Presented at National Conference on Communications, NCC 2002, Jan 26 -28, IIT Bombay 2002

BUILDING A MELODY RETRIEVAL SYSTEM

M. Anand Raju and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
prao@ee.iitb.ac.in

ABSTRACT

This paper describes some early attempts at developing a music indexing and retrieval system based on melody, or tune, of songs. In the envisaged system, the "query", a song fragment whistled or sung by the user into a microphone, is used to search a database of soundtracks to find the entry that is best matched to it in tune. The challenging issues that this project raises are described. Signal processing tools suitable for melody detection are presented, and finally some experimentally obtained results are discussed.

1 INTRODUCTION

Databases which, traditionally, have held only text and numeric data, are increasingly faced with audiovisual data, thanks to growing multimedia applications. Such audiovisual content will have to be indexed and queried using methods natural to the new data types in order to enable intelligent navigation and search. For example, a natural way of searching a database of songs would be by humming a fragment of the desired song.

This paper addresses the problem of building a music indexing and retrieval system based on melody, or the "tune", of the music. While melody is only one of many aspects of a piece of music, it is certainly among its most salient features. This is especially true of songs (vocal music). A melody retrieval system based on acoustic querying would allow a user to hum or sing a short fragment of a song into a microphone and then search and retrieve the "best matched" song from the database.

Figure 1 shows the functional blocks of a basic melody retrieval system. The melody database is essentially an indexed set of soundtracks. The acoustic query, which is typically a few notes whistled, hummed or sung by the user, is processed to identify its melody features. The database is searched to find those that best match the query. The system returns a ranked set of matching melodies, which can be used to retrieve the desired original soundtrack. While the overall task is one that is easily performed by humans, many challenging problems arise in the implementation of an automatic system. These include the signal processing needed for extracting the melody from the stored audio and from the acoustic query, and the pattern matching algorithms to achieve proper ranked retrieval. Further, a robust system must be able to account for inaccuracies in the user's singing. The system will typically operate on a substantial database and must respond within seconds. The recent growth of interest in melody retrieval research is evident in the efforts of major audio research groups including MIT Media Labs [2], Cornell University [3] and Waikato Univ. in New Zealand[1]. The New Zealand group has developed a prototype system with a database of 9600 folk songs.

In this paper, we explore the feasibility of building a melody retrieval system for Hindi film songs (undeniably a significant segment of the audio entertainment industry in India). We begin with a brief introduction to the elements of music. Signal processing methods to extract musically relevant information from the acoustic signal are described next. Finally, we present our exploratory studies on constructing a melody retrieval system for Hindi film songs based on queries that are either whistled or sung by the user.

