CREATING AND EVALUATING MULTI-PHRASE MUSIC SUMMARIES

Konstantinos A. Meintanis and Frank M. Shipman

Center for the Study of Digital Libraries & Department of Computer Science
Texas A&M University
{kam2959, shipman}@cs.tamu.edu

ABSTRACT

Music summarization involves the process of identifying and presenting melody snippets carrying sufficient information for highlighting and remembering a song. In many commercial applications, the problem of finding those snippets is addressed by having humans select the most salient parts of the song or by extracting a few seconds from the song's introduction. Research in the automatic creation of music summaries has focused mainly on the extraction of one or more highly repetitive phrases to represent the whole song. This paper explores whether the composition of multiple "characteristic" phrases that are selected to be highly dissimilar to one another will increase the summary's effectiveness. This paper presents three variations of this multi-phrase music summarization approach and a human-centered evaluation comparing these algorithms. Results showed that the resulting multi-phrase summaries performed well in describing the songs. People preferred the multi-phrase summaries over presentations of the introductions of the songs.

1. INTRODUCTION

People use summaries to concisely describe or highlight the major points of the genuine object. In text, for example, the authors of a scientific paper summarize the key points of their presentation in an abstract, a paragraph briefly describing the topic and their achievements or ideas. Accordingly, in music, vendors of CDs and mp3s (like Amazon.com) provide small snippets of songs to help potential customers become familiar with the contents of an album or to find songs they can only recall by melody. Similarly, radio stations remind listeners of the top-ten hits of the week by playing the refrains of the respective songs.

As the examples above indicate, a music summary consists of the part(s) of the song that are short in duration but rich enough in information to describe and identify the total for the given current task. Such a conclusion implies that the location or the duration of those parts is not fixed for all music since their selection depends on factors like the song, the user's perception of music and the task at hand (e.g. selecting from known music or deciding whether to buy unknown music). Finding a summarization approach that takes into account all three factors requires a model of the music, a model of the user, and a model of the task. Most commercial on-line music stores preview their songs by either the introduction of the song, a randomly selected phrase, or the (often manually selected) refrain. The simplicity of these approaches has two main problems. On one hand, there is no guarantee that the selected phrase is sufficient for becoming familiar with or recognizing the song. On the other hand, using human resources to find the refrain for thousands of songs is costly in terms of time, effort and money.

Much of the existing work in music summarization focuses on the selection of the most repeated phrase(s). When more than one phrase is selected, it is generally because the desired summary length is longer than the identified refrain and the added segment is the phrase identified as the next most frequent regardless of its similarity to the already selected refrain. As a step beyond refrain selection, we explore summaries designed to include more parts than the most salient phrase or the introduction of the song. To examine the design space for such algorithms, we compare algorithms that compose a summary from a fixed number of components (three) but vary the selection of those components between preferring phrases that are sonically different and phrases that are repeated more often.

In this paper we present three summarization algorithms that follow the principles above and an evaluation comparing their performance in the context of pop and rock music. Section 2 discusses the related work in automatic summarization and a comparison of techniques. Section 3 describes the summarization algorithms. Section 4 introduces the study and the procedures followed. An analysis of the results is presented in section 5. Finally, section 6 summarizes the results and presents directions for possible improvements of the current algorithms and follow-on studies.

2. RELATED WORK

Research into techniques for the extraction of sound / music features (i.e. tempo, brightness, fundamental frequencies, bounds of phrases) is quite fertile. This work has expanded into research for developing music summaries that tend to focus on the problem of identifying musical phrases and, in particular, the refrain. Hence, the success of summarization algorithms has been typically

evaluated based on how accurately they can determine the most repeated phrase. There are a variety of approaches to identifying the refrain.

A number of algorithms [1, 2] use a pattern matching approach where the structure of the content, and more specifically the most salient phrase, is determined by comparing candidate segments (a fixed sequence of frames) with the whole song. Cooper and Foote [3], after the parameterization of the signal with the calculation of the Mel Frequency Cepstrum Coefficients (MFCCs), find the distance in the parameter vectors of all frame combinations and store the results in a two-dimensional self-similarity matrix. To select the segment (sequence of frames) that best represents the entire song, they calculate the similarity of each segment to the whole and choose the one with the maximum value. If the phrase is not as long as the desired summary, they add the next highest ranking phrase(s).

