HMM-BASED MUSIC RETRIEVAL USING STEREOPHONIC FEATURE INFORMATION AND FRAMELENGTH ADAPTATION

Björn Schuller, Gerhard Rigoll, and Manfred Lang

Institute for Human-Computer Communication
Technische Universität München
D-80290 München, Germany
(schuller | rigoll | lang)@ei.tum.de

ABSTRACT

Music retrieval methods are in the focus of recent interest due to the increasing size of music databases as e.g. in the internet. Among different query methods content-based media retrieval analyzing intrinsic characteristics of the source seems to form the most intuitive access. The key-melody in a song can be regarded as the major characteristic in music and leads to a query by humming or singing. In this paper we turn our attention to both, the features and the algorithm of matching in audio music retrieval. Nowadays approaches propagate the use of Dynamic Time Warping for the matching process. As reference mostly midi-data or humming itself is used. However, first attempts matching humming to polyphonic audio exist. In this contribution we introduce hidden Markov models as an alternative for humming queries matching humming itself, mobile phone ring tones and polyphonic audio. The second object of our research is the introduction of a new way of melody enhancement prior to a latter feature extraction by use of stereophonic information. Further an adaptation throughout the extraction process of the frame length to the tempo of a musical piece helps improving similarity matching performance. The paper addresses the design of a working recognition engine and results achieved with respect to the alluded methods. A test database consisting of polyphonic audio clips, ring tones, and sung user data is described in detail.

1. INTRODUCTION

Due to the growing amount of music data bases and the larger number of available titles per media efficient methods of music retrieval are a key necessity. Recently efforts in content based retrieval can be observed as it seems to enable the most intuitive access. Letting users hum or sing a melody to find a desired audio track leads to two general challenges: On the one hand the correct corresponding part has to be found in the reference signal. On the other hand the correct monophonic melody line has to be extracted. Most actual approaches deal with the recognition of hummed melodies matching midi data or humming itself. Humming matching is easy due to its monophonic nature, what means that no note starts before another note has finished sounding. Especially matching to midi-data showed great potential for large databases. However, users often need a query for songs that do not exist in a midi-notation yet. These are songs of the actual charts or personally favored pieces off the main-stream. We turn our attention to solutions for selection of such titles. As matching to monophonic sources is easy, we consider matching humming queries to ring-tones for mobile-phones, since these are provided in a regular way especially for actual songs. However the final aim is matching to polyphonic audio itself [1]. A user could easily access titles in a personal databank without the need of humming each once as a reference or downloading an appropriate ring-tone before selection. Retrieval in audio file databases could be realized without extra efforts. Therefore we propagate novel methods to efficiently process audio data for a similarity matching on an intermediate level to humming input. In the following sections also the matching process itself will be object of the paper. Finally a test-database is introduced, and results achieved with our advances are discussed.

2. SYSTEM OVERVIEW

While humming and ring-tones are processed in the same way, polyphonic audio is processed in a parallel computation. In a successive stage incoming retrieval requests are matched to a database.

3. GENERAL FEATURE EXTRACTION

As features we chose the rather commonly used pitch contour and its first and second order deviations. Further harmonic sum based on partial enhancements is integrated in the feature vector. Finally we discuss the use of energy related features.

3.1. Pitch related features

Pitch is extracted using the normalized cross correlation function and dynamic programming due to the noise robust characteristic of the algorithm. The first and second order derivatives are integrated in the feature vector, which leads to an improvement of minimum two up to four percent recognition rate throughout our tests which will be discussed later on in this paper. Additionally the harmonic sum based on partial enhancement as described in [5] is calculated. First the predominance of a partial considering the amplitudes of surrounding signals in the FFT-spectrum within a frequency range is calculated. Next the

average over the enhancement of a partial and its higher order harmonic partials is computed.

3.2. Energy related features

While the use of pitch information seems to be the only reasonable stand-alone features, the addition of first and higher order derivatives of the energy contour aids matching humming especially to humming itself. Energy represents the loudness of a musical phrase. Rhythmic information and expression is transferred mostly by the energy of the signal. We calculate the instantaneous values of energy by the logarithmic mean energy in a frame. To eliminate influences of the absolute energy level we use the first, second, and third order derivatives. However smoothening of the contour seems to be a must to avoid dependence of the lyrical content of a sung melody and filter slight tremolo. In the work presented we use a symmetrical moving average filter. The contours are freed of their mean value and normalized to their overall standard deviation. The addition of energy derivatives leads to an improvement of up to six percent correct recognition rate.

