Audio Key Finding Using Low-Dimensional Spaces

Özgür İzmirli
Center for Arts and Technology
Computer Science
Connecticut College
New London, Connecticut, 06320
oizm@conncoll.edu

Abstract

This paper presents two models of audio key finding: a template based correlational model and a template based model that uses a low-dimensional tonal representation. The first model uses a confidence weighted correlation to find the most probable key. The second model is distance based and employs dimensionality reduction to the tonal representation before generating a key estimate. Experiments to determine the dependence of key finding accuracy on dimensionality are presented. Results show that low dimensional representations, compared to commonly used 12 dimensions, may be utilized for key finding without sacrificing accuracy. The first model's independently verified performance enabled it to be used as a benchmark for evaluation of the second model. Key finding accuracies for both models are given together with detailed results of the second model's performance as a function of the number of dimensions used.

Keywords: Key finding, chroma based representations, dimensionality reduction.

1. Introduction

Audio key finding is the problem of estimating the key of a musical piece in terms of the most stable pitch and the mode of the musical scale used. Finding the key of a piece using general polyphonic audio as input is one of the important problems in content analysis for music information retrieval. A robust solution to this problem is essential to higher levels of processing and analysis, and therefore, study of models that attempt to solve this problem is of great interest. For example, in tonal music, most high level music analyses require the determination of key as the first step. Other applications include musical style modeling, modulation detection and similarity modeling by tonal evolution.

This paper describes two key finding models and presents a comparative evaluation of the two. Both models work on recorded polyphonic audio input and produce a key estimate for each file. They are designed to operate on short fragments of audio taken from the beginnings of musical works. The first model participated in the Music Information Retrieval Evaluation Exchange in 2005 (MIREX 2005). This model scored the best composite score among the 7 participating models on the unreleased audio data set. Therefore, this model serves as a good reference to evaluate the second model. The second model explores the effect of dimensionality of tonal representation in the context of the key finding problem.

The remainder of the paper is organized as follows: Section 2 refers to related work. Section 3 describes the first model in detail and summarizes the results of the independent evaluation carried out in MIREX 2005. Section 4 describes the second model. Evaluation results for both models are given in Section 5. Section 6 concludes the paper.

2. Related Work

Many audio key finding models use a chroma based representation. A chroma based representation is a compact form of spectral representation obtained by a many-to-one mapping from the short-time spectrum of audio. The most commonly used mapping is the Pitch Class Profile (PCP) originally proposed by Fujishima [1] for recognizing chords. Izmirli [2] compared pure spectral and chroma representations for key finding and reported significantly higher accuracy with chroma based representations weighted by pitch distribution profiles. Gomez [3] used a chroma based representation called the Harmonic Pitch Class Profile, which used the peaks in the spectrum. Cabral et al. [4] studied the effects of weighting the contribution of FFT bins by their distance to the closest note. Pauws [5] used a chromagram that models chroma as a decaying spectral impulse train and arrived at a collection of chroma likelihoods in a single octave.

Models for key finding and others that use some form of tonal description have utilized classification and machine learning techniques to learn from existing data and to achieve higher rates of accuracy. Purwins et al. [6] used various classifiers to classify composers using Constant Q profiles and reported on a method for finding the degree of major/minor ambiguity. Gómez and Herrera [7] studied machine learning methods for key finding using the Harmonic Pitch Class Profile. Izmirli [8] studied the dimensionality reduction of spectra for major and minor pitch sets. Sheh and Ellis [9] used a chroma based representation together with Hidden Markov Models for finding chord boundaries. Chai and Vercoe [10] proposed an HMM-based model to segment musical pieces according to points of key change.

In general, regarding the solutions to the audio key finding problem, many models have been reported in the literature (e.g.[2] [3] [5][11] [12] [13][14].)

3. Model I

This section describes the key finding model that participated in MIREX 2005. The model is designed with the following assumptions: The input to the algorithm is a sound file that contains the musical work for which the key is to be estimated. The algorithm analyzes fragments of polyphonic audio taken from the beginnings of musical works. It is assumed that pieces input to this algorithm start in the same key as the one designated by the composer. The output consists of a single key estimate that is one of 24 possibilities – 12 for major and 12 for minor. The model has three stages: chroma template calculation using monophonic instrument sounds, chroma summary calculation from the input audio file and overall key estimation using the chroma templates and chroma summary information. Every file is processed independently – there is no learning across files.