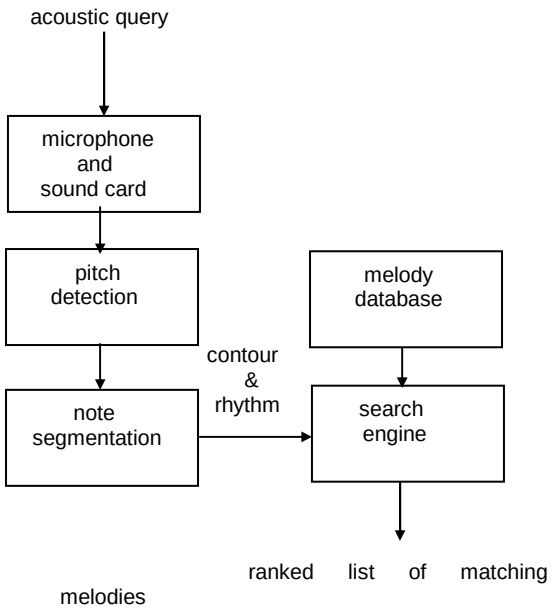


Figure 1. Basic blocks of a melody retrieval system

2. COMPONENTS OF MUSIC

Music, in nearly all cultures, can be described as the structured evolution of rhythm and pitch. Practically all instruments (except perhaps percussion) generate sounds that have a clear pitch, associated with the periodicity of the sound. The pitch or fundamental frequency allows the arranging of sounds ranked low to high on a musical scale. A time sequence of changing pitches (notes) makes up a sound that is musically interesting. That is, what we perceive in music is not only the pitch of individual notes but also how they correspond to particular moments in time. This attribute of music, known as rhythm (or beat), is important enough that even a "noisy" sound played with a strong rhythm sounds "musical". Other attributes that distinguish one piece of music from another are the dynamics (loud/soft), tempo (fast/slow), texture (timbre or voices) and lyrics (for vocal music). However, it is the *sequence of pitches together with the timing information* of each note which specifies the melody line or "tune", the most commonly employed identification of a piece of music. Although the melody is described by the time sequence of pitches, it is evident that people are able to recognize melodies even after pitch transposition (as the same tune played in a different key). For this reason, more characteristic than the absolute pitches of the successive notes are the relative frequency intervals between the notes. This relative variation of pitch in time is known as the "pitch contour", and it provides a dimension which is invariant to key transposition. Apart from pitch contour, the only other dimension in which melodies in general cannot be transformed is the rhythm [2]. The pitch contour and rhythmic information (note duration in beats) thus constitute the two most significant attributes of the melody feature of music. It has been suggested in the literature, however, that instead of representing each pitch interval exactly, the *direction of change*, is the more important element in melody recognition [2]. In this representation, the symbol "U", "D" or "S" is used to indicate whether a note is higher (Up), lower (Down) or same (S) in pitch as the previous note. For example, the pitch contour of the phrase "me-ra-joo-tha-hai-ja-pa-ni" (first 8 notes from the song of the same title) is given by "*,U,U,S,S,D,U,S". The notes are separated by commas, and "*" indicates the first note. In this phrase the notes are equal in time with one note falling on each beat. The trained listener would note that some pitch frequency intervals are larger than others – a fact not evident in the above "direction only" representation. However while capturing the most prominent aspect of the tune, this representation frees the untrained user from singing exact intervals and points to the potential of a more robust system [2].

2 SIGNAL PROCESSING FOR MELODY RETRIEVAL

From the previous section we see that a major issue in melody recognition is to identify the sequence of notes of a music signal and to label each note with a pitch value and a rhythmic value. Only after this step, can the pitch contour representation be derived. Pitch tracking and note segmentation are typically carried out based on the estimation of the instantaneous fundamental frequency and amplitude of the

acoustic signal. Many instruments produce only discrete pitches. (In the Western tonal system there are 12 equal pitch intervals per octave.) The human voice, on the other hand, produces notes with continuously varying pitches and can show considerable pitch variation even within a sung note. In this case assigning a discrete pitch value to a note involves smoothing as well as quantisation to the nearest available discrete pitch value.

Pitch estimation of the human voice, or even of signals generated by real musical instruments, is not a trivial problem due mainly to: (i) period-to-period variations of pitch, and (ii) the fact that the fundamental frequency component may not be strong relative to the other harmonics. There exist several algorithms for detecting the pitch of an acoustic signal. These may be broadly classified into time domain and frequency domain algorithms. The former are computationally simpler, and operate directly on the speech waveform using measurements of peaks, zero-crossings or autocorrelation.

In the present work we have used an autocorrelation-based pitch estimator computed for non-overlapping frames of fixed duration (corresponding to 3 times that of the lowest anticipated pitch period). The local maxima of the autocorrelation versus lag function are identified. To minimize pitch errors arising from selecting a pitch period multiple, the smallest lag with autocorrelation peak greater that 0.3 times the highest autocorrelation peak is taken to be the pitch period estimate for the frame. Spurious peaks, especially near note boundaries, are smoothed out by considering neighbouring estimates. This algorithm, developed initially [4] for use on tonal instruments with wide pitch ranges, has been used successfully for generating pitch tracks from the whistled and sung queries considered in the present project.

3 CREATING A TUNE DATABASE

In the envisaged system, the database is expected to contain melody information of a large set of songs. In order to understand the issues that arise in creating such an indexed database, we undertook the small-scale exercise of computing the pitch contour and rhythm information for a set of 10 prominent phrases from eight well known Hindi film songs. We extracted the phrases of interest from MIDI files (instrument music only) of Hindi film songs available on internet. The MIDI files were converted to acoustic (.wav) files and then processed using the automatic pitch tracker accompanied by manual verification. Table 1 shows the contour and rhythm information for each phrase. The duration of the shortest note in the phrase is taken to indicate one beat. The duration of each of the notes is then represented by a number of beats by adding the needed number of “S” symbols in each note. As an alternative representation that has also been used by some researchers, the contour only representation is provided in Table 2 for the same items.

We observe that for the given set both the representations (contour+rhythm and contour only) provide unique signatures for each of the 10 phrases. That is, any one of the two can be used to uniquely identify the phrase. The contour-only representation is more compact and therefore expected to require simpler matching strategies. However, it is also of interest to consider the minimum number of notes needed to uniquely identify a phrase from the set. We observe that if we consider identification from the first 5 notes only, phrase #5 and phrase #10 cannot be differentiated from the contour-only representation but can from contour+rhythm. This implies that fewer notes will be needed to identify a song if contour+rhythm information is utilized rather than contour only. This has been noted also in [2].

