TONAL SIMILARITY FROM AUDIO USING A TEMPLATE BASED ATTRACTOR MODEL

Özgür İzmirli
Center for Arts and Technology
Connecticut College
270 Mohegan Ave.
New London CT, USA
oizm@conncoll.edu

ABSTRACT

A model that calculates similarity of tonal evolution among pieces in an audio database is presented. The model employs a template based key finding algorithm. This algorithm is used in a sliding window fashion to obtain a sequence of tonal center estimates that delineate the trajectory of tonal evolution in tonal space. A chroma based representation is used to capture tonality information. Templates are formed from instrument sounds weighted according to pitch distribution profiles. For each window in the input audio, the chroma based representation is interpreted with respect to the precalculated templates that serve as attractor points in tonal space. This leads to a discretization in both time and tonal space making the output representation compact. Local and global variations in tempo are accounted for using dynamic time warping that employs a special type of music theoretical distance measure. Evaluation is given in two stages. The first is evaluation of the key finding model to assess its performance in key finding for raw audio input. The second is based on cross validation testing for pieces that have multiple performances in the database to determine the success of recall by distance.

Keywords: Tonal similarity, key finding, dynamic time warping, tonal space.

1 INTRODUCTION

In the field of MIR, the importance of time series representations is well recognized since listeners can only experience music through time. Recently, in this field, many methods dealing with similarity of time series have been either revisited and reinterpreted or new approaches have been proposed. These methods focus on factors such as efficiency of the representation, algorithm complexity and processing load. Selection of representative features and a resulting efficient and effective representation are important factors in model design. This paper, introduces a method for similarity calculation of an aspect of music cognition: tonal evolution. The representation used for tonal evolution is a sequence of symbols that enables application of fast and efficient string processing algorithms. This can be viewed in contrast to other methods dealing with similarity that generally use rich features and consequently have higher processing demands. The method presented here first finds a sequence of position estimates in tonal space and then uses the time series to calculate similarity by warping one sequence onto another.

This paper explores the problem of similarity from a tonality standpoint. The method utilizes a template based key finding model to estimate the position in tonal space at regular intervals throughout a piece. The sequence of symbols representing the tonal evolution is used in similarity calculations across pieces in a database. In music, this kind of similarity is understood as a more abstract and high-level similarity when compared to similarity of more direct musical attributes such as rhythm or melody. Nevertheless, in the context of Western tonal music the induction of tonality is central to the interpretation of music. The compositional process addresses the interplay between the elements of time and pitch inducing the sense of tonality. A tonal center can be defined as the most stable pitch in a fragment of music sometimes also referred to as the tonic. Tonality is ubiquitous and most listeners musically trained or untrained can identify the most stable pitch while listening to tonal music. Furthermore, this process is continuous and remains in action throughout the listening experience. As a musical work unfolds, the stable pitch might change as a result of the music modulating from one key into another. In simple terms, the mode of the musical scale together and the tonic signify the key of a piece. The main key can also be viewed as the global key. Musical works in the tonal tradition generally start and end with the global key, moving through multiple other keys throughout the piece. On the other hand, a localized key estimate can be viewed as an estimate of the tonal center given only a fragment of a larger musical work. In this paper, tonal evolution is represented by a sequence of symbols obtained by the application of a localized key finding model on adjacent fragments of music.

To estimate the key from an audio recording, one might look at the beginning or end of the piece and develop heuristics to arrive at a decision. A more complicated problem is the calculation of tonality evolution over time which has been addressed in a limited number

of works. This comes closer to harmonic analysis where chords or at least tonal regions need to be identified as the piece unfolds. The evolution of the tonal center characterizes the piece in an abstract and general way. Several levels that would be useful to music information retrieval can be identified. First, key finding would group an entire database into 12 or 24 classes (or major-minor). Second, identification of modulations would give more information about pieces and lead to a further division of the database based on a more detailed representation that contains a sequence of keys. Third, and most useful level, would extract information regarding tonal evolution so as to be used in applications ranging from functional analysis to segmentation into musical sections and even to transcription.