In the first stage, templates are formed using monophonic instrument sounds spanning several octaves. For this, initially, average spectra are calculated for each monophonic note. Next, spectral templates are formed as a weighted sum of the average spectra obtained from individual instrument sounds. Two types of weighting are performed. The first is done according to a pitch distribution profile. In general, the profile can be chosen to be one of Krumhansl's probe tone ratings [15], Temperley's profiles [16], flat diatonic profiles or combinations of these. The second is a weighting that is a function of the contributing note's (MIDI pitch) value. The first weighting is used to model the pitch class distribution and the second weighting is used to account for the registral distribution of notes. The resulting spectral templates are then collapsed into chroma templates. This process comprises a many-to-one frequency mapping for each chroma in order to form a 12-element chroma template. As a result 24 chroma templates are formed. These templates act as prototypes of chroma vectors for major and minor keys. It should be mentioned that the 12 element chroma representation is most common but other divisions are also possible.

In the second stage, spectra are calculated from the input audio file and then mapped to chroma vectors. A summary chroma vector is obtained by averaging the chroma vectors in a window of fixed length. Windows of different lengths are used to obtain a range of localities. All windows start from the beginning of the piece and therefore longer windows contain information in the shorter windows as well as the new information in the later parts of their span. The lengths of the windows start from a single frame and progressively increase up to a maximum time into the piece.

The key is estimated in the third stage using the precalculated chroma templates and the summary chroma vectors calculated from the input file. For each window, correlation coefficients are calculated between the summary chroma vector and all chroma templates. The template index with the highest correlation is regarded as the estimate of the key for that window. In order to find the most prevalent key estimate for a piece, the confidence of the estimate for each window is also found. Next, the total confidence over all windows is calculated for each plausible key. The key with the maximum total confidence is reported as the overall key estimate.

3.1 Template Calculation

Templates act as prototypes to which information obtained from the audio input is compared. The purpose of constructing templates is to have an ideal set of reference chroma patterns for all possible keys. This section outlines the calculation of templates and the following section describes how the input audio is processed in order to perform the key estimation.

3.1.1 Instrument Sounds

In this algorithm, templates are obtained from recordings of real instruments, but, they could equivalently be obtained from synthetically generated harmonic spectra. A collection of sound files is used. Each file contains a single note and is appropriately labelled to reflect its note content. For this algorithm, piano sounds from the University of Iowa Musical Instrument Samples online database have been used. The sounds were converted to mono from stereo and down sampled to a sampling rate of 11025 Hz. The frequency analysis is carried out using 50% overlapping 4096-point FFTs with Hann windows. The analysis frequency range for this algorithm is fixed at 55 Hz on the low end and 2000 Hz on the high end. In general, depending on the spectral content of the input a wider frequency range may be used.

3.1.2 Pitch Distribution Profiles

Pitch distribution profiles may be used to represent tonal hierarchies in music. Krumhansl [15] suggested that tonal hierarchies for Western tonal music could be represented by probe tone profiles. Her method of key finding is based on the assumption that a pattern matching mechanism between the tonal hierarchies and the distribution of pitches in a musical piece models the way listeners arrive at a sense of key. Although she formulated this key finding method on symbolic data, many key finding models, symbolic and audio, rely on this assumption and several extensions have been proposed. In one such extension, beside other additions, Temperley [16] proposed a modification to this pitch distribution profile. We utilize this profile in combination with a diatonic profile as this combination results in the best performance of the current model. The diatonic profile can be viewed as a flat profile which responds to presence or absence of pitches but is not sensitive to the relative importance of pitches. Application of this profile alone would resemble earlier approaches to key finding in which pattern matching approaches had been used. Figure 1 shows the normalized composite profile used in this model together with Temperley's and the diatonic profiles.

Profiles are incorporated into the calculation of templates to approximate the distribution of pitches in the spectrum and the resulting chroma representation. The base profile for a reference key (A in this case) has 12 elements, represents weights of individual chroma values and is used to model pitch distribution for that key. Given that this distribution is invariant under transposition, the profiles for all other keys are obtained by rotating this base profile.

