Competence-Based Song Recommendation: Matching Songs to One's Singing Skill

Kuang Mao, Lidan Shou, Ju Fan, Gang Chen, and Mohan S. Kankanhalli, *Fellow, IEEE*

***Abstract*—Singing is a popular social activity and a pleasant way of expressing one's feelings. One important reason for unsuccessful singing performance is because the singer fails to choose a suitable song. In this paper, we propose a novel competence-based song recommendation framework for the purpose of singing. It is distinguished from most existing music recommendation systems which rely on the computation of listeners' interests or similarity. We model a singer's vocal competence as a singer profile, which takes voice pitch, intensity, and quality into consideration. Then we propose techniques to acquire singer profiles. We also present a song profile model which is used to construct a human annotated song database. Then we propose a learning-to-rank scheme for recommending songs by a singer profile. Finally, we introduce a reduced singer profile which can greatly simplify the vocal competence modelling process. The experimental study on real singers demonstrates the effectiveness of our approach and its advantages over two baseline methods.**

***Index Terms*— Learning-to-rank, singing competence, song recommendation.**

I. Introduction

SINGING is a popular social activity and a good way of expressing one's feelings. While some people enjoy the experience of rendering a wonderful solo in a karaoke party, many others are upset by their own singing skill due to an unpleasant performance in the past. Many times, this is due to a poor choice of song rather than the singing ability. It is extremely hard for a girl with a soft voice to sing like Mariah Carey whose songs require a strong voice to express strong emotions. It is equally hard for a bass singer to perform Tristan in tenor. A good performance is only possible if a song is carefully chosen with regard to the singer's vocal competence.

However, song recommendation for singers appears to be a task comprehensible to professionals only. Experienced singing teachers listen to find the advantages in one's voice and choose suitable songs matching one's vocal competence. Typically, they choose *challenging* songs in order to distinguish the singer from others. In other words, they tend to recommend songs which secure the best singing performance. Such selection is different from the traditional scenario of song recommendation, which typically selects songs based on the singer's interests. With the development of computational acoustic analysis, it is possible to study the vocal competence from a singer's digitized voice, and then make automatic song recommendation based on the singer's "performance caliber".

In this paper, we report our work on human *competence-based song recommendation* (CBSR). The main objective is to computationally simulate the know-how of a singing teacher–To recommend challenging but manageable songs according to the singer's vocal competence. Specifically, we develop a system which takes a singer's digitized voice recording as the input, and then recommends a list of songs relying on analysis of the singer's personal vocal competence and a subsequent search process in a song database. Although the general procedures of our approach appear similar to Music Retrieval By Humming [20], the underlying ideas and techniques are totally different from it. Our research purpose is significantly different from most existing song retrieval and recommendation systems, which focus on matching the listener's tastes or interests. To the best of our knowledge, it is the first work to study singing-song recommendation using singer's own voicing capabilities.

Competence-based song recommendation faces three main technical challenges:

First, how should the singing competence be modeled? If we consider the singer's voicing input as a query, then a next question would be, what is the query like? As we all know, different people produce different ranges of pitches and intensity in their singing. Even for the same person, the singing performance may vary significantly depending on the pitch and intensity. The competence model and the query method must take such variations into consideration.

Second, a song database should be constructed. Likewise, we should ask, what model can be used to represent each song for the recommendation? Unlike previous work which focuses on transcription [26], [3], we attempt to discover the voice characteristics of each song, which in turn pose different requirements

Manuscript received July 03, 2014; revised December 29, 2014; accepted January 04, 2015. Date of publication January 15, 2015; date of current version February 12, 2015. This work was supported by the National Basic Research Program (973 Program) under Grant 2015CB352400, the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO, the National Science Foundation of China under Grant 61170034 and Grant 61472348, the National High Technology Research and Development Program of China under Grant SS2013AA040601, the National Key Technology R&D Program of the Ministry of Science and Technology of China under Grant 2013BAG06B01, and the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

K. Mao and G. Chen are with the Database Lab, College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: mbill@zju.edu.cn; cg@zju.edu.cn).

