-BASED HIDDEN MARKOV MODELS

J. Ashley Burgoyne and Lawrence K. Saul
Department of Computer and Information Science
University of Pennsylvania, Philadelphia, PA 19104 USA
{burgoyne, lsaul}@cis.upenn.edu

ABSTRACT

Harmonic analysis is a standard musicological tool for understanding many pieces of Western classical music and making comparisons among them. Traditionally, this analysis is done on paper scores, and most past research in machine-assisted analysis has begun with digital representations of them. Human music students are also taught to hear their musical analyses, however, in both musical recordings and performances. Our approach attempts to teach machines to do the same, beginning with a corpus of recorded Mozart symphonies. The audio files are first transformed into an ordered series of normalized pitch class profile (PCP) vectors. Simplified rules of tonal harmony are encoded in a transition matrix. Classical music tends to change key more frequently than popular music, and so these rules account not only for chords, as most previous work has done, but also for the keys in which they function. A hidden Markov model (HMM) is used with this transition matrix to train Dirichlet distributions for major and minor keys on the PCP vectors. The system tracks chords and keys successfully and shows promise for a real-time implementation.

Keywords: Dirichlet, harmony, PCP, HMM, Mozart

1 INTRODUCTION

Machine-assisted harmonic analysis has a long history and continues to spawn active research [1, 2]. Much of that research, however, has focused on digital analogues to the symbols on a paper score [1, 3, 4, 5]. We choose to begin with sound, using models drawn from the speech and signal processing communities.

The project shares ground with Christopher Raphael’s work on automatic transcription of piano music [6], although the “harmonic” space of his problem domain enjoys less *a priori* structure. Its inspiration is Alexan-

der Sheh and Daniel Ellis’s chord recognition project for songs by the Beatles [7]. We begin with a hidden Markov model, as do Sheh and Ellis, but replace the more traditional Gaussian emission distributions with Dirichlet distributions. Dirichlet distributions have properties particularly well suited to recognizing chords and are the key advance made in our work. They also bring the project closer to several recent key tracking algorithms for audio data [8, 9], and in a second important departure from Sheh and Ellis, we choose a more complex harmonic model, rooted in traditional music theory, that enables the system to track key simultaneously with chord.

The second section of this paper outlines the theoretical background necessary for understanding the model, the third discusses the details of implementation and the results, and the fourth offers ideas for further development.

2 THEORETICAL BACKGROUND

2.1 Hidden Markov Models

Hidden Markov models (HMMs) are a family of statistical models that have proven very useful for speech recognition and certain tasks in robotics and are growing increasingly popular for musical problems. They are defined by a discrete state space \mathcal{S} that cannot be observed directly but is assumed to generate a set of possibly multidimensional and continuous observations, a complete set of transition probabilities between these states as time passes, and a formula for computing the probability of any observation given some state in \mathcal{S} . The simplest of these models – and ours – make the first-order Markov assumption that for a time-ordered set of random state variables $S_1, \dots, S_T \in \mathcal{S}$,

$$P(S_{t+1}|S_1, \dots, S_t) = P(S_{t+1}|S_t) \quad . \quad (1)$$

Under this assumption, the transition probabilities can be stored in matrix form.

2.2 Harmonic Space

Much previous work either assumes that the music analyzed will remain in a single key from start to finish or disregards the notion of key altogether, tackling the chord recognition problem explicitly or considering transitions between chords independent of their tonal context. Al-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

though it is frequently necessary to make simplifying assumptions, these ones are especially limiting. Western classical music modulates frequently, and contemporary tonal theory rests on the assumption that it is not merely local configurations of pitches that create musical meaning but also the contexts in which these configurations arise. A D major chord in the key of G major should yield a PCP similar to that of a E $\flat\flat$ major chord in the key of D \flat major, but quite different harmonies (and corresponding PCPs) are likely to follow. Our model thus considers chord and key to be inseparable properties of any given harmony.

Another group of previous work attempts only to identify the prevailing key. Most of these models are based on Carol Krumhansl’s probe-tone key profiles or Vos and van Geenen’s derivative algorithm [10, 5], although there have been interesting alternatives using Elaine Chew’s spiral model [11]. By choosing a different model, we are in some sense attempting to relearn profiles from the ground up: see also [12].