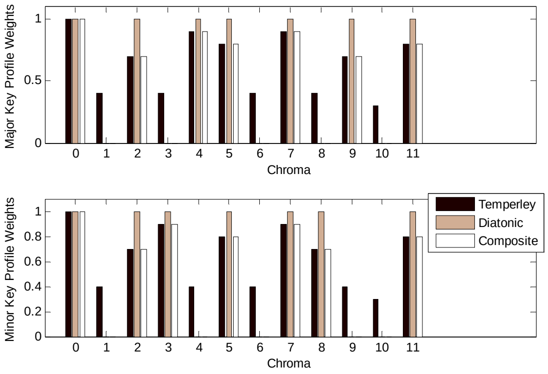

Figure 1. Pitch distribution profiles used in Model I: Diatonic major (D_M), Temperley major (T_M), composite major (P_M) (top); harmonic minor (D_m), Temperley minor (T_m), and composite minor (P_m) (bottom).

3.2 Chroma Templates

The average spectrum of an individual monophonic sound with index *i*, X_i, is computed by averaging the spectra, obtained from windows that have significant energy, over the duration of the sound. The average spectrum is then scaled by its mean value. Here, *i*=0 refers to the note A in the lowest octave, *i*=1 refers to Bb a semitone higher etc. *R* is the total number of notes within the instrument's pitch range used in the calculation of the templates. For this algorithm *R* is chosen to be 51. The lowest note is A1 and the highest is B5.

Templates are obtained by weighted sums of the average spectra calculated for individual notes. A template for a certain scale type and chroma value is the sum of X_i weighted by the profile element for the corresponding chroma and by the second weighting that is a function of the note index (*i*). A template is calculated for each scale type and chroma pair resulting in a total of 24 templates as given in Equation (1). The first 12 are major, starting from reference chroma 'A', and last 12 are minor.

$$C_n = \begin{cases} \Psi\left[\sum_{i=0}^{R-1} X_i\, f(i)\, P_M(\,(i-n+12) \bmod 12)\right] \\ \qquad if \quad 0 \le n \le 11 \\ \Psi\left[\sum_{i=0}^{R-1} X_i\, f(i)\, P_m(\,(i-n+24) \bmod 12)\right] \\ \qquad if \quad 12 \le n \le 23 \end{cases} \qquad (1)$$

$P_e(k)$ is the profile weight as shown in Figure 1, where *e* denotes the scale type (M:major or m:minor) and *k* denotes the chroma. In this work, the profile is given by the elementwise product of the diatonic and Temperley profiles: $P_e(k)=D_e(k)T_e(k)$. *f(i)* is the secondary weighting function that accounts for the registral distribution of notes. Here, it is chosen to be a simple decreasing function: $f(i)= 1\text{-}0.14i^{0.5}$. Ψ is a function that maps the spectrum into chroma bins. The mapping is performed by dividing the analysis frequency range into 1/12th octave regions with respect to the reference A=440 Hz. Each chroma element in the template is found by a summation of the magnitudes of the FFT bins over all regions that have the same chroma value.

3.3 Summary Calculation

Once the profiles are calculated they become part of the model and are used to determine the key estimates for all audio input. i.e. one set of templates is used for all audio files in a dataset. The second stage of the method involves calculation of chroma summary vectors.

Initially, a chroma vector is calculated for each FFT frame from the audio input with the same analysis parameters used for calculating the templates. Next, the actual starting point of the music is found by comparing the signal energy to a threshold. This frame is made the pivot point for the remainder of the analysis. A summary chroma vector is defined to be the average of individual chroma vectors within a window of given length. All windows start from the pivot frame with the first window containing a single frame. Window length is progressively increased in succeeding windows until the maximum analysis duration is reached. The maximum length of the audio to be analyzed is chosen to be approximately 30 seconds in this particular implementation. This results in a sequence of summary chroma vectors where each summary vector corresponds to a window of specific length.

3.4 Estimation of Key

The key estimate for an input sound file is determined from the individual key estimates corresponding to the various size windows and their associated confidence values. These two entities are determined as follows: For each window a key estimate is produced by computing correlation coefficients between the summary chroma vector and the 24 precalculated chroma templates and then picking the one with the maximum value. The confidence for an individual key estimate is given by the difference between the highest and second highest correlation value divided by the highest value. At this point each window has a key estimate and an associated confidence value. Finally, the total confidence for each plausible key is found by summing confidence values over all windows. A key is plausible if it has appeared at least once as an individual estimate in one of the windows. The key with the maximum total confidence is selected as the key estimate.