L. Shou is with the CAD and CG Lab, College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: should@zju.edu.cn).

J. Fan and M. Kankanhalli are with the School of Computing, National University of Singapore, 119077, Singapore (e-mail: fanj@comp.nus.edu.sg; mohan@comp.nus.edu.sg).

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/TMM.2015.2392562

to the singer. For example, some songs must be sung in a soft voice while some others need to be delivered in a loud one. A good song model has to capture these features properly.

Third, a search mechanism must be provided for the database to bridge the gap between the singer's competence model and the songs. Meanwhile, a ranking method is needed to provide relevance-like ordering for the recommended songs.

To solve the first challenge, we tackle it by proposing a novel singing competence model which is instantiated as a *singer profile*. To construct a singer profile, we first consider an existing vocal capability model called Vocal Range Profile (VRP), which has been proposed in the literature of medical acoustics for clinical assessment of voice diagnosis [29], voice treatment [30] and vocal training [31]. Specifically, the VRP of a person is a two-dimensional bounded area in the (pitch,intensity) space. For each pitch within the person's voicing capability, the range of intensity produced by her/him is depicted. Unfortunately, the VRP model cannot sufficiently describe one's singing competence. The main reason is that VRP overlooks a singer's *voice quality*, which largely determines how nicely a voice is produced. Our primary observation here is that, due to the fact that a person has variable performance (quality) when producing voice at different pitch and intensity, the voice quality for a person should be defined as a numerical function on the (pitch, intensity) space. As a result, the singer profile consists of two components: the singer's *VRP* and the respective *voice quality function* defined on her/his VRP area.

However, modeling a complete singer profile requires a lot of recording tasks, we study the relative importance of different parts of one's singer profile in recommendation and propose a reduced singer profile to model one's singing competence. The reduced model uses only a small part of VRP to model one's vocal competence while not losing much recommendation accuracy. During the reduced VRP recording, we perform a binary search recording strategy to quickly locate subjects' singing limitation and collect the voice samples used for building the reduced singer profile.

The above competence model (singer profile/reduced singer profile) creates a new problem–the voice quality function of a singer is not readily available. In fact, singing voice quality is an empirical value and its mathematical formulation has not been adequately studied in the acoustics community. The only obvious way to acquire a person's voice quality is manual annotation on various (pitch, intensity)-pairs. However, manual annotation at query time is obviously unacceptable. In our solution, we avoid the mathematical formulation of the voice quality function. Instead, we "learn" the function from empirical values of the population given by experts. This leads to a supervised learning method which automatically computes the voice quality function at query time.

For the second challenge, we introduce the notion of *song profile*. Like a singer profile, each song profile in the database must also be annotated by the pitches of its notes and their respective intensities. While the pitches of a song are typically available, the intensity of each note cannot be easily acquired. To the best of our knowledge, extracting the singing intensity from polyphonic songs still remains an unexplored problem. We employ a number of professionals to annotate each song with a *piecewise intensity sequence* using a software tool. This process is feasible as it can be done during an offline phase.

The third challenge can seemingly be solved with a naive approach–that is to recommend songs whose pitch and intensity ranges are completely contained in one's vocal range with good quality. However, this approach tends to prioritize only "easy" songs and therefore contradict our motivation. In contrast, we propose a *competence-based song ranking* scheme to rank songs in the database for the singers. These criteria include the pitch and intensity. Nevertheless, it is possible to extend the scheme by adding other criteria. In our scheme, we extract features from singer and song profiles as well as the respective rankings of experts to train a Listnet model. This model is cross-validated on our datasets

Our main contributions are summarized as follows.

