HUMMING METHOD FOR CONTENT-BASED MUSIC INFORMATION RETRIEVAL

Cristina de la Bandera, Ana M. Barbancho, Lorenzo J. Tardón, Simone Sammartino and Isabel Barbancho
Dept. Ingeniería de Comunicaciones, E.T.S. Ingeniería de Telecomunicación
Universidad de Málaga, Campus Universitario de Teatinos s/n, 29071, Málaga, Spain
{cdelabandera, abp, lorenzo, ssammartino, ibp}@ic.uma.es

ABSTRACT

In this paper a humming method for music information retrieval is presented. The system uses a database with real songs and does not need another type of symbolic representation of them. The system employs an original fingerprint based on chroma vectors to characterize the humming and the references songs. With this fingerprint, it is possible to get the hummed songs without needed of transcription of the notes of the humming or of the songs. The system showed a good performance on Pop/Rock and Spanish folk music.

1. INTRODUCTION

In recent years, along with the development of Internet, people can access to a huge amount of contents like music. The traditional information retrieval systems are text-based but this might not be the best approach for music. There is a need for retrieving the music based on its musical content, such as humming the melody, which is the most natural way for users to make a melody based query [3].

Query by humming systems are having a great expansion and their use is integrated not only in computer but also in small devices like mobile phones [10]. A query by humming system can be considered as an integration of three main stages: construction of songs database, transcription of users' melodic information query and matching the queries with songs in the database [5].

From the first query by humming system [3] to nowadays, many systems have appeared. Most of these systems use Midi representation of the songs [2], [6], [9] or they process the songs to obtain a symbolic representation of the main voice [8] or, also, these systems may use special formats such as karaoke music [11] or other hummings [7] to obtain the Midi or other symbolic representation [9] of the main voice of the songs in the database. In all the cases the main voice or main melody must be obtained because it is the normal content of the humming. Somehow, the normal query by humming systems are based on the melody transcription of the humming queries [5], [7], [11] to be compared with the main voice melody obtained from the songs in the database.

The approach employed in this paper is rather different from other proposals that can be found in the literature. The database contains real stereo songs (CD quality). These songs are processed in order to enhance the main voice. Then, the humming as well as the signal with the main voice enhanced, follow the same process: fingerprints of the humming and of the main voice are obtained. In this process, it is not necessary to obtain the onset or the exact tone of the sound, so, this fingerprint is a robust representation for the imprecise humming or main voice enhancement.

The paper is organized as follows. Section 2 will present a general overview of the proposed method. Section 3 will present the method of enhancement of the main voice of a stereo sound file. Next, section 4 will propose the fingerprint used to compare the humming and the songs. Section 5 will present the comparison and search methods used in the proposed system. Section 6 will present some performance results and finally, Section 7 draws some conclusions.

2. OVERVIEW OF THE PROPOSED METHOD

In this section, a general overview of the structure of the humming method for MIR is given. Figure 1 shows the general structure of the proposed method in which both the humming and the songs with the main voice enhanced follow the same process.

As Figure 1 shows, a phrase fragmentation is needed for the songs. The reason for this is the following: when people sing or hum after hearing a song, they normally sing certain musical phrases, not random parts of the songs [11]. So, the main voice enhancement will be performed in the phrases of the songs. The result of the main voice enhancement of the phrases of the songs and the humming pass through a preprocessing stage that obtains a representation of these

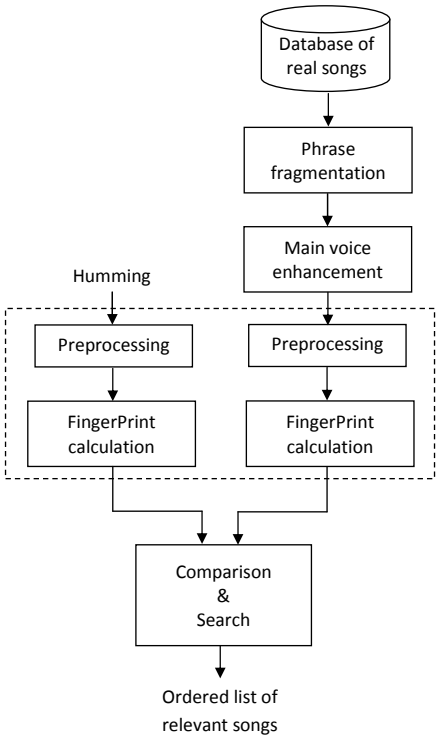


Figure 1. General structure of the proposed method.

