Improving Query-by-Singing/Humming by Combining Melody and Lyric Information

Chung-Che Wang and Jyh-Shing Roger Jang, *Member, IEEE*

***Abstract*—This paper proposes a novel method for improving query-by-singing/humming systems by using both melody and lyric information. First, singing/humming discrimination is performed to distinguish between singing and humming queries, which is achieved by considering the similarity between acoustic models. For the humming queries, a pitch-only melody recognition method that was ranked first among the MIREX (Music Information Retrieval Evaluation eXchange) query-by-singing/humming task submissions is applied. For the singing queries, a lyric similarity is computed using speech recognition techniques; the computed similarity is subsequently combined with the melody distance to exploit additional information in the lyrics. Several methods for combining melody distance and lyric similarity are investigated. Under the optimal experimental settings, the proposed query-by-singing/humming system achieves 51.19% error rate reduction for the top-10 retrieved results, indicating the feasibility of the proposed method.**

***Index Terms*—Combined melody distance and lyric similarity, query-by-singing/humming (QBSH), singing voice recognition, singing/humming discrimination (SHD).**

I. Introduction

QUERY-BY-SINGING/HUMMING (QBSH) is an intuitive, natural user interface for music retrieval; in this interface, a user can retrieve a song by singing or humming a portion of the song. The aim of this study was to distinguish singing from humming and to extract lyric information from the singing input to achieve optimal accuracy. The related works are described in the following subsections.

A. Related Work: Melody and Textual Lyric Information

Recent studies on QBSH have used melody information as the only cue for retrieval [1]–[3], [7]. Ghias *et al.* [1] proposed a query-by-humming method that involves using three characters (U, D, and S) to indicate whether the pitch of a note is higher than, lower than, or identical to the preceding note, and an approximate string-matching algorithm subsequently identifies potential song matches. McNab *et al.* [2] enhanced the representation by considering rhythm information obtained from segmented notes. Jang and Gao [3] proposed the first QBSH system using Dynamic Time-Warping (DTW) over frame-based pitch contours, which improved the retrieval performance by accommodating natural singing/humming. Jang *et al.* [7] proposed a QBSH system that uses Linear Scaling (LS), which is a simpler yet effective method compared with DTW.

Lyrics are a critical identifier of a song, and they can also indicate its mood or genre. However, the use of lyrics for content-based music analysis did not begin until much later. Mayer *et al.* [19] used song lyrics to improve music classification and similarity ranking systems. Chi *et al.* [4] and Chen [13] used textual lyric input to enhance music mood estimation. Wang *et al.* [5] proposed a music information retrieval system that used both lyric and melody information; however, instead of extracting lyric information from an acoustic input, the system requires users to input the queried lyrics manually. Xu *et al.* [6] indicated that acoustic distance must be considered if an acoustic input approximates the lyric query in performing a lyric search. The method proposed in this study involves exploiting additional information in the lyrics and decoding the queried lyrics directly from the singing input, thus improving user convenience.

B. Related Work: Singing Voice Recognition

Suzuki *et al.* [12], [18] proposed a system in which singing is used as an input for singing voice recognition. Furthermore, the system verifies the candidates identified by the singing voice recognition module by analyzing time-alignment information and comparing the relation between music scores and recognized word sequences. However, the system cannot handle humming inputs, which are potentially crucial mode of retrieval in QBSH systems. Papiotis and Purwins [14] attempted to locate the exact position of a query input within a single music piece by combining the warping cost of DTW based on pitch contour, Mel-Frequency Cepstral Coefficients (MFCCs), and root mean square energy. However, this approach is potentially unsuitable for large-scale systems where storing a large number of solo vocal clips in a database would be difficult. Mesaros and Virtanen [20], [21] used n-gram language models to perform singing voice recognition from singing inputs. However, that approach involved using lyrics to directly build up the recognition network, which is similar to those proposed by other studies [12], [18]. McVicar *et al.* [33] improved the accuracy of lyric transcription by evaluating repetitions in music to transcribe the whole song. However, in the system proposed in the current study, only a short query clip must be transcribed and repetitions seldom occur.

Manuscript received August 15, 2014; revised December 25, 2014; accepted February 23, 2015. Date of publication March 06, 2015; date of current version March 16, 2015. This work was supported in part by the National Science Council, Taiwan, under Grant NSC 102-2221-E-002-164 -MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thushara Abhayapala.

C.-C Wang is with the Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan (e-mail: geniusturtle@mirlab.org).

J.-S. R. Jang is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan (e-mail: jang@mirlab.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2409735

C. Related Work: Singing/Humming Discrimination

Because humming contains no lyric information, developing a QBSH system using both melody and lyric information requires a Singing/Humming Discrimination (SHD) function to distinguish between them. Many studies have attempted to discriminate among various types of audio signal [15]–[17]. Schuller *et al.* [15] applied a Support Vector Machine (SVM) method to discriminate among noise, speech, and monophonic singing by analyzing the statistical features of pitch contours and duration of voiced sound/silence in continuous audio streams. Ohishi *et al.* [16] combined short-term feature measures (i.e., MFCCs and their derivatives modeled by Gaussian mixture models) and long-term feature measures based on F0 contours for discriminating between singing and speech voices, achieving more than 90% accuracy for signals approximately 2 s in duration. Gärtner [17] proposed a system for classifying singing/rap by using features derived from speech/singing classification and speech emotion recognition. However, few studies focusing on SHD were identified in a thorough review of relevant research.

D. Related Work: Melody/Lyric Information Combination

Guo *et al.* [22] proposed a system similar to that proposed in this study that accepts both singing and humming inputs. The system uses an SVM to classify the input query as singing or humming, and then selects candidates from a database by using melody recognition module, such as LS, DTW, recursive alignment [23], and earth mover's distance [24]. The melody recognition module locates the input query to the candidate songs. If the input query is classified as singing, the system uses the alignment information to dynamically build up the recognition network. The final result of a singing query, however, was highly dependent on the accuracy of the melody recognition module.

Regarding information combination, Suzuki *et al.* [12], [18] and Guo *et al.* [22] used a two-stage approach to combine melody and lyric information. Kao *et al.* [30] compared a two-stage approach with Borda Count [31], which is a single-winner election method. Degani *et al.* [32] normalized the distances of each method and then used l^2 norm for combination. In this study, various methods were attempted for conducting distance/similarity combination, including different methods of normalization and weight adjustment.