The methods dealing with similarity that work directly from audio generally use features with many dimensions such as the 12 dimensional chromagram or Mel-frequency Cepstral Coefficients. The method presented here uses an additional step to further reduce the dimensionality of the representation prior to similarity calculation.

The organization of the remainder of the paper is as follows: Section 2 outlines related work in key finding, feature extraction from audio, similarity and time alignment. Section 3 explains the procedure for obtaining templates which represent attractor points in tonal space. Section 4 discusses the determination of localized key estimates which are found with respect to the templates. The result of this stage is a sequence of symbols representing a trajectory in tonal space. Section 5 explains the alignment process that deals with tempo differences between the pieces being compared. An evaluation of the method is given in Section 6.

2 RELATED WORK

In this section an outline of related work in several areas is given. 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. Chroma based representations have been used in key finding (İzmirli 2005; Gómez and Herrera 2004; Pauws 2004), discovering similarity and repetition in audio recordings (Bartsch and Wakefield, 2001) and chord segmentation, recognition and alignment in audio (Sheh and Ellis, 2003). Fujishima (1999) originally proposed the Pitch Class Profile (PCP) for use in chord recognition. This chroma based spectral representation is widely used because it effectively summarizes chroma information and harmonic structure in the spectrum using a manageable number of dimensions. However, the mapping is not unique, octave information is ambiguous and fine spectral detail is lost as a result of this mapping.

Many approaches to extracting tonal center information have been reported in the literature. Leman (1992) proposed a method inspired by cognition that uses an ear model front-end for determination of tonal context and tone centers. İzmirli and Bilgen (1996) reported on a model that has a pitch-class note recognition front-end followed by a stage that consists of leaky integrators to model recency effects and decay. In this model, as musical events are encountered, leaky integrators are charged according to respective strengths of pitch events. Huron and Parncutt (1993) use a psychoacoustic model of pitch perception that employs echoic memory and pitch salience to model key perception. Chuan and Chew's model (2005) estimates pitch strength using peaks in the spectrum which are then used by the Spiral Array model to estimate key. Purwins, Blankertz and Obermayer (2001) proposed a model for tonal center and modulation tracking which collapses the spectrum into constant Q (CQ) profiles and calculates distances using a fuzzy distance measure between the profiles and reference CQ sets. Gómez and Herrera (2004) presented a comparison of cognition-inspired models based on Krumhansl's method and feature-based machine learning methods for key finding from polyphonic audio. One of the features they use is the Harmonic Pitch Class Profile which is a specialized version of PCP that uses the peaks in the spectrum. Pauws' model (2004) uses an auditory perception inspired front-end to compute a chromagram which is then used to compute the correlations with the Krumhansl and Kessler profiles (1982). Zhu, Kankanhalli and Gao (2005) first find the tuning frequency of the input, perform partial tracking, apply consonance filtering, obtain a pitch profile, and determine the scale root and key separately.

Similarity within a single audio recording has been subject to much research. Finding thumbnails or repeating sections are of interest for systems that perform automatic summarization. Dannenberg and Hu (2002) describe and compare three methods that find repetition of segments within musical pieces. Cooper and Foote (2002) describe a method to determine the most representative segment in a piece by maximizing the average segment similarity over the piece. Bartsch and Wakefield (2001) perform similarity analysis on chroma based representations of audio to identify chorus sections. İzmirli (2002) uses spectra of diatonic collections, as references, to calculate tonal context vectors indicating relative strengths of tonal centers which in turn are used to calculate similarity of tonal evolution in fragments within and across audio recordings in a database.

