A PERPLEXITY BASED COVER SONG MATCHING SYSTEM FOR SHORT LENGTH QUERIES

Erdem Unal[1] Elaine Chew[2] Panayiotis Georgiou[3] Shrikanth S. Narayanan[3]

[1]TÜBİTAK BİLGEM
[2]Queen Mary, University of London
[3]University of Southern California
[1]unal@uekae.tubitak.gov.tr
[2]elaine.chew@eecs.qmul.ac.uk
[3]{georgiou,shri}@sipi.usc.edu

ABSTRACT

A music retrieval system that matches a short length music query with its variations in a database is proposed. In order to avoid the negative effects of different orchestration and performance style and tempo on transcription and matching, a mid-level representation schema and a tonal modeling approach is used. The mid-level representation approach transcribes the music pieces into a sequence of music tags corresponding to major and minor triad labels. From the transcribed sequence, *n*-gram models are built to statistically represent the harmonic progression. For retrieval, a perplexity based similarity score is calculated between each *n*-gram in the database and that for the query. The retrieval performance of the system is presented for a dataset of 2000 classical music pieces modeled using *n*-grams of sizes 2 through 6. We observe improvements in retrieval performance with increasing query length and *n*-gram order. The improvement converges to a little over one for all query lengths tested when *n* reaches 6.

1. INTRODUCTION

Due to advances in computer and network technologies, the development of efficient multimedia data storage and retrieval applications have received much attention in recent years. In the music domain, motivations for such systems can vary from industry objectives such as royalty rights management to individual use such as personal database organization, music preference list creation, etc. Due to the wide range of expressive and instrumental variations possible in music pieces, in order for such systems to have the necessary performance reliability as to be useful in the industrial domain, music variation matching must be addressed. A number of challenges such as feature extraction, representation, tempo and key variability, need to be handled with high precision in order to achieve reasonable performances.

To eliminate the kinds of differences caused by expressive variations or instrumental arrangements of the same music piece, researchers have focused on accurately extracting the types of musical content in which such variations have minimal or no effect.

A considerable amount of research focused on the transcription of music signal to MIDI or piano roll representation for accurate understanding of the note sequence of the music. Numerous researchers have modeled sound events with known machine learning techniques, in order to detect musical notes and their onset and offset times [1,2,3,4 and 5]. Their results are promising, although not accurate enough to provide an extension to a general solution for music variations matching.

Since accurate transcription of multi channel audio is not easy, a mid level representation of music is desired. Recent research attempts in [6,7 and 8] showed that different representation techniques such as extracting the salient melody or a chord progression from the music piece could be a feasible solutions for polyphonic representation since harmonic structure tends note to change dramatically with expressive and instrumental deviations.

On the other hand, some researchers focused on extracting fingerprints that carry information about the acoustic feature distribution of the music piece over time. [9 and 10] used chroma based features to directly represent music pieces, without labeling and used simple cross correlation of chroma vectors for measuring similarity. Kim also adopted delta features that represent general movement in the harmonic structure of music pieces for more accurate representation and retrieval [11].

Pickens et. al [13] used existing polyphonic transcription systems in the literature to abstract note features from music audio. The transcription was then mapped to the harmonic domain. A bi-gram (2-gram) representation, namely a 24 × 24 triad (three-note chord) transition matrix was used to represent both the query and the music pieces in the database. A distance metric between an input transition matrix and the transition matrices available in the database was calculated to determine similarity.

Our study differs from other researchers' who use some kind of mid level representation in the similarity metric we use, and in that we use a sliding window approach in our transcription independent of the exact locations of note onsets and offsets. While our strategy loses note level details in the audio, it makes our representation more robust to

note transcription errors. In contrast to the retrieval methods reported in [12 and 13] we tested our model on not only bi grams but also higher order *n*-grams, for *n* up to and including 6, and observed a major boost in the retrieval performance with increasing Markov chain order.

In later studies, Lavrenko & Pickens [14] used random fields to model polyphonic music pieces from MIDI files. Using random fields, they automatically induced new high level features from the music pieces, such as consonant and dissonant chords, progressions and repetitions, to efficiently model polyphonic music information.

The F-measure, Correct Retrieval Accuracy, and Mean Reciprocal Rank are used to measure the performance of the systems available in the literature. The reported results vary with respect to the database selected, its size and the complexity of the variations available. Since the algorithms used are generally computationally expensive, the experimental databases tend not to be larger than a couple of thousand songs. For a more detailed overview of the systems available in the literature, please refer to [18].

