An Improved Query by Singing/Humming System Using Melody and Lyrics Information

Chung-Che Wang
MIR Lab, Dept. of CS,
Tsing Hua Univ., Taiwan
geniusturtle
@mirlab.org

Jyh-Shing Roger Jang
MIR Lab, Dept. of CS,
Tsing Hua Univ., Taiwan
jang@mirlab.org

Wennen Wang
Institute for Information
Industry, Taiwan
wennen@iii.org.tw

ABSTRACT

This paper proposes an improved query by singing/humming (QBSH) system using both melody and lyrics information for achieving better performance. Singing/humming discrimination (SHD) is first performed to distinguish singing from humming queries. For a humming query, we apply a pitch-only melody recognition method that has been used for QBSH task at MIREX with rank-1 performance. For a singing query, we combine the scores from melody recognition and lyrics recognition to take advantage of the extra lyrics information. Lyrics recognition is based on a modified tree lexicon that is commonly used in speech recognition. The performance of the overall QBSH system achieves 39.01% and 23.53% error reduction rates, respectively, for top-20 recognition under two experimental settings, indicating the feasibility of the proposed method.

1. INTRODUCTION

Query by singing/humming (QBSH) is an intuitive method for music retrieval. With a QBSH system, users are able to retrieve intended songs by singing or humming a portion of the intended song in order to retrieve it. Most of the QBSH researches so far utilize melody information as the only cue for retrieval [1-3]. Ghias et al. [1] proposed one of the early papers of query by humming, which used three different characters ('U', 'D', and 'S') are to represent pitch contours. McNab et al. [2] enhanced the representation by considering rhythm information of segmented notes. Jang and Gao [3] proposed the first QBSH system using dynamic time warping (DTW) over frame-based pitch contours, which accommodates natural singing/humming for better performance. More recently, QBSH task is held in MIREX since 2006 and quite a few related methods and corresponding performance can be found therein [12].

Lyrics are also an important part of a song which serve the cue for detecting the song's identity, or its mood or genre. However, the use of lyrics for content-based music analysis appears much later. Wu et al. [4] and Chen [14] used lyrics to enhance music mood estimation. Wang et al. [5] proposed one of the few music information retrieval systems which used both lyrics and melody information. However, queried lyrics were input by the user instead of decoded from the user's acoustic input. Xu et al. [6] suggested that acoustic distance must be considered for a lyrics search when the user input approximate lyrics query. Our method also takes advantage of the extra information provided by the lyrics, except that we attempt to decode the queried lyrics from the user's singing input directly, which does not impose extra efforts on the user. Suzuki et al. [13] also proposed a similar system which took singing input for lyrics recognition, and the results were verified by the corresponding melody information. However, their system could not handle humming input, which is likely to happen in a music retrieval system. Moreover, the corpus used for their experiments is too small to justify the method's feasibility.

The proposed QBSH system uses singing/humming discrimination (SHD) to detect whether there exists lyrics information. If yes, we apply speech recognition to decode the lyrics information and come up with a lyrics score. The lyrics score is combined with the melody score to enhance the recognition performance.

The remainder of this paper is organized as follows. The proposed QBSH system is introduced in section 2. Experimental results are shown in section 3. We conclude this paper and address directions for future work in section 4.

2. SYSTEM OVERVIEW

Figure 1 shows the schematic diagram of the proposed system, in which blocks enclosed by thicker lines are the methods proposed by this paper. In the offline part, acoustic models and test corpus are used to obtain the model similarity, where each model is characterized by an RCD (right-context dependent) bi-phone. Phone-level similarity is used for SHD, and syllable-level similarity is used for lyric scoring. Lexicon network is also created in this part for lyrics recognition. In the online part, SHD is first performed to decide if the acoustic input is singing or humming. If the input is classified as humming, the result is based on melody recognition alone. On the other hand, if the input is classified as singing, lyrics recognition is performed to obtain a decoded string of lyrics. The output of our system then uses the combined scores of melody and lyrics. The melody recognition module uses UPDUDP [11] for pitch extraction and linear scaling (LS) [7] for comparison, which achieved the

best performance during QBSH task in MIREX2009 [12]. The other components will be explained in detail in the following subsections.