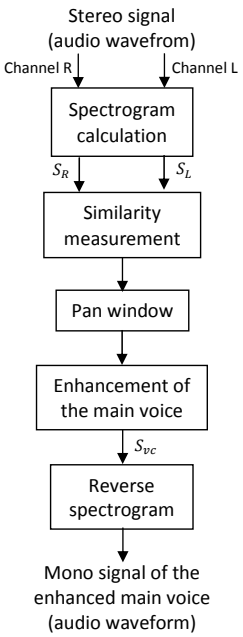


Figure 2. General structure of the process of enhancement of the main voice.

signals in the frequency domain. Then, the fingerprints are calculated. The fingerprints are the representation used for the comparison and search of the humming songs and humming. Note that, the proposed method does not perform any conversion to Midi or other symbolic music representation. Finally, the system provides a list of songs ordered by their similitude with the humming entry.

3. ENHANCEMENT OF THE MAIN VOICE

The reference method selected to enhance the main voice is based on the previous knowledge of the pan of the signal to enhance [1]. The database considered contains international Pop/Rock and Spanish folk music. In this type of music the main voice or melody of the songs is performed by a singer and this voice is placed in the center of the audio mix [4].

In Fig. 2, the general structure of the algorithm of enhancement of the main voice is presented. The base of this algorithm is the definition of the stereo signal produced by a recording studio. A model for this signal is as follows:

$$x_c(t) = \left[\sum_{i=1}^{N} a_{cj} s_j(t)\right] \quad (1)$$

where: N is the number of sources of the mix, the subscript c indicates the channel (1-left and 2-right), a_{cj} are the amplitude-panning coefficients and $s_j(t)$ are the different audio sources. For amplitude-panned sources it can be assumed that the sinusoidal energy-preserving panning law is $a_{2j} = \sqrt{(1 - a_{1j}^2)}$, with $a_{1j} < 1$.

The spectrogram is calculated in temporal windows of 8192 samples for signals sampled to $44100Hz$. This selection is a balance between temporal resolution ($0.18s$) and frequency resolution ($5Hz$).

The panning mask, $\Psi(m,k)$, is estimated using the method proposed in [1], based on the difference of the amplitude of the spectrograms of the left channel ($S_L(m,k)$) and right channel ($S_R(m,k)$). The values of $\Psi(m,k)$ vary from -1 to 1. To avoid distortions due to abrupt changes in amplitude between adjacent points of the spectrogram produced by the panning mask, $\Psi(m,k)$, a Gaussian window function is applied to $\Psi(m,k)$ [1]:

$$\Theta(m,k) = \nu + (1-\nu) \cdot e^{-\frac{1}{2\xi}(\Psi(m,k)-\Psi_o)^2} \quad (2)$$

where Ψ_o is the panning factor to locate (from -1 totally left and 1 totally right), ξ controls the width of the window that has an influence in the distortion/interference allowed, that is, the wider the window, the lower distortion but the larger the interference between other sources and vice versa. ν is a floor value to avoid setting spectrogram values to 0.

The enhancement of the main voice is made as:

$$S_{vc}(m,k) = (S_L(m,k) + S_R(m,k)) \cdot \Theta(m,k) \cdot \beta \quad (3)$$

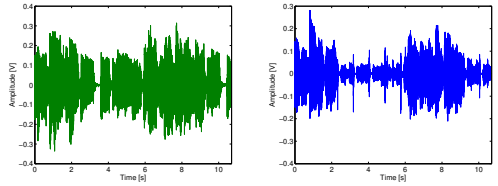
where $S_{vc}(m,k)$ is the spectrogram of the signal with the main voice enhanced. Once the spectrogram $S_{vc}(m,k)$ is obtained, the reverse spectrogram is calculated to obtain the waveform of the enhanced main voice (Figure 2).

The parameters of equation 2, have been set experimentally to achieve a good result in our humming method. The selected values are: $\nu = 0.15$, $\Psi_o = 0$ due to the fact that the desired source is in the center of the mix and ξ is calculated with the following equation:

$$\xi = -\frac{\Psi_c - \Psi_o^2}{20logA} \quad (4)$$

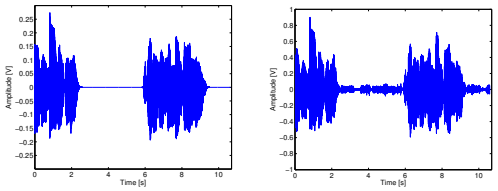
where $\Psi_c = 0.2$ is the margin around Ψ_o where the mask will have an amplitude A such that $20logA = -60dB$ [1].