For ease of implementation and training, the model is restricted to major and minor triads only, ignoring augmented and diminished triads as well as sevenths, ninths, and other additions. A full range of chromatic alterations, however, are available within each key. Following Aldwell and Schachter [13], we divide the triads into four groups based on their degree of “mixture.” The diatonic chords within each key employ no mixture. Primary mixture accounts for chords that are “borrowed” from the parallel major or minor.¹ Secondary mixture describes other chromatic alterations of the third or fifth of the diatonic triads. Double mixture, as its name implies, includes both a borrowing and a further alteration of the third or fifth. Table 1 lists the chords included in the model and the types of mixture necessary to produce them.

2.3 Pitch Class Profiles

One of the primary challenges of the project is to work with digitized audio rather than a score-like format. Among the many useful features that can be computed from audio are pitch class profiles (PCPs) [14], which Sheh and Ellis found to perform significantly better than several popular alternatives. The computation of PCPs begins with a windowed short-term Fourier transform (STFT):

$$X_{\text{STFT}}[k, n] = \sum_{m=0}^{N-1} x[n-m]w[m]e^{-j(2\pi/N)km} \quad (2)$$

where n is the index of the edge of a window of length N in the discrete time series $x[\cdot]$, w is a discrete windowing function of length N , and k indexes the frequency axis from DC to the Nyquist frequency. The values of the STFT for each bin are squared to generate a power spectrum and then mapped to the musical pitch class closest to the frequency of the bin. For each window, these squared

¹Aldwell and Schachter consider IV and V in minor to be instances of primary mixture but note that they are essential to the the key; we consider them to be native to minor keys for the purposes of our model.

	Major	Minor
I	none	primary
i	primary	none
\flat II	primary	none
\flat ii	double	secondary
II	secondary	double
ii	none	primary
\flat III	primary	none
\flat iii	double	secondary
III	secondary	double
iii	none	primary
IV	none	none
iv	primary	none
V	none	none
v	primary	none
\flat VI	primary	none
\flat vi	double	secondary
VI	secondary	double
vi	none	primary
\flat VII	primary	none
\flat vii	double	secondary
VII	secondary	double
vii	secondary	double

Table 1: Tonal vocabulary and classification of mixture.

values are summed over the pitch class labels to generate a twelve-dimensional PCP. The PCPs then represent the total amount of spectral energy in each musical pitch class at regularly sampled points in time. They are similar to Tzanetakis, Ermolinskyi, and Cook’s pitch histograms [15] but with a more localized scope.

An alternative calculation of PCPs might begin with the constant Q transform, which has had some success in chord identification and the related task of polyphonic pitch tracking [16, 17].

2.4 Dirichlet Distributions

A Dirichlet distribution is a probability distribution over a set of discrete probability distributions. It is the conjugate prior of the multinomial distribution, which is a generalization of the binomial distribution from a binary decision to a set of n alternatives. If we label the probabilities of choosing each of these alternatives $\vec{p} = p_1, \dots, p_n$, $\sum_{i=1}^n p_i = 1$ and $p_i > 0 \forall i$, then a Dirichlet distribution with parameters $\vec{u} = u_1, \dots, u_n$, $u_i > 0 \forall i$ is defined as

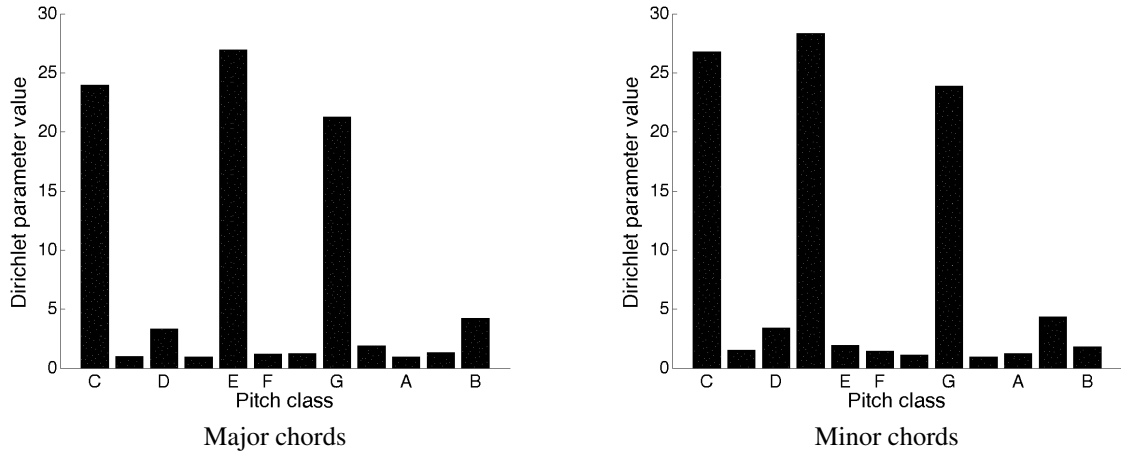
$$\text{Dir}(\vec{p}, \vec{u}) \triangleq \frac{1}{Z(\vec{u})} \prod_{i=1}^n p_i^{u_i-1} \quad (3)$$