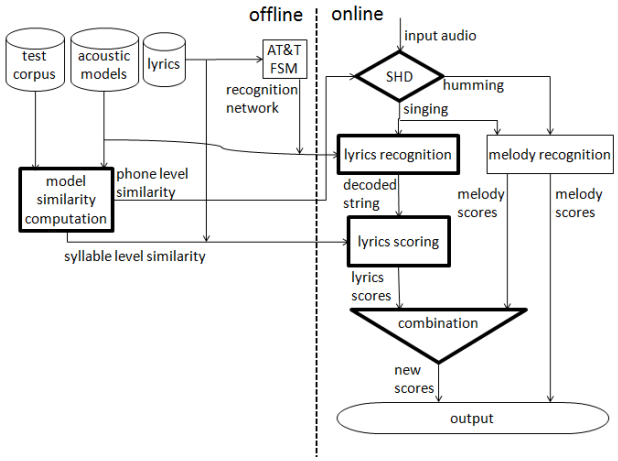

Figure 1. The proposed system.

2.1 Phone and Syllable Similarity

An intuitive approach to SHD is based on the number of distinct phones decoded in the user's acoustic input. The more distinct phones in an acoustic query, the more likely the query is singing instead of humming. In counting distinct phones, we also need to take phone similarity into consideration for achieve most robust results. Moreover, we also need to have similarity between syllables for obtaining lyrics score. The procedure for computing phone and syllable similarity is explained next.

Firstly, we can obtain the confusion matrix of 156 biphone models by performing free phone decoding on a speech corpus by HTK [8]. Then we can use the accuracy rates of the confusion matrix as the similarity measure between any two phone models. It should be noted that the similarity is not symmetric, but this does not affect the proposed methods.

After defining phone similarity K, the similarity matrix of 423 Mandarin syllables can be computed via dynamic programming (DP) as follows. Considering two syllables Syl_A and Syl_B, with phone sequence $a_1, a_2, \dots, a_m$ and $b_1, b_2, \dots, b_n$, respectively, the definition of the similarity between Syl_A and Syl_B is:

$$sim(Syl_A, Syl_B) = \frac{t_{A,B}(m,n)}{\max(m,n)}, \quad (1)$$

where the recursive formula of $t_{A,B}$ is:

$$t_{A,B}(i,j) = \max\begin{pmatrix} t_{A,B}(i-1,j) \\ t_{A,B}(i,j-1) \\ K(a_i,b_j)+t_{A,B}(i-1,j-1) \end{pmatrix}, \quad (2)$$

with boundary conditions :

$$t_{A,B}(i,j) = 0, \ if\ i = 0\ or\ j = 0. \quad (3)$$

Figure 2 shows the image of 423 x 423 similarity matrix in gray scale, where white points represent 1, and black points represent 0.

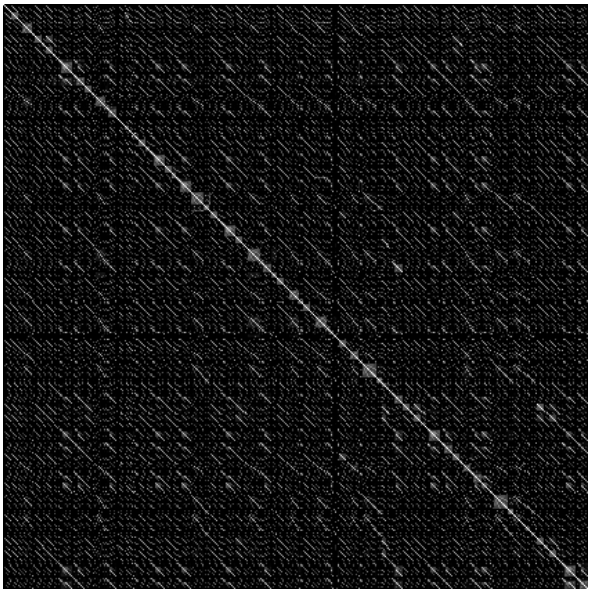

Figure 2. The similarity matrix of 423 Mandarin syllables displayed as a gray-scale image.