1) We propose a novel competence-based song recommendation framework.
2) We present a singer profile to model singing competence. We illustrate the process of generating singer profiles.
3) We study the importance of singer profile areas and present a reduced singer profile to simplify the modeling process of one's vocal competence.
4) We also present the song profile and describe the method of generating the respective song profile from a database.
5) The song recommendation is implemented using a multiple criteria learning-to-rank scheme.
6) Our experiments on a group of users show promising results of the proposed framework.

The rest of our work is organized as follows: Section II introduces the related work. Section III conducts an overview of the framework. Section IV, V presents the singer profile, song profile models and the techniques to acquire these profiles. Section VI describes the learning-to-rank recommendation scheme. Section VII presents the reduced singer profile modeling techniques. The experiments are detailed in Section VIII. Finally, Section IX concludes the paper.

II. Background And Related Work

In this section, we shall discuss the related work in the literature and introduce some important concepts. We will look at previous studies in vocal range profile, voice quality, and song recommendation.

A. Vocal Range Profile

As shown in Fig. 1, a *vocal range profile* (VRP), also called phonetogram, is a two-dimensional map in the pitch-intensity space (in acoustic terms, it is also called the frequency-amplitude space), where each point represents the phonation of a human being. This map depicts all possible (pitch, intensity)-pairs that one can produce. The projection of a VRP map on the pitch axis, which defines the range of pitches that one can ever produce, is called the *pitch range*. Specifically, the VRP characterizes one's voicing capability by defining the *maximum* and *minimum* vocal intensity at each pitch value across the entire pitch range.

The concept of VRP was first introduced by Wolf *et al.*[1] in 1935. Since then, VRP has been widely applied in objective

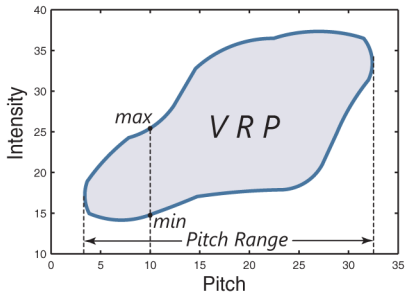


Fig. 1. Vocal range profile (VRP) of a singer.

clinical voice diagnosis and singer's vocal training. Many papers [4], [2] have studied the variation of VRP with regard to gender, age, voice training and so forth. It has been found that the VRPs of different people usually differ significantly. Therefore, it can be used as a voice signature for human being.

The recording process of VRP has been standardized and recommended by the Union of European Phoniatricians [5]. To describe it simply, the process requires the singer to traverse each pitch in her/his pitch range from the loudest to the softest through voicing vowel /a/. In our work, we employ a similar process to acquire each singer's VRP. The result is used as a basis for computing one's singer profile.

B. Voice Quality

The technique of *objective voice quality measurement* has been widely used in voice illness diagnosis. Such techniques usually extract sound sampling features to represent voice characteristics, for example *period perturbation*, *amplitude perturbation* etc. In the field of vocal music, there are other measures that describe the voice quality of sounds. For example, singing power ratio [6] is defined based on the spectral analysis of voice samples. This measure differs a lot between trained and untrained singers. The other similar examples include tilt [18], and ltasSlope [8]. The last two are meant to discover the singer's singing talent [6]. The above mentioned measures reveal many characteristics of the voice. However, these measures cannot adequately solve our problem, which requires a detailed voice quality evaluation on a singer's VRP map.

As described in the previous subsection, VRP describes the singer's voicing area in the pitch-intensity space. Some previous studies on proprietary voice quality measures reveal that each measure may vary significantly across VRP area. [9]evaluates quality parameters such as jitter, shimmer, and crest factor over VRP, and finds that each of these quantities differs significantly across VRP. Another work in [28]analyzes the distribution of three separate acoustic voice quality parameters on VRP, and has reached a similar conclusion. In our work, we do not evaluate each single parameter. Instead, we model the voice quality as an overall function on VRP.