There are several conditions that are going to negatively affect the localization of the main voice; the overlapping of sources with the same panning and the addition of digital effects, like reverberation. However, since the aim of the proposed method is just the enhancement of the main voice, certain level of interference can be allowed to avoid distortions in the waveform of the main voice.



(a) Left channel of an stereo signal

(b) Right channel of an stereo signal



(c) Original main voice signal without any mixer

(d) Signal with the enhacement of the main voice

Figure 3. Waveforms of the (a) left channel and the (b) right channel of an stereo signal. (c) Original main voice without any mixer. (d) Waveform obtained after the process of enhancement of the main voice

As an example of the performance of the enhancement process of the main voice, Figure 3 shows the waveform of the two channels of a stereo signal (Figure 3(a) and Figure 3(b)), the original main voice (Figure 3(c)) and the waveform obtained after our main voice enhancement process (Figure 3(d)). Theses figures show how the main voice is extracted from the mix although some distortion appears. This happens because the gaussian window selected is designed to avoid audio distortion but it allows some interference.

4. FINGERPRINT CALCULATION

Figure 4, shows the block diagram of the fingerprint calculation procedure for the humming and the music in the database. Two main stages can be observed: the preprocessing and the chroma matrix calculation. In subsection 4.1, the preprocessing stage is presented and then, in subsection 4.2, the estimation of the chroma matrix, the fingerprint, is presented.

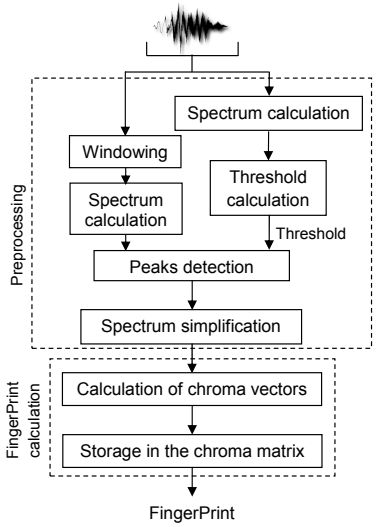


Figure 4. Block diagram of the fingerprint calculation.

4.1 Preprocessing of humming and music database

In the preprocessing, the first step consists on calculate the spectrum of the whole signal, to determine the threshold. The threshold is fixed to the $75th$ percentile of the values of the power spectrum. This threshold determines the spectral components with enough power to belong to a voice fragment. Now, the signal is windowed without overlapping with a Hamming window of 8192 samples. For each window the spectrum is computed. Then, we select the frequency range from $82Hz$ to $1046Hz$, that corresponds to $E2$ to $C6$, because this is a normal range for signing voice.

In this range, a peaks detection procedure is performed. The local maxima and minima are located and the ascending and descending slopes are calculated. We consider significative peaks the maxima detected over the threshold that present an ascending or descending slope larger than or equal to the 25% of the maximum slope found. Between these peaks, the four peaks with larger power are selected to represent the tonal distribution of the window. Ideally, the four peaks selected should correspond to the fundamental frequency and the first three harmonics of the signing note. The number of peaks has been restricted to four because the objective is just to gather information of the main voice (monophonic sound), which has several interferences from other sound sources, or because of the enhacement process of the main voice (Section 3). If we selected more peaks, these peaks would corresponding to other notes different from the notes sung by the main voice and then, the comparison with

the humming would be worse. In Fig. 5, an example of this process is shown.

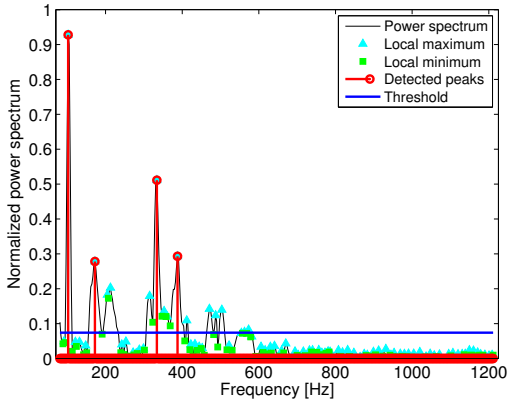


Figure 5. Example of peaks selected.