2.2 Singing/Humming Discrimination

The basic rationale behind SHD is that the number of distinct phones occurring in humming is often less than that in singing. Thus, free phone decoding is performed on the singing input to obtain a phone sequence. If these phones are similarity in acoustics, then the effective count of distinct phones should be less. Assume there are n distinct phones $a_1, a_2, \dots, a_n$ in the decoded phone sequence, then the effective count is defined as follows. Let P to be a sub matrix of sim for $a_1 a_2, \dots, a_n$, then we can calculate the effective phone count in the sequence:

$$r = n + 1 - median(s) \quad (4)$$

where r is the effective phone count, and s is the column sum of P. A lower effective phone count indicates that the acoustic query is more likely to be humming instead of singing. In particular, when $a_1 a_2, \dots, a_n$ are very different in pronunciation, then $median(s)$ is close to 1 and r is close to n. On the other hand, if these phones are very similar in pronunciation, then $median(s)$ is close to n and r is close to 1.

Thus an optimum threshold of effective phone count can be set for SHD for minimizing classification error.

2.3 Lyrics Recognition

If an acoustic query is classified as singing, we can apply lyrics recognition for better performance. Since the aver-

age length of a singing query is about 7 seconds, the first 30 syllables of each song are used to build up the recognition network. (Without loss of generality, here we assume the anchor position of each query is the beginning of a song. If not, then we can simply use the phrase onset as the beginning to select the first 30 syllables for building the network.) By considering the recognition network as a finite state machine, this network is determinized and minimized by AT&T FSM tool [9]. Moreover, to handle the case of "stop in the middle", an epsilon transition is added between each internal state and the terminal state. Figure 3 shows an example of the network, consisting of the first 3 syllables of 2 songs, where Song-i-Syl-j denotes the j-th syllable of the i-th song, and <eps> denotes the epsilon transition.

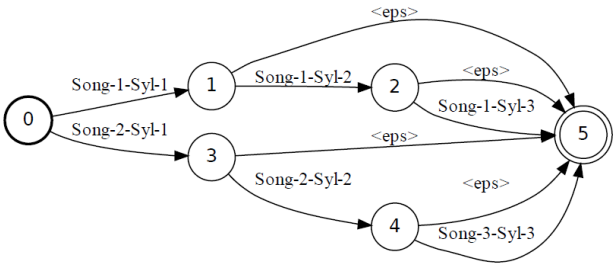

Figure 3. An example of the recognition network.

2.4 Lyrics Scoring and Combination

After running Viterbi search over the recognition network, we can decode a syllable sequence that has the maximum likelihood. To obtain the similarity score based on lyrics, we then compare the decoded syllable sequence with the 30 syllables of each song. This is again achieved by DP instead of using exact string matching since we want to have a score indicating the similarity. Consider two syllables sequences, Seq_A and Seq_B containing syllables $A_1, A_2, \dots, A_m$ and $B_1, B_2, \dots, B_n$ respectively. The recursive formula for DP is:

$$t(i,j) = max\begin{pmatrix} t(i-1,j) \\ t(i,j-1) \\ sim(A_i, B_j) + t(i-1,j-1) \end{pmatrix}, \quad (5)$$

with boundary conditions :

$$t(i,j) = 0, if\ i = 0\ or\ j = 0 \quad (6)$$

Thus, $t(m,n)$ can be taken as a similarity score between the decoded string from the query and the lyrics of each song in the database. In implementation, we let Seq_A to be the decoded string, and the first k syllables of the lyrics (with k equal to the length of Seq_A) to be Seq_B for computing the score.

For score combination, we need to normalize each individual score. For a given query to a song database of 2000 songs, we will eventually obtain vector L of size 2000 representing the lyrics raw similarity scores, and vector M of size 2000 representing the melody raw distance measures (computed by LS). We than linearly normalize M and $-L$, respectively, to the range [0, 1]. The normalized distance vectors M' and L' are then combined via the following formula:

$$C = p \times \log M' + (1-p) \times \log L' \quad (7)$$

Then the minimum entry in C corresponds to the most likely song considering both lyrics and melody. Note that if p is equal to 0, only the lyric information is considered. On the other hand, if p is equal to 1, only the melody information is considered. The value of p was set to 0.5 empirically in our experiments.