Most systems, including the ones described above, were designed assuming the availability of the entire query and target songs from beginning to end. To our knowledge, no tests were reported when only short length queries are present. In this work, a mid level tonal representation of audio and a statistical tonal modeling method for performing retrieval of short length audio queries is proposed.

In order to ensure robust transcription against musical variations, a 3 dimensional Tonal Space (TS), a toroidal version of the Spiral Array model [15] is used. The details are explained in Section 2. 12 dimensional Pitch Class Profile (PCP) features are mapped onto the TS and a *centroid* (center of weight) is calculated in order to find the representative position of each audio frame in 3D space. A 1-nearest neighborhood classifier is used for identifying the *centroid*s of each frame with respect to triad chord classes. A key and tempo invariant time series of triad chord labels are then acquired, from which we derive *n*-gram representations of each music piece in the database. The similarity between the extracted triad series and the *n*-gram models is calculated using the perplexity measure. The flowchart of the proposed system can be seen in Figure 1. The paper concludes with the explanation of the experimental setup, the results and the discussion on future work.

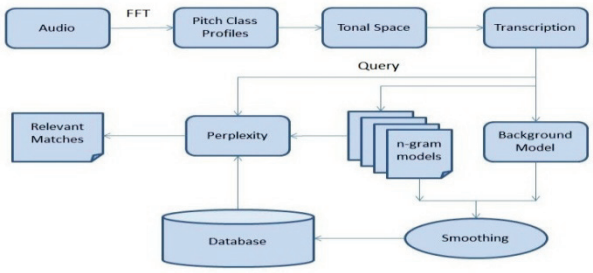

Figure 1. Flowchart for the proposed system.

2. TONAL MUSIC SPACE

There exists an illustrious history of mathematical and music theoretic work on geometric modeling of tonal relationships between pitches, intervals, chords, and keys. A review of these models can be found in [16].

We use a toroidal version of the Spiral Array for a number of reasons. We are interested in a flexible tonal representation that combines different tonal features in the same space. The Spiral Array clusters tonal objects that are harmonically close; this is especially important for robust analysis of audio without exact transcription.

The model consists of a series of nested helices in three-dimensional space. The outermost spiral consists of pitch classes that form the line or circle of fifths. Pitch classes are placed at each quarter turn of the spiral, so that vertically aligned pitch classes are a major third apart. This network of pitches is identical to the neo-Riemannian *tonnetz* shown in Figure 2. Pitch classes that are in the same triads are closely clustered, as are those that are in the same key. Chord representations are generated as weighted combinations, a kind of centroid, of their component pitch classes, and key representations are constructed from their I, IV, and V chords. The details and applications of the Spiral Array model are explained in [15][17].

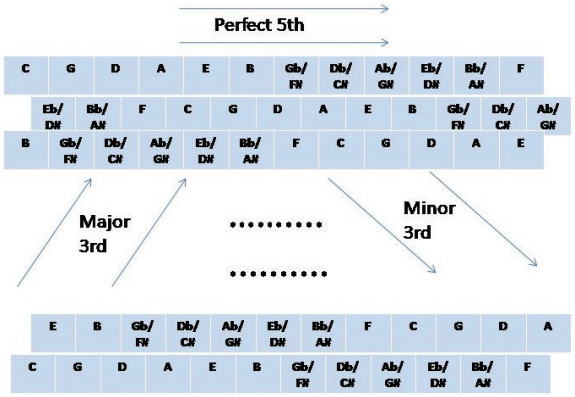

Figure 2. The tonnetz. Perfect 5th, *Major* 3rd and *Minor* 3rd distances

The Spiral Array model assumes a cylindrical form to preserve enharmonic spellings. In contrast, we wrap the model into a torus so as to ignore pitch spelling. The resulting pitch class torus is shown in Figure 3. The 24 chord representations are then defined by constructing the triangle outlined by each chord's root, fifth, and third, and calculating the centroid of these vertex points. A chord representation is illustrated in Figure 3. While the toroid model no longer has the same kinds of symmetries and invariance in the cylindrical model, the chord and key regions remain

sufficiently distinct for geometric discrimination between different chords.