3.5 MIREX Evaluation

MIREX 2005 provided the opportunity for empirical evaluation and comparison of algorithms in many areas related to music information retrieval. Algorithms participating in MIREX 2005 were submitted directly to the MIREX committee and the evaluations were run without intervention of the participants. The results of the MIREX 2005 evaluations were reported for all participating algorithms [17]. Beside other contests that took place during the exchange, a closely related category was symbolic key finding. The MIREX evaluation framework for audio key finding and symbolic key finding used the same dataset containing 1252 pieces. The symbolic key finding algorithms used data directly from MIDI note files whereas sound files were synthesized from the same set of MIDI files for use in audio key finding. This enabled, for the first time, a performance comparison of symbolic key finding and audio key finding methods. It should be stressed, however, that because the audio material in the dataset was synthesized, the results of the audio key finding cannot be generalized to actual audio recordings containing the same pieces.

Prior to evaluation, a test set of 96 pieces were made available to the participants for testing and calibrating their algorithms. The performance evaluation criteria were established before the actual evaluation started. According to these the performance of an algorithm was determined by the percentage of correctly identified keys as well as closely related keys. In order not to severely penalize closely related key estimates the following fractional allocations were used: correct key, 1 point; perfect fifth, 0.5; relative major/minor, 0.3; parallel major/minor, 0.2 points. This was determined by the proposers of the contest at an early stage of the audio key finding contest proposal.

The audio dataset was reported to have two versions. Different synthesizers were used to generate the different versions - Winamp and Timidity. A percentage score was calculated for each version of the dataset taking into account the fractional allocations mentioned above. The composite percentage score was the average performance of the algorithms on the two datasets.

The algorithm explained in this paper performed as follows: Using the Winamp database, 1086 pieces were estimated correctly. Furthermore, an additional 36 estimates were perfect fifths of the correct key, 38 were relative major/minors and 17 were parallel major/minors. 75 of the estimated keys were considered unrelated. The percentage score for this database was 89.4 percent. Using the Timidity database, the algorithm found the correct key for 1089 pieces. For this database, an additional 42 estimates were perfect fifths of the correct key, 31 were relative major/minors and 18 were parallel major/minors. 72 of the estimated keys were considered unrelated. The percentage score for this database was 89.7 percent. The resulting composite percentage score was 89.55 percent.

This algorithm performed slightly better than the other algorithms in this evaluation exchange for the given dataset. The performances of the 7 participating algorithms ranged from 79.1 percent to 89.55 percent in their composite percentage scores.

4. Model II

This section describes a second model for key finding that operates on a representation with fewer dimensions compared to Model I. The motivation behind this is to find the optimal number of dimensions for a specific problem - key finding in this case - rather than to decrease the computational cost. Dimensionality reduction is performed on the data prior to a distance based calculation to determine the most probable key. The model takes in a parameter indicating the number of dimensions to be used in the process of key finding.

4.1 Dimensionality Reduction

Chroma based representations have been used extensively with great success in models that deal with tonal content analyses such as chord recognition, major/minor key detection and key finding. These chroma based representations are obtained by a fairly straightforward many-to-one mapping from the spectrum to a low-dimensional vector. This vector often has 12 elements - hence the name chroma mapping or chromagram - although the same calculation can be carried out for different vector sizes. Performing this mapping from many bins in the FFT to 12 bins can be viewed as a significant reduction in dimensionality but the question remains as to whether further reduction is possible. We explore the relationship between the number of dimensions used in the representation of tonal content and recognition performance in key finding using Model II.

In order to find the relationship between the number of dimensions and key finding accuracy, one could start from a spectral representation and perform dimensionality reduction. This leads to a compact tonal representation by preserving the cognitive distances between keys. Izmirli [8] demonstrated that cyclic distance patterns can be obtained through dimensionality reduction of raw spectral data pertaining to diatonic sets. In [2] a comparison of spectral and chroma representations of tonal content showed that chroma representations were more effective in correlation based key recognition. Therefore, we have chosen to start with the 12-dimensional PCP chroma representation and explore the effects of dimensionality reduction on the performance of key finding.

This model uses Principle Components Analysis (PCA) for performing dimensionality reduction. In PCA, the eigenvectors and eigenvalues of the covariance matrix give the rotation and scaling of the axes. This is based on maximization of the variance for each principle component. PCA automatically orders the principle components in order of importance and the variance of the data projected onto each principle component monotonically decreases by index. The first has the highest variance. As PCA performs rotation and scaling on the original data to obtain the transformed data it maintains all linear relationships of the data points.

