Multimodal Music Retrieval for Large Databases

Submitted to Special Session: Novel Techniques for Browsing in Large Multimedia Collections

Björn Schuller, Gerhard Rigoll, and Manfred Lang
Institute for Human-Machine Communication
Technische Universität München
D-80290 München, Germany
(schuller | rigoll | lang)@ei.tum.de

Abstract

In this contribution we present a novel multi-modal access to large MP3 music data-bases. Retrieval can be either fulfilled in a content-based manner or by keywords. As input modalities speech by natural language utterances or singing, and manual interaction by handwriting, typing or hardkeys are used. In order to achieve especially robust retrieval results and automatically suggest music to the user contextual knowledge of the time, date, season, user emotion, and listening habits is integrated in the retrieval process. The system communicates with the user by speech or visual reactions. The concepts shown are especially designed for home and mobile access on Tablet-PCs, PDAs, and similar PC solutions. The paper discusses the concept and a working prototype called Shangrila. An evaluation by a user study leads to an impression of the capabilities of the suggested approach to multimodal music retrieval.

1. Introduction

Recently an enormous increase in sizes of personal digital music databases can be observed. Due to high compression standards more than 10^4 tracks can be stored even on small portable devices or living-room PC-based music players. The user group and spectrum accessing such large archives grow rapidly. On the other hand hard-key based navigation like in common CD players cannot cope with such enormous quantities. Therefore it is considered as a fact among Music Information Retrieval experts [1] that the alluded trends clearly claim for new intuitive retrieval methods for the search of musical pieces in such large databases. Traditionally keyword-based approaches are used benefiting of the fact that digital music often comes labeled with information as the interpret and song name. However, not all tracks carry this information, and especially in Internet databases such information may be erroneous. Furthermore some situations as media players in automotive environments or very small portable devices do not allow for such a keyword search due to possible distraction or limited display and entry abilities. Finally the multilingual character of international music titles and artist names leads to problems considering different character sets for typed input or speech recognition engines that have to cope with a variety of languages. Content based retrieval methods as query by humming or singing seem to help, as the melody is always available and correct within a song. However, ambiguity still exists as many versions and interpretations may exist of a melody. And in some public or noisy situations humming may not be appropriate or possible. In order to combine the advantages and bring efficient and intuitive retrieval to all user groups we consider multimodal access providing keyword and content based approaches by different input channels as the most promising solution. Therefore we introduce *Shangrila*, a multimodal music browser combining spoken, sung, handwritten, typed and hard-key input with touch-screen or mouse navigation. Possible scenarios using all or a selection of the modalities comprise desktop PC’s, tablet-PC’s, PDA devices, or automotive use. In order to constrain the hypothesis sphere of possible songs and to lead to suggestive retrieval contextual knowledge is used. The system profiles the listening habits according to time, date, season, and the actual user emotion recognized by the speech signal. In the following sections the input processing, the contextual integration and multimodal fusion will be presented. Finally user evaluation by means of usability engineering will be discussed.

0-7803-8603-5/04/$20.00 ©2004 IEEE.

2. Speech input

As we want to keep the interaction as comfortable as possible, we use a keyword initialization in an open microphone manner. No push-to-talk button is needed prior to an interaction. The user is allowed to talk or sing a melody that she wants to hear at any time to the interface. The open microphone mode is also needed to spot for the user emotion in the speech signal. In a first step a frame-wise decision is made by an energy threshold, if a signal is present. After the detection of a minimum number of consecutive signal frames comprising pauses we have to decide whether the spotted segment contains speech or singing or just background music or noise. As we combine spoken and sung input a further discrimination between these two is needed. A Multi-Layer Support Vector Machine classification [2] is used for the overall discrimination between noise, singing, or speech. On the first layer we apply Mel Frequency Cepstral Coefficients (MFCC) and δMFCCs for the discrimination of singing or speech, and noise or impostor speakers. A speaker verification is necessary considering the high background noise produced by playing music, which might be interpreted as singing of the user. On the second layer we use 10 derived static features of the pitch and energy contour including higher level features as the rate of voiced sounds, the mean silence durations, and the absolute pitch area. In case of speech a Hidden Markov Model based automatic speech recognition engine and a Belief Network (BN) based spotting engine for the natural language interpretation process the spoken content as described in [3]. If certain keywords as *Shangrila*, the name of the system, are contained within the utterance, and a high acoustic and semantic confidence in the interpretation exists, the user intention will be fulfilled. 27 basic intentions for retrieval, navigation and controlling *Shangrila* are implemented. In any case an interpretation of the underlying user emotion by the spoken content and the acoustic characteristics is made as described in [4]. In case of sung input the melody is recognized by a Query-by-humming engine as described in the following section.