3. EXPERIMENT

3.1 The Dataset

The public corpus MIR-QBSH [10] is used extensively in our experiment, where the anchor positions for all queries are from the beginning. Since our speech recognition engine is for Mandarin, therefore we selected 2469 clips from the corpus which correspond to 23 Mandarin songs. To increase the complexity of the comparison, we added 2131 noise songs to the database, such that the total size of the database is 2154.

First of all, 80 clips are selected from the corpus with evenly distributed gender and types (singing/humming). These clips were hand labeled to have the ground truth, and then used to train the threshold-based classifier for SHD. The remaining 2389 clips are used for testing the overall performance of our QBSH system.

Acoustic models of RCD bi-phones for computing phone/syllable similarity were obtained by training over a normal Mandarin speech of 80 subjects.

3.2 Experimental Results

3.2.1 Results of SHD

Figure 4 shows the detection error tradeoff (DET) curve of SHD using the training data of 80 clips, where singing clips are viewed as positive while humming clips negative. Based on this plot, the threshold of effective count is set to 20.4958 for SHD to achieve equal error rates of false positive and false negative. Figure 5 gives the distribution of the effective phone counts of the training data, together with the identified Gaussian models via maximum likelihood estimate. The Gaussian models for positive data (of size n_1) and negative data (of size n_2) are denoted as g_1 and g_1, respectively, in the figure.

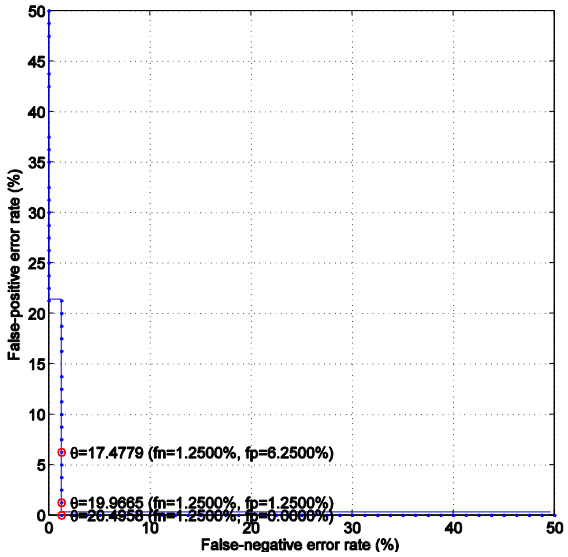

Figure 4. The DET curve of the training data for SHD.

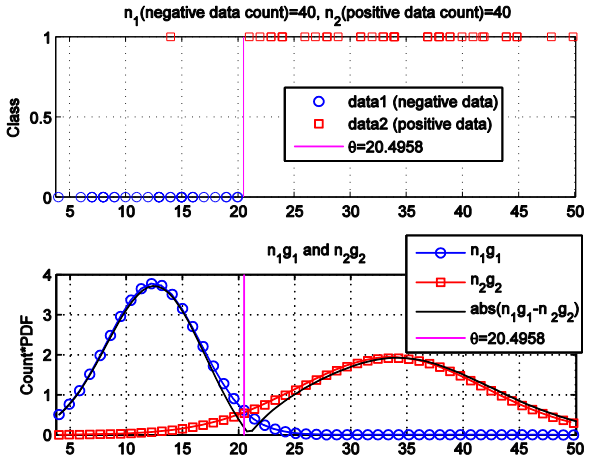

Figure 5. The distribution of the effective phone counts of the training data of SHD. The Gaussian models for positive data (of size n_1) and negative data (of size n_2) are g_1 and g_1, respectively.