Other algorithms develop more domain-specific models of the music in order to identify the most repeated phrase. Logan and Chu [5] use a three step process for extracting the key phrase. After segmenting the song, they cluster the resulting segments using a modified cross-entropy or Kullback Leibler (KL) distance to infer the structure of the song and label its different parts. The key phrase is then selected based on the frequency of those labels. Lu and Zhang [6] use the frequency, energy and position to detect the boundaries of musical phrases by analyzing each frame's estimated tempo and computing a confidence value of the frame being a phrase boundary. Depending on the type of music (instrumental or including vocals), Xu and Maddage [12] first extract the features that better catch the attributes of the segmented signal (e.g. MFCCs and amplitude envelope for instrumental music; linear prediction coefficients (LPCs) and derived cepstrum coefficients (LLPCs) [10] for vocal music). Those features are then used for content based clustering, and the output is used for the extraction of the most representative theme. Kim et al. [4] take changes in tempo as a primary indicator for summarization. They first segment the signal based on changes in tempo and then cluster segments based on their MFCCs. Shao et al. [11] analyze a song's structure based on the rhythm and note the onset of the signal and then cluster the segments according to their melody-based (chord contours) and content-based (chord contours and vocal content) similarity. The earliest segments containing the chorus together with some directly preceding and succeeding phrases are used for the creation of the final summary. Mardirossian and Chew [7] generate music thumbnails using the sequence of the keys in time and the average time in each key to detect the most prominent melody.

Peeters et al. [9] generate a state representation of the song to discover its structural components. After discovering the potential states of the signal, they apply k-means clustering to associate each frame to one of the discovered states and a Hidden Markov Model (HMM) to identify the state sequence. The state representation is then used for the creation of the summary by choosing states and transitions according to user needs. They describe four different possible ways to generate a multi-phrase summary based on the signal analysis.

As described, most of the work on music summarization has focused on the identification of music phrases. Our work is complementary in that it explores the design of multi-phrase summaries once phrases have been identified.

3. ALGORITHMS

Augmenting the refrain by compositing music phrases that are repeated in the music yet significantly different from one another can enhance the value of a summary. There are many examples where frequently occurring phrases other than the refrain are effective for recognizing a song. A highly repeated instrumental motif or a dominant verse can be as characteristic as the most salient phrase of a melody. In Lynyrd Skynyrd's "Sweet Home Alabama", for example, the introductory theme, which appears several times in the song, is almost as recognizable as the refrain itself. To explore how choices in the selection of additional phrases affects users perceptions of the summary, our summaries consist of three parts: the most salient phrase (usually the refrain) and two additional phrases. We compare three algorithms for selecting the two additional phrases. These algorithms vary the bias between phrases that are repeated and phrases that are sonically distinct.

3.1 Most Salient Phrase Detection

Phrase detection is not the focus of our work and, indeed, many of the algorithms found in related work could be used instead to identify phrase boundaries and determine repetitions. All three of our algorithms follow a common approach for the detection of the most salient (or key) phrase. In the preprocessing phase, the signal, after the removal of its first and last 10 seconds (that often carry non-useful information), is segmented into fixed, non overlapping blocks of 0.75 second each. A Hamming window is applied on each block to prepare the signal for the Fast Fourier Transform (FFT) which in turn returns the frequency components of the signal. Afterwards, the algorithm calculates the MFCCs of each block as they provide a better estimation of how humans perceive frequencies.

In the next phase, groups of eight successive blocks are formed where successive groups have a 50% overlap (i.e., 4 blocks). For each group we determine its MFCCs by taking the average of the MFCCs of its blocks. The Euclidean distance between the MFCCs of each pair of groups is then calculated and normalized. Starting with a strict (restrictive) distance threshold, clusters are computed using each group as a centroid. The largest resulting cluster is then selected. Clusters that include only

contiguous segments of the music are not considered. If the threshold is too strict to generate any non-contiguous clusters, the process repeats with a more relaxed threshold.

Once the largest cluster is identified, the key-phrase is selected by identifying the block with the smallest amplitude (lowest sound level) within a range of eight blocks (6 seconds) before the starting block of each group in the cluster. The group with the smallest corresponding amplitude is selected due to the likelihood that the block is near the start of a music phrase. The start of the key-phrase is chosen to be 3 seconds prior to the selected group and the key-phrase lasts for 8 seconds.

The next subsection presents a high-level description of the three algorithms for selecting the complementary parts of the summary. This is followed by a more detailed description of how the algorithms are instantiated.