3.3. Quantization by semitones

Human singers slightly vary the pitch even though they still feel the same note. Instruments do also slightly vary in pitch and are often played with a vibrato. A quantization of frequencies in semitone-step intervals seems appropriate to solve these fluctuations and has been introduced in many works [2] dealing with music information retrieval. It substitutes contour smoothening in a suggestive way by integration of general musical frame conditions and reduces the dimensionality of the feature vectors. However in some foreign music or musical styles an adequate different tonal structure exists and claims for an adapted quantization scheme. For example in blues music blue notes resemble bends by quarter notes. This characteristic sound will be lost by the propagated 12-tone diatonic intervals shown in the next equation (1) where x represents any note and $x\#$ an increase in frequency by a semi-tone. The pre-factor respects the doubled frequency of an octave shift and twelve semitones steps within an octave in western music.

$$f(x\#) = \sqrt[12]{2} \cdot f(x) \quad -(1)$$

Instead of triangular filters as found in general speech processing applications we use rectangular band filters without overlap considering musical smear bends as in the alluded blue-notes which are assigned the lower neighboring note. The propagated quantization improves recognition results by 5.8% compared to equidistant narrow filters.

3.4. Transposition into a neutral key

Since the reference and the query pattern are in general in different musical keys they are both transposed into a neutral key. Initially we used the first note as a anchor-note for the transposition. But the fact that it cannot be assumed that all matching patterns start at the same position leads to either the minimum occurring or average pitch as an anchor. The minimum occurring note is susceptible to false pitch estimation. Using the average pitch results in freeing the musical piece of its mean pitch value and raises the recognition rate by an average of 4.7% throughout our test series.

3.5. Musical pitch constraints

It is said that melodic perception reaches from 30 Hz to 4000 Hz. However, our considered spectral band of interest is assumed between the low D (73.416 Hz) , the lowest note a well trained human bass singer is able to sing, and the high c''' (1046.502 Hz) , the highest note an equally experienced human soprano singer is able to perform. By this we achieve a stronger limitation and a faster recognition by reduction of the dimensions of the feature vector to 47 spectral coefficients in the case of harmonic sum computation. Higher frequencies are respected in the calculation of harmonicity by higher order partials. We assume a maximum raise of one octave between two adjacent notes. After finding the average pitch a maximum window of overall three octaves around this note is assumed.

3.6. Elimination of pauses

When matching humming against humming we experienced that an elimination of pauses leads to a decrease in recognition rate. The pauses seem to carry important rhythmical information. However in polyphonic audio the main melody is often interrupted and only the surrounding orchestration is audible. It proved reasonable to search for such breaks by the alluded pitch constraints and set these to zero pitch and energy level analog to the values computed out of the hummed phrase.

4. PROCESSING OF POLYPHONIC AUDIO

While most of the current work in music information retrieval is based on monophonic music [3] the most challenging task seems to be matching a query directly to an formerly unknown polyphonic audio track. Other than in similarity measurement of two polyphonic audio tracks matching to a hummed input demands for the extraction of the key-melody line. In the following two novel advances for an optimized feature extraction are proposed.

4.1. Melody enhancement by stereophonic information

The aim of the propagated approach is to increase the signal to noise ratio between the melody of interest and the accompanying backing instruments. Stereophonic recordings in general pan the main melody in the center position. The center lays a certain stress or importance to a phrase. In such stereophonic arrangements accompanying further instruments will be mostly panned outside the center. Only the bass-phrase can often be found in the middle of the stereophonic spectrum as well. This is due to the fact that low frequencies cannot be located easily by human listeners but more bass presence will be provided by using both stereophonic speakers. Therefore we suggest an extraction of monophonic parts in polyphonic audio to enhance the main melodic phrases. In the following the term monophonic will be used opposing stereophonic and shall not be confused with the counterpart of polyphony. The inverse processing can be found in karaoke applications: phase cancellation helps to eliminate or weaken the voice of the original lead singer and leaves space for the karaoke singer. However reflections from the side of the original recording are still left which seems disturbing for a karaoke application. Since we aim to keep only the lead-melody we have to cope with extracting instead of

eliminating the monophonic part of a recording. Especially in true multi-channel surround-sound recordings this seems to be an easy task since the center-speaker channel is stored in an extra track. In general stereo recordings already by addition of the left and right channel information a raise of 6 dB in level is obtained. This effect is also used in surround solutions which use only two channels for compatibility reasons with former cinematic systems [4]. The center signal is derived by a simple addition of the two channels. However, this results only in a raise of level, but the outside panned information will still be left. Since we aim to extract only the monophonic parts we suggest a fast approximation for a pseudo-monophonic signal $s_{mono}(t)$ according to the following equation (2) where $s_r(t)$ represents the signal of the right, and $s_l(t)$ the signal of the left channel at a time instant t.