3. Melody based retrieval

In this paper we aim to introduce the basic matching of hummed patterns by use of a Dynamic Time Warp (DTW) algorithm, and provide a short comparison of features for this task and performances achieved. The *Shangrila* system offers the user two retrieval methods. The more robust is user training of the system by humming or singing his most favored tunes as a shortcut once to the system. If there is only a very restricted number of around 50 polyphonic original digital music pieces, comparable to the capacity of small portable sports MP3 devices, a search can be performed matching the monophonic query to the completely unknown polyphonic tracks automatically. However, using large databases as the 15k database considered, matching to the unknown polyphonic audio tracks takes too long and does not lead to reasonable results. Therefore the user is allowed to train the system by humming the personally most important 6 seconds of each song that he wants to find again later on once. The chosen duration is a compromise between recognition performance and the time users are willing to sing. As features we use besides pitch, energy and MFCC’s as described in section 2, the spectral harmonic sum based on partial enhancement [5] calculated in the FFT of each 20ms music frame. The spectrum is filtered by rectangular filters according to the distances of two musical notes in the western 12-tone scale. Starting at the note D up to c’’’ we achieve a reduction to 47 semitone-bands. In order to enhance the melodic part in stereophonic polyphonic audio we concentrate on the center panned parts where in general mostly the key melody can be found. In the databank references for all previously hummed or sung songs or the processed polyphonic songs in the database are stored. The matching is done by use of a DTW with Euclidean distance metric and Itakura constraints. The endpoints constraints are loose to allow for beginning or ending at another note or bar within the intended song. The following results obtained with a test-set of 50 songs hummed by 9 users, 2 of them female, give an impression of the obtained results. As mentioned, larger sets are only reasonable using user training.

Table 1. Retrieval performance sung query

Top x	1	-2	1-3	1-5	1-10
Mono, UD	98%	98%	98%	100%	100%
Mono, UI	74%	86%	90%	90%	92%
Polyphonic	50%	56%	62%	72%	78%
Max. human	86%				

To the right the performance according to the top n best hits is shown. Downwards the different types of references in the database are listed. *Mono* abbreviates monophonic data trained either by the user himself *UD*, or independent of the user *UI*. *Polyphonic* respectively stands for references extracted out of the original polyphonic song. In the last column the maximum human performance is given as a basis of comparison. Therefore probands were allowed to listen to the original music clips at any time of the 50 songs and had to recognize the same hummed melodies as used for the evaluation.

Regarding these results it seems to be a reasonable trade-off between recognition rate and extra-effort for the users having the best hit played and the five best alternatives provided to the user on the screen for selection, e.g. by pointing on the touch-screen.

The optimal feature configurations for matching monophonic to sung monophonic references were using the pitch contour quantized to musical semitone intervals and normalized to the average pitch in order to become independent of the musical key hummed in, plus the first and second order derivatives. Furthermore the signal energy and its first and second order derivatives were used to include rhythmic information. Finally 7 MFCC's and their first order derivatives were integrated to model occurring lyrical content. In order to match the humming queries to the original polyphonic music we used the alluded enhanced partials based harmonic sum with three harmonics per semitone and a width of 8 for neighboring frequencies.

4. Typed and handwritten input

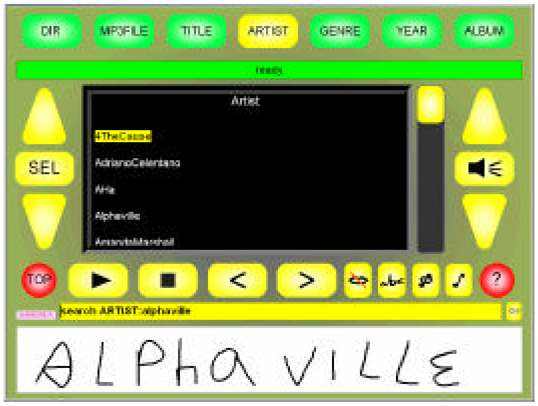
For a content based query either the keyboard or handwritten input in printed characters via a touch screen can be used. As there are different general set-ups of the systems we have in mind as a PDA providing no keyboard, but handwriting, or a living-room PC providing a keyboard but handwritten input only via mouse, both alternatives have been considered. As keywords we decided for the MP3 ID3 tags artist, title, year, album, and genre. As this information lacks in many MP3 files, we additionally use the filename and directory name in which the file is stored. This information is always available and often offers the artist, title and album information. Considering false recognition of single characters and misspellings both of the user and the tags a soft string matching algorithm seems obligatory and is done in a Belief Network spotting approach, which is described in detail elsewhere [6]. As features for the handwritten character recognition the planar Cartesian x- and y-coordinates (plus δ and δδ) are used. For invariance of the size and location they are normalized to their bounding box and the starting point of a stroke. Each character is modeled by one continuous Bakis hidden Markov model (HMM). Up to four Gaussian mixtures and a variable optimized number of states were used. The HMM scores are fed forward to the soft string matching algorithm for more robustness. A rule-based post-processing increases the recognition rate for an alphanumeric alphabet.