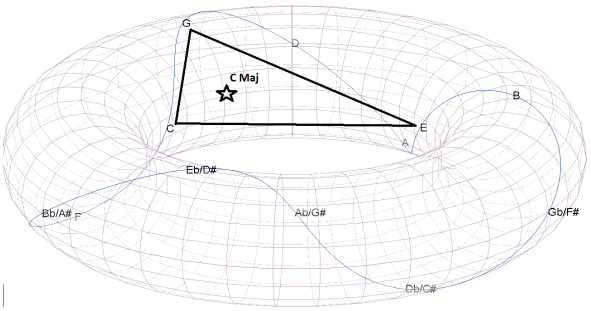

Figure 3. Tonal Space: positions of the 12 pitch classes and construction of the C Maj triad chord using C, G and E.

3. FEATURE EXTRACTION

As discussed earlier, to overcome the effects of incorrect transcription, we use a mid level transcription approach for the transcription task. The goal is to accurately label each frame of music audio with *major* or *minor* triad chords. For this, we use the tonal space described in Section 2. We now present our feature extraction process. This process is outlined in the top row of boxes in Figure 1.

Audio Input Frames: 250 ms audio frames with 90% overlap is used. A large window with a wide margin of overlap is preferred because our goal is to track the general harmonic movement and not instantaneous local changes that would be expected to be sensitive to variations in instruments and expression and thus pose problems for the retrieval system's similarity calculations.

Pitch Class Profile: 12 dimensional Pitch Class Profile (PCP) features are collected from each audio frame. The pitch classes extracted range from A0 (27.5 Hz) to A7 (3520 Hz). From the PCP's, the note weights are mapped to pitch class positions in the tonal space, and a centroid is calculated in 3D space as shown in Fig 4 (red star).

Chord Labels: The centroid derived in the fashion described above represents a kind of tonal center of the particular frame. The system aims to capture and record the movement of centroids, after they are marked with the most appropriate chord label. First, the system classifies the centroid as one of the triads located in the Tonal Space, using a straightforward 1-NN algorithm, like in [15]. The classification boundaries are not calculated from training data, but deterministically defined as described in Section 2. This transcription strategy compensates for variations in spectral characteristics and intensity levels when the same notes and harmonies are played on different instruments.

4. N-GRAM MODEL OF HARMONIC SEQUENCES

We use *n*-grams to model the harmonic progressions of the music pieces. The output of the feature extraction process is an *L* length chord sequence. We describe here the normalization process to make the sequence tempo and key invariant. Such normalization is required because the queries and the matching music in the database may be in different keys and tempi.

To ensure key invariance, relative chord changes are extracted from the transcribed sequence, an approach that has also been used by other researchers [19].

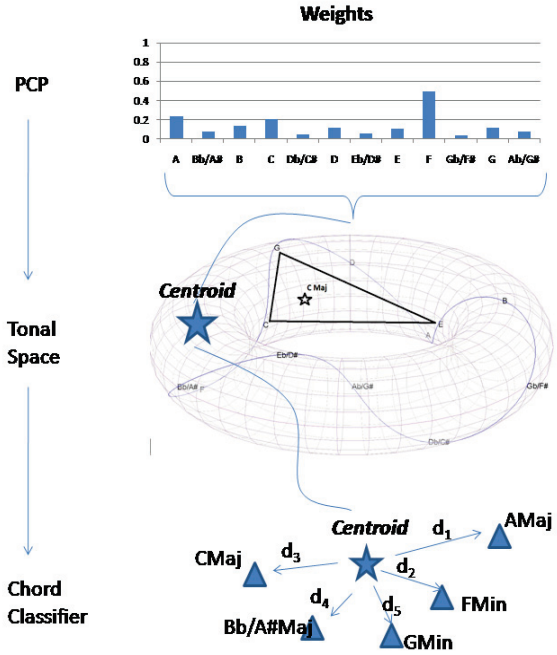

Figure 4. Mapping from PCP to the Tonal Space. Calculation of the tonal *centroid* and its distance to the triad chords.

Since the window length and overlap rate is high (250ms and 90%, respectively), the transcription of the harmonic progression contains many chord repetitions. We remove these repetitions so as to focus on harmonic changes, rather than harmonically stable parts of the music sequence. By doing so, tempo variations are also eliminated. The resulting harmony sequences thus carry more distinct information about the harmonic progression.

In our experiments, *n*-grams were selected for modeling harmonic progressions. Results for different *n*-grams are reported in Section 6. The audio coverage range of a 6-gram in our experiments is between 0.8 seconds and 2.3 seconds. On average 1.5 seconds of music audio is represented by a 6-gram feature set.