Work related to processing of time series information generally deals with time alignment, segmentation and sequence recognition. Hu, Dannenberg and Tzanetakis (2003) describe a method to align polyphonic audio to symbolic score information. They use a chroma based representation and align the chroma vectors obtained from the query of the polyphonic input to those obtained from symbolic information. Work by Sheh and Ellis (2003) demonstrates chord recognition from music recordings. They use an HMM model for sequence recognition and report that PCP features are more effective than cepstral coefficients. Although the octave is usually divided into 12 they use a higher resolution PCP by dividing the octave into 24. Yoshioka et al. (2004) re-

port on a system for chord recognition that simultaneously detects chords and chord boundaries in the input audio. Adams et al. (2004) describe dynamic alignment procedures for various time series representations of sung queries.

3 TEMPLATES

A template based key finding model is described in Izmirli (2005). This model uses short fragments from the beginnings of polyphonic audio recordings that contain classical music including symphonic, vocal, solo and ensemble recordings. The model has been found to produce 86% correct labelling of the key using a database of 85 recordings. In the mentioned work, various spectral representations and profiles are compared with one another. The model operates on the assumption that a piece starts in the key that appears in its label designated by the composer. Given the viability of the model, here, we choose to utilize it in a sliding window fashion to estimate the position in tonal space at a given time in the piece. The model that results in the best performance will be described here. This will constitute the basis for the estimation of position in tonal space. In this paper however, the model is used with a different parameter selection to make it suitable for the current purpose.

Pitch distribution profiles may be used to represent tonal hierarchies in music. Krumhansl (1990) suggested that tonal hierarchies for Western tonal music could be represented by the probe tone profiles found experimentally in an earlier study (Krumhansl and Kessler, 1982). 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 model the way listeners arrive at a sense of key. Many key finding models rely on this assumption and several extensions have been proposed. In one such extension, beside other additions, Temperley (2001) has proposed a pitch distribution profile. We utilize this profile in combination with a diatonic profile as this combination results in the best performance. 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.

Templates are obtained using recordings from monophonic instrument sounds. These sounds, for example, could be piano sounds from the McGill Master Samples or from the University of Iowa Musical Instrument Samples. Templates represent a prototype spectrum according to a distribution determined by the chosen profile. The sounds are low pass filtered and then sampled at 5512.5 Hz. The analysis is carried out using 50% overlapping 2048-point FFTs with a Hann window. Analysis frequency range is taken to be from 50Hz to 2000 Hz. The spectrum of an individual monophonic sound with index i, X_i, is computed by averaging windows that have significant energy over the duration of each sound and then scaling the average spectrum 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.

Chroma	Diatonic Major D_M	Diatonic Minor D_m	Temperley Major T_M	Temperley Minor T_m
0	1	1	5.0	5.0
1	0	0	2.0	2.0
2	1	1	3.5	3.5
3	0	1	2.0	4.5
4	1	0	4.5	2.0
5	1	1	4.0	4.0
6	0	0	2.0	2.0
7	1	1	4.5	4.5
8	0	1	2.0	3.5
9	1	0	3.5	2.0
10	0	0	1.5	1.5
11	1	1	4.0	4.0

Table 1. Two profiles used in this study: major and minor profiles for Temperley and diatonic.

Using the spectra obtained for each individual note, templates are calculated by weighted sums. A template for a certain mode and chroma value is the sum of X_i weighted by the profile element that has the corresponding chroma value. A template is calculated for each mode-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.

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

X_i denotes the averaged amplitude spectra of the sound corresponding to note i. $P_e(k)$ is the profile weight as given in Table 1, where e denotes the mode (M:major or m:minor) and k denotes the chroma. In this work, the profile is given by the product of the diatonic and Temperley profiles: $P_e(k)=D_e(k)T_e(k)$. 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 weighted magnitudes of the FFT bins over all regions that have the same chroma value.