The $Z(\vec{u})$ term is a normalization constant defined as

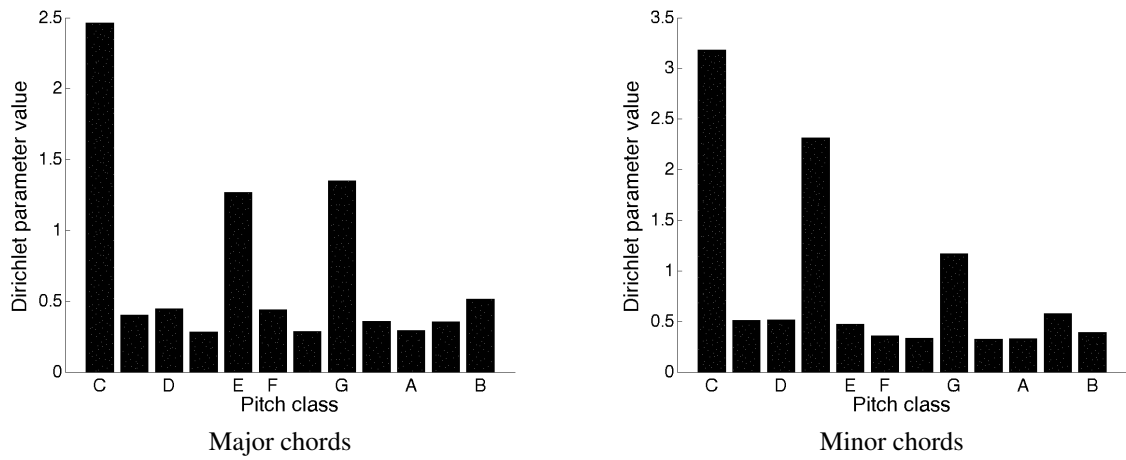
$$Z(\vec{u}) \triangleq \frac{\prod_{i=1}^n \Gamma(u_i)}{\Gamma(\sum_{i=1}^n u_i)} \quad (4)$$

where Γ represents the standard gamma function

$$\Gamma(x) \triangleq \int_0^{\infty} dt e^{-t} t^{x-1} \quad (5)$$



(a) Initial Dirichlet parameters, rotated for C major and minor. These parameters were estimated from synthesized chords of four sawtooth waves within the standard playing frequency range of the Western classical orchestra. The shapes of the distributions illustrate the mean distribution (see Eq. 6) while the parameter values are inversely proportional to the variance.



(b) Trained Dirichlet parameters, rotated for C major and minor. These parameters were estimated from the training corpus using the EM algorithm seeded with the parameters above. The distribution shapes remain unmistakably major and minor triads; variance has increased considerably.

Figure 1: Dirichlet parameters before and after training.

The mean of a Dirichlet distribution is the normalization of its parameters:

$$\langle \vec{p} \rangle = \frac{\vec{u}}{\sum_{i=1}^n u_i} \quad (6)$$

Its variance decreases as the magnitude of \vec{u} increases.

Dirichlet distributions are more attractive than Gaussian models for systems where the relations among outputs are more important than their magnitude. PCP vectors are a good example of this sort of system: what defines a C major chord is not that C, E, and G are sounding loudly but that they are relatively louder than any other pitch classes sounding at the time. Although there have been successes with Gaussian models for chord recognition, because Dirichlet distributions model the underlying phenomenon more accurately, we expect an improvement in performance. In order to use them, all PCP vectors must be normalized such that their components sum to unity.