To evaluate the performance of SHD over unseen data, 1183 clips were selected (out of the 2389 clips) and hand labeled as singing or humming. Table 1 shows confusion matrix of SHD, with a recognition rate of 78.61%. In particular, 10.99% of the humming clips are misclassified as singing, which may generate erroneous output in lyrics recognition. Initial error analysis indicates that some of the misclassified humming clips are caused by a variety of pronunciation during the humming. On the other hand, 24.51% of the singing clips are misclassified as humming, which is not so detrimental to the overall performance since the accuracy of melody recognition is already high.

ground truth \ recognition results	singing	humming
singing	75.49% (687)	24.51% (223)
humming	10.99% (30)	89.01% (243)

Table 1. Recognition result of SHD.

3.2.2 Lyrics recognition and Combined Results

When we applied SHD to 2389 clips, 1477 of them were classified as singing. Figure 6 shows the top-20 recognition rate of these 1477 clips over a song database of size 2154. The top-20 recognition rate is 72.99%.

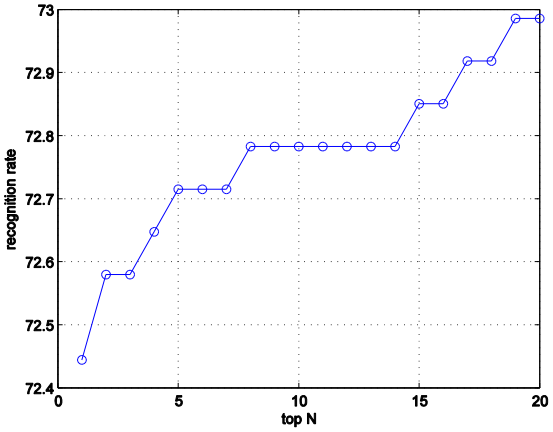

Figure 6. The top-20 lyrics recognition rate of 1477 clips classified as singing.

The value of p in Eq. (7) was set to 0.5 empirically. Now it is time to do a post analysis by plotting the overall recognition rates versus the values of p, as shown in Figure 7. Apparently the performance stays much the same for these two cases of LS resolution equal to 11 and 51, respectively, as long as the value of p lies within [0.3, 0.9]. This confirms our selection of p value of 0.5.

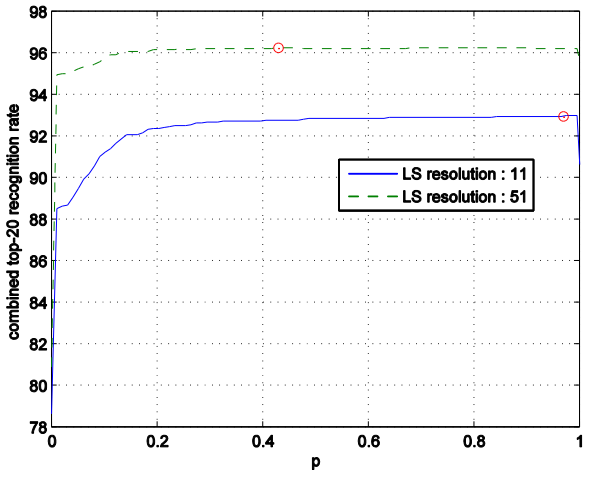

Figure 7. The plots of overall recognition rates with respect to the value of p, for two cases of different LS resolutions. The maximum of these two curves are labeled with circles.

Figure 8 shows the overall performance of the proposed QBSH system. Different values of resolution in LS, 11 and 51, were used in this experiment. The lower and upper ratios of LS were set to be 0.5 and 2. The top-20 recognition rates of resolution 11 and 51 are 92.80% and 96.19%, respectively, which outperform the baseline system (88.20% and 95.02%). The error reduction rates are 39.01% and 23.53%, respectively.