$$s_{mono}(t) \approx \frac{1}{2} \cdot sign(s_r(t) + s_l(t)) \cdot \max\{|s_r(t) + s_l(t)| - |s_r(t) - s_l(t)|, 0\} \quad (2)$$

The addition of the two channels resembles the normal conversion of stereo information into monophonic representation. The subtraction results in phase cancellation of the parts panned in the center, in general the melody of interest. The subtraction of the center freed reminiscent has to be calculated by the absolute values to avoid preserving only the information of one channel. The phase information is restored afterwards by a multiplication with the original sign of the monophonic transformed information. If the outside panned parts show little correlation and no center signal is present, the pseudo-monophonic signal is set to zero, or the value can be simply skipped. This solution does not deliver the true monophonic information but seems appropriate for the latter calculation of the feature contours of interest. Band pass filtering helps in a next process to eliminate remaining bass parts. The extraction of the monophonic parts leads to a significant improvement of 8.1% in performance throughout our tests.

4.2. Frame length adaptation

We consider an adaptation of the frame length according to the intermediate temporal resolution of a musical piece on the note level. In a first approach we introduce a static adaptation providing an individual length for each polyphonic audio track throughout the song. This seems reasonable mostly for tracks with less or no temporal dynamics such as in modern pop and dance music where drum computers are used. The advantage is a stable segmentation over time resulting in further smoothening of the contours and data reduction over time. Too short frames are influenced strongly by the sung syllabic contour while melodic development takes significantly longer periods. Different singers tend to ornament the original melody. These variations are in general very short note successions around the tonal center like vibrato as a simple form. We are interested in the average pitch over the length of e.g. 8th notes. Considering the temporal resolution one has to take triplets into account which possess a timing off the general dual basis. Since windowing with a soft, e.g. Hamming, function is reasonably done with an overlap, we decided for a common amount of 50% overlap. To provide an impression of the length of interest 166 ms would be the minimum length if 180 beats per minimum are considered as fastest tempo. These rather long periods in comparison to the usual speech related frame lengths lead to a reduction of data by a factor larger than ten. Therefore we calculate the mean over an appropriate number (e.g. 8) of adapted shorter frames (e.g. 20.75 ms) according to the tempo of a song. The averaged short frames build a larger master-frame of the length of e.g. an 8th note. As musicians regard temporal shifts of more than 10ms as unbearable disturbance synchronization is obligatory if frame length adaptation is used to obtain the actual note information instead of an average between two notes. Therefore it seems reasonable to trigger the beginning of a master frame to a non syncopated on-beat when adapting the frame length to the assumed beats per minute. E.g. kicks of a bass-drum can be easily detected by low-pass filtering for this purpose. In our work the tempo information was achieved by human labeling. Methods of high performance for the estimation exist, and will be integrated in a next step. Frame length adaptation raised recognition results by 4.2% in average throughout our tests.

5. MATCHING

The final 100 dimensional feature vector consisting of energy (first, second, and third order derivative), harmonic sum by partial enhancement (47 note steps, first order derivative) and pitch (plus first and second order derivatives) is matched in a second stage. Since the humming input will differ in tempo compared to the reference and singers lack perfect timing throughout their humming, methods of dynamic programming seem reasonable for elastically matching a sequence of observations to a reference pattern. While other works propagate the use of Dynamic Programming using a Dynamic Time Warping (DTW) algorithm for the matching process, we introduce hidden Markov models (HMM) as an alternative and draw a comparison to the use of a DTW. Hidden Markov models are well known as a standard probabilistic reasoning method in speech processing [6].

5.1. Dynamic Time Warping

In a first step a DTW-algorithm with Itakura constraints and an Euclidean distance metric is used. For each melody a set of references was created. The minimum distance to a representative of each class was calculated to finally select the overall minimum distance of the assumed song. The endpoint constraints were set loose to cope with different starting or ending points.

5.2. Hidden Markov Models

In a second step we use HMMs. Each melody is represented by one single HMM. The HMM types were chosen as Bakis-models, such as in usual speech processing. The number of states is set according to the length of the reference pieces. Using frame length adaptation leads to a quantization into 8th notes. Considering the fact that most melodies span around eight bars, we use around 64 states per model. A jump constraint for the increase in the index was chosen in respect that little parts of the melody may be skipped. In the recognition phase the maximum-likelihood model is chosen. The initial state distribution is set to an equally distribution among the first states to enable a different starting point. Up to four mixtures are used when matching to humming. The HMMs are trained using Baum

Welch re-estimation with 10^5 iterations or an abruption criterion of a change in model parameters of $\varepsilon < 10^{-4}$. Due to the fact that only one reference exists for the training when matching to polyphonic audio or ring-tones - namely the extracted melody of the original sound source - a sparse data approach for the training is used here: one Gaussian mixture component is laid around the observed vector with a constant standard deviation throughout the models and states. The state transition coefficients are equally artificially assumed constantly.