3 IMPLEMENTATION AND RESULTS

The implemented HMM draws its states from the harmonic space outlined in Section 2.1, derives a transition matrix from the mixture classifications, accepts ordered sets of normalized PCP vectors as observation data, and parameterizes the observation distributions as Dirichlet distributions tied to the underlying states.

The transition matrix is defined by five hand-tuned parameters: p_k , the probability of remaining within the current key, and $p_{d_1} > p_{d_2} > p_{d_3} > p_{d_4}$, the probabilities of remaining within a harmony when it employs no, primary, secondary, and double mixture. All other entries are set uniformly with respect to these constraints and the “pivot region” constraints outlined by Fred Lerdahl in *Tonal Pitch Space* [18]. Lerdahl’s constraints allow major keys to modulate only to the related keys i, ii, iii, IV, V, vi and minor keys only to I, bIII, iv, v, bVI, and bVII, and even in these cases, at least one of the chords must be a tonic.

In order to maximize the utility of the training data, our system defines only two Dirichlet parameter vectors, \vec{u}_{major} and \vec{u}_{minor} . To get the expected observation distribution for any given harmony, the base parameter vector for its mode is rotated until the root of the parameter vector matches the root of the harmony. Note that triads share the same observation distribution regardless of their key: all the information for tracking keys must be encoded in the transition matrix. Figure 1(a) shows our initial parameters for the Dirichlet vectors, standardized on C major and minor. Their overall shapes illustrate the expected distributions, as per Equation 6, and the high magnitude represents a high confidence level in these initial selections.

These initial estimates were estimated from a randomly generated four-note chords composed of sawtooth waves from across the orchestral range of frequencies. Thirds and fifths were alternately doubled. Modulo the downsampling to 11,025 Hz and conversion to mono necessary to make computation on the CD audio files in our corpus tractable, the random samples were processed exactly like the training data. First, they were broken into windows of 2765 samples (250 ms) with a 50 percent

overlap. We tuned a Gaussian windowing function

$$w[k] = e^{-\frac{1}{2} \left(\alpha \frac{k-N/2}{N/2} \right)^2} \quad (7)$$

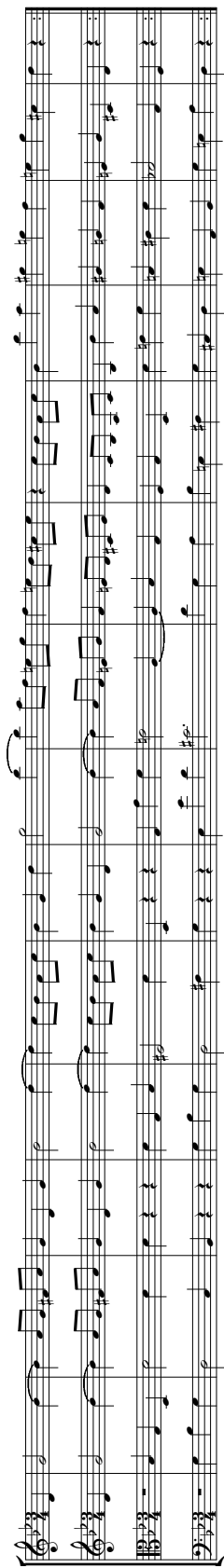
and zero-padded the STFT to 4096 samples. Setting α to 1.3 yields a main lobe width of 3.7 Hz with a leakage factor of 2.3 percent, allowing us to discard only frequency information below MIDI note 36 (65.4 Hz) while preserving the integrity of the PCP vectors. Any vectors whose total energy after this processing fell below 1e-6 were removed – in most cases, the leading and trailing samples of each track.

The training corpus comprised professional compact disc recordings of five Mozart symphonies in fifteen movements altogether with a total duration of 70’39’: Symphony No. 21 in A (K. 134), Symphony No. 22 in C (K. 162), Symphony No. 23 in D (K. 181), Symphony No. 24 in B \flat (K. 182), and Symphony No. 25 in G minor (K. 183) [19]. After processing, the expectation maximization (EM) algorithm was used to tune the Dirichlet parameters. The algorithm converges very quickly on these data, and to avoid overtraining, we limited the algorithm to five iterations.

