Warping Indexes with Envelope Transforms for Query by Humming

Yunyue Zhu
Dennis Shasha

Aarti Bindlish
Bindlish@usc.edu

Know the song but can't remember its name?

- You just sing into a microphone which is connected to a computer base station and the computer then resynthesises what you have been singing and gives back result
  - Consider music as Time series
- Exploit and improve well developed techniques from time series databases to index the music for fast similarity queries.

Query by humming (QBH)

- Particular case of query by content
  - Symbolic database of melodies
  - General acoustic database

Contribution of this paper

- Assertion → Time series database techniques especially 'Dynamic Time Warping Index' can be applied to build a fast and robust query by humming database system.

Melodic Contours

- Contour → Sequence of Relative differences in pitch between successive notes
- Represented by few letters
  - "U", "D" or "S" represent a note is above, below or same as the previous one.
  - 'u', 'd' or 's' give a finer measure

Difficulty with contours

- Contour information alone is not enough.
- Hard to segment a user's humming into discrete notes.
  - User should be intelligent enough to clearly hum the notes of the melody using syllables 'ta', 'la' and 'da'.

Strategy Highlights

- Treat both melody in the database as well as humming input as time series
- Indexing scheme invariant to shifting, time scaling and local time warping
- Improve on the State-of-the-art indexing technique by providing a better lower-bound distance for DTW
- A general method for dimensionality reduction transforms on time series envelopes
- Scalable to large music databases

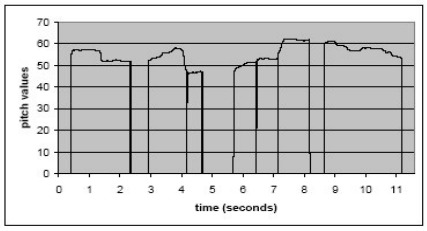
Architecture of Humming System

- User Humming –input hum query
- Database of music
- Database indexing scheme for efficient retrieval of hum-query

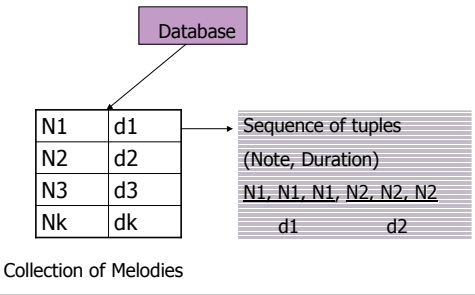
Input hum-query

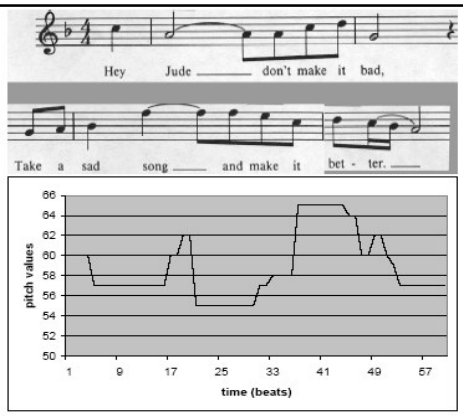
- Acoustic input is segmented into frames of 10ms each and each frame is resolved into a pitch
- Problem → Hard for human being to mark the borders between notes.

Pitch time series



Music database





What should we expect from the mediocre performer?

- Flexibility in Absolute Pitch
  - People of different gender, age and in different moods will hum with different pitch.
  - We need shift invariant technique
    - Subtract average pitches from the time series before matching

Variation in Tempo

- Normally a melody will be hummed at a tempo that ranges from half to double the original tempo
- As tempo changes, the duration of each note changes proportionally
- We need uniform stretching/squeezing of time axis called Time Scaling

Variation in Relative Pitch

- Nearest Neighbor Query Approach
  - Distance between humming and candidate time series = sum of differences at each sampling moment

Local time Variation

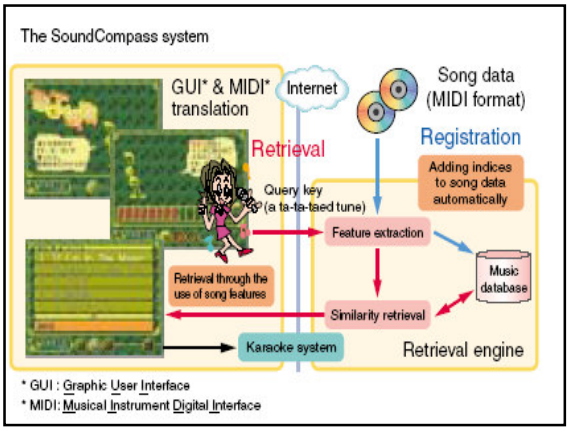
- Idea is to stretch and squeeze the time axis locally to minimize the point to point distance of two time series.