4 ESTIMATION OF POSITION IN TONAL SPACE

Templates can be viewed as attractor or focal points in tonal space that represent the ideal locations of tonal centers. Once the profiles and scales are chosen and templates are formed, they become part of the model to which incoming information is compared. Summary vectors are obtained from the raw audio input using the same method to obtain the templates with two exceptions: the first is that each summary vector is obtained from a window of fixed duration (instead of the entire sound) as the window is slid through the entire audio input. Note that this window spans a much larger duration compared to the FFT window. In this work a window size of 2.5 seconds has been utilized. This has been determined experimentally to balance the averaging over time and sluggishness of the estimation. A longer window covers more notes and tends to be more stable in the estimation whereas a shorter window will make the estimation more adaptive. The window is noncausal and has a time registration point at the center. The hop size is 35 percent of window length. The second difference is that at each hop a new summary vector is calculated and compared to the templates as described below.

For each window the position in tonal space is estimated by computing correlation coefficients between the summary vector and all 24 spectral templates, and picking the one with the maximum value. The index of the template with the maximum correlation is then recorded for that time step. For the entire audio, a sequence of indices which fall in the range of 1-24 are calculated and recorded. This results in a sequence represented by $S_k=(s_{1,k},s_{2,k},s_{3,k},...s_{N,k})$ where $s_{n,k}$ represents the mode and chroma value (tonic) estimate for the n'th window in audio file k.

5 ALIGNMENT

The extracted sequence of indices S_k, represents a sampling of the tonal evolution in both time and tonal space. The resulting discrete representation can be used to efficiently compare the similarity of tonal evolution between pieces. Given two performances of the same piece, direct similarity comparison using, for example, the Euclidean distance is not possible due to tempo differences in the performances that lead to misalignment of the two sequences. For this reason, Euclidean distance and similar distance measures that do not allow warping of the source sequence toward the target sequence fail to serve as viable indicators of similarity. We therefore use dynamic time warping to reduce the effects of local tempo differences between performances.

As explained above, the duration of analysis used to determine elements of S is on the order of seconds. If the model had been operating with a shorter window, say at the note level, then grouping would have been necessary, for example, to convert arpeggios into chords or obtain roman numeral analysis from fixed size time spans. At this level of analysis each symbol is calculated from a window that spans a sizeable duration which performs the necessary averaging. Therefore, we can use a method that assumes monotonic unfolding of both sequences to find an optimal warping path and a resulting distance.

The sequence S consists of elements that represent mode and tonic information. This means that the dynamic time warping algorithm cannot use a geometric distance measure directly on the values themselves. As such, an absolute value of the difference could not be used due to the more complicated distance relationships between the indices. For example, index 4 represents C major and index 5 represents Db major (or C# major). Although the numerical difference between these two keys is 1 the distance in tonal space should be one of the maximal distances. Even using a simple circle-of-fifths distance Db should be 5 steps away while F would be 1 step away from C. To perform the dynamic time warping a distance measure needs to be defined that models distances in tonal space. Lerdahl's regional space (2001) is a tonal space in which these distances can be calculated. The regional space is created by combining the circle-of-fifths with the parallel and relative major-minor cycle. Lerdahl defines this generalized tonal pitch space to calculate distances between chords when pitches are either chosen from a single diatonic collection or from a different one as a result of a shift in the diatonic set. In this work we approximate the tonal distance between elements in the sequence S using Lerdahl's regional distance. Table 2 shows the distances in tabular form as given in (Lerdahl, 2001). Readers are referred to the original source for the geometrical representation and history of tonal pitch space.

Region	Distance	Region	Distance
1	0	13	7
2	23	14	23
3	14	15	10
4	14	16	21
5	16	17	9
6	7	18	14
7	30	19	21
8	7	20	14
9	16	21	23
10	14	22	7
11	14	23	21
12	23	24	16

Table 2. Lerdahl's regional distances. Regions are given in semitones with respect to region 1.