5. Context and information fusion

The frequency of listening to a song by the user, the actual season (winter, summer) and time and actual user emotion (happy, sad, angry, neutral) are chosen as contextual variables. In our opinion the selection of a musical style, or even a concrete song, can depend on these aspects. As an example it seems more probable that someone listens to Christmas Carols in the winter. However, the system learns the typical behavior patterns of the user, if any can be observed. This can even lead to a suggestive query. In order to integrate this contextual knowledge and all input modalities we chose a fast singly connected Belief Network (BN) architecture, where each song is modeled in a BN. The primary attributes of a song are the mentioned ID3 tags plus the filename and directory name as described in section 3, and if available a melody print. They allow for a fast play-list creation by demanding music of a specific year, genre, etc. Each attribute consists of a content and a context related variable. Each evidence in a content related attribute raises the belief in the song. The fact that the content state depends on every modality allows parallel and complementary use of the modalities. For example a user can write a title. After a first query result showing several interprets she can select by speech. On the other hand more conventional modalities less prone to errors as typing can be used as fall-back solution. If e.g. a query by humming fails, the user can write down information related to the song. In each final node the confidences of the recognition engines and contextual observations are fed into the net. A song is finally selected by a maximum likelihood decision.

6. Usability Analysis

In order to evaluate the propagated methods the functional *Shangrila* system was provided to five test users which are listening daily to MP3 music on their standard PC. Two of them were female, with an average of 26 years. For two months their conventional MP3-player software was replaced by the full *Shangrila* system. The databases used were private collections of the probands each exceeding a total number of 5k titles with a maximum of 15k tiles. In a short initialization phase the speech and singing discrimination and emotion models were trained user specifically. The following figure provides an impression of the look of the GUI.



Artist

4TheCause
AdrianoCelentano
A-Ha
Alphaville
AmandaMarshall

search ARTIST:alphaville

Figure 1: Shangrila GUI

The system automatically logs all usage data as the chosen modality, the listening habits and the user emotion at a retrieval request. The users are from time to time automatically asked about their underlying emotion, the correctness of the system reaction, and their opinion in order to achieve an impression of the performance. Table 3 shows the average recognition rates and modality specific relative usage. The distribution of the modalities used refers to complete user actions.

Table 3. Performances of Shangrila

Modality	**Mean Perf.**	**Mean Usage**
Handwriting and matching	0.98	0.11
Typing and matching	0.99	0.10
Mouse navigation	-	0.44
Natural language	0.96	0.19
Humming/singing	0.98	0.16
Signal type discrimination	0.99	-
Emotion recognition	0.92	-

The system was rated as very intuitive and the multimodal access as very comfortable. The user could at any time ask Shangrila to suggest a song if she was not sure what to listen to. The suggested tracks by the system were never stopped and the ideas was judged as a funny feature. However, it seems sure that this feature was used especially due to its novelty. This aspect has to be taken also in consideration when deriving conclusions out of the presented very first results of the usage.

7. Conclusion

We believe that the shown principle of multimodal music retrieval greatly improves usability for the everyday user of large music databases and provides access for a larger spectrum of users. The key advantages of the multimodal access are the opportunity for each user to select his personal favorite modality and navigation style, and the use of different retrieval methods according to the surrounding conditions. More innovative query methods as speech or singing found broad acceptance throughout the first logging of usage. However, not in all scenarios each modality may be useful or needed. Long-term analysis of the logging files with more users and tests by means of specific objective measures will show lead to further insights. In future works we aim to investigate musical mood recognition in order to better find music suiting the actual user's emotion.

8. References

[1] J. Reiss, M. Sandler: *"Benchmarking Music Information Retrieval Systems,"* JCDL Workshop Creation of Standardized Test Collections, Tasks, and Metrics for MIR and MDL Evaluation, Portland, 2002.

[2] C. Xu, N. Maddage, X. Shao, F. Cao, Q. Tian: "Musical Genre Classification Using Support Vector Machines," *Proc. of the ICASSP 2003 Vol. V*, pp. 429-432, Hong Kong, China, 2003.

[3] B. Schuller, et al.: "A Hybrid Music Retrieval System using Belief Networks to Integrate Queries and Contextual Knowledge," *ICME 2003*: *Multimedia Human-Machine Interface and Interaction Vol. I*, pp. 57-60, Baltimore, MD, USA, 2003.

[4] B. Schuller, et al.: "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture", ICASSP 2004, IEEE, Canada, 2004.

[5] J. Song, S. Bae, K. Yoon: "Query by humming: Matching humming query to polyphonic audio," ICME 2002, IEEE, Lausanne, Switzerland, 2002.

[6] B. Schuller, et al.: "Applying Bayesian Belief Networks in Approximate String Matching for Robust Keyword-based Retrieval," ICME 2004, CD-Rom Proceedings, Taipei, Taiwan, 2004.