3.2 Complementary Parts Selection (Overview)

The three algorithms presented here vary the selection of the two complementary parts (segments or clusters) of the summary based on a combination of the segments' musical similarity (distance between MFCCs), the number of identified repetitions (size of cluster), and the temporal location in the musical piece. Conceptually, the first algorithm follows an approach oriented more in finding complementary parts according to their frequency of occurrence in the song. In comparison, the second algorithm increases the importance of the sonic distance in the selection process while the third algorithm places most of the emphasis on the sonic distance.

The first algorithm (Repetition Emphasis Algorithm - REA) selects the complementary phrases by placing an upper bound on the similarity between the three phrases but otherwise picks the most repeated phrases prior to and after the identified key phrase. The second algorithm (Intermediate Algorithm - IA) again selects the first complementary phrase by selecting the most repeated phrase prior to the key phrase that differs by more than a threshold. It selects the second complementary phrase to maximize the minimum of 1) the similarity between the second phrase and the refrain and 2) the similarity between the second phrase and the first selected phrase. In this way, the IA puts a higher precedence on ensuring difference between all three of the selected musical segments than it puts on the second phrase's repetition. The third algorithm (Sonic Difference Emphasis Algorithm - SDEA) goes a step further by selecting the two complementary segments that minimize the musical similarity between the three segments without considering whether the complementary phrases were repeated or not.

3.3 Complementary Parts Selection (Details)

The first two algorithms share their approach to selecting the first complementary part. They also steer the selection of the first and second complementary parts towards earlier and later potions of the song, respectively.

After the selection of the key-phrase from the largest cluster of blocks, the first complementary part is selected from the next largest cluster that resides, if possible, in the interval between the start of the song and the key-phrase and differs by more than a minimum threshold. The difference between the two clusters is the mean distance between the MFCCs of its groups. To be a candidate for selection of a complementary part, the mean distance must be greater than a predefined threshold and the variance of the distances must be lower than a specific limit. These thresholds reduce the likelihood that the algorithm will choose a cluster of phrases that sound very similar to the key phrase (i.e. the refrain without the voice or a variation of it). Once the next-largest cluster that meets the MFCC distance requirements has been found, the group of blocks that occurs prior to the key-phrase and is temporally most distant from the key-phrase is chosen as the first complementary part. If no groups of blocks are prior to the key-phrase, then the first complementary part is chosen to be the one closest to the end of the song (furthest from the key-phrase).

The selection of the second complementary part in the REA proceeds similarly. Again, the algorithm selects the next largest cluster with significant differences in MFCC means from both the key-phrase and the first complementary part. Once the cluster is identified, the group of blocks closest to the end of song is selected (assuming the first complementary part was chosen from before the key-phrase, otherwise it will select the group of blocks closest to the start of the song).

In the IA, the selection of the first complementary part uses the technique described in the first algorithm. However, the selection process of the second complementary part deviates significantly except that the search is still focused on the interval between the key-phrase and the end of the song. The second complementary part is chosen as the group that maximizes the minimum of the value of F(i) in formula (1):

$$F(i) = \min (DK(i), DP(i)) \text{ where} \quad (1)$$
$$DK(i) = \Delta (V(group_i), V(\text{key-phrase})) \quad (2)$$
$$DP(i) = \Delta (V(group_i), V(\text{first complementary part})) \quad (3)$$

where $\Delta(V_1,V_2)$ is the Euclidean distance of the MFCC vectors V_1,V_2 and i is the number of the groups for the portion of the song being examined.

The SDEA differs substantially. After the extraction of the key-phrase, the ten least similar groups of blocks (in terms of MFCC vector similarity) to the key-phrase group of blocks are used as candidates for the selection of the phrases. From the ten candidates, the pair with the minimum similarity (maximum Euclidean distance) is used for the extraction of the two complementary parts. Thus, this algorithm places greater emphasis on sonic difference.

3.4 Summary Creation

To create the final summary, we take an eight-second slice from the key phrase and a six-second slice from each of the two complementary parts (total twenty seconds) and order them temporally. A one-second silence is introduced between the segments to diminish the effects of the abrupt switches. Fading in and out is not currently used since it "steals" potentially valuable time from the summary. However, the evaluation indicated that smoothing the transitions between phrases is important to users so crossfades will be part of our future efforts. The final summary has length twenty two seconds which is similar to many of the commercial summaries found.