4 EXPERIMENTAL RESULTS AND DISCUSSION

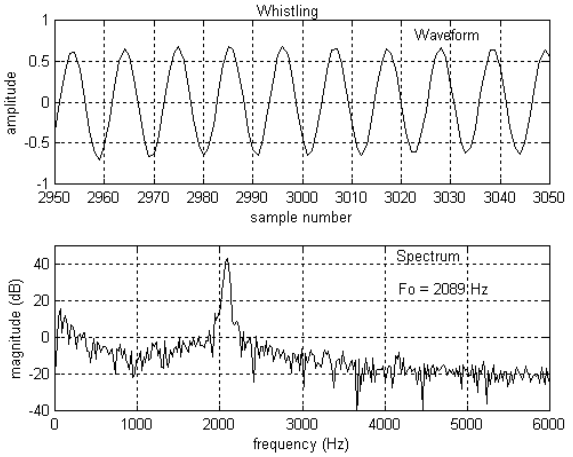
So far we have recorded voice (singing or whistling) data from 3 participants for the set of 10 phrases. While only one of them has had formal training in music, all three were familiar with the songs in the database and were able to sing the desired phrase from memory. Pitch contours and rhythm were extracted automatically for all the queries. Figure 2 shows some typical waveforms and spectra for single

notes of whistling and singing each note with the syllable “la”. The waveforms were recorded via a microphone and PC sound card at the sampling frequency of 22.05 kHz. We see that both waveforms are periodic with the whistling waveform being nearly sinusoidal. On the other hand, the singing voice displays a large number of harmonics. The total pitch range was whistling was observed to be 1-2.6 kHz and that for the singing 180-400 Hz.

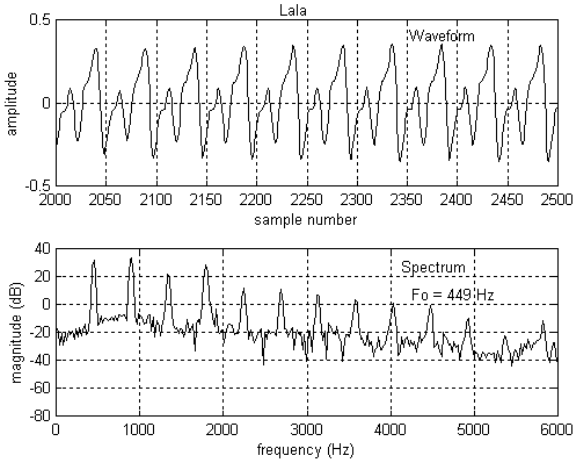
In Fig.3 we show the energy and pitch contours for whistled and sung (“la”) first phrase (7 notes) of the song “Amar, Akbar, Anthony”. The input acoustic signal is divided into frames as described in Sec. 2 and a fundamental frequency estimate is computed for each frame. While rapid changes of pitch can be used to signal note boundaries, in the case of whistling added waveform energy estimation has been used to aid in the detection of note boundaries. We see that the “la-la” waveform shows no discernible amplitude variation at note boundaries. A smoothing operation was carried out to identify a single pitch value for each note. From the two pitch versus time plots, we see that the same tune appears with pitch transposition. However, the “U/D/S” (direction of change) contours are identical! The rhythmic information is also preserved.

Table 2. compares the contour-only (i.e. note duration is neglected) representations of the database items with those extracted from corresponding “queries by whistling” of one of the participants. We observe that the contours match perfectly except for small discrepancies in items 6 and 9. While other participants’ data showed similar or better matching, we noted that or item 6, all the participants deviated from the stored phrase contour in the same manner - prompting us to believe that there is an error in the music of the stored phrase. Rhythm information can potentially serve to improve identification.

In summary, we have demonstrated by a small-scale experiment, the feasibility of developing an automatic system for melody based retrieval from a database of Hindi film songs. The issues raised by this problem have been discussed. We see that much of the work that lies ahead in building a prototype system involves the testing, and enhancing the scale, of the present scheme. Automatic pattern matching algorithms for the database search task remain to be developed. Experimental data from a much larger set of participants needs to be investigated in order to come up with correct classification strategies which must also take into account subjective similarity criteria. It is anticipated that continuing the work on these lines will lead to a demonstrable prototype by the date of the Conference.



