ISSN 2320 - 2602
Volume 2, No.10, October 2013
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2013/ijacst052102013.pdf

Experiments on Vietnamese Folk Songs Content-based Searching based on Pitch Estimation

Thi-Thu-Hien Phung Thai Nguyen University, Vietnam, pthientng@gmail.com

ABSTRACT

Feature extraction in content-based song searching is required not only to efficiently represent the musical information but also to reduce the redundant information. It leads to the need of choosing the feature vector of music signal that should be close as much as possible with the musical sounds source and the human auditory perception models. Pitch represents excitation source of periodic musical signal. Human are sensitive with the pitch changes rather than that of other acoustic features. Therefore, pitch is an efficient feature in content-based music retrieval. In this paper, we experiment state-of-the-art pitch estimation methods and apply them in a Vietnamese folk song searching system for comparison. The experimental results show that the Cepstral-based method outperforms all other methods. Therefore, we suggest that pitch estimated by the Cepstral-based method is appropriate feature vector in Vietnamese folk song content-based searching in which each song has many word versions but same melody.

Key words: Pitch estimation, content-based music retrieval, ceptrum analysis, dynamic time wrapping

1. INTRODUCTION

Content-based music retrieval is an interesting topic which has been considered by many researchers. One kind of its applications is the song searching in multimedia database. Feature extraction is an essential step in content-based song searching systems which is required not only to well represent the musical information but also to reduce the redundant information. To efficiently characterize the musical information, feature extraction needs to represent closely with the musical sound source models. To reduce the redundant of musical information, feature extraction needs to be built based on the human auditory perception models which keep most of perceptible sounds and discard most of unperceivable sounds. Pitch represents periodic excitation source corresponded with melody of musical signal and it is one of the most important parameters of the musical sound source. Human are sensitive with the pitch changes rather than that of other acoustic features. Therefore, pitch is an efficient feature in content-based music retrieval. In this paper, we experimented state-of-the-art pitch estimation methods and apply in a Vietnamese folk song content-based searching. The database includes several well-known Vietnamese folk songs in which some songs have many word versions with same melodies. Structure of the paper is as follow, section 2 presents pitch contour concepts and methods; section 3 describes the Dynamic Time Wrapping (DTW) algorithm which is used for temporal alignment and to compare the two pitch vectors with different sizes; section 4 presents the Vietnamese folk song database used in our experiments and experimental results, section 5 draws conclusions and gives discussions.

2. PITCH ESTIMATION OF MUSICAL SIGNAL

Fundamentally, there is a fact that audio signal is quasi-periodic signal. Although audio signal is not a pure sine wave, they will be similar from one period to the next, this smallest period called as pitch period or pitch. Pitch is inversely proportional to the fundamental frequency of audio signal which is defined as the lowest frequency component in Fourier analysis of the signal. Perceived pitch is an important characteristic of audio signal and an appropriate estimation of pitch would be valuable when characterizing audio files. There are many pitch estimation methods which can be done in time or frequency domain. In this research, we choose three most success and popular pitch estimation algorithms for experimental implementation. These are the ACF (Autocorrelation Function), the AMDF (Average Magnitude Difference Function) in time domain and the Cepstrum Analysis in frequency domain.

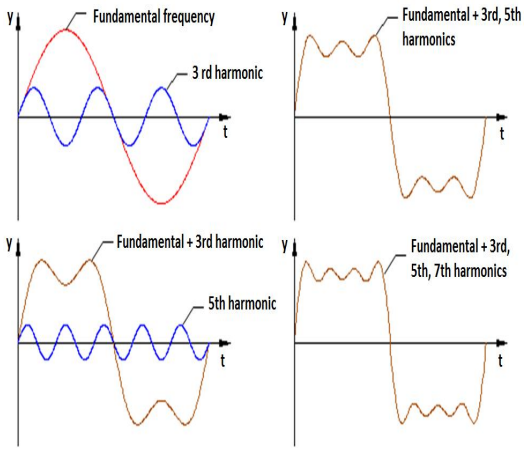

Figure 1. Fundamental frequency of quasi-periodic audio signal

2.1 PITCH ESTIMATION USING AUTO CORRELATION FUNCTION

The correlation between two waveforms is a measure of their similarity. The waveforms are compared at different time intervals, and their similarity is calculated at each interval.