Given two sequences S_k and S_m we find the warping path $R=(r_1, r_2, r_3,\ldots r_N)$ with N being the length of the path and $r_n=(i,j)$ holding the association between element i in sequence S_k and element j in sequence S_m. Dynamic time warping is implemented using the recur-

sion given in Equation 2. The conventional path constraint that chooses between a single step of the diagonal, vertical or horizontal moves was initially tested. This led to many successive vertical or horizontal moves in the optimal path when the sequences were uncorrelated. Therefore, another path constraint was used to prevent two non-diagonal moves to occur in sequence.

$$D(i,j) = d(i,j) + \min \begin{cases} D(i-1,j-1) \\ D(i-1,j-2) + d(i,j-1) \\ D(i-2,j-1) + d(i-1,j) \end{cases} \quad (2)$$

D is the global distance up to the point in the recursion, with D(1,1)=d(1,1) as the initial condition. d(i,j) is Lerdahl's distance as given in Table 2 between element i in sequence S_k and element j in sequence S_m. After the dynamic programming algorithm is run, the minimum global distance found is divided by the length of the trace-back path to eliminate the dependence on duration of the recordings. Although the local constraint given in Equation 2 can be interpreted as a global constraint that prevents some points in the grid from being reached, this constraint does not lead to any speed-up until explicitly stated as a global constraint and implemented in the recursion. We therefore use an Itakura parallelogram to prevent extreme warping and to attain speedup.

6 EVALUATION

The key finding model described in this paper scored 86 percent correct labelling on a set of 85 general polyphonic audio files containing short fragments of classical music recordings. The fragments were taken from the beginnings of audio recordings that contained symphonic, vocal, solo and ensemble music. A single window of duration 7.5 seconds was used. The works in the database collection were chosen to approximately have uniform distribution across the 24 keys. The key labelling in the titles of the pieces were used as ground truth. The results showed that the output of the key finding model was able to produce a good estimate of the key in a fragment of audio and could be used as a front-end to higher level processing.

This database was composed of the beginnings of the pieces and did not contain performances of the same piece by different performers. Each file contained approximately 1 minute into the piece. Next, different performances of 5 pieces that were already in the database were recorded and added to the collection. All 5 queries using the new pieces returned the correct files as being most similar.

Next another database of 125 recordings of Chopin Mazurkas by Vladimir Ashkenazy, Ignaz Friedman, Arthur Rubinstein, Vladimir Sofronitsky and Jean-Marc Luisada were used (some historical recordings were very noisy.) Only 12 of these recordings were unique and the remaining were all played by multiple pianists. The sliding window key finding model and the dynamic time warping algorithm were tested using this set. Cross validation was applied by using each piece as a query and testing it against the rest of the database. For this, a similarity matrix, M, was constructed based on the minimum cost path found by the dynamic time warping algorithm between all pairs of recordings. The diagonal elements were not calculated and were assigned large values to prevent self matches. The most similar tonal evolution for a piece with index k, was given by the index of the minimum element in row k in matrix M.

The first measure of performance was the retrieval accuracy of the most similar item for a piece. If in response to a query, the piece with the minimum distance to that query had the same name then it was considered a successful retrieval. The measure was calculated as the sum of all successful recalls divided by the total number of pieces that had multiple versions. For the 125 pieces this measure yielded 88.8%. The second measure looked for a successful retrieval in the first two similar pieces. That is, two pieces with the smallest distances to the query were checked to see if any one of them was a successful retrieval. Again, only pieces that had multiple versions were considered. This measure yielded 100%. A third measure was used to understand the overall performance by calculating the average ratio of all successful recalls in the top 5 to the possible successful recalls. For example, if a piece had 4 different versions and only 3 were retrieved in the first 5 then the ratio would be 3/4. The ratios for all pieces were averaged. This yielded 92.5%.

7 CONCLUSIONS

The method presented here has been shown to produce an efficient representation of tonal evolution in the form of a time series. The time series is a one-dimensional sequence of symbols representing tonal centers in tonal space at discrete points in time. A sliding window version of a key finding model has been shown to work in this context with encouraging results. The alignment of the sequences obtained from recordings in the database are performed using dynamic time warping with a tonal space distance measure. To demonstrate the effectiveness of this representation in finding a piece that has the most similar tonal evolution in a database, a similarity matrix is constructed. Finding the most similar N pieces is a matter of finding the N smallest distances in a row of the similarity matrix. The resulting performance indicates that the method shows potential for use in MIR applications.