Next, the new signal spectrum that contains just the selected peaks, is simplified making use of the Midi numbers. The frequency axis is converted to Midi numbers, using:

$$MIDI = 69 + 12log_2\left(\frac{f}{440}\right) \quad (5)$$

where MIDI is the Midi number corresponding to the frequency f. The simplification consists of assigning to each of the selected peaks the nearest Midi number. When two or more peaks are fixed to the same Midi number, only the peak with the largest value is taken into account. The simplified spectrum is represented by $X_s(n)$. In our case, the first element of the simplified spectrum, $X_s(1)$, represents the spectral amplitude of the note $E2$, that corresponds with the frequency $82Hz$ (Midi number 40). Likewise, the last element of the simplified spectrum, $X_s(45)$, represent the spectral amplitude of then note $C6$, that corresponds to the frequency $1046Hz$ (Midi number 84).

4.2 Chroma matrix

Now, to obtain the fingerprint of each signal, the chroma matrix, the chroma vector is computed for each temporal window. The chroma vector is a 12-dimensional vector (from C to B) obtained by the sum of the spectral amplitudes for each tone, spawning through the notes considered (from $E2$ to $C6$). Each $k-th$ element of the chroma vector, with $k \in \{1, 2, \cdots, 12\}$ of the window, t, is computed as follows:

$$chroma_t(k) = \sum_{i=0}^{3} X_s((k+7)_{mod\ 12} + 12 \cdot i + 1) \quad (6)$$

The chroma vectors for each temporal window t are computed and stored in a matrix denominated chroma matrix, $\mathbf{C}$. The chroma matrix has 12 rows and a column for each of the temporal windows of the signal analyzed.

In order to unify the dimensions of all the chroma matrices of all the phrase fragments of the songs and humming, the matrix is interpolated. To perform the interpolation, the number of selected columns is 86, this value corresponds, approximately, to 16 seconds. This number of columns has been selected taking into account the length of the phrase fragments of the songs in the database and the reasonable duration of the humming. Let $\mathbf{C} = \left[\bar{F}_1, \bar{F}_2, \cdots, \bar{F}_{86}\right]$, denote this matrix, where $\bar{F}_i$, represents the column i in the interpolated chroma matrix that represents the fingerprint.

In Figure 6, an example of a chroma matrix with interpolation is represented.

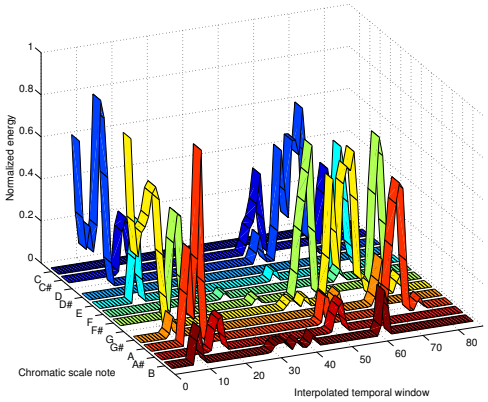


Figure 6. Chroma matrix with interpolation.

5. COMPARISON AND SEARCH METHOD

Once the fingerprint has been defined, the fingerprints for each phrase fragment of the songs in the database are computed. Now, the task is to find the song in the database that is the most similar to a certain humming. To this end, the fingerprint of the humming is obtained, then, the search for the most similar fingerprint is made. This search is based on the definition of the distance between the fingerprint of the humming signal and the fingerprints of the songs in the database.

The objective is to create a distance vector with length equal to the number of phrase fragments in the database. Then, a list of ordered songs from the most similar song to the less similar one can be obtained. The distance between fingerprints is computed using:

$$Dst_k(\mathbf{C}^{humm}, \mathbf{C}^k) = median\left(\{d_{kj}\}\right) \quad (7)$$

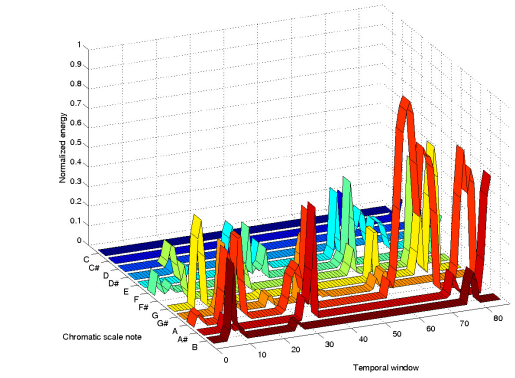
$$d_{kj} = \left\|\bar{F}_j^{humm} - \bar{F}_j^k\right\| \quad (8)$$

where Dst_k is the distance of the humming to a phrase fragment k, k is the index of all the phrase fragments in the database. $\mathbf{C}^{humm}$ is the fingerprint of the humming and $\mathbf{C}^k$