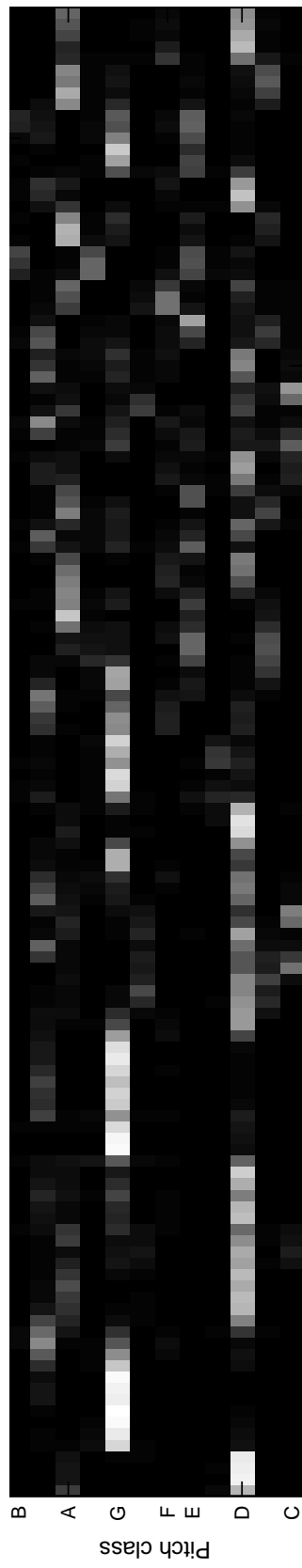
Graphs of the trained Dirichlet parameters are in Figure 1(b). Although the variance has grown much greater for experience with noisier, real-world data, the triadic profiles are clear. The third of minor chords stands out, suggesting correctly that in the absence of defining information, major chords should be the default prediction. The relative lack of prominence of the fifth of minor chords can be explained by the predominance of diminished triads in the corpus. There is no corresponding state in the model, and the system usually guesses that they are minor triads instead, enabled by this weaker emphasis on having a perfect fifth. Unfortunately, these guesses can make poor harmonic sense – when vii $^\circ$ is functioning as a dominant, for example – and it will be worthwhile to extend the model to account for diminished triads directly.

The system was tested on a recording of the Minuet from Mozart’s Symphony No. 40 in G minor (K. 550) from the same boxed set. Figure 2(a) includes a condensed excerpt from the score. Figure 2(b) underneath displays a matrix of the corresponding PCP distributions from the first time through the repeat; it is scaled linearly in time, not necessarily with the score above it. Figure 2(c) includes a ground truth harmonic analysis of the excerpt given the limitations of our model to major and minor triads; chords are in the top row in roman type and the keys underneath them in boldface. The phrase modulates from G minor to D minor with some chromatic trickery in the second half as well as a number of suspensions and other non-harmonic tones. Also notice the paucity of well articulated triads in the PCP plot despite the fully orchestrated texture above.

Underneath the ground truth in Figure 2(c) are the chord and key sequences predicted by the model for both the first and second passes through repeat in the recording (without timing data). The free chord identification results are excellent. The model fails to identify only 4 of the 24 harmonies in this example (17 percent), comparable with previous work, and these are “good” mistakes. The first error is g for E \flat in m. 7, which is almost a viable alter-



(a) Reduced musical score.



(b) PCP distributions (first repeat only), scaled linearly in time. Note the lack of obvious triads.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ground truth:	D g	D g	g g	g g	D g	D g	E ^b g/d	A d	g d	A B ^b	g d	E A	e d	A d
First repeat:	D g	D g	G d	g d	D d	D d	G d	A c [#]	d d	A C	d d	E A	e d	A D
Second repeat:	D d	D d	G d	g d	D d	D d	G d	e g	d g	A C	d g	E A	e g	A D

(c) Ground truth harmonic labels and machine analysis. The results are especially impressive considering the few well defined triads above: clearly, the harmonic model is working.