One study worth mentioning is [12], which incorporates the knowledge of voice diagnosis experts to train a linear model, and then predicts the overall voice quality of a patient for clinical voice diagnosis. Our method for computing voice quality on VRP area is motivated by this work. But our underlying problem and expert knowledge of singing voice quality is very different from the previous study.

C. Song Recommendation

Traditional song/music recommendation focuses on recommending songs by user's listening interests. The earlier studies such as [13], [7], [10] explore techniques in the domain of *content based song recommendation*. These techniques aim at discovering user's favorite music in terms of music content similarity such as moods and rhythms. However, this kind of methods has its limitation because typically the low-level features cannot fully represent the user's interests. A more effective way is to employ the so-called collaborative methods [14], [11], [17], [25] which recommend songs among a group of users who have similar interests.

Our work is different from the above studies as it recommends songs by singer's performance needs rather than interests. It also differs from post-singing performance appraisal [32] which requires singing to be performed in the first place. In our preliminary studies [21], [22], we formulated the scientific problem of competence-based song recommendation, proposing a novel solution and demonstrated a system for karaoke song recommendation. However, the proposed singing competence model requires too many expensive human recordings and is very complex to model. This paper extends [22] by introducing a simplified singing competence model, called *reduced singer profile*. This model can reduce half of the recording task while not losing much accuracy in recommendation.

Recently, we have presented a song recommendation framework for a social singing community [23]. It recommends songs for singing through a set of pre-built difficulty orderings. The difficulty ordering between two songs indicates their relative ease in terms of rendering a good performance. However, the used difficulty orderings may not fit everybody. In this work, we build an accurate individual singing model for each singer. The recommendation result is therefore more reliable.

III. Overview of CBSR Framework

As Fig. 2 shows, our competence-based song recommendation framework works in two phases, namely training phase and testing phase. During the training phase, we employ a group of singers as the subjects and a number of music experts to train a competence-based ranking function. The main procedures of training phase are listed as follows.

1) *Data Preparation*: We first record the voice of a group of singers, and generate the VRP for each singer. Meanwhile, a song database is annotated with pitch and intensity information by a few vocal music experts.
2) *Singer Profile Generation*: Each singer's voice is used to construct a singer profile which depicts (i) the singer's vocal area by a *VRP* and (ii) the singer's competence by a *voice quality function* on her/his VRP.
3) *Song Profile Generation*: The song database together with its annotated data are used to generate song profiles, which contain its note distribution and other statistical information.

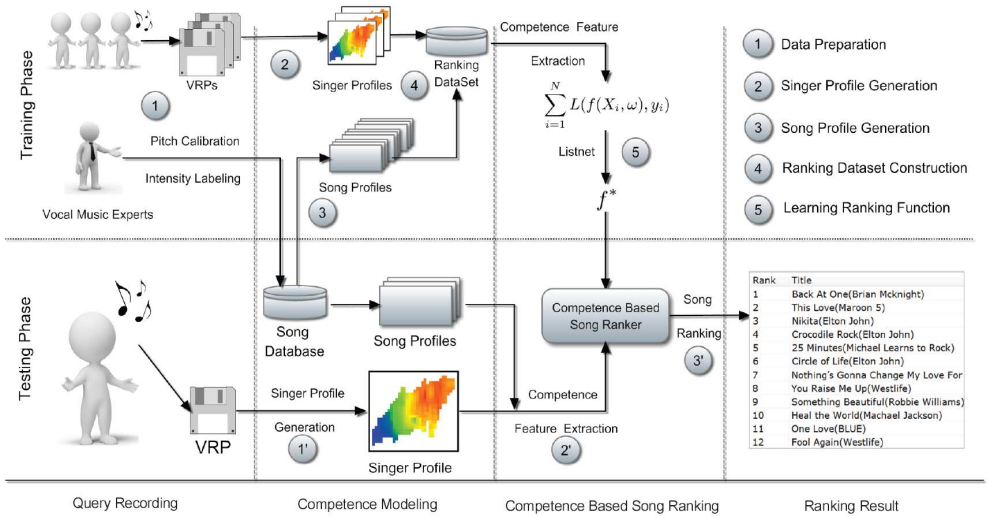


Fig. 2. Overview of the competence-based song recommendation framework.