E. Proposed System

The QBSH system proposed in this study uses lyric and melody information independently to improve the system accuracy. An SHD function is used to detect the presence of lyric information. If an acoustic input is classified as singing, lyric information is used to produce a lyric similarity that is independent of the corresponding melody distance. Subsequently, the lyric similarity and melody distance are combined to enhance the recognition performance.

The remainder of this paper is organized as follows. The proposed QBSH system is introduced in Section II. Section III reports the experimental results, and Section IV offers a conclusion and recommendations for future research.

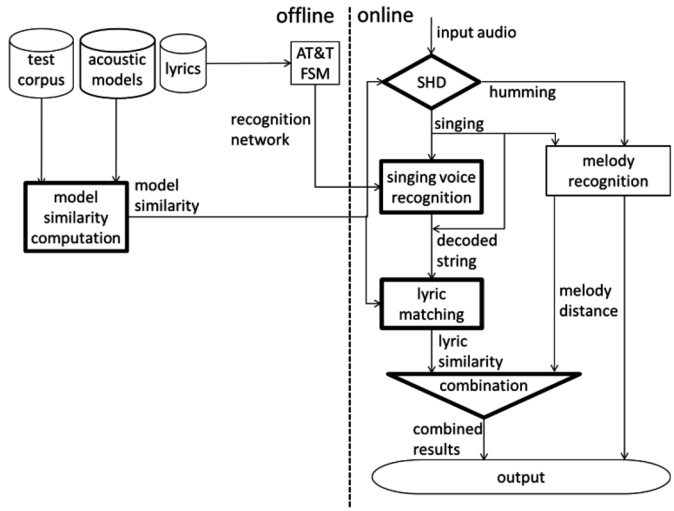


Fig. 1. The proposed system.

II. System Overview

Fig. 1 shows a schematic diagram of the proposed QBSH system, where the components enclosed by thicker lines represent the proposed methods. In the offline section (left-hand side), the model similarity is assessed using acoustic models and a test corpus, where each model is characterized using a right-context-dependent biphone. Phone-level similarity is estimated directly based on the decoded result, whereas syllable-level similarity is computed based on phone-level similarity by using Dynamic Programming (DP). For singing voice recognition, a lexicon network is created according to the lyric database. A Finite-State Machine (FSM) tool proposed by AT&T [8] is used to determinize and minimize the lexicon network. In the online section (right-hand side), SHD is first performed to determine whether the acoustic input is singing or humming. When the input is classified as humming, the output is determined by melody recognition alone. However, when the input is classified as singing, the corresponding lyric information is used to compute a lyric similarity. Subsequently, in the system output, potentially matching songs are ranked using the combined melody distance and lyric similarity. The following subsections describe the system components.

A. Melody Recognition

The melody recognition module contains two main methods; pitch extraction (from the input query) and database comparison. For the pitch extraction, the system uses the unbroken pitch determination using DP method proposed in [10], which produces a smooth continuous pitch contour that is more robust than using autocorrelation alone.

For the database comparison method, key transposition and tempo variation must be addressed. For key transposition, the average pitch of the input query and each song in the database are first computed; during comparison to a database song, the input query is then pitch-shifted to match the database song. For tempo variation (which is typically linear), this study applied LS [7], which was ranked first among the MIREX2009 QBSH tasks [11]. Assume that the input pitch vector has a duration of d seconds. The vector must be compressed or stretched to obtain

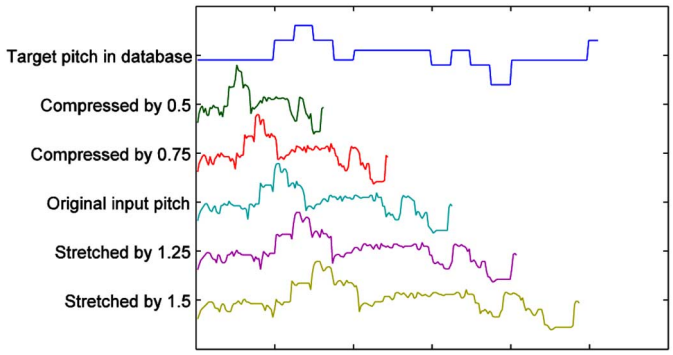


Fig. 2. A typical example of LS.

r versions of the original vector, with durations equally spaced between $s_{\min} \times d$ and $s_{\max} \times d$, where $s_{\min}$ (< 1) and $s_{\max}$ (> 1) are the minimal and maximal scaling factors, respectively. The distance between the input pitch vector and a particular song is thus the minimal distance between each vector and the song. In Fig. 2, a d-second vector is compressed/stretched to obtain five vectors of decreasing/increasing duration based on equally distributed increments between $0.5 \times d$ and $1.5 \times d$. The optimal result is obtained when the scaling factor is 1.25. Typically, for fixed $s_{\min}$ and $s_{\max}$, increasing r results in higher accuracy, but longer computation time.

B. Phone and Syllable Similarity

An intuitive approach to SHD is based on the number of distinct phones or syllables decoded in the acoustic input. The more distinct phones or syllables occurring in an acoustic query, the higher the probability is that the query is singing rather than humming. When counting the number of distinct phones or syllables, phone or syllable similarity must be considered to achieve more robust results [6]. The procedure for computing phone and syllable similarity is explained as follows.

In this study, the Mandarin speech corpus TCC300 [25] and other privately collected corpora were used to train a set of acoustic models based on right-context-dependent biphone HMMs (hidden Markov models). Here, 34 phones [34] were used to construct 155 right-context-dependent biphones and 446 base syllables without tones. Acoustic models thus obtained were adapted according to singing voice in this study. The effect of adaptation is discussed in Section III.

First, a confusion matrix comprising 155 biphone models was obtained by performing free-phone decoding on a speech corpus [26], where element (i, j) represents the number of times phone i was identified as phone j. Subsequently, each row is divided by its maximum to obtain a normalized confusion matrix. The phone similarity matrix $\boldsymbol{K}$ is then defined as the mean of the normalized confusion matrix and its transpose. Accordingly, a syllable similarity matrix composed of 446 Mandarin syllables can be computed using a DP method, as follows. Considering two syllables Syl_A and Syl_B, with respective phone sequences $a_1, a_2, \ldots a_m$ and $b_1, b_2, \ldots b_n$, the similarity between Syl_A and Syl_B is:

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

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

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

with boundary conditions

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

The recursive formula of $t_{A,B}$ is based on the concept of the longest common subsequence; thus, $t_{A,B}(m,n) \leq \min(m,n)$. However, if $a_1, a_2, \ldots a_m$ is a subset of $b_1, b_2, \ldots b_n$ (or vice versa), $sim(Syl_A, Syl_B)$ is 1 if the denominator is $\min(m,n)$; this result is unreasonable because Syl_A and Syl_B are unequal. Consequently, $t_{A,B}(m,n)$ must be divided by $\max(m,n)$ for normalization.