To enable the efficient use of this strategy, smoothing of the *n*-gram models is required. Smoothing is widely used to eliminate computational problems caused by non-existing *n*-grams in natural language processing applications. A Universal Background Model (UBM) is produced using the entire music database and mixed with each individual *n*-gram model using a low weight for smoothing (0.9 vs 0.1). Finally, the collection of the smoothed *n*-grams constitutes

our database. We use the SRILM toolkit [20] to create the *n*-gram models, to perform smoothing, and to evaluate the model.

5. RETRIEVAL METHOD

We use the perplexity measure to evaluate the similarity between the *n*-gram model of each music piece in the database and that of the short-length query sequence. The perplexity measure gives the likelihood the query was generated by a specific probability distribution, namely one of the *n*-gram harmonic progression models in the database.

The perplexity of a discrete probability distribution *p* can be defined as:

$$2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)},$$

where *H(p)* is the entropy of the distribution. Suppose *p* is unknown. One can model the unknown distribution *p* using a training sample drawn from *p*. Given a proposed model *q*, one can evaluate how successfully *q* predicts the sample set $\{x_1, x_2, x_3, \ldots x_N\}$ drawn from *p* using the perplexity measure. The perplexity of the model *q* can be defined as:

$$P_{(p,q)} = 2^{-\sum_{i=1}^{N} \frac{1}{N} \log_2 q(x_i)}.$$

A model *q* that better predicts the unknown distribution *p* gives higher probabilities of $q(x_i)$, which leads to lower perplexity.

Our system first builds *n*-gram models of the query and of each piece in the database. It then uses the perplexity measure to determine which of the harmonic progression models of the pieces in the database best fits the query sequence. The system then returns an N-best list of the most likely candidates.

6. EXPERIMENTS

A list of 1000 classical music pieces from famous composers is selected. For each piece in the list, 2 recordings are acquired (one termed the "original" and the other a variation). The variation can be a different instrumental arrangement of the piece or a recording of the same piece by another artist. We replace the ones for which we cannot find an additional audio recordings (CD or mp3) with audio synthesized from the MIDI version as the variation (about 250 such MIDI variations are created). All files are converted to 16 kHz 16-bit wav format. All 2000 files (1000 originals and 1000 variations) are converted to strings of chord labels using the method explained in Section 3. The original recordings are used to train *n*-gram harmonic progression models that constitute the database. The short length test queries are extracted from random parts of each music piece. For each of the query pieces, the system aims to retrieve the original recording of the target piece in the N-best list.

		Length of the query			
		15s	25s	35s	Full
Top-1 match	Accuracy	37.6	41.6	42.9	51.6
	MRR	-	-	-	-
Top-5 match	Accuracy	55.8	57.4	59.6	63.7
	MRR	60.4	63.6	67.4	70.1
Top-20 match	Accuracy	56.8	59.6	62.6	71.5
	MRR	71.8	75.4	73.2	79.8

Table 2. Retrieval results (%) for the 6-gram model over different query lengths.

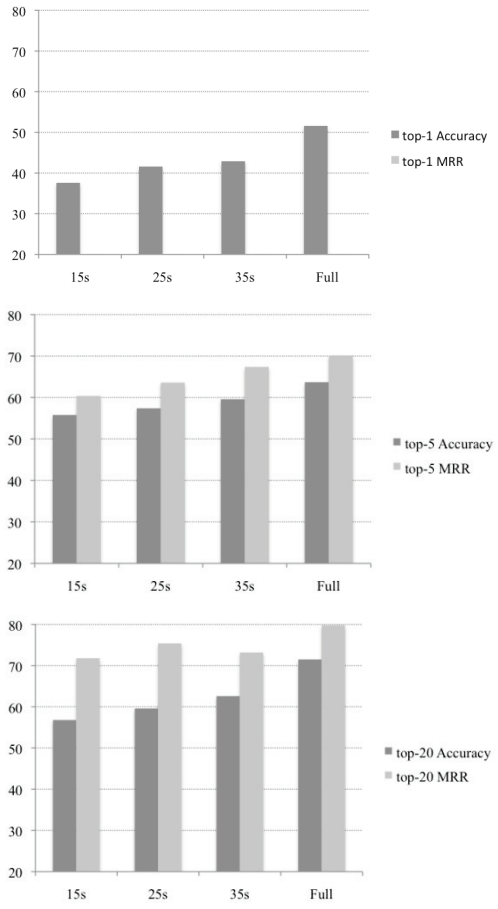

Figure 5: Graph showing the effect of query length on the top-N match correct retrieval accuracy for N = 1, 5, and 20 (actual numbers given in Table 2).