4.2 Distance-Based Key Finding

In this model, the precomputed chroma templates and the summary chroma vectors calculated for Model I are used. For each summary chroma vector, PCA is applied to that vector and the 24 chroma templates. The 25 data points are then projected onto the new axes. A new PCA is calculated for every summary chroma point (instead of applying the same projection repeatedly) to ensure that the new data point contributes to the process. For a given number of dimensions, *m*, the chroma template point with the minimum Euclidean distance to the summary chroma point is found in the *m*-dimensional space. The key label of the nearest point is regarded as the key estimate. Here, *m* is the parameter to the model that determines how many dimensions of the transformed data will be used. The remaining components are ignored.

As in Model I, a confidence value is calculated for each chroma summary vector. In this case, the confidence is given by the proximity of the estimate to a template point compared to its proximity to the next nearest template point. The confidence value is given by the difference between the two distances divided by the nearest distance. Similar to Model I, the confidence values are used as weights in determining the key estimate.

5. Evaluation

The key finding accuracies of the two models were evaluated using an audio collection consisting of 152 classical pieces recorded from the naxos collection [18]. The first 30 seconds of each piece was processed. Pieces were chosen randomly among those with key information in their labels. The collection had music in all keys but the distribution was not uniform. The number of files in the same key ranged from 3 to 11. Works by the following composers were used: Albinoni, Albrechtsberger, Alkan, Bach, C.P.E. Bach, Beethoven, Bella, Brahms, Chopin, Clementi, Corelli, Dvorak, Grieg, Handel, Haydn, Hofmann, Kraus, Liszt, Mendelssohn, Mozart, Pachelbel, Paganini, Prokofiev, Rachmaninov, Scarlatti, Schubert, Scriabin, Telemann, Tchaikovsky, Vivaldi.

Table 1. Raw and composite scores for Model I and Model II.

Model	Raw Score (%)	Composite Score (%)	Variance (%)
Model I	86.2	88.9	-
Model II – 12 components	85.5	88.4	100.0
Model II – 8 components	84.2	87.2	98.7
Model II – 6 components	85.5	88.7	94.8
Model II – 4 components	78.9	83.7	87.6
Model II – 3 components	76.3	81.9	79.6
Model II – 2 components	32.9	44.9	71.1

Model I was run on this audio set to obtain a reference for key finding performance. Model II was run on the same audio set with different parameters. Table 1 gives the results of the evaluation. The raw score is the percentage of the correctly identified keys. The composite score reflects the weighted contributions due to closely related keys as those used in the MIREX 2005 audio key finding evaluation explained in Section 3.5.

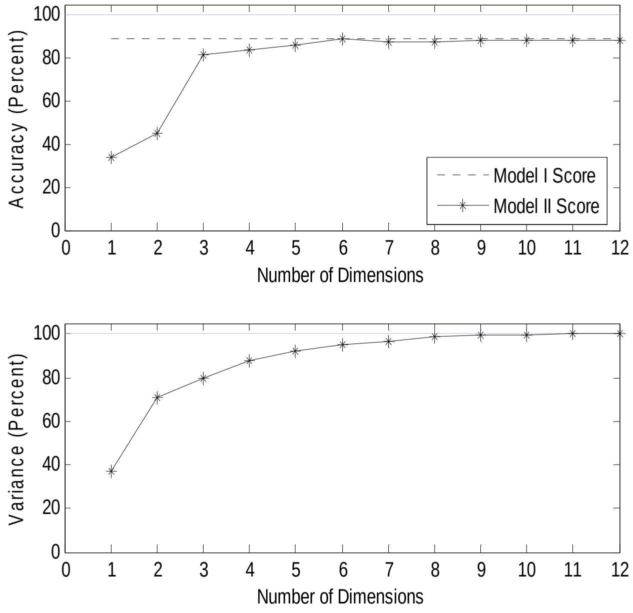

Figure 2. Key finding accuracy (top) and the percentage of variance accounted for in the projected data (bottom) versus number of dimensions.

The top plot in Figure 2 shows the key finding accuracy versus the number of dimensions. This is indicated with the '*' symbols and the continuous line. The dashed line shows the performance of Model I for comparison. The bottom plot shows the percentage of total variance accounted for in the portion of the projected data used as a function of the number of dimensions. This percentage can be viewed as the explanatory power of the data being used with the selected number of components. It should be noted that although there is not much difference in the percentage variance between 2 and 3 components, the accuracy is drastically poorer with 2 components. Again, for accuracy, it can be seen that the performance remains almost constant from 12 dimensions down to 6. This suggests that a tonal representation with at least 6 dimensions can be used to capture the essential portion of information. It should be noted that accuracy plunges going from 3 dimensions down to 2 dimensions. This may be interpreted as a corroboration of toroidal models of tonal space over planar models (see for example [19]).