Examples of syllable-level similarity are presented as follows:

1) "*Pao*" and "*pao*" are identical; thus, the similarity is 1.
2) "*Huang*" and "*wang*" differ by one consonant; thus, the similarity is 0.75.
3) "*Han*" and "*hou*" differ by one vowel. Because vowels typically yield higher recognition rates than consonants do, the similarity is relatively low; 0.0204.
4) "*Huang*" and "*min*" differ distinctly; thus, the similarity is almost zero; 0.0007.

C. Singing/Humming Discrimination

The basic rationale of SHD is that humming typically produces fewer unique phones or syllables than singing does. Thus, free-phone/syllable decoding is performed on the singing input to obtain a sequence of phones or syllables. If these phones or syllables are acoustically similar, then the effective count of unique phones or syllables is reduced. For a decoded phone sequence (excluding silence) comprising m unique phones $a_1, a_2, \ldots a_m$, the Effective Unique Phone Count (EUPC) implemented in this study can be defined as follows:

$$EUPC = m + 1 - median\left(s\right), \tag{4}$$

where s is the column sum of the submatrix of $\boldsymbol{K}$ corresponding to $a_1, a_2, \ldots a_m$. The Effective Unique Syllable Count (EUSC) is defined similarly, except that the submatrix is extracted from $\boldsymbol{sim}$. A lower EUPC/EUSC indicates a relatively higher probability that an acoustic query is humming rather than singing. In particular, if $a_1, a_2, \ldots a_m$ are similar in pronunciation, then $median(s)$ is close to m and EUPC/EUSC is close to 1. By contrast, if these phones differ markedly in pronunciation, then $median(s)$ is close to 1 and EUPC/EUSC is close to m Fig. 3 depicts two examples of EUPC computation. Where the phones are pronounced similarly (left-hand side), $median(s)$ is high and EUPC is low. Where the difference in pronunciation is distinct (right-hand side), $median(s)$ is close to 1 and the EUPC is almost equal to m.

Thus, an optimal value of EUPC/EUSC can be set as a threshold for SHD in order to minimize classification errors.

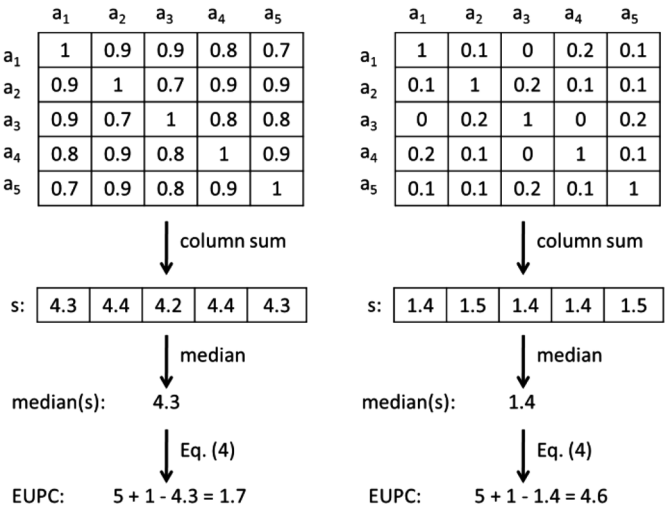


Fig. 3. Examples of computing EUPC. The left example is the case where phones are similar in pronunciation; the right example is the case where phones are very different in pronunciation.

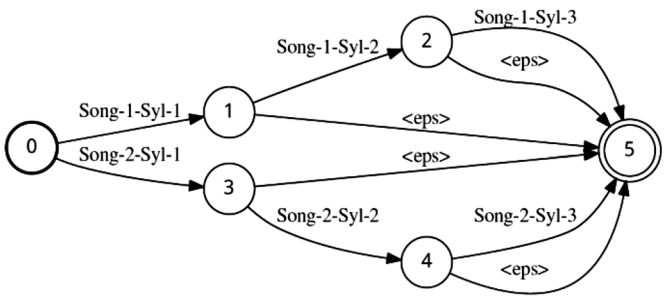


Fig. 4. Example of the recognition network for singing voice recognition.

D. Singing Voice Recognition

When an acoustic query is classified as singing, the accuracy of a QBSH system can be improved by applying singing voice recognition. Because the duration of query clips is 8 s in the corpus (MIR-QBSH) [9], the first 30 syllables of each song are sufficient for establishing the recognition network. (Without loss of generality, the anchor position of each query is assumed to be the beginning of a song. When this is not the case, the anchor positions can be set at the onsets of phrases or notes, and a brute-force search can be initiated.) The recognition network is considered as an FSM, and the network is thus determinized and minimized using the FSM tool proposed by AT&T [8]. Moreover, to handle the case of "stop in the middle of a sentence," an epsilon transition is inserted between each internal state and the terminal state, as described in [12], [18]. Fig. 4 shows depicts a network composed of the first three syllables of two songs, where " $\mathrm{Song}-i-\mathrm{Syl}-j$ " denotes the jth syllable of the ith song, and " <eps> " denotes the epsilon transition.