The autocorrelation function is the correlation of a waveform with a time shifted version of itself. For a finite discrete function $s(m)$ of size N, where k is the shifted interval, the mathematical definition of the autocorrelation function is shown as:

$$r(k) = \sum_{m=0}^{N-1-k} s(m)s(m+k) \qquad (1)$$

The first peak in the autocorrelation indicates the pitch period of the waveform.To detect the pitch, we take a window of the signal, with a length at least twice as long as the longest period that we might detect.

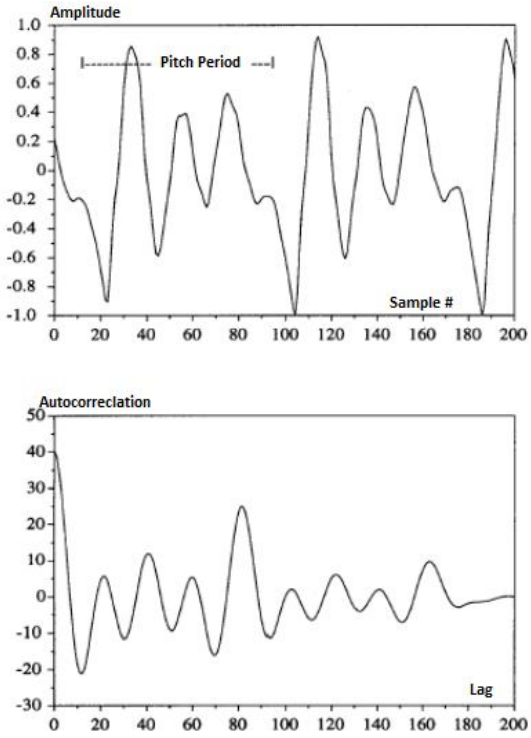

Figure 2. Waveform (top panel) and Autocorrelation (bottom panel) in the time domain

2.2 PITCH ESTIMATION USING AMDF

AMDF is a modified version of ACF in which, we use the difference of a framed signal a time shifted version of it self instead of multiply them as in the original auto-correlation function. This modification helps to optimize the ACF algorithm which uses the subtraction instead of multiplication.
The AMDF is defined in an audio frame sized N as in [5]:

$$d(k) = \sum_{m=0}^{N-1-k} \left| s(n) - s(n+k) \right| \qquad (2)$$

The pitch period k_0 is chosen when the $d(k_0)$ is minimum.

2.3 PITCH ESTIMATION USING CEPTRUM ANALYSIS

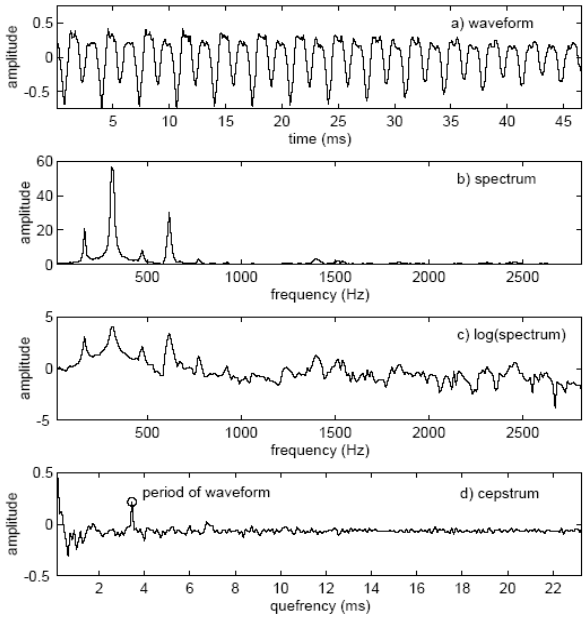

Figure 3. Pitch estimation based on ceptrum analysis

Cepstrum analysis is a kind of spectral analysis where the output is the Fourier transform of the log of the magnitude spectrum of the input waveform [1].
Supposed that $x(n)$ and $X(e^{j\omega})$ are the time-domain waveform and its spectrum. The ceptral $c(n)$ is computed as:

$$c(n) = \frac{1}{s\pi} \int_{-\pi}^{\pi} \log \left| X(e^{j\omega}) e^{j\omega n} d\omega \right| \qquad (3)$$

Naturally occurring partials in a frequency spectrum are often slightly inharmonic, and the cepstrum attempts to mediate this effect by using the log spectrum. The independent variable related to the cepstrum transform has been called “quefrency”, and since this variable is very closely related to time [2] it is acceptable to refer to this variable as time.
This method is based on the fact that the Fourier transform of a pitched signal usually has a number of regularly spaced peaks, representing the harmonic spectrum of the signal. When the log magnitude of a spectrum is taken, these peaks are reduced, their amplitude brought into a usable scale, and the result is a periodic waveform in the frequency domain, the period of which (the distance between the peaks) is related to the fundamental frequency of the original signal. The Fourier transform of this waveform has a peak at the period of the original waveform.
Figure 3 shows the progress of the cepstral algorithm.

3. DYNAMIC TIME WRAPPING

In this paper, the DTW algorithm which is used for temporal alignment and to compare the two pitch vectors with different sizes. Although there are many other method to solve this problem, this paper just uses on DTW since we focus on the feature extraction rather than the methods of marching two pitch vectors.

In [3], the time warping problem is stated as follows: Given two signals X and Y, of lengths |X| and |Y|,

$$X = x_1, x_2, .., x_i, ..x_{|X|}$$

$$Y = y_1, y_2, .., y_i, ..y_{|Y|}$$

construct a warp path W, $W = w_1, w_2, .., w_k$,

$$\max(|X|, |Y|) \le K < |X| + |Y|$$

where K is the length of the warp path, and the k^{th} element of the warp path is

$$w_k = (i, j) \quad (4)$$

where i is an index of signal X, and j is an index of signal Y. The warp path starts at the beginning of each time series at w_1=(1, 1) and finishes at the end of both time series at w_K=(|X|, |Y|).

There is also a constraint on the warp path that forces i and j to be monotonically increasing in the warp path, which is why the lines representing the warp path in Figure 4 do not overlap. Every index of both signals must be used. Stated more formally:

$$w_k = (i, j), w_{k+1} = (i', j'), i \le i' \le i+1, j \le j' \le j+1 \quad (5)$$

An optimal warp path is a minimum distance warp path, where the distance (or cost) of a warp path W is:

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj}) \quad (6)$$

$Dist(w_{ki}, w_{kj})$ is the distance between the two data point indexes (one from X and one from Y) in the k^{th} element of the warp path.Dynamic programming is used to find this minimum-distance warp path between two signals.

A two-dimensional |X| by |Y| cost matrix D, is created where the value at D(i, j) is the minimum distance of a warp path for the two signals X'=x1,...,xi and Y'=y1,...,yj. D(|X|, |Y|) contains the minimum distance of a warp path between signals X and Y. Both axes of D represent time. The x-axis is the time of signal X, and the y-axis is the time of signal Y.

Figure 5 shows an example of a cost matrix and a minimum distance warp path traced through it from D(1, 1) to D(|X|, |Y|).

If the warp path passes through cell D(i, j) of the cost matrix, then the i^{th} point in signal X is warped to the j^{th} point in signal Y. If X and Y were identical, the warp path would be a a linear warp path.

Single points in one signal can map to several points in the other. Since a single point may map to multiple points in the other signals, the signals do not need to be of equal length.

