Matching Monophonic Audio Clips to Polyphonic Recordings

Björn Schuller, Gerhard Rigoll, Manfred Lang
Institute for Human-Machine Communication, Technische Universität München, D-80333 München, Germany
Email: (Schuller | Rigoll | Lang)@tum.de

Introduction

Growing sizes of nowadays digital music archives and a broadening user spectrum clearly claim for novel retrieval and browsing techniques [1]. Among such content based retrieval by presentation of the key melody as query seems most intuitive [2]. Thereby matching to complete polyphonic tracks as found e.g. on CDs or in MP3 databases tends to be most comfortable, yet challenging. In this work we introduce Hidden Markov Models for this task. Furthermore diverse feature sets for an utmost performance including pitch, energy, harmonic sum based on partial enhancement, MFCCs and LPCCs are compared. We also deal with a reasonable pre-processing respecting characteristics of the musical nature of the signal streams. Such comprise stereophonic aspects, and temporal development. Considering the sparse data of only one reference for model training, namely the song to be retrieved, we extract dominant repetitions within the polyphonic signal to obtain several data using multiple model instances. A working implementation is introduced and results are shown. As test set we chose the MTV Most Wanted collection of the years 1981 - 2000. The query requests were sung by several test users and a recognition performance by humans as basis of comparison is introduced.

Database

The Top 5 Titles of the MTV Europe Most Wanted collection of the years 1981-2000 were chosen as polyphonic music database in MP3 format with 128kps, 44kHz, stereo. Monophonic query clips were assembled of 11 persons, 3 of them female. The clips were hummed or sung freely. The probands were only instructed to retrieve each clip in the polyphonic database. 1,100 clips in total were collected with an average length of 7.46 sec.

Features

Different feature combinations will be evaluated once for matching monophonic query clips to monophonic clips, and once for matching them to polyphonic clips. We firstly chose the rather commonly used pitch contour respectively its first and second order derivatives. As pitch detection algorithm AMDF is applied. Human singers slightly vary the pitch even though they still feel the same note. A quantization of frequencies in semitone-step intervals seems appropriate to solve these fluctuations and has been introduced in many works. It substitutes contour smoothening in a suggestive way by integration of general musical background conditions and reduces the dimensionality of the feature vectors. However, in some foreign or non-well tempered music an adequate different tonal structure exists and claims for an adapted quantization scheme. For example in blues music blue notes resemble smear 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 pitch 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. We use rectangular band filters without overlap.

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

Alternatively the harmonic sum based on partial enhancement is calculated. First the predominance of a partial considering the amplitudes of surrounding signals in the FFT- spectrum within a frequency range is calculated. Thereby eight neighboring bands turned out to be an optimum. We use dB(A) correction in order to model human loudness perception dependant on the frequency. Next the sum over the enhancement of a partial and its higher order harmonic partials is computed [3]. Thereby the sum over three higher order partials turned out to be optimal in our case. It is said that melodic perception reaches from 30 Hz to 4000 Hz. However, our considered spectral band of interest consists of the human singing range from the low D (73.4 Hz) to the high c''' (1046.5 Hz). Likewise we limit to 47 spectral band coefficients in the case of harmonic sum computation. However, higher frequencies are applied in the calculation of harmonicity respecting higher order partials. In order to eliminate invariant parts as harmonic accompaniment we compute differences between the harmonic sum vectors. As the reference and the query pattern are in general in different musical keys, we fulfill exponential frequency scaling for the query clip [4]. Matching in different keys is fulfilled until the optimal key is found. We next consider energy related information as features. It represents the loudness of a musical phrase, and rhythmic information and expression is transferred by it. We calculate the logarithmic mean energy in a frame. To eliminate influences of the absolute energy level we use the first, and second order derivatives. SMA filtering of the contour helps to smoothen slight tremolo. Finally we use 12 MFCC and 17 LPCC coefficients and their derivatives throughout our evaluations.