Alongside the N-best list scores, the Mean Reciprocal Rank (MRR) measure, which gives the average rank of the correct matches in the top-N retrieved results (by percentage), are also calculated. Table 2 shows the retrieval results for the 6-gram model as it varies with different query lengths and different N-best list lengths. The numbers are graphed in Figure 5.

One can see from the results that one of the main determinants of retrieval performance is the length of the query.

Since the system retrieves similar songs based on the relative frequency of *n*-length subsequences, the longer the query, the more its *n*-gram model resembles that of the target song. The number of distinct harmonic progressions that identifies the target song is also directly increased with query length.

	Length of the query			
	15s	25s	35s	full
n = 6	37.6	41.6	42.9	51.6
n = 5	36.4	40.2	40.9	49.7
n = 4	32.5	35.4	37.1	43.4
n = 3	28.8	35	36.2	42.3
n = 2	22.6	30.2	33.1	40.2

Table 3. Top-1 match retrieval accuracy (%) over different order *n*-gram models and different query lengths.

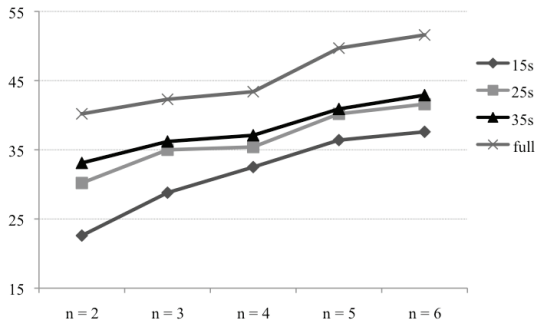

Figure 6: Graph showing the effect of query length and *n*-gram size on the top-1 match correct retrieval accuracy (actual numbers given in Table 3).

Table 3 and Figure 6 present results for different length *n*-grams. It illustrates how the use of higher order *n*-grams (n>2) boosts the system's performance. For all query lengths, larger *n*-grams yield better results. For all *n*, longer queries yield higher accuracies.

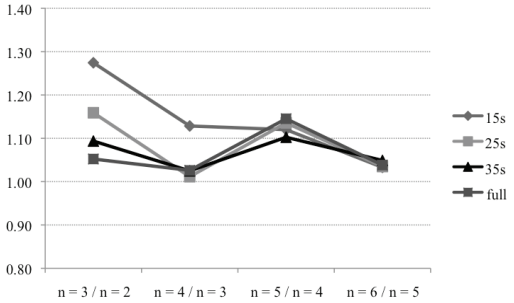

Figure 7: Graph of retrieval accuracy ratios as *n* is increased by one.

Figure 7 shows the graph of the accuracy ratios (an indicator of performance improvement) as the *n*-gram order is increased by one. All numbers are above one, indicating that performance improves by increasing the *n*-gram order. It is interesting to note that the ratio of the accuracy for *n* = 6 over that for *n* = 5 converges between 1.03 and 1.05 for all query lengths. As shown by these numbers, the performance difference between 5-grams and 6-grams is small with respect to accuracy. This may be because 5-grams become sufficiently sparse for capturing the unique harmonic features of the music pieces. Thus, building 6-gram and higher models will likely not have a strongly positive effect on retrieval performance for this particular dataset. The tradeoff between computation time and retrieval accuracy should also be a consideration since building models and calculating perplexity for larger *n*-grams takes more computational power and time.

7. CONCLUSION

In this work, a perplexity based audio music retrieval system that is robust to instrumental variation is proposed. PCP features are extracted from overlapping frames and mapped to a 3-dimensional tonal space. A1-NN classifier decides the harmonic identity of the particular frame based on pre-defined positions of the 24 major and minor triads in the tonal space. Key normalization is performed. From the classifier output, repetitions are removed so as to focus on changes in the series of harmonies. From the resulting harmonic sequence, *n*-gram statistics are acquired and a database is constructed. Given a music query, the transcription is completed using the same strategy and the similarity between the transcribed input and the database models is computed using the perplexity measure.

The algorithm is tested on a database of 2000 music pieces. While there is room for improvement, the results show that, for short length queries, the perplexity-based approach is capable of finding the target piece. The work could be strengthened by testing on a larger dataset with more versions of each song.