Figure 2: Symphony No. 40 (K. 550), III, mm. 1–14.

native analysis. It misses two other deceptive cadence figures, one in m. 10 and other in m. 12, mistaking $\flat VI$ for the tonic in both cases. These mistakes are again reasonable: the submediant in deceptive cadences is meant to mimic a tonic. It would have been almost impossible for our model to figure out that it could label the $c\sharp^{o7}$ in m. 11 as A, and its guesses of $c\sharp$ the first time and $\flat b$ the second each make the diminished-fifth compromise discussed earlier.

The key finding results are less impressive but a marked improvement from before training. G minor and D minor are very closely related keys, and with so much chromaticism in the second half of the phrase, it is understandable that the repeat would have confused the system. Moreover, because of the way in which harmonic states are tied to the Dirichlet parameters, the system almost has to determine the chord sequence *a priori* and then try to match a smooth progression of keys beneath it using just the rudimentary information contained in the transition matrix. A more advanced harmonic model is needed.

4 CONCLUSIONS AND FUTURE WORK

Dirichlet distributions on PCP vectors are an efficient and effective means for chord recognition in recorded performances of symphonic music, and their accuracy should improve further with more sophisticated harmonic transition models like Lerdahl's or Raphael's. Our current system could be adapted for real-time performance easily using a belief-updating network. Because it is an unsupervised framework, it also lends itself to incorporating more training data in the future, perhaps enough data to learn more sophisticated harmonic models on its own.

ACKNOWLEDGEMENTS

This work was supported by NSF Award 0238323.

REFERENCES

- [1] H. Christopher Longuet-Higgins and Mark J. Steedman. On interpreting Bach. *Machine Intelligence*, 6:221–41, 1971.
- [2] Robert Rowe. *Machine Musicianship*. MIT Press, Cambridge, MA, 2001.
- [3] Christopher Raphael. Harmonic analysis with probabilistic graphical models. In *Proceedings of the International Symposium on Music Information Retrieval*, Baltimore, MD, 2003.
- [4] David Temperley. *The Cognition of Basic Musical Structures*. MIT Press, Cambridge, MA, 2001.
- [5] Piet G. Vos and E. W. van Geenen. A parallel-processing key-finding model. *Music Perception*, 14:185–224, 1996.
- [6] Christopher Raphael. Automatic transcription of piano music. In *Proceedings of the International Symposium on Music Information Retrieval*, Paris, France, 2002.
- [7] Alexander Sheh and Daniel P. W. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proceedings of the International Symposium on Music Information Retrieval*, Baltimore, MD, 2003.
- [8] Hendrik Purwins, Benjamin Blankertz, and Klaus Obermayer. A new method for tracking modulations in tonal music in audio data format. In C. L. Giles, M. Gori, and V. Piuri, editors, *Proceedings of the International Joint Conference on Neural Networks*, volume 6, 2000.
- [9] Steffen Pauws. Musical key extraction from audio. In *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.
- [10] Carol L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York, 1990.
- [11] Ching-Hua Chan and Elaine Chew. Polyphonic audio key-finding using the spiral array CEG algorithm. In *Proceedings of the International Conference on Multimedia and Expo*, Amsterdam, Netherlands, 2004.
- [12] Samer A. Abdallah. *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*. Doctoral thesis, Department of Electronic Engineering, King's College London, 2002.
- [13] Edward Aldwell and Carl Schachter. *Harmony and Voice Leading*. Harcourt Brace Javanovich, New York, 1978.
- [14] Takuya Fujishima. Real-time chord recognition of musical sound: A system using Common Lisp Music. In *Proceedings of the International Computer Music Conference*, Beijing, 1999.
- [15] George Tzanetakis, Andrey Ermolinskyi, and Perry Cook. Pitch histograms in symbolic and audio music information retrieval. *Journal of New Music Research*, 32(2):143–52, 2003.
- [16] S. Hamid Nawab, Salma Abu Ayyash, and Robert Wotiz. Identification of musical chords using constant-Q spectra. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001.
- [17] Ozgur Izmirli. A hierarchical constant Q transform for partial tracking in musical signals. In *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects*, Trondheim, Norway, 1999.
- [18] Fred Lerdahl. *Tonal Pitch Space*. Oxford University Press, New York, 2001.
- [19] Neville Marriner and the Academy of St. Martin-in-the Fields. Mozart: Early symphonies. Compact disc recording, 1990. Philips 4225012.