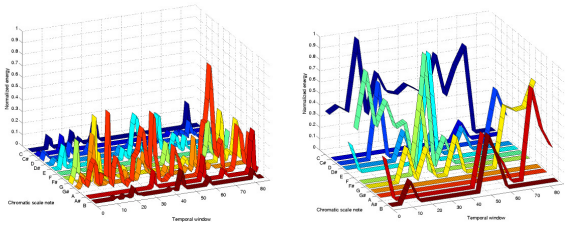
is the fingerprint of each phrase fragment. The euclidean distance between columns of the fingerprints d_{kj}, is calculated. Afterwards, the median of the set of euclidean distances, $\{d_{kj}\}$, is stored in Dst_k.

The distance values Dst_k are ordered from the smallest value to the largest value. Now, since for each song several phrase fragments have been considered, the phrase closest to the humming is selected to define the closest songs. The list of similar songs is created likewise.

An illustration of the utilization of the fingerprints to find similar songs to a given humming is shown in Figure 7. The fingerprint of a humming (Figure 7(a)), the nearest song, that is, the corresponding song (Figure 7(b)) and the farthest song (Figure 7(c)) are presented. It can be observed how the fingerprint of the humming and the corresponding song look very similar. On the contrary, the fingerprint of the farthest song looks totally different.



(a) Fingerprint of a humming



(b) Fingerprint of the nearest song

(c) Fingerprint of the farthest song

Figure 7. Fingerprint of (a) a humming, (b) the nearest song and (c) the farthest song.

6. RESULTS

The music database used in this study contained 140 songs extracted from commercial CDs of different genres: Pop/-Rock and Spanish folk music. The selected phrase fragments of each song are segments of 5 to 20 seconds, depending on the predominant melodic line of each song.

For the evaluation of the system, we have used 70 hummings from three male and three female users, whose ages are between 25 and 57 years, and 50% of the users have musical knowledge. The hummings were recorded at a sampling rate of $44.1kHz$ and the duration of each humming ranges from 5 to 20 seconds.

The retrieval performance was measured on the basis of *Song accuracy*. In general, we computed the *Top-N accuracy*, that is the percentage of humming whose target songs were among the $Top-N$ ranked songs. The *Top-N accuracy* is defined as:

$$Top-N\ accuracy(\%) = \frac{\#Songs\ in\ Top-N}{\#hummings} \times 100\% \quad (9)$$

Different experiments have been made to test the system effectiveness as a function of the musical genre. The musical genre has influence on the harmonic complexity of the songs, the number of musical instruments played, the kind of accompaniment and the presence of rhythm instruments such as drums. All these musical aspects affect in the main voice enhancement process.

In Table 1, the evaluation of the proposed method in the complete database, for all hummings, for 5 different ranking are presented. These results are rather similar to the ones presented in [7] and [8], with the difference that our method uses real songs instead of other hummings [7] and our method does not need to obtain the symbolic notation of neither the database nor the humming [8]. Thus, a mathematical comparison against other systems has not been possible since other systems found do not use real audio waveforms. The Table 1 also includes the *Top-N accuracy* for musical genres: Pop/Rock and Spanish folk. It can be observed that the performance of the system is better for the Spanish folk music. This is due to the fact that in this type of music the main voice is the most important part in the music and does not have digital audio effects like reverberation, therefore the main voice enhancement process performs better.

Table 1. Evaluation of the proposed method in the complete database for all hummings, Pop/Rock and Spanish folk.

	Top-N accuracy (%)		
Ranking	All	Pop/Rock	Spanish folk
Top-1	37.12	33.33	47.33
Top-5	52.86	43.14	78.95
Top-10	55.71	45.10	84.21
Top-20	60.00	49.02	89.47
Top-30	61.43	49.02	94.74

In Table 2, the evaluation of the proposed method is done with the database divided into two music collection: one corresponding to Pop/Rock music (70% of songs in the database) and other corresponding to Spanish folk music (30% of songs in the database). The hummings are divided in the same percentages as the music in the database. In Table 2, it

Table 2. Evaluation of the proposed method with the database divided into two music collections: Pop/Rock and Spanish folk.