4) *Construction of the Ranking Dataset*: Each training subject is asked to sing a number of songs in the song database in front of the vocal music experts. The latter will rate the song with a score for the subject. The (i) singer profiles, (ii) song profiles in the database, and (iii) the rankings given by the experts, comprise the ranking dataset.
5) *Learning the Ranking Function*: We extract features from the ranking dataset. These features are fed into a listwise learning-to-rank algorithm called *Listnet* to learn the ranking function.

In the testing phase, (1') a subject is asked to record voices for singer profile generation. After (2') extracting features from the tester subject's singer profile and the song profiles in the database, we can (3') make recommendation using the ranking function learnt from the training phase.

Our main technical contributions focus on procedure 2, 3, and 5. We will give the details of the other procedures in the experimental study.

IV. Singer Profiles

In this section, we first propose a vocal competence model called the singer profile. Then we detail the process of generating a singer profile. Finally, we present a simple method for per-profile analysis, which extracts some important singer profile characteristics.

A. Singer Profile Modeling

In our model, a singer profile contains two components: (1) VRP of the singer, and (2) a voice quality function defined over the VRP area. Given the definition of VRP in Section II-A, we shall now formulate the definition of voice quality. If we consider each (pitch, intensity) point in VRP a *vocal point*, denoted by vp, then voice quality is defined as a function of vp.

Definition 1: Voice Quality: Given the VRP of a singer, voice quality is a scalar function $\psi(vp) > 0$ for any vocal point $vp \in VRP$.

Practically, voice quality indicates a quantity measuring whether the singing voice at a particular vocal point is fair-sounding.

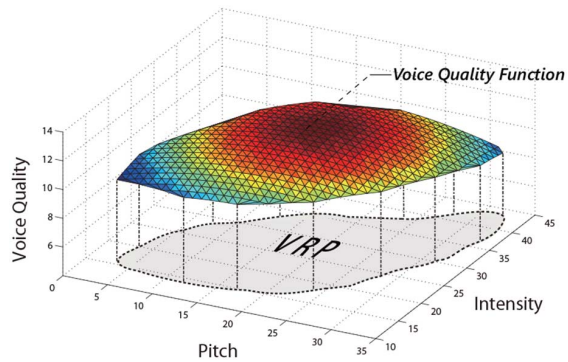


Fig. 3. Singer profile. Colors on the surface indicate the voice quality.

Now a singer profile can be defined as a tuple of $< VRP, \psi >$, where VRP is the VRP of the singer and ψ is her/his respective voice quality function. In practice, however, we prefer a discretized form of singer profile, where all vocal points in a VRP are enumerated, as being defined in the following:

Definition 2: Singer Profile: A singer profile is a set of tuples, written as $< vp, \psi(vp) >$, where $vp \in VRP$ is a vocal point that the singer can voice.

Fig. 3 is a schematic diagram of a singer profile. If the VRP becomes discretized on both pitch and intensity dimensions, then the total number of vocal points in a VRP will be finite. Thus the singer profile will become a finite array of the tuples.

In our system, we discretize pitches into semitone scale and intensity into units of 2 dB. This is consistent with most vocal music requirements. However, it is a trivial task to use finer scales if necessary.

B. Singer Profile Generation

Generating the singer profile includes two major steps: *VRP generation* and *voice quality computation*. The first one is quite standard and straightforward, but the second is much more complicated.

1) Step 1: VRP Generation: Before the VRP recording, the singer has to perform "warm-up" exercises such as singing.