To find a minimum distance warp path, every cell in the cost matrix must be filled. We use dynamic programming because the solutions are already known for all slightly smaller portions of that signals that are a single data point away from lengths i and j, then the value at D(i, j) is the minimum distance for all these smaller signals, plus the distance between the points ii and jj.

Since the warp path must either increase by one or stay the same along the i and j axes, the distances of the optimal warp paths one data point smaller than lengths i and j are contained in the matrix at D(i-1, j), D(i, j-1), and D(i-1, j-1). So the value of a cell in the cost matrix is

$$D(i, j) = Dist(i, j) + \min[D(i-1, j), D(i, j-1), D(i-1, j-1)] \quad (7)$$

The warp path to D(i, j) must pass through one of those three cells, and since the minimum warp path distance is already known for them, all that is needed is to add the distance between the current pair of points, Dist(i, j), to the smallest value in those three cells.

The cost matrix is filled one column at a time from the bottom up, from left to right. After the entire matrix is filled, a warp path must be found from D(1, 1) to D(|X|, |Y|). The warp path is calculated backwards, starting at D(|X|, |Y|).

A greedy search evaluates three nearby cells: to the left, below, and diagonally to the bottom-left. Whichever of these three cells has the smallest value is then added to the beginning of the warp path, and the search continues from that cell until D(1, 1) is reached.

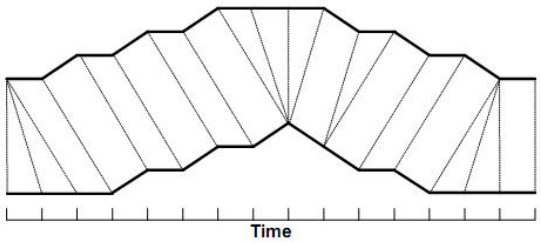

Figure 4. A warping between two signals.

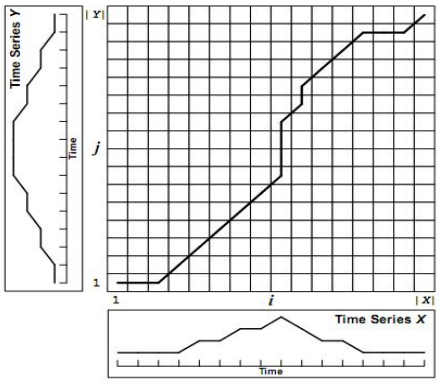

Figure 5. A cost matrix with a warp path

4. EXPERIMENTS

4.1 VIETNAMESE FOLK SONGS

Vietnamese have a long lasting culture with thousands folk songs. Each Vietnamese ethnic group has its own kinds of folk songs. Each sub-region of the nation has its own folk songs too. In social view, the need of collecting Vietnamese folk songs into a large multimedia database and building a convenient searching tool for people are indispensable to preserve the Vietnamese ancient culture as well as to popularize Vietnamese culture to people all over the world.

The Vietnamese folk songs are classified into two kinds: the original and the adapted songs. The songs adapted from an original song almost keep the original melody and just modify the word [4]. Many Vietnamese people are familiar with the core melody of some famous folk songs but just a few ones can remember the names and the lyrics of the songs. Therefore, content-based music retrieval is appropriate to apply in Vietnamese folk songs searching system.

4.2 VIETNAMESE FOLK SONGS DATABASE

Our database for experimenting was collected from the public website http://dancavietnam.net/. This database provides approximately 1000 songs cover most kinds of Vietnamese folk songs of most Vietnamese ethnic minority groups from all sub-regions of Vietnam. All original songs which use different audio formats were changed to the standard PCM wave format with the following parameters: the sampling frequency was 44 KHz, the number of bit per sample wais 16, using both left and right channel in stereo mode.

The database was managed in categories indexed by name, singer, folk typeof the song, and name of the original folk song.

4.3 EXPERIMENTS

For fast implementation, we changed the audio mode to mono which used only 1 channel. The audio files were resampled at 16 KHz, using 16 bits to decode one sample. Because of time limitation, we only used 100 songs selected from our database in which all songs had different melody with each others. These whole signals of songs were used for training; for testing, we chose only one stable part from each song with the duration approximately 5 s. Thus, we had 100 musical short samples correspondents with 100 trained songs.