	Top-N accuracy (%)	
Ranking	Pop/Rock	Spanish folk
Top-1	35.29	57.89
Top-5	45.10	84.21
Top-10	47.06	89.47
Top-20	52.94	94.74

can be observed that the performace of the system is better for the Spanish folk music, like in the previous experiment shown in Table 1.

In Figure 8, the evolution of the *Top-N accuracy (%)* as a function of N as percentage of the music collection in which the humming is expected to be found, is shown. This evolution is presented for the complete database, the Pop/Rock music collection and the Spanish folk music collection. Figure 8 shows that the Spanish folk music obtains the best results, as presented the Table 2. This figure also shows that if the user or the system have some knowledge of the musical genre, the humming method becames more effective.

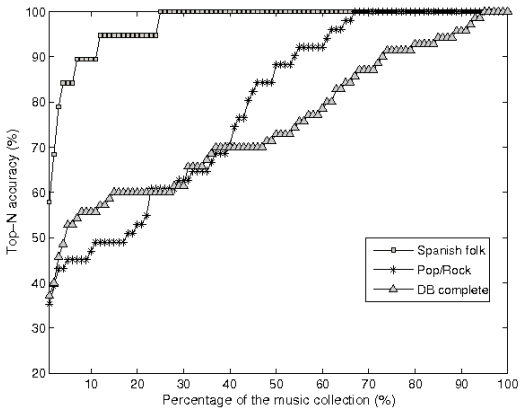


Figure 8. Evolution of the *Top-N accuracy (%)* as a function of N as percentage of the music collection in which the humming is expected to be found.

7. CONCLUSIONS

In this paper a humming method for content-based music information retrieval has been presented. The system employs an original fingerprint based on chroma vectors to characterize the humming and the reference songs. With this fingerprint, it is possible to find songs similar to humming without any transcription or Midi data. The performance of the method is better in Spanish folk music, due to the main voice enhancement procedure in relation with the mixing style used in this type of music, than in Pop/Rock music. The method performance could be improved if an estimation of the musical genre is included. Also, the parameters of the panning window could be tuned for each musical genre to improve the performance of the main voice enhancement. Finally, the system could also be made robust to transposed hummings, employing a set of transposed chroma matrices for each humming.

8. ACKNOWLEDGMENTS

This work has been funded by the Ministerio de Ciencia e Innovación of the Spanish Government under Project No. TIN2010-21089-C03-02.

9. REFERENCES

[1] C. Avendano: "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 55–58, 2003.

[2] L. Chen and B.-G. Hu: "An implementation of web based query by humming system," *International Conference on Multimedia and Expo* (ICME2007), pp. 1467–1470, 2007.

[3] A. Ghias, J. Logan and D. Chamberlin: "Query by humming-musical information retrieval in an audio database," *Proceedings of ACM Multimedia* (ACM1995), pp. 231–236, 1995.

[4] D. Gibson: *The art of mixing*, MixBooks, Michigan, 1997.

[5] J. Li, J. Han, Z. Shi and J. Li: "An efficient approach to humming transcription for Query-by-Humming System," *3rd International Congress on Image and Signal Processing* (CSIP2010), pp. 3746–3749, 2010.

[6] J. Li, L.-m. Zheng, L. Yang, L.-j. Tian, P. Wu and H. Zhu: "Improved Dynamic Time Warping Algorithm the research and application of Query by Humming," *Sixth International Conference on Natural Computation* (ICNC2010), pp. 3349–3353, 2010.

[7] T. Liu, X. Huang, L. Yang and P.Zhang: "Query by Humming: Comparing Voices to Voices," *International Conference on Management and Service Science* (MASS'09), pp. 1–4, 2009.

[8] J. Song, S.-Y. Bae and K. Yoon: "Query by Humming: Matching humming query to polyphonic audio," *IEEE International Conference on Multimedia and Expo* (ICME02), Vol. 1, pp. 329–332, 2002.

[9] E. Unal, E. Chew, P.G. Georgiou and S.S. Narayanan: "Challenging Uncertainty in Query by Humming Systems: A Fingerprinting approach," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16, No. 2, pp. 359–371, 2008.

[10] X. Xie, L. Lu, M. Jia, H. Li, F. Seide and W.-Y. Ma: "Mobile search with multimodal queries," *Proceedings of the IEEE*, Vol. 96, No. 4, pp. 589–601, 2008.

[11] H.-M. Yu, W.-H. Tsai and H.-M. Wang: "A Query-by-Singing System for Retrieving Karaoke Music," *IEEE Transactions on multimedia*, Vol. 10, No. 8, pp. 1626–1637, 2008.