(a)



(b)

Figure 2. Waveform and magnitude spectrum for steady sounds from (a) whistling (b) singing / a /

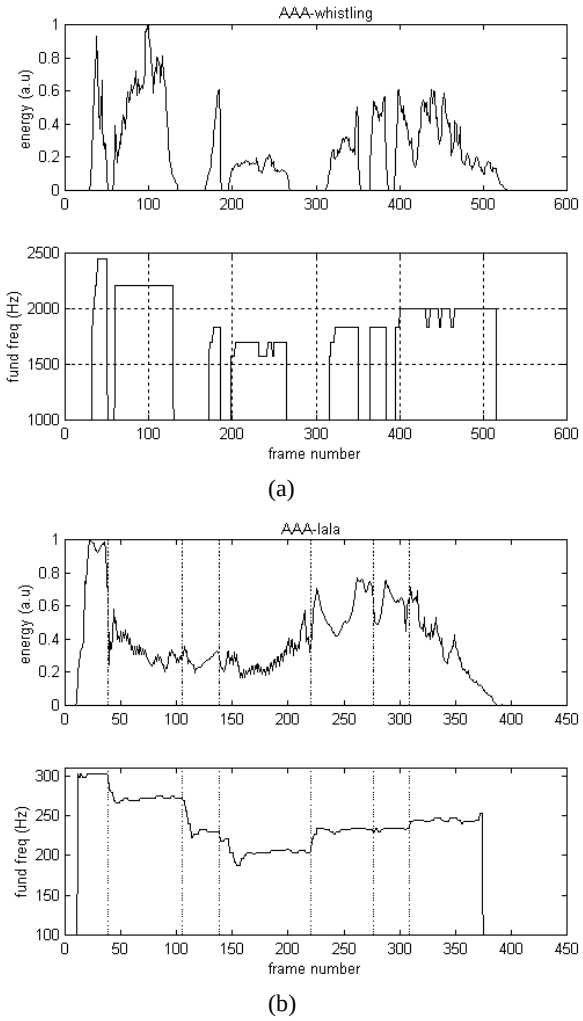


Figure 3. Energy and pitch contour for the song " AAA " as (a) whistled (b) sung with " la " (manually inserted note boundaries in (b))

5. REFERENCES

[1] McNab.R.J, Smith.L.A, Witten.I.H, Henderson.C.L, Cunningham.S.J, " Towards the Digital Music Library: Tune Titrieval form Acoustic Input", *Proc. ACM Digital Libraries*,Bethesda, 1996.

[2] Kim.Y.E, Chai.W, Garcia.R, Vercoe.B, " Analysis of a Contour-based Representation for Melody ", *Proc. International Symposium on Music Information Retrieval*, Oct 2000.

[3] Ghias A, Logan J, Chamberlin D, Smith B.C, " Query By Humming ", *Proc. ACM Multimedia*, San Francisco, 1995.

[4] A.Malot, P. Rao, and V.M.Gadre, " Spectrum Interpolation Synthesis for the Compression of Music Signals, " to appear in *Proc. of the COST* G-6 Conf. On Digital Audio Effects, Ireland,2001

	Name & # of notes	Instrument Contour + Rhythm
1	AAA (7)	*, DSS, D, DSS, US, S, US
2	Ajib (8)	*, DS, U, DS, D, DS, D, US
3	Bazigar (8)	*, U, U, US, D, D, U, DS
4	Churaliya (10)	**, SSS, U, DSS, DSS, SSSS, USSS, DSS, D, DS
5	Kisiki_1 (12)	*, US, D, US, D, US, D, US, D, US, D, US
6	Kisiki_2 (11)	**, D, US, D, US, D, DS, D, US, U, US
7	Mera_joota (8)	*, U, U, S, S, D, U, S
8	Mera_naam_1 (7)	**, SS, DSS, U, US, US, U
9	Mera_naam _2 (8)	**, S, US, S, SS, DS, SS, US
10	Pukarta (9)	*, US, D, USSSSSSS, D, U, U, D, DS

Table 1. Melodic information for the 10 database phrases; Individual notes separated by comas

	Contour from instrument	Contour from whistling
1	*, D, D, D, U, S, U	*, D, D, D, U, S, U
2	*, D, U, D, D, D, D, U	*, D, U, D, D, D, D, U
3	*, U, U, U, D, D, U, D	*, U, U, U, D, D, U, D
4	*, S, U, D, D, S, U, D, D, D	*, S, U, D, D, S, U, D, D, D
5	*, U, D, U, D, U, D, U, D, U, D, U	*, U, D, U, D, U, D, U, D, U, D, U
6	*, D, U, D, U, D, D, D, U, U, U	*, S, U, D, D, D, U, D, U, U, U
7	*, U, U, S, S, D, U, S	*, U, U, S, S, D, U, S
8	*, S, D, U, U, U, U	*, S, D, U, U, U, U
9	*, S, U, S, S, D, S, U,	*, S, U, S, S, D, D, U
10	*, U, D, U, D, U, U, D, D	*, U, D, U, D, U, U, D, D

Table 2. Contour (only) from database and that obtained from whistled query for each phrase of Table 1

.