6. Conclusion

Two audio key finding models that produce successful results are presented and a comparative evaluation of the two is given. Model I that performed well in the MIREX 2005 audio key finding evaluation is used as reference for performance evaluation of Model II. The second model implements key finding using a low-dimensional space. This model is run with a range of parameters in order to determine the effect of dimensionality for tonal representation in key finding. It was found that the key finding performance did not significantly change from 12 dimensions to 6 dimensions, dropped slighted between 5 to 3 dimensions, and dropped significantly using 2 dimensions. It can be concluded that excellent performance is obtained with 6 and higher dimensions and 3 to 5 dimensions provide acceptable performance.

Future work will concentrate on further analysis of properties of the new low-dimensional space and explore mappings from spectral and chroma representations to low-dimensional features. Segmentation with respect to modulation points constitutes a direct application of these models. Chord boundary detection combined with key context may be used for tonal analysis. A front-end for tuning adjustment will prove useful for these models to cater to arbitrary reference frequencies.

References

[1] T. Fujishima, "Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music," *Proceedings of the International Computer Music Conference*, Beijing, China, pp. 464–467, 1999.

[2] Ö. İzmirli, "Template Based Key Finding From Audio," *Proceedings of the International Computer Music Conference* (ICMC2005), pp. 211-214, Barcelona, Spain, 2005.

[3] E. Gómez, "Tonal Description of Polyphonic Audio for Music Content Processing," *INFORMS Journal on Computing, Special Cluster on Computation in Music,* Vol.18 .3, 2006.

[4] G. Cabral, J.-P. Briot, F. Pachet, "Impact of Distance in Pitch Class Profile Computation," *Proceedings of the 10th Brazilian Symposium on Computer Music* (SBCM2005), Belo Horizonte, Brazil, 2005.

[5] S. Pauws, "Musical Key Extraction from Audio," *Proceedings of the Fifth International Conference on Music Information Retrieval*, pp. 96-99, Barcelona, Spain, 2004.

[6] H. Purwins, B. Blankertz, G. Dornhege and K. Obermayer, "Scale Degree Profiles from Audio Investigated with Machine Learning Techniques," *Audio Engineering Society 116th Convention*, Berlin, 2004.

[7] E. Gómez and P. Herrera, "Estimating the Tonality of Polyphonic Audio Files: Cognitive versus Machine Learning Modelling Strategies", *Proceedings of the Fifth International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[8] Ö. İzmirli, "The Emergence of Cyclic Patterns among Spectra of Diatonic Sets," to appear *in Computing in Musicology, MIT Press,* Vol. 15.

[9] A. Sheh and D. P. W. Ellis, "Chord Segmentation and Recognition using EM-Trained Hidden Markov Models," *Proceedings of the International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, 2003.

[10] W. Chai and B. Vercoe, "Detection of Key Change in Classical Piano Music," *Proceedings of the International Conference on Music Information Retrieval* (ISMIR2005), pp. 468-474, London, UK, 2005

[11] C.-H. Chuan and E. Chew, "Fuzzy Analysis in Pitch Class Determination for Polyphonic Audio Key Finding," *Proc. of the International Conference on Music Information Retrieval* (ISMIR2005), pp. 296-303, London, UK, 2005.

[12] Ö. İzmirli and S. Bilgen, "A Model for Tonal Context Time Course Calculation from Acoustical Input," *Journal of New Music Research,* Vol.25, No. 3, pp. 276-288, 1996.

[13] H. Purwins, B. Blankertz, and K. Obermayer, "Constant Q Profiles for Tracking Modulations in Audio Data," *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[14] Y. Zhu, M. S. Kankanhalli, and S. Gao, "Music Key Detection for Musical Audio," *Proceedings of the 11th International Multimedia Modelling Conference*, Melbourne, Australia, 2005.

[15] C. Krumhansl, *Cognitive Foundations of Musical Pitch*, Oxford University Press, New York, 1990.

[16] D. Temperley, *The Cognition of Basic Musical Structures*, Cambridge, MA: MIT Press, 2001.

[17] http://www.music-ir.org/evaluation/mirex-results/

[18] http://www.naxos.com.

[19] F. Lerdahl, *Tonal Pitch Space*, New York: Oxford University Press, 2001.