6. RESULTS

In the following the results obtained with a working implementation of the suggested methods and the used training and test corpora are introduced and explained in detail.

6.1. Database

100 clips of stereo- and polyphonic audio have been carefully selected. They were chosen from different musical styles as Rock, Pop, Techno, Jazz, and Classic. Among the tracks instrumental and also rather unknown pieces could be found. The use of clips seems a reasonable approach [5] stressing the problem of extraction of the melody. The key-parts were cut liberal leaving time in front of and after them to manifest the use of loose starting and endpoint constraints or equally distributed initial state probabilities. For test purposes the corresponding ring-tones for mobile phones were provided. A humming and singing corpus of the same melodies was obtained by tests with five amateur singers, two of them female. The test subjects were allowed to either hum or sing the melodies. The samples have been recorded at 44.1 kHz, 16 bit, using an active AKG-C1000S MK II condenser microphone in an acoustically isolated room.

6.2. Signal replay

Throughout our work it proved comfortable to be able to play the extracted pitches and energy values of the polyphonic audio data. By applying this method also human deciders could judge the quality of an extracted melody line as an upper benchmark considering them as excellent classifiers.

6.3. Recognition performance

In the test 16 bit, 16kHz samples were used. In the following the overall correct recognition results in percentage using optimal configurations can be found. The rates allude to the number of hits among the top n tracks, where *HMH* stands for humming matching humming, *HMR* for humming matching ring-tones, *HMP* for humming matching polyphonic audio, and *RMP* for ring-tones matching polyphonic audio. HMH has been evaluated with a training disjunctive user.

Figure 1: Table of recognition results

%	Top 1	Top 2	Top 3	Top 5	Top 10
HMH	94.2	95.0	97.1	98.3	100.0
HMR	91.8	94.2	95.5	96.3	98.8
HMP	46.5	52.9	57.8	65.1	81.3
RMP	47.2	53.2	58.9	68.3	83.8

7. DISCUSSION

While matching hummed or sung input to hummed references is recognized robust and independent of the reference performer, polyphonic audio matching is still in its beginning. Due to the fact that humming lengths of the users differ greatly from a fraction of the melody up to several repetitions and users tend to expect a fast result, we use a constant humming length of five seconds for the final implementation. This seemed to be a reasonable compromise between a longer data stream for better recognition and a time most users are willing to hum. The HMMs outperform the DTW by 5.3% when matching hummed references. Matching polyphonic audio, they increase recognition rates by 2.8%. The most gain is achieved by use of the stereophonic information. Adaptation of the frame length leads to a further improvement, but can decrease recognition results if the synchronization fails. A dynamic instead of static frame length adaptation allowing changes of tempi will be considered in our future research. We believe that the introduced methods form an important contribution to the field of query by humming matching polyphonic audio. Novel methods of feature extraction and matching could be shown. Our further future research will deal with finding parts within a longer song by suited methods of clustering. In our opinion further improvement in music information retrieval is achieved by integration of more modalities for the query. A multimodal query system using Bayesian Belief networks to combine typed, handwritten, spoken and hummed or sung input on the symbolic layer will therefore be evaluated in view of acceptance by the users.

8. REFERENCES

[1] J. Reiss, M. Sandler: *"Benchmarking Music Information Retrieval Systems,"* JCDL Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retreival (MIR) and Music Digital Library (MDL) Evaluation, Portland, Oregon, July 18th, 2002.

[2] H. Nagano, K. Kashino, H. Murase: *"Fast Music Retrieval using polyphonic binary feature vectors,"* ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, CD-Rom Proceedings.

[3] J. Pickens: *"A Survey of Feature Selection Techniques for Music Information Retrieval,"* Technical report, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA, 2001.

[4] R. Dressler: "Dolby Surround Pro Logic Decoder Principles Of Operation," www.dolby.com.

[5] J. Song, S. Bae, K. Yoon: *"Query by humming: Matching humming query to polyphonic audio,"* ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, CD-Rom Proceedings.

[6] L. Rabiner: *"A tutorial on Hidden Markov Models and selected applications in Speech Recognition,"* Proceedings IEEE, pp. 257-284, February 1989.