4. STUDY DESIGN

We designed an experimental study to evaluate and compare our three summarization approaches and to test their performance over a widely used technique. The study was conducted in the Center for the Study of Digital Libraries at Texas A&M University. Fifteen participants over 18 years old, mainly students, were recruited to take part including 12 men and 3 women. The majority (67%) had some kind of music education and more than half of them (67%) had a personal music collection of at least 50 songs (8 participants had more than 200 songs).

Participants were asked to listen carefully to the summaries of twenty popular rock and pop songs and choose the summary that best represents each song. In contrast to the study conducted by Ong [8], our evaluation criteria focused on user preference and summary completeness and not on metrics measuring the ability of subjects to assign song titles. There were four 22 second-summaries per song, three generated with our algorithms and one that was merely the first 22 seconds of the song. Participants answered a series of multiple-choice questions about the quality of the selected summary and their familiarity with the song before proceeding to the next one. The summaries, as well as the songs (in their full version), were accessible through a web-based interface. Participants were able to navigate through the songs and listen to the summaries as many times they wanted. There was no time limit for the completion of the task. The order in which the songs and the summaries were presented to the participants was balanced across participants.

Demographic data about the participants was collected via a pre-task questionnaire. Post-task, semi-structured interviews were used to gather information about the participants' perceptions of the task, their experience with the algorithms and their ideas for future improvements.

5. STUDY RESULTS

To get an idea of what users really appreciate in a music summary we asked them to name (pre-task questionnaire) the parts or features of songs they consider fundamental for becoming familiar with and recalling music.

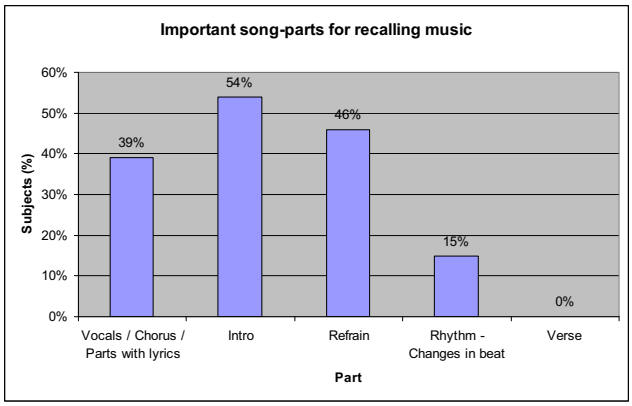

Figure 1. Important parts for recalling music

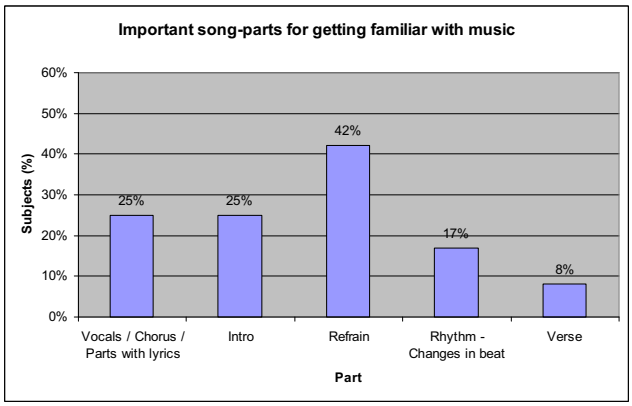

Figure 2. Important parts to become familiar with music

The results confirmed that both the introduction and the refrain are believed to have an important role in the process of understanding and recognizing music (see Figures 1 & 2). However, there was a distinction between the two cases. The introduction of the song was indicated most important for remembering (although with small difference from the refrain that was second) while the refrain was ranked best for becoming acquainted with the music. The higher score of the introduction and the vocals / chorus / lyrics in recalling music matches the fact that a few words or notes can be sufficient for identifying a song we know but provide little information for a song we do not know. Finally, other musical parts like bridges and verses scored very low in the preference ranking or were not mentioned at all by the participants.

Figure 3 shows the distribution of participants' selections for their favored summary. Participants chose the introduction summary in only 13% of the cases and the REA, the most popular, in 35% of the cases. Analysis of the data shows the difference in the selection of the four