To our knowledge, no other study in the literature reports results from short length queries. Our motivation here is that royalty rights management systems usually work with short length queries and we would like to apply our system in such scenarios. The MRR and top-N best list scores suggest that a more fine-grained representation may be needed in order to more successfully retrieve the target piece. Ideally, we would like a retrieval system for which the target piece tops the results list, an important criterion for royalty rights management applications.

Future work includes systematically isolating components of our system for evaluation and improvements. We have used a straightforward feature extraction strategy, which should be compared against other methods. We can substitute chord labeling algorithms in the literature for the particular method used to extract harmonic labels to examine the impact of chord labeling technique on retrieval success. Other further work includes implementing multi

stage search algorithms, in order to improve search performance with respect to time and accuracy.

8. ACKNOWLEDGEMENTS

This work was supported in part by a National Science Foundation (NSF) Grant No. 0219912, and in part by a TÜBİTAK Career Grant 3501-109E196. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect those of the NSF or the TÜBİTAK.

9. REFERENCES

[1] C. Raphael: "Automatic Transcription of Piano Music," *Proceedings of the International Conference on Music Information Retrieval*, 2002.

[2] A. Pertusa and J. M. Inesta: "Polyphonic Music Transcription Through Dynamic Networks and Spectral Pattern Identification," *Proceedings of the International Workshop on Artificial Neural Networks in Pattern Recognition*, 2003.

[3] P. Smargadis and J. C. Brown: "Non-Negative Matrix Factorization for Polyphonic Music Transcription," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[4] M. Ryyananen and A. Klapuri: "Polyphonic Music Transcription Using Note Event Modeling," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[5] G. E. Poliner, and D. P. W. Ellis: "A Discriminative Model for Polyphonic Piano Transcription," *EURASIP Journal on Advances in Signal Processing,* Vol. 2007.

[6] W.-H. Tsai, H.-M.Yu, and H.-M. Wang: "Query by example technique for retrieving cover versions of popular songs with similar melodies," *Proceedings of the International Conference on Music Information Retrieval*, 2002.

[7] M. Marolt: "A mid level melody based representation for calculating audio similarity," *Proceedings of the International Conference on Music Information Retrieval*, 2006.

[8] E. Unal, P. Georgiou, E. Chew, and S. Narayanan: "Statistical modeling and retrieval of polyphonic music," *Proceedings of the IEEE Multimedia Signal Processing Workshop, 2007.*

[9] J. Serra and E. Gomez: "Audio cover song identification based on tonal sequence alignment," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2008.

[10] D. P. W. Ellis, C. V. Cotton, and M. I. Mandel: "Cross-correlation of beat-synchronous representations for music similarity," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2008.

[11] S. Kim, E. Unal, and S. Narayanan: "Fingerprint extraction for classical music cover song identification," *Proceedings of the IEEE International Conference on Multimedia Expo*, 2008.

[12] "S. Doraisamy and S. Ruger: "Robust Polyphonic Retrieval with N-grams," *Journal or Intelligent Information Systems*, Vol. 21, No. 1, pp. 53-70, 2003.

[13] J. Pickens, J. B. G. Monti, M. Sandler, T. Crawford, M. Dovey, and D. Byrd: "Polyphonic Score Retrieval Using Polyphonic Audio Queries: A Harmonic Modeling Approach," *Journal of New Music Research,* Vol. 32, 2003.

[14] V. Lavrenko, and J. Pickens: "Polyphonic Music Modeling with Random Fields," *Proceedings of ACM Multimedia, 2003.*

[15] E. Chew: *Towards A Mathematical Model of Tonality*, Ph.D. dissertation, Massachusetts Insitute of Technology, Cambridge, MA, 2000.

[16] C. L. Krumhansl: "The geometry of musical structure: a brief introduction and history," *ACM Computers in Entertainment*, Vol. 3, No. 4, 14 pages, 2005.

[17] E. Chew, Elaine: "Slicing It All Ways: Mathematical Models for Tonal Induction, Approximation and Segmentation Using the Spiral Array," *INFORMS Journal on Computing*, Vol. 18, No. 3, pp.305–320, 2006.

[18] J. Serra, E. Gomez and P. Herrera: "Audio cover song identification and similarity: background, approaches, evaluation, and beyond", *Studies in Computational Intelligence*, 2010.

[19] T.E. Ahonen and K. Lemstrom: "Identifying cover songs using the normalized compression distance", *Proceedings of the International Workshop on Machine Learning and Music,* 2008.

[20] A. Stolcke: "Srilm – an Extensible Language Modeling Toolkit," *Proceedings of the International Conference on Spoken Language Processing,* 2002.