The general diagram of our song searching system is shown in figure 6.Three pitch estimation methods mentioned above were used to extract the pitch vectors used as a signature of musical waveform for training and testing. The pitch vectors were aligned by DTW and stored in training steps. In testing step, trained pitch vectors were loaded, we used DTW again to compare the input pitch vector and each of trained pitch vectors. Finally, the system returned the most similar result.

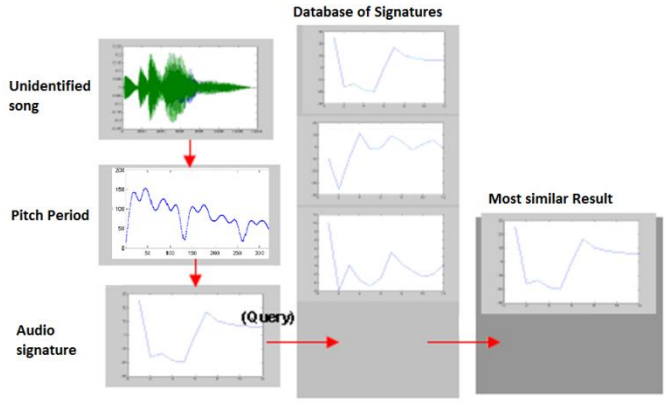

Figure 6. *General diagram of the song searching system*

We used the DTW for aligning and matching the pitch sequence as same in all three experiments. The first, second and third experiment used the ACF, AMDF and Ceptral analysis relatively to estimate the pitch vectors. In all experiments, the frame-size was fixed as 256 ms.
After training all 100 songs, we searched each of 100 musical samples correspondent with the trained files in turn. The correct searching rates and the computation time run on MATLAB 7.0 were depicted in the Table 1. These results show that, the ACF and AMDF were simple but less accurate than the Cepstral algorithm. The AMDF algorithm had the less computation cost while that of the Cepstral and ACF algorithms were almost approximate. Thus we conclude that the Cepstral method outperformed the two mentioned time-domain methods and might be appropriate in content-based music retrieval.

TABLE 1. EXPERIMENTAL RESULTS

Pitch Estimation Method	Recognition Rates (%)	Average Searching Time (s)
ACF	81	9.8
AMDF	83	7.5
Cepstral	94	10.2

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we presented the role of pitch as a feature vector for content-based music retrieval. We investigated the three pitch estimation methods built on time and frequency domain. After that, we conducted some experiments to evaluate the performance of each method. The Cepstral method seems the most accurate method with acceptable computation cost.

In this research, we also investigated the Vietnamese folk songs to suggest that content-based song searching is indispensable for building a multimedia database of this song as well as building a searching tool for users.

In the next research, we will study the human auditory models to estimate the pitch more natural with human perception. We will investigate the pitch estimations in time-frequency

domain in order to propose an efficient pitch estimation method used for content-based music retrieval, and we will also develop a Vietnamese folk song searching system based on the method studied.

REFERENCES

1. B. P. Bogert, M. J. R. Healy, and J. W. Tukey: **The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking, Proceedings of the Symposium on Time Series Analysis** *(M. Rosenblatt, Ed)* Chapter 15, 209-243. New York: Wiley (1963)
2. Curtis Roads. **The Computer Music Tutorial**, MIT *Press, Cambridge* (1996)
3. Sakoe, H. and Chiba, S. **Dynamic programming algorithm optimization for spoken word recognition**, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1) pp. 43- 49, ISSN: 0096-3518 (1978)
4. Vinh Phuc, **Correlation between folk and scientific factors in Hue folk songs**, *Vietsciences,* 03, (2008) (Vietnamese)
5. W. HESS, **Pitch Determination of Speech Signals**, *Springer-Verlag Publisher* (1983).