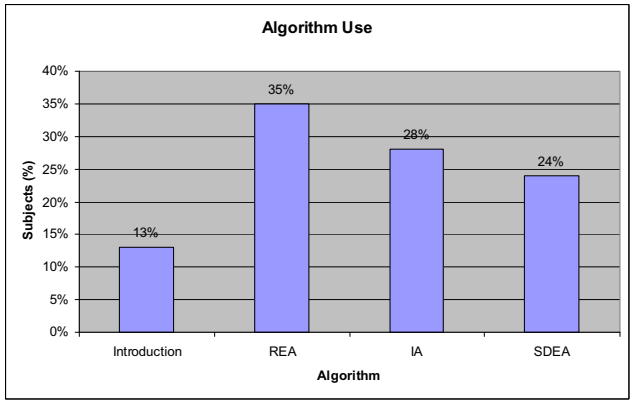

Figure 3. Users' algorithm choices

(I) Algorithm	(J) Algorithm	Mean Diff. (I-J)	Std. Error	Sig.
Introduction	REA	-4.4286	1.03067	**.001**
	IA	-2.8571	1.03067	**.041**
	SDEA	-2.1429	1.03067	.178
REA	Introduction	4.4286	1.03067	**.001**
	IA	1.5714	1.03067	.433
	SDEA	2.2857	1.03067	.136
IA	Introduction	2.8571	1.03067	**.041**
	REA	-1.5714	1.03067	.433
	SDEA	.7143	1.03067	.899
SDEA	Introduction	2.1429	1.03067	.178
	REA	-2.2857	1.03067	.136
	IA	-.7143	1.03067	.899

Table 1. Pairwise comparison of the four algorithms

algorithms was statistically significant (F-test, P=0.0013, α=0.05). Table 1 presents the results from the pairwise comparison of the algorithms with the Tukey HSD test. The numbers show a statistically significant difference between the introduction-based algorithm and the REA and IA (P=0.001 and P=0.041 respectively, α=0.05).

While there was no statistically significant difference between the three multi-phrase algorithms, the trends in the data show that identifying repeated phrases is likely to add value to the resulting multi-phrase summary.

The correlation between the algorithm choice and how good the summaries were (see Figure 4) wasn't statistically significant. However, the numbers look promising considering that 13 of the participants evaluated their choice as at least a good representation of the song, and this choice was one of our algorithms in 87% of the songs. However, the most interesting point about the weakness of the song introduction as summary came out from the post-task interviews. The analysis showed that, while a respective number of the participants (10) listen to the song previews provided by the on-line music stores, only 3 of them are confident that these previews describe the songs sufficiently. Since most online stores preview music using a single contiguous snippet, this indicates a need for an alternative to current summaries.

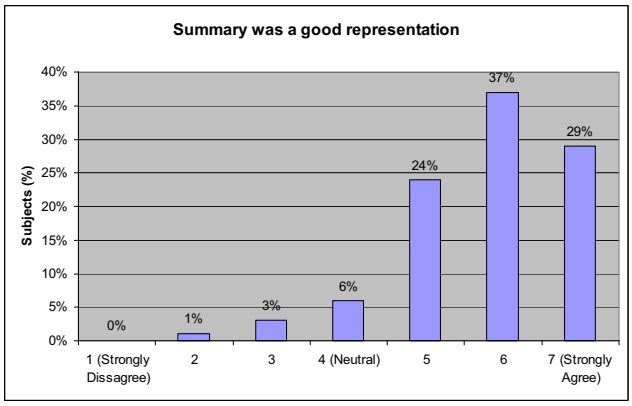

Figure 4. Evaluation of summaries performance

Algorithm \ Knowledge of Songs	I know it well	I don't know it well	I haven't heard the song
Introduction	32 (16%)	4 (11%)	1 (2%)
REA	66 (33%)	10 (28%)	23 (54%)
IA	53 (26%)	14 (39%)	10 (23%)
SDEA	50 (25%)	8 (22%)	9 (21%)

Table 2. Algorithm selection and familiarity with music

Participants reported knowing 71% of the songs well, not knowing 16% of the songs, and having limited knowledge of 13% of the songs. Analysis of the data showed that there is no statistically-significant correlation between the choice of the algorithm and how familiar the participants were with the music. Table 2 shows the number of times participants selected each algorithm based on their knowledge of the song. The proposed techniques are superior in effectiveness over the traditional introduction-based approach no matter whether the user is a customer browsing a new album or a person trying to retrieve an already known mp3 from a personal collection.

Strongly related to participants' knowledge of the songs is how recently they had heard them. Participants had listened on average to about 59% of the songs within a year (see Figure 5). However, the statistics again did not show a significant correlation with the algorithm choice.

One of our concerns when we were designing the proposed summarization techniques was that the segments in each summary would be too short to sufficiently describe the section of the song that had been extracted. A music phrase (especially in classical music) can have duration much longer than the six seconds selected as the length for complementary parts. However, for this collection of pop and rock songs, the results showed that only 20% of the selected summaries were considered "too short" while 53% evaluated were considered "good" and 27% were viewed as "too long".