In order to increase the signal to noise ratio between the melody of interest and the accompanying backing instruments, we consider use of stereophonic spectral arrangement information [5]. 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

Quelle: DAGA 05, München, DEGA, Berlin

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. As we aim to keep only the lead-melody we have to cope with extracting the monophonic part of a recording. Especially in true multi-channel surround-sound recordings this is an easy task as the center-speaker channel is stored separately. We suggest a fast approximation for a pseudo-monophonic signal s_{mon} according to the following equation (2) where s_r represents the signal of the right, and s_l the signal of the left channel at a time instant t. λ resembles a dampening factor.

$$s_{mon} = \lambda \cdot sign(s_r + s_l) \cdot \max\left(\left(|s_r + s_l| - |s_r - s_l|\right), 0\right) \quad (2)$$

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. 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.

HMM Classification

Each melody is represented by one single continuous left-right Hidden Markov Model (HMM). The number of states is set according to the length of the reference pieces. In the recognition phase the maximum-likelihood model is chosen. 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: We aim at finding dominant repetitions within the original audio [6]. The basing assumption is that users will likely hum or sing such often repeated parts of the audio, as chorus or verse rather than sparsely appearing parts as short fills or a bridge. Likewise we construct multiple HMMs per song for these different parts each, and train them with repetitions of tracked parts. As further advantage we need not search through the whole audio in the recognition phase, but only through found parts.

Dominant Repetitions

In order to find dominant repetitions, we first fulfill a frame-wise pre-segmentation based on energy, spectral centroid, roll-off point and MFCCs. By distance metrics we search for instants of significant changes. For each such point we search the adjacent such within a range of 5 and 12 seconds. Afterwards a threefold dynamic programming based comparison takes place: Firstly, we search for the 20 clips with minimum cumulative distance. Secondly, we search the clips with maximum cumulative distance among these. Thereby we find the different parts as chorus, bridge, etc. By thresh holding we decide upon the total number of diverse parts contained within the audio. Finally, we search for the clips with minimum distance to each part, which we consider the repetitions of these parts.

Results and Conclusion

We let eight probands, two of them female, which were not among the singers test, whether they would recognize the searched audio. They knew the database well, and could listen to the original song at any time. Their mean accuracy was 53.6% ± 6%, while the maximum was 86.0%. Considering features, pitch only information alone equaled harmonic sum only information, but was found 2% below inclusion of energy information, and 6% below additional inclusion of MFCC when matching monophonic to monophonic clips *(MM)*. As pitch detection cannot be easily fulfilled within polyphonic music, the harmonic sum features proved optimal when matching monophonic to polyphonic music *(MP)*. MFCCs were thereby outperformed. LPCC use clearly fell behind any of these. When splitting the set in twice 50 songs, resembling the amount of MP3 songs stored on a 256 MB device, the following mean matching performance was obtained for MM matching: Top1: 96%, Top5: 98%. For MP matching we achieved: Top1: 42%, Top3: 70%, Top5: 95%. These results clearly show the challenge of MP compared to MM matching. Still, we believe that Top5 MP matching shows considerable results [3, 4], especially considering that users were allowed to freely hum or sing. In future research we will include search for parts that include singing in the polyphonic audio. Furthermore we want to compare the introduced features to MDCT coefficients used in MP3 audio coding [4].

Literature

[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] B. Schuller, G. Rigoll, M. Lang: *"Multimodal Music Retrieval for Large Databases",* Proc. of ICME 2004, Special Session "Novel Techniques for Browsing in Large Multimedia Collections", Taipei, Taiwan, 2004.

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

[4] W.-N. Lie, C.-K. Su: *Content-based Retrieval of MP3 Songs Based on Query by Singing,* ICASSP 2004, Vol. V, pp. 929-932, Montreal, Kanada, 2004.

[5] B. Schuller, G. Rigoll, M. Lang: *"HMM-Based Music Retrieval Using Stereophonic Feature Information and Framelength Adaptation",* Proc. of ICME 2003, Vol. II, pp. 713-716, Baltimore, MD, USA, 2003.

[6] S. Sood, A. Krishnamurthy: *Extraction of Characteristic Music Textures (Eigen-Textures) via Graph Spectra and Eigenclusters,* IEEE International Conference on Speech, Acoustics, and Signal Processing, Philadelphia, USA, 2004.