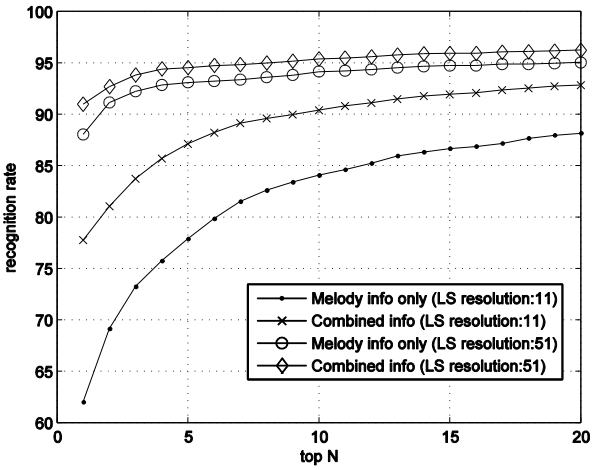

Figure 8. The top-N recognition rate of 2389 clips.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an improved QBSH system that distinguishes singing queries from humming ones, and then applies different procedures in order to take advantage of the lyric information of singing input. The experimental results demonstrate the effectiveness of the proposed system, with error reduction rates of 39.01% and 23.53% for LS with resolutions of 11 and 51, respectively.

Several directions for immediate future work are under way. Currently, our acoustic models were obtained by training on normal speech corpus. This can be improved by training or simply adapting using singing corpora instead. Moreover, it would be desirable to incorporate multi-lingual speech recognition since there are quite a few famous songs with the same tune but different lyrics in different languages.

5. ACKNOWLEDGEMENT

This study is conducted under the "III Innovative and Prospective Technologies Project" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

6. REFERENCES

[1] A. J. Ghias, D. C. Logan, and B. C. Smith, "Query by humming-musical information retrieval in an audio database," in *Proc. ACM Multimedia'95*, San Francisco, 1995, pp. 216–221.

[2] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham, "Toward the digital music library: Tune retrieval from acoustic input," in *Proc. ACM Digital Libraries*, 1996, pp. 11–18.

[3] J.-S. R. Jang and M.-Y. Gao, "A query-by-Singing system based on dynamic programming," in *Proc. Int. Workshop Intell. Syst. Resolutions (8th Bellman Continuum)*, Hsinchu, Taiwan, R.O.C., Dec. 2000, pp. 85–89.

[4] Y.-S. Wu, W.-r. Chu, C.-Y. Chi, D. C. Wu, R. T.-H. Tsai, and J Y.-j Hsu, "The Power of Words: Enhancing Music Mood Estimation with Textual Input of Lyrics," *International Conference on Affective Computing & Intelligent Interaction*, pp. 1–6, 2009.

[5] T. Wang, D.-J. Kim, K.-S. Hong, and J.-S. Youn, "Music Information Retrieval System using Lyrics and Melody Information," *Asia-Pacific Conference on Information Processing*, pp. 601–604, 2009.

[6] X. Xu, M. Naito, T. Kato, and H. Kawai, "Robust and Fast Lyric Search Based on Phonetic Confusion Matrix," *Proceedings of the International Symposium on Music Information Retrieval*, pp. 417–422, 2009.

[7] J.-S. R. Jang, H.-R. Lee, M.-Y. Kao, "Content-based Music Retrieval Using Linear Scaling and Branch-and-Bound Tree search," in *Proc. of IEEE*

International Conference on Multimedia and Expo, August 2001.

[8] Cambridge University Engineering Department , HTK Web-Site, http://htk.eng.cam.ac.uk/, 2006

[9] AT&T Labs Research , AT&T Labs Research - FSM Library , http://www2.research.att.com/~fsmtools/fsm/, 2008

[10] J.-S. R. Jang, "MIR-QBSH Corpus", MIR Lab, CS Dept, Tsing Hua Univ, Taiwan. Available at the "MIR-QBSH Corpus" link at http://www.cs.nthu.edu.tw/~jang.

[11] J.-C. Chen, J.-S. R. Jang, "TRUES: Tone Recognition Using Extended Segments", *ACM Transactions on Asian Language Information Processing, No. 10, Vol. 7*, Aug 2008.

[12] MIREX 2009, http://www.music-ir.org/mirexwiki/index.php/Main_Page, 2009

[13] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 38727, 8 pages, 2007. doi:10.1155/2007/38727

[14] J.-H Chen, "Content-based Music Emotion Analysis and Recognition", Master Thesis, CS Dept., National Tsing Hua University, June 2006