The key finding model is not meant to be employed as a music analysis tool such as a chord recognizer. It should be regarded as a tractable approach to tonal evolution modeling and one that captures the essential gist of tonal structure of a musical work as it unfolds over time. The model employs a structural approach to tonality by operating solely on pitch information, disregarding note order and only indirectly using information relating to time structure of the input. For future work, incorporating time structure will be considered. Of particular interest is finding the longest common sequences within elements in the database. This will serve to per-

form intelligent segmentations and find repetitions. Another direction will be to adapt the method to use a relative representation to account for transpositions.

REFERENCES

Adams, N. H., Bartsch, M.A., Shifrin, J. B. and Wakefield, G. H. (2004) Time Series Alignment For Music Information Retrieval. *Proceedings of the Fifth International Conference on Music Information Retrieval*, Barcelona, Spain.

Bartsch, M. A. and Wakefield, G. H. (2001) To Catch a Chorus: Using Chroma-based Representations for Audio. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY.

Ching-Hua, C. and Elaine, C. (2005) Polyphonic Audio Key-Finding Using the Spiral Array CEG Algorithm. *Proceedings of the International Conference on Multimedia and Expo* (ICME), Amsterdam, Netherlands, July 6-8.

Cooper M. and Foote, J. (2002) Automatic Music Summarization via Similarity Analysis. *Proceedings of the Third International Conference on Music Information Retrieval*, Paris: IRCAM.

Dannenberg, R.B. and Hu, N. (2002) Pattern Discovery Techniques for Music Audio. *Proceedings of the Third International Conference on Music Information Retrieval*, Paris: IRCAM.

Fujishima, T. (1999) Realtime chord recognition of musical sound: A system using common lisp music. *Proceedings of the International Computer Music Conference*, Beijing, China, 464–467.

Gómez, E. and Herrera, P. (2004) 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.

Hu, N., Dannenberg, R.B. and Tzanetakis, G. (2003) Polyphonic Audio Matching and Alignment for Music Retrieval. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA.

Huron, D. and Parncutt R. (1993) An Improved Model of Tonality Perception Incorporating Pitch Salience and Echoic Memory. *Psychomusicology*, 12(2), 154-171.

Leman, M. (1992) Tonal Context by Pattern Integration Over Time. *In D. Baggi (Ed.), Readings in Computer-Generated Music*, Los Altimos, CA: IEEE Computer Society Press. pp. 117-137.

Lerdahl, F. (2001) *Tonal Pitch Space*, New York: Oxford University Press.

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

İzmirli, Ö. (2002) Determination of Tonal Similarity Based on Spectral Diatonic Bases**.** in electronic proceedings in the *Informatics Report Series*, University of Edinburgh, II International Conference on Music and Artificial Intelligence (ICMAI02), September 12-14, Edinburgh, United Kingdom.

İzmirli, Ö. (2005) Template Based Key Finding From Audio. *Proceedings of the International Computer Music Conference* (ICMC), Barcelona, Spain.

Krumhansl, C. and Kessler, E. (1982) Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89, 334–368.

Krumhansl, C. (1990) *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York.

Pauws, S. (2004) Musical Key Extraction from Audio. *Proceedings of the Fifth International Conference on Music Information Retrieval*, Barcelona, Spain.

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

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

Temperley, D. (2001) *The Cognition of Basic Musical Structures*, Cambridge, MA: MIT Press.

Yoshioka, T., Kitahara, T., Komatani, K., Ogata, T. and Okuno, H. G. (2004) Automatic Chord Transcription with Concurrent Recognition of Chord Symbols and Boundaries. *Proceedings of the Fifth International Conference on Music Information Retrieval*, Barcelona, Spain.

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