Typical results of determinization and minimization are shown in Fig. 5, where the recognition network is composed of the following three phrases: 1) " " ("*qing hua da xue*"; i.e., "Tsing Hua University"); 2) " " ("*jiao tong da xue*"; i.e., "Chiao Tung University"); and 3) " " ("*qing chu*"; i.e., "clear"). The upper recognition network is the original one. After determinization, the middle recognition network is obtained where the identical subpaths originating from node 0 merge. After minimization, the lower recognition network is obtained where

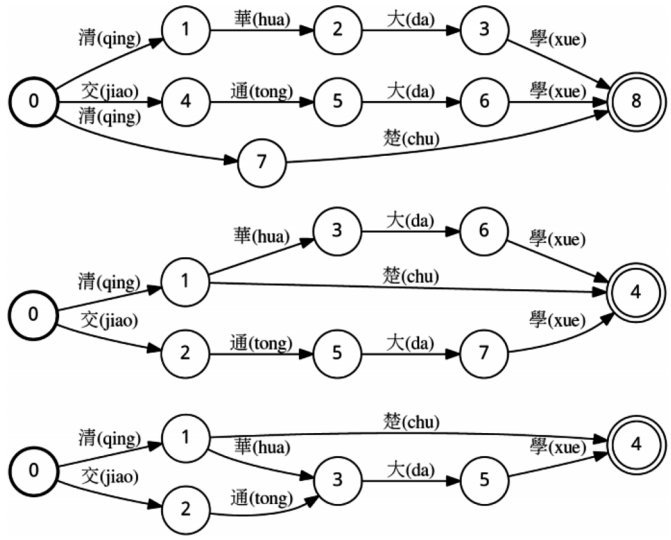


Fig. 5. The effect of determinization and minimization.

the identical subpaths ending at node 4 merge. Subsequently, the number of nodes and transitions is reduced considerably, thereby reducing the memory usage during implementation.

E. Lyric Matching and Distance/Similarity Combination

The Viterbi search can be executed over the recognition network to obtain a decoded syllable sequence with the maximum likelihood. To obtain the lyric similarity, the decoded syllable sequence is compared with the first 30 syllables of each song; this procedure was achieved through DP instead of using exact string matching. Specifically, the DP formula for computing the similarity between two syllables sequences $Seq_{A=}A_1, A_2, \ldots A_m$ and $Seq_B = B_1, B_2, \ldots B_n$ can be expressed as

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

where $t(i,j)$ is the similarity between $A_1, A_2, \ldots A_i$ and $B_1, B_2, \ldots B_j$; and $\boldsymbol{sim}$ is the similarity matrix of syllables defined in (1). The boundary conditions are:

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

Thus, $t(m,n)$ can be taken as a similarity between the decoded string from the query and the lyrics of each song in the database. In a previous study [28], the decoded string was obtained from the singing voice recognition extracted from linear lexicons. In this study, the decoded sequence is extracted directly from the SHD to minimize the required computation.

Fig. 6 illustrates an example of distributions of melody distance and three types of lyric similarity, where the singing lyrics are " " ("*hao jiu hao jiu de hu shi shi ma ma gao su wo*"; i.e., "mommy told me a story about long long time ago"); furthermore, the decoded strings for lyric matching are extracted from the singing voice recognition module or SHD module (using free-phone or free-syllable decoding). Because the decoded string extracted from the singing voice recognition module is an exact subset of the corresponding song in

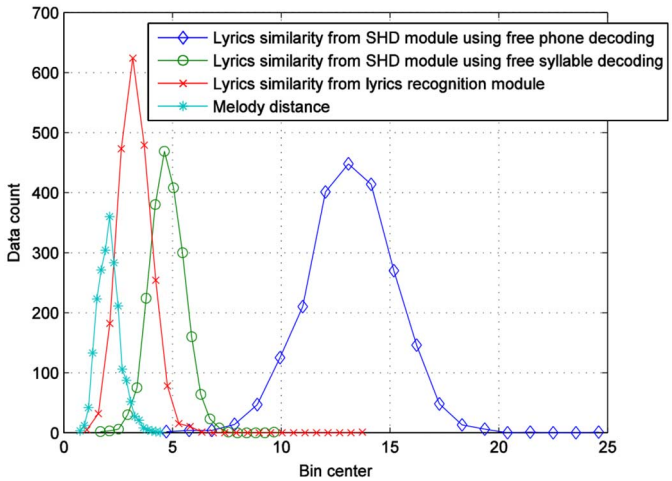


Fig. 6. The distribution of distance and similarity of the decoded string " " ("hao jiu hao jiu de hu shi shi ma ma gao su wo"; i.e., "mommy told me a story about long long time ago").

the database, the length of this string is the highest similarity value (which is 13 in this example). However, regarding the similarities computed using free-phone or free-syllable decoding, because insertions occur frequently, the length of the decoded string is typically longer, and the obtained similarity values are generally higher. Higher values are favorable for lyric similarity, whereas lower values are favorable for melody distance. Thus, this study must design a method to combine them consistently.

For a database comprising N songs, the vectors $\boldsymbol{L}$ (size = N) and $\boldsymbol{M}$ (size = N), were defined to represent the raw lyric similarity and raw melody distance measures, respectively. Accordingly, this study investigated several methods for combining $\boldsymbol{M}$ and $\boldsymbol{L}$:

$$C=\begin{cases} p\times M-(1-p)\times L \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{method } 1 \\ p\times LN(M)-(1-p)\times LN(L) \ldots\ldots\ldots\ldots \textit{method } 2 \\ p\times ZN(M)-(1-p)\times ZN(L) \ldots\ldots\ldots\ldots \textit{method } 3 \\ LN(M)^{p}\times LN(-L)^{1-p} \ldots\ldots\ldots\ldots\ldots\ldots \textit{method } 4 \\ p\times M+(1-p)\times L^{-1} \ldots\ldots\ldots\ldots\ldots\ldots \textit{method } 5 \\ p\times LN(M)+(1-p)\times LN\left(L^{-1}\right) \ldots\ldots\ldots \textit{method } 6 \\ p\times ZN(M)+(1-p)\times ZN\left(L^{-1}\right) \ldots\ldots\ldots \textit{method } 7 \\ M^{p}\times\left(L^{-1}\right)^{1-p} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{method } 8 \\ LN(M)^{p}\times LN\left(L^{-1}\right)^{1-p} \ldots\ldots\ldots\ldots\ldots\ldots \textit{method } 9 \end{cases} \tag{7}$$

For a vector $\boldsymbol{\nu}$ in (7), $\boldsymbol{LN}(\boldsymbol{\nu})$ indicates linear normalization, where $\boldsymbol{\nu}$ is linearly mapped to the range [0,1] $\boldsymbol{ZN}(\boldsymbol{\nu})$ indicates z-normalization, where $\boldsymbol{\nu}$ is normalized to have zero mean and unit variance; $\boldsymbol{\nu}^{k}$ is used to raise each element in $\boldsymbol{\nu}$ to the th power; and the vector $\boldsymbol{C}$ (size = x) represents the combined results of one of the methods. Because L indicates similarity, $-\boldsymbol{L}$ or $\boldsymbol{L}^{-1}$ must be used in the combined formulas to convert $\boldsymbol{L}$ into a distance-like quantity.

The minimal entry in $\boldsymbol{C}$ corresponds to the most likely candidate song when both lyric and melody information are considered. If $p=0$, then only the lyric information is considered; however, if $p=1$, then only the melody information is considered. For the experiments conducted in this study, the value of p was empirically set to 0.5.

III. Experiments

A. Experimental Setup

In this study, a public corpus MIR-QBSH [9] was used for the experiments, where the anchor positions of all queries were assumed to be the beginning of a song. The corpus comprised 5460 query clips, including 959 humming clips, 4299 singing in Mandarin clips, and 202 singing in English clips. Because the proposed speech recognition engine was designed to recognize Mandarin, in this study, 5023 clips were selected from the corpus corresponding to 35 songs in Mandarin (4299 singing in Mandarin clips in addition to 724 humming clips belonging to one of the 35 Mandarin songs). All clips were manually labeled as either singing or humming. To increase the complexity of the comparison, 2119 noise songs from the Essen collection [29] were added to the database; consequently, the database contained 2154 songs.

First, 200 humming clips and 200 singing clips were selected from the corpus (5023 query clips) and used as an SHD training set. The remaining 4623 clips (4084 singing clips and 539 humming clips) were used to test the overall performance of the proposed QBSH system.

Acoustic models were constructed by training with more than 150 h of Mandarin speech corpora, including TCC300 [25] and some privately collected corpora. The test corpus for obtaining the model similarity was Tang Poetry corpus 2002 [26] (approximately 4 h in length). The Cepstral mean normalized 12 MFCCs combined with log energy as well as their delta and acceleration (MFCC_E_D_A_Z in HTK terminology) were used as acoustic features. To facilitate the comparison of the results with those reported in [28], the acoustic models (for SHD and singing voice recognition) were adapted using maximum likelihood linear regression, where the adapted corpus is the vocal component of the MIR-1 K data set [27] (approximately 2 h in length). The detailed parameters for training and adapting acoustic models were set empirically.

The following experiments were performed on a laptop with i5-3230M CPU and 12 GB RAM. The experiments were implemented in C.

B. Experimental Results of SHD

Fig. 7 shows the SHD Detection Error Tradeoff (DET) curve of the training data at various EUPC/EUSC thresholds. In the figure, the free-phone and free-syllable decoding results are denoted as "FreePhn" and "FreeSyl," respectively. Moreover, "-Adapt" is appended to indicate usages of the adapted acoustic models. The minimal error rate (defined as the ratio between the number of misclassifications and total number of cases) of each case is identified by the circles on the curve. Occasionally, humming may contain certain pronunciations that are incomparable to any syllable in speech or singing; consequently, the performance of free-syllable decoding is inferior to that of free-phone decoding. Because the adapted corpus contains only singing voices, predicting the overall effect on SHD is

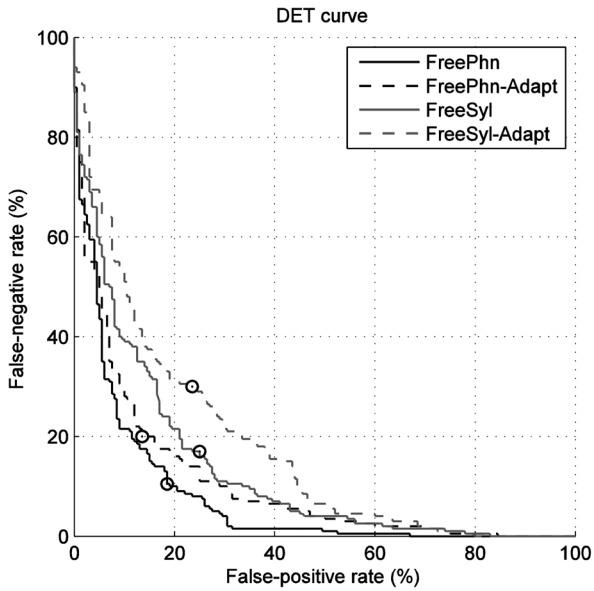


Fig. 7. The DET curve for SHD.

TABLE I
CONFUSION MATRIX OF SHD OVER THE TEST DATA

Ground truth	Computed	Singing	Humming
		percentage (# of clips)	
Singing	FreePhn	82.93% (3387)	17.07% (697)
	FreePhn-Adpat	76.64% (3130)	23.36% (954)
	FreeSyl	78.01% (3186)	21.99% (898)
	FreeSyl-Adapt	72.28% (2952)	27.72% (1132)
Humming	FreePhn	15.96% (86)	84.04% (453)
	FreePhn-Adpat	18.18% (98)	81.82% (441)
	FreeSyl	18.92% (102)	81.08% (437)
	FreeSyl-Adapt	20.78% (112)	79.22% (427)

difficult because the adapted models are suitable for singing but unsuitable for humming. The optimal training result was obtained through free-phone decoding using the original (i.e., nonadapted) acoustic models; the minimal error rate occurs when the EUPC threshold was 28.8541.

The thresholds that resulted in minimal errors in the training stage were used to evaluate the test data (different thresholds were used in different methods). Table I shows the confusion matrix of the test data based on four decoding processes. For "FreePhn" (threshold = 28.8541), the recognition rate is 83.06%. In particular, 15.96% of the humming clips were misclassified as singing, which could result in erroneous outputs in singing voice recognition. An initial error analysis indicated that the misclassification of some humming clips probably resulted from contrasting variations in pronunciation (i.e., mixtures of "da," "la," "deng," and so on) in the humming. Regarding the singing clips, 17.07% of them were misclassified as humming, which probably resulted from repetitive or similar syllables in the lyrics, or from the slow tempo of the song.

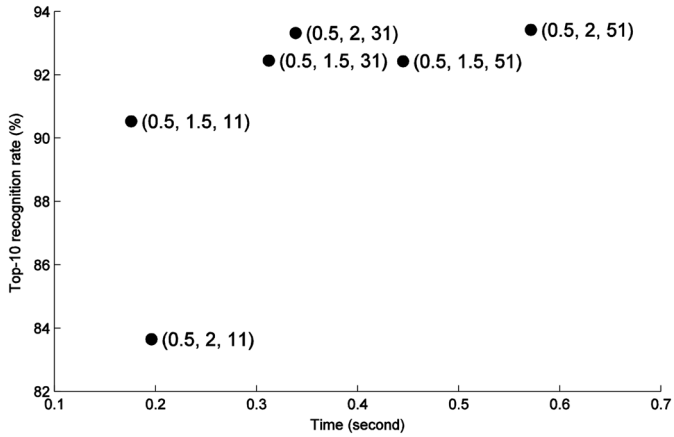


Fig. 8. The top-10 melody recognition rates versus computation time.

However, the accuracy of melody recognition is already high; thus, the misclassification of singing clips had a relatively minor impact on the overall performance.

C. Melody Recognition Results

Both the melody recognition and lyric matching performance levels were assessed according to the top-n recognition rate, as defined in (8): Fig. 8 depicts the top-10 melody recognition rates versus computation time of the test data. The parameters are denoted as ($s_{\min}$, $s_{\max}$, r) on the right-hand side of the points. As shown in this figure, for fixed $s_{\min}$ and $s_{\max}$, increasing r results in higher accuracy but longer computation time. The parameter sets, (0.5, 2, 11) and (0.5, 2, 51), that resulted in the lowest and highest recognition rates, respectively, were used for the following experiments. Because $s_{\min}$ and $s_{\max}$ of these two parameter sets are identical, only the resolution was used to distinguish between these two sets in the subsequent discussions.)

D. Lyric Matching Results

Table II shows the six settings for combining the SHD and lyric string decoding methods used in this study, and Fig. 9 illustrates the top-10 recognition rates for lyric matching versus computation time of the singing clips corresponding to the six settings shown in Table II.

As shown in this figure, because two decoding processes (SHD and singing voice recognition) are required for "Lin" and "Lin-Adapt," these two settings are slower than the other four settings. However, because the output strings from the free-phone/syllable decoding contain more noises than the results from using linear lexicon, the recognition rates yielded by "Lin" and "Lin-Adapt" are considerably higher than those obtained under the other four settings. Moreover, because the target utterances are singing, the adapted acoustic models are more accurate than the original (nonadapted) ones. Using the adapted models, the recognition rates are 72.30% for "Lin-Adapt," 40.71% for "FreePhn-Adapt," and 39.30% for

$$\text{top}-n\text{recognition rate} = \frac{\#\text{of clips with its ground truth appearing in the top-}n\text{ candidates}}{\#\text{of clips}}. \quad (8)$$

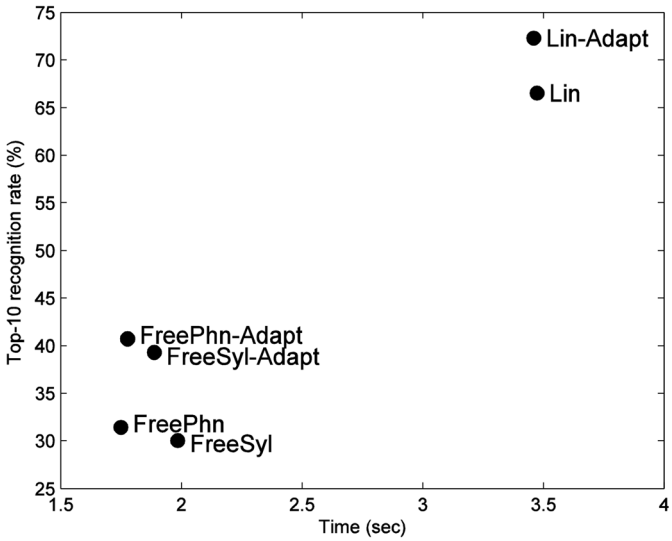


Fig. 9. Top-10 recognition rates of lyric matching versus computation time.

TABLE II
DESCRIPTION OF SETTINGS FOR LYRIC MATCHING

Name of settings	SHD methods	Decoded string for lyrics matching	No. of clips that are classified as singing
Lin	FreePhn	Singing voice recognition module, using original acoustic models	3473
Lin-Adapt	FreePhn	Singing voice recognition module, using adapted acoustic models	3473
FreePhn	FreePhn	SHD module	3473
FreePhn-Adapt	FreePhn-Adapt	SHD module	3228
FreeSyl	FreeSyl	SHD module	3288
FreeSyl-Adapt	FreeSyl-Adapt	SHD module	3064

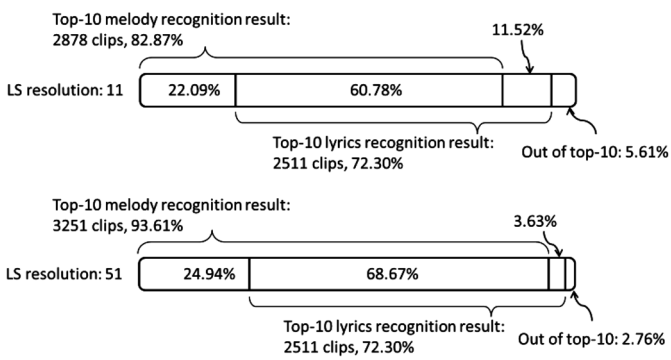


Fig. 10. Distribution of recognition results of 3473 clips for two different LS resolutions.

"FreeSyl-Adapt." Accordingly, these three settings were selected for further experimentation.

Fig. 10 illustrates the distribution of the recognition results of the 3473 clips classified as singing when the "Lin-Adapt" setting was used. The top-10 accuracy of the melody recognition is 82.87% and 93.61% when the LS resolutions were set at 11 and 51, respectively. The minimum and maximum of

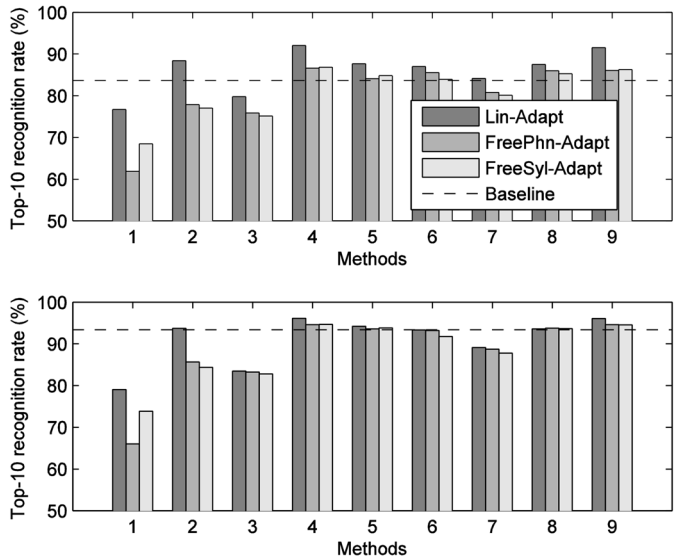


Fig. 11. The top-10 recognition rates of the baseline system and the nine different distance/similarity combination methods. The upper and lower plots correspond to LS resolution equal to 11 and 51, respectively.

the scaling factors $s_{\min}$ and $s_{\max}$ were set to 0.5 and 2, respectively. Furthermore, when the LS resolution was set at 11, 11.52% of the clips in top-10 candidates that could not be obtained using melody recognition could be obtained using lyric matching. Moreover, even when the accuracy of melody recognition was high (93.61% when the LS resolution was 51), 3.63% of the clips that could not be retrieved using melody recognition could be retrieved using lyric matching, despite most of the clips containing out-of-tune singing. These observations demonstrate the potential advantage of combining melody and lyric information.

E. Combined Result

Fig. 11 depicts the top-10 recognition rates of the baseline system (which uses only melody information) and the nine distance/similarity combination methods at LS resolutions of 11 and 51, respectively. The dashed line represents the baseline system, followed by distance/similarity combination methods 1-9 for three settings (i.e., "Lin-Adapt," "FreePhn-Adapt," and "FreeSyl-Adapt"). The value of p in (7) was empirically set to 0.5. According to these results, method 4 was selected for distance/similarity combination in the proposed system.

A post analysis was subsequently conducted for method 4 by plotting the overall recognition rates against the values (Fig. 12). At an identical LS resolution, the recognition rate of Lin-Adapt is consistently higher than those of FreePhn-Adapt and FreeSyl-Adapt. Apparently, the performance remained relatively similar at LS resolutions equal of 11 and 51, provided that the value of p is within [0.4, 0.9], thereby validating the decision set the value for p at 0.5.

Fig. 13 illustrates the overall performance of the proposed QBSH system. The top-10 recognition rates of the combined results at an LS resolution of 11 (51) are 92.02% (96.15%) for "Lin-Adapt," 86.63% (94.61%) for "FreePhn-Adapt," and 86.83% (94.70%) for "FreeSyl-Adapt," which outperforms the baseline system using melody information alone, which

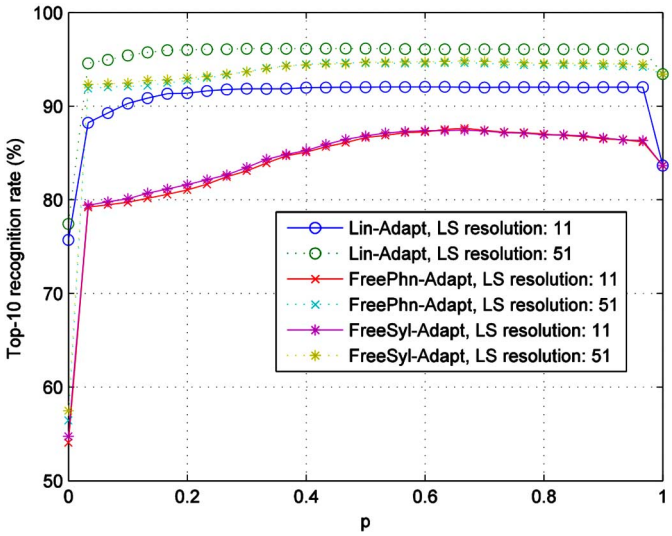


Fig. 12. Plots of overall recognition rates with respect to the values of p, for two values of LS resolution.

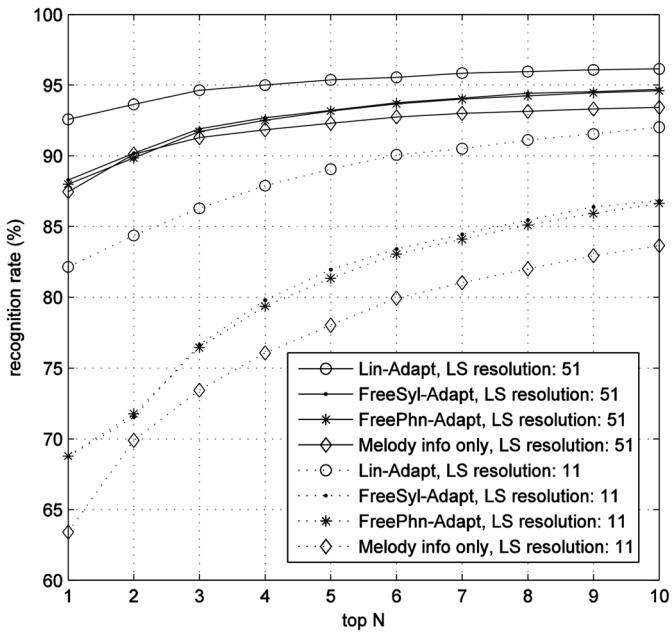


Fig. 13. The top-N recognition rate of 4623 clips.

achieved a recognition rate of 83.65% (93.42%). The error reduction rates at an LS resolution of 11 (51) are 51.19% (41.45%) for "Lin-Adapt," 18.25% (18.09%) for "FreePhn-Adapt," and 19.44% (19.41%) for "FreeSyl-Adapt." The results obtained from a sign-rank test showed that the improvement of the top-10 hit results for all the methods were at the 0.01 level, indicating that the improvement is statistically significant. These results show that the improvement from combining melody and lyric information in a QBSH system.

IV. Conclusion and Future Work

This paper proposes an improved QBSH system that distinguishes between singing and humming queries, and subsequently applies various procedures to exploit the lyric information in the singing input. The experimental results demonstrate the effectiveness of the proposed system, with error reduction rates ranging from 18.09% to 51.19%, depending on the parameter settings.

Several directions for future research are currently being examined. A potentially desirable direction for future research would be to incorporate multilingual speech recognition, particularly because a considerable number of famous songs in various languages have the same melody but different lyrics. Moreover, using graphics processing unit computation methods could assist in improving the system performance, which would allow the comparison of input queries to start from any point in a song in the database.

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, CA, USA, 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, Dec. 2000, pp. 85–89, R.O.C..

[4] C.-Y. Chi, Y.-S. Wu, W.-r. Chu, D. C. Wu, J Y.-j Hsu, and R. T.-H. Tsai, "The power of words: Enhancing music mood estimation with textual input of lyrics," in *Proc. Int. Conf. Affective Comput. Intell. Interact.*, 2009, pp. 1–6.

[5] T. Wang, D.-J. Kim, K.-S. Hong, and J.-S. Youn, "Music information retrieval system using lyrics and melody information," in *Asia-Pacific Conf. Inf. Process.*, 2009, pp. 601–604.

[6] X. Xu, M. Naito, T. Kato, and H. Kawai, "Robust and fast lyric search based on phonetic confusion matrix," in *Proc. Int. Symp. Music Inf. Retrieval*, 2009, pp. 417–422.

[7] J.-S. R. Jang, H.-R. Lee, and M.-Y. Kao, "Content-based music retrieval using linear scaling and branch-and-bound tree search," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2001.

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

[9] J.-S. R. Jang, "MIR-QBSH Corpus," *MIR Lab, CS Dept, Tsing Hua Univ, Taiwan* [Online]. Available: http://mirlab.org/jang, Available at the "MIR-QBSH Corpus" link at

[10] J.-C. Chen and J.-S. R. Jang, "TRUES: Tone Recognition Using Extended Segments," *ACM Trans. Asian Lang. Inf. Process.*, vol. 7, no. 3, pp. 1–23, Aug. 2008, Article 10.

[11] MIREX 2009, 2009. [Online]. Available: http://www.music-ir.org/mirex/wiki/2009:Query-by-Singing/Humming_Results

[12] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice based on verification of recognized hypotheses," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR'06)*, 2006.

[13] J.-H. Chen, "Content-based music emotion analysis and recognition," M.S. thesis, Comput. Sci. Dept., National Tsing Hua Univ., Hsinchu , Taiwan, Jun. 2006.

[14] P. Papiotis and H. Purwins, "A lyrics-matching QBH system for interactive environments," in *Proc. Sound and Music Comput. Conf.*, 2010.

[15] B. Schuller, G. Rigoll, and M. Lang, "Discrimination of speech and monophonic singing in continuous audio streams applying multi-layer support vector machines," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2004, pp. 1655–1658.

[16] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," in *Proc. INTERSPEECH*, 2005, pp. 1141–1144.

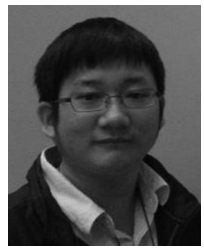
[17] D. Gärtner, "Singing/Rap classification of isolated vocal tracks," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR'10)*, 2010, pp. 519–524.

[18] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP J. Adv. Signal Process.*, vol. 2007, 10.1155/2007/38727, Article ID 38727, 8 p..

[19] R. Mayer, R. Neumayer, and A. Rauber, "Rhythm and style features for musical genre classification by song lyrics," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR 2008)*, 2008.

[20] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," in *Proc. 35th Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, 2010.

[21] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, 2010.
[22] Z. Guo, Q. Wang, G. Liu, J. Guo, and Y. Lu, "A music retrieval system using melody and lyric," in *Proc. Multimedia Expo Workshops (ICMEW)*, 2012.
[23] X. Wu, M. Li, J. Liu, J. Yang, and Y. Yan, "A top-down approach to melody match in pitch contour for query by humming," in *Proc. Int. Conf. Chinese Spoken Lang. Process.*, 2006.
[24] S. Huang, L. Wang, S. Hu, H. Jiang, and B. Xu, "Query by humming via multiscale transportation distance in random query occurrence context," in *Proc. ICME*, 2008.
[25] Mandarin microphone speech corpus - TCC300, [Online]. Available: http://www.aclclp.org.tw/use_mat.php\#tcc300edu
[26] Tang Poetry Corpus, pp. 2002–2006 [Online]. Available: http://mirlab.org/research/corpus/tangpoetry
[27] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1 K dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb. 2010.
[28] C.-C. Wang, J.-S. R. Jang, and W. Wang, "An improved query by singing / humming system using melody and lyrics information," in *Proc. 11th Int. Society Music Inf. Retrieval Conf.*, 2010.
[29] ESAC Data Homepage, [Online]. Available: http://www.esac-data.org/, 2014
[30] W.-T. Kao, C.-C. Wang, K. C. K. Chang, J.-S. R. Jang, and W. S. Liou, "A two-stage query by singing/humming system on GPU," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, 2013.
[31] T. K. Ho, J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Patter Anal. Mach. Intell.*, vol. 16, no. 1, pp. 66–75, 1 1994.
[32] A. Degani, M. Dalai, R. Leonardi, and P. Migliorati, "A heuristic for distance fusion in cover song identification," in *Proc 14th Int. Workshop Image Anal. Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.
[33] M. McVicar, D. P. W. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. 39th Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 3117–3121.
[34] X. Zhao and P. Li, "An online database of phonological representations for mandarin chinese," in *Proc. Behavior Res. Meth.*, May 2009, vol. 41, no. 2, pp. 575–583.



Chung-Che Wang is a Ph.D. candidate in computer science at National Tsing Hua University (Hsinchu, Taiwan). His research interests include query-by-singing/humming and audio fingerprinting.

Jyh-Shing Roger Jang (M'93) received his Ph.D. from the EECS Department at the University of California, Berkeley. He studied fuzzy logic and artificial neural networks with Prof. Lotfi Zadeh, the father of fuzzy logic. As of Dec. 2014, Google Scholar shows over 9000 citations for Dr. Jang's seminal paper on adaptive neuro-fuzzy inference systems (ANFIS), published in 1993. After obtaining his Ph.D., he joined MathWorks to coauthor the Fuzzy Logic Toolbox (for MATLAB). He has since cultivated a keen interest in implementing industrial software for pattern recognition and computational intelligence. He was a professor in the CS Dept. of National Tsing Hua Univ., Taiwan, from 1995 to 2012. Since August 2012, he has been a professor in the CSIE Dept. of National Taiwan Univ., Taiwan. He has published one book entitled *Neuro-Fuzzy and Soft Computing*, two books on MATLAB programming, and one book on JavaScript programming. He has also maintained toolboxes for Machine Learning and Speech/Audio Processing, and online tutorials on *Data Clustering and Pattern Recognition* and *Audio Signal Processing and Recognition*. He is the conference co-chair of ISMIR (International Society for Music Information Retrieval), Taipei, Oct 2014.

His research interests include machine learning and pattern recognition, with applications to speech recognition/assessment/synthesis, music analysis/retrieval, image identification/retrieval, and semiconductor manufacturing intelligence. For further information on Prof. Jang, visit http://mirlab.org/jang.