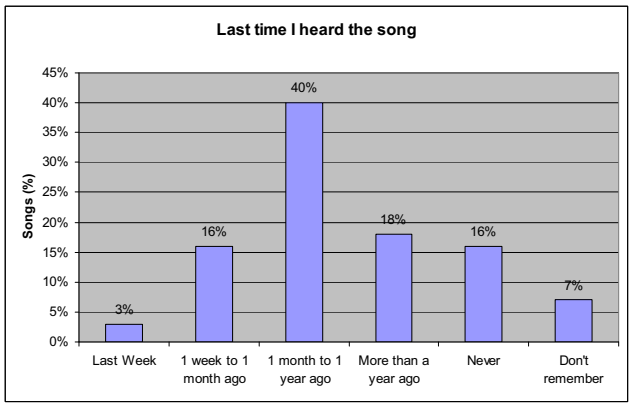

Figure 5. Participants' last time of listening to the song

6. DISCUSSION

Three algorithms for creating multi-phrase music summaries in are presented and evaluated. The design of the algorithms reflects a range of approaches that vary between emphasizing the selection of repeated phrases and the selection of sonically different phrases. The study showed that participants believed that the multi-phrase summaries better represented the song than the introduction to the song. While the difference between the three algorithms was not significant, the results indicate a likely preference for algorithms that emphasize the selection of repeated phrases, at least in the genre of pop and rock where the structural components of the melody are more standardized and identifiable.

There are several things we want to test and improve about our algorithms in the future. One of the complaints participants had during the task was that the switch from part to part in the summaries was too abrupt and hence distractive or even annoying. Use of phrase bounds detection for selecting the start of phrases could help as could the use of fade-in and fade-out effects. Based on participants' feedback about which parts / features of the songs are important, it would be interesting to examine if the integration of the introduction in our summaries can improve or accelerate the process of becoming familiar with new music. A comparison of the best of our techniques with summaries containing only the most salient phrase of the song is also in our future plans. Finally, we would like to improve the accuracy of our summarization approach to be applicable in genres like classical music and jazz where identification of the various themes and important components is more challenging.

7. REFERENCES

[1] Bartsch, M. and Wakefield, G. "To Catch a Chorus: Using Chroma-Based Representation for Audio Thumbnailing", *Proc. of the Int. Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 2001.

[2] Chai, W. and Vercoe, B. "Music Thumbnailing via Structural Analysis", *Proc. of ACM Multimedia*, Berkeley, USA, 2003.

[3] Cooper, M. and Foote, J. "Automatic Music Summarization via Similarity Analysis, *Proc. of the Int. Symposium on Music Information Retrieval*, Paris, France, 2002.

[4] Kim, S., Kim, S., Kwon, S., and Kim, H. "A Music Summarization Scheme using Tempo Tracking and Two Stage Clustering", *Proc. of the IEEE 8th Workshop on MMSP2006*, BC, Canada, 2006.

[5] Logan, B. and Chu, S. "Music summarization using key phrases", *Proc. of the IEEE Int. Conf. on Audio, Speech and Signal Processing*, Orlando, USA, 2000.

[6] Lu, L., and Zhang, H. "Automated Extraction of Music Snippets", *Proc. of the ACM Multimedia*, Berkeley, USA, 2003.

[7] Mardirossian, A., and Chew, K. "Music Summarization via Key Distributions: Analyses of Similarity Assessment Across Variations", *Proc. of the Int. Conf. on Music Information Retrieval*, Victoria, British Columbia, Canada, 2006.

[8] Ong, B. "Structural Analysis and Segmentation of Music Signals", Dissertation submitted to the Department of Technology of Universitat Pompeu Fabra, 2006.

[9] Peeters, G., Burthe, A., and Rodet, X. "Toward automatic music audio summary generation from signal analysis", *Proc. of the Int. Conf. on Music Information Retrieval*, Paris, France, 2002.

[10] Rabiner, L. and Juang, B. *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[11] Shao, X., Maddage, N., Xu, C. and Kankanhalli, M. "Automatic Music Summarization Based on Music Structure Analysis", *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, USA, 2005.

[12] Xu, C., Maddage, N., and Shao, X. "Automatic Music Classification and Summarization", *IEEE Transactions on Speech & Audio*, 13 (3), May, 2005.