Dimensionality reduction

- Hard to index time series data because of their high dimension
  - Given a time series, a dimensionality reduction transform T will reduce it to a lower dimension

$$\vec{X}^N = T(\vec{x}^n), N << n. \quad \vec{X}^N$$

This is called Feature Extraction



DTW Distance

$$D^2_{DTW}(\vec{x},\vec{y}) = D^2(First(\vec{x}), First(\vec{y}))$$

$$+min\begin{cases} D^2_{DTW}(\vec{x}, Rest(\vec{y})) \\ D^2_{DTW}(Rest(\vec{x}), \vec{y}) \\ D^2_{DTW}(Rest(\vec{x}), Rest(\vec{y})) \end{cases}$$

Uniform Time Warping (UTW)

- Special Case of DTW
  - Warping path must be diagonal
  - The uniform warping distance between two series =

$$D^2_{UTW}(\vec{x}^n, \vec{y}^m) = \frac{\sum_{i=1}^{mn}(x_{\lceil i/m \rceil} - y_{\lceil i/n \rceil})^2}{mn}$$

Stretch both axis to be mn

Using UTW, we can compute distance between time series of different lengths

Local Dynamic Time Warping

- Two step transform
  - Stretch the two time seriws globally to same length
  - Compare them locally point to point with some warping within small neighborhood

$$D^2_{LDTW(k)}(\vec{x},\vec{y}) = D^2_{constraint(k)}(First(\vec{x}), First(\vec{y}))$$

$$+min\begin{cases} D^2_{LDTW(k)}(\vec{x}, Rest(\vec{y})) \\ D^2_{LDTW(k)}(Rest(\vec{x}), \vec{y}) \\ D^2_{LDTW(k)}(Rest(\vec{x}), Rest(\vec{y})) \end{cases}$$

$$D^2_{constraint(k)}(x_i, y_j) = \begin{cases} D^2(x_i, y_j) & if |i-j| \le k \\ \infty & if |i-j| > k \end{cases}$$

Lower bound Technique

- Container Invariant transformation

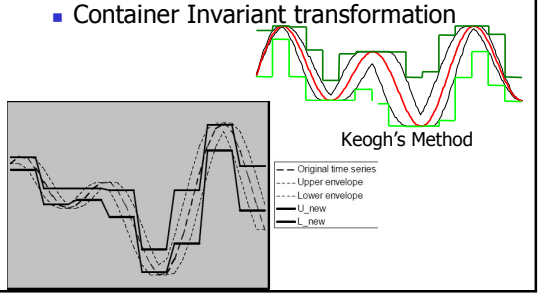


Table 2: The number of melodies correctly retrieved using different approaches

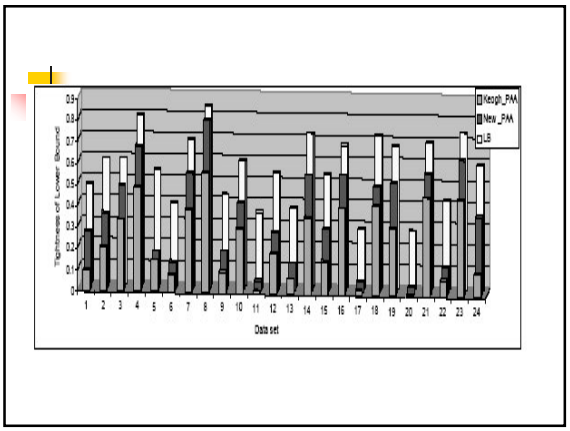
Rank	Time series Approach	Contour Approach
1	16	2
2-3	2	0
4-5	2	0
6-10	0	4
10-	0	14

Table 3: The number of melodies correctly retrieved by poor singers using different warping widths

Rank	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.2$
1	2	4	2
2-3	2	3	5
4-5	4	5	7
6-10	3	5	4
10-	9	3	2

Tightness measure Comparison

- Tightness of lower bound T:
  - (Lower bound of DTW based on dimensionality reduction)/True DTW distance


Keogh_PAA
New_PAA
LB
Tightness of Lower Bound
Data set


So it actually can make you a better singer?
It can make you to train and to improve your singing, yes.
Doh