Then the singer is asked to stand 1 meter away from the microphone and start the recording procedure. The recording procedure requires the singer to vocalise each pitch in her/his pitch range from the softest intensity to the loudest. Meanwhile, a singing teacher is present to help the singer locate the pitch and guide the singer to increment the intensity while keeping the pitch steady. To help stabilizing the voice, we also provide the singer real-time visual cue of the singing pitch and intensity. However, this practice is optional.

For an untrained singer, it is difficult to increase the pitch by semitones. Therefore, singers are only requested to increase pitch by the whole tone scale. Actually, by voicing each whole tone, the neighboring semitones will also be sufficiently covered. For each singer, an average number of 24 semitones are recorded in the recording procedure. Each piece of voicing is stored in a separate WAV file. The average time for recording is around 10 minutes.

Note the above procedure is in fact a sampling process in the pitch-intensity space, which results in a discrete VRP (with a number of vocal points). After this, we segment all voice files into *voice pieces* with a time duration of 0.2 second. The reason for splitting voice into short pieces is that the voice pitch, intensity, and quality can be regarded as invariant in each piece. Thus, each voice piece finds its respective (pitch, intensity) value and gets associated with a vocal point in the VRP. Now the VRP can be seen as a set of vocal points, each associated with one or more voice pieces.

2) Step 2: Voice Quality Computation: As mentioned before, there exists no prior work on the mathematical formulation of the voice quality function, even though we need the value of this function on different vocal points. Considering the aggregated voice pieces that we collected for each vocal point in the previous step, we can take such pieces as input and manually label them with a quality value. This idea motivates a supervised learning method to learn a *quality evaluation function* from empirical voice quality annotation given by the experts. The input of this function is a voice piece, and the output is the voice quality of this voice piece (coupled by its respective vocal point, as each voice piece can be uniquely mapped to a vocal point). Thus, the quality evaluation function generates in effect a vocal point sampling for the voice quality function.

Note that the learning technique discussed here is only for generating intermediate data–the voice quality function. The reader should differentiate it from the learning-to-rank scheme proposed in Section VI which is aimed at recommending songs. In the following, we will first present the method of training the quality evaluation function, and then describe how to utilize it for voice quality computation (prediction).

C. Supervised Learning

In order to train the quality evaluation function, a number of vocal music experts are requested to annotate the quality of voice pieces in each VRP recording using a software tool called *Praat* [27]. Each expert listens to the recorded WAV files and annotates the voice quality of different parts in each file based on the steadiness and clearness of the sound. The possible annotation scores range from 1 to 5 (the lower the better quality). Table I shows some criteria for voice quality rating in each grade. After an entire file becomes annotated, it will be split into voice pieces for training.

TABLE I
VOICE QUALITY RATING CRITERIA

Quality Grade	Criteria Description
1-Perfect	Excellent controllability of steadiness, excellent clearness
2-Satisfactory	Good controllability of steadiness, good clearness
3-Normal	Hard to decide good or bad
4-Uncontrollable	Hard to control the steadiness or hoarseness
5-Noise	Unacceptable voice for singing

TABLE II
FEATURES FOR VOICE QUALITY EVALUATION

Feature Category	Feature Names
Pitch Features	*medianPitch*, *meanPitch*,*sdPitch*,*minPitch*, *maxPitch*,*nPulses*,*meanPeriod*,*sdPeriod*
Frequency Perturbations	*jitter_loc_abs*[24], *jitter_loc*[24], *jitter_rap*[24],*jitter_ppq5*[24]
Amplitude Perturbations	*shimmer_loc*[24], *shimmer_loc_dB*[24], *shimmer_apq3*[24],*shimmer_apq5*[24], *shimmer_apq11*[24]
Spectrum Features	*mean_nhr*[18], *mean_hnr*[19], *singing power ratio*[6],*tilt*[18],*ltasSlope*[8]

The quality evaluation function is trained as follows. First several acoustic features are extracted for each voice piece. Table II shows these features classified in four categories.

- The pitch related features describe the global pitch level change of the voice piece.
- The frequency and amplitude perturbation features reflect local period's pitch perturbation and local period's amplitude perturbation within one voice piece respectively. These two classes of features indicate the sound WAV form variation with respect to pitch and intensity.
- The spectrum related features are those defined on spectral analysis results and reflect the energy of sound along the frequency. For example, the hoarseness of the voice can be measured by HNR and NHR.

Second, we use the linear regression model to learn the quality evaluation function.

D. Voice Quality Prediction

The above trained linear regression model can be used for computing the voice quality of a new recorded VRP. We first split the testing sound file into voice pieces as what we did for the training phase. Each voice piece is mapped to a vocal point vp. Meanwhile, the voice piece is fed into the regression model to obtain a voice quality value. Note that there could be multiple voice pieces being mapped to the same vocal point. In such case, the multiple predicted values will be averaged to give the final voice quality value for vp.

E. Singer Profile Analysis

A singer profile SP computed from the above method consists a list of tuples $t =< vp, vq >$, where each vp indicates a vocal point, vq indicates its respective voice quality. Suppose

the pitch range of SP is PR, we can perform a simple profile partitioning algorithm described as following: (1) First, the vocal points whose $vq > \theta$ are marked as *good* points and those whose $vq \leq \theta$ are marked as *bad* ones. θ is an empirically determined threshold for the voice quality evaluation. (2) Second, we look at all good points for a pitch $pt \in PR$. The one with the maximum intensity is denoted by vp_{max}, and the one with the minimum intensity is denoted by vp_{min}. Then, vocal points on pt whose intensity lie between the maximum and minimum are all marked as good ones. It is easy to see that the rest vocal points on pt are all bad ones.

The output of the above partitioning algorithm will be used to derive some characteristics of a singer profile. These characteristics are important for understanding the singer's competence and learning the recommendation function in Section VI.

We first define the controllable and uncontrollable areas for a singer profile.

Definition 3: Controllable Area and Uncontrollable Area: The controllable area of a singer profile is the VRP region comprised of all good vocal points; while the uncontrollable area is the region made up of all bad vocal points.

This definition is consistent with the fact that a singer performs good quality when the vocal point is under her/his control. A typical controllable area is a continuous region inside the VRP. This is reasonable because the voice quality produced by human vocal cords is continuous. The boundary vocal points in VRP are always voiced in one's extreme condition (e.g. highest possible pitch, strongest possible intensity), and therefore uncontrollable.

The controllable area deserves particular attention. When we look at the few leftmost or rightmost pitches of the controllable area, we find that these "pitch edges" have strong implication for singing performance. Many people feel uneasy when singing notes in these edges, as they feel themselves to be close to extreme voicing positions. However, they can actually finish a performance successfully if the song is retained within the controllable boundary. Therefore, we shall further split the controllable area into two, namely the *challenging area* and *well-performed area*.

Definition 4: Challenging Area and Well-Performed Area: Given a singer profile, the *challenging area* is a subset of the controllable area, whose vocal points lie on either the β leftmost semitones or the β rightmost semitones of the controllable area, where β is an empirical number. The *well-performed area* is defined as the complement of the challenging area in the controllable area, or *(controllable area–challenging area)*.

In our implementation, $\beta = 4$. Fig. 4 shows a schematic diagram of the defined areas. The challenging area indicates the "boundary pitches" which could be challenging but manageable for the singer. In contrast, the well-performed area contains vocal points which even an untrained singer would confidently produce.

V. Song Profiles

In our solution to competence-based song recommendation, the pitch and intensity information of voices made by each singer is taken as input to generate a singer profile. Similarly, we need to build song profiles that contain singing pitch and intensity information in order to retrieve suitable singing songs for the singer. In this section, we first present the model for song profile and then describe the song profile acquisition process.

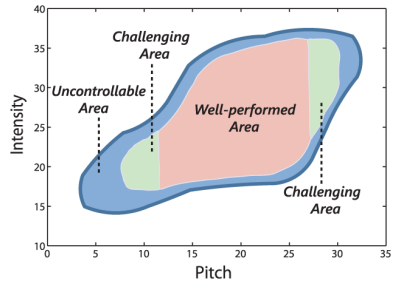


Fig. 4. Singer profile partitioning.

A. Song Profile Modeling

In our model, each song in the database contains a list of *notes*. Each note is a tuple in the form of $< pitch, duration, intensity >$, where $duration$ indicates the temporal length of the note, $intensity$ is the singing intensity of the note. Each $(pitch, intensity)$ pair defines a $term$. In other words, notes with the same (pitch, intensity) pair are regarded as having the same $term$. For each song, we count the numbers of occurrences and aggregate the durations by terms. This results in the following definition of song profile:

Definition 5: Song Profile: Song profile is a list of term-related quadruples as $< term\_pitch, term\_intensity, term\_freq, agg\_duration >$, where $term\_freq$ is the number of occurrences of the term and $agg\_duration$ is the aggregated (sum) duration of the term.

It should be noted that each term actually determines a (pitch, intensity) pair. Therefore, the song recommendation problem is transformed to that of matching the singer profile to the set of terms.

B. Song Profile Acquisition

Obtaining the profile of a song mainly involves two steps: (i) to acquire the singing melody and then (ii) to obtain the singing intensity for each note. As state-of-the-art techniques in music transcription cannot accurately extract the singing melody from a polyphonic song, we choose to rely on the MIDI databases available online. A typical MIDI file contains not only the singing melody but also its accompaniment. Most melodies in MIDI files are not on the same tune with the ground-truth music scores. We perform a cleaning procedure to extract only the singing melody from a MIDI file. Then we compare some pitch characteristics (e.g. lowest/highest pitch, starting pitch etc.) of the melody against ground-truth numerical musical notation to diminish the differences in their tunes.

The singing intensity data has to be annotated manually by professionals. Each expert listens to the original song and annotates a piecewise intensity sequence using the graphical interface provided by the Cubase 5 software. The software allows one to easily annotate the intensity sequence by drawing a few lines aside the notes. Given a song melody with a note sequence in the form of $< pitch_1, duration_1 >$,

$< pitch_2, duration_2 >, < pitch_3, duration_3 >, < pitch_4, duration_4 >, \ldots$, its respective piecewise intensity sequence is $\{< intensity_1, num_1 >, < intensity_2, num_2 >, \ldots, < intensity_n, num_n >\}$, where $num_i \geq 1$ indicates the number of notes that each piece of intensity covers. These intensity values are stored in the "velocity" attribute of the MIDI file and can be extracted later for constructing the song profile. The intensity values annotated by multiple experts can be averaged to give the final intensity value. Due to the simplicity of the process, the labor cost of the offline manual annotation in song profile acquisition is limited.

VI. Competence-Based Song Ranking

We apply *Listnet*, a listwise learning-to-rank approach, to learn our *competence-based song ranker*. In this section, we first present the Listnet-based learning method. Then we describe the features to be used in learning.

A. Listwise Approach

In the song ranking problem we treat a singer profile as a query, and song profiles as documents. Our aim is to learn a ranking function f which takes feature vector $\mathbf{X}$ defined on each $<$ singer profile, song profile $>$ pair as input and ω as parameter, and produces ranking scores of the songs. The target can be written in the form

$$y = f(X, \omega) \tag{1}$$

The goal of the learning task is to find a function f^* that minimizes the following loss function:

$$f^* = \arg\min_f \sum_{i=1}^{N} L(f(X_i, \omega), y_i) \tag{2}$$

where N is the number of singer profiles in the training set, y_i is the human annotated relevance scores for each song profile with the i-th singer profile, X_i is the feature vector for the i-th singer profile.

We decide to learn the target function employing a listwise approach. In a listwise approach, the feature vector is extracted from all possible pairs (cross-product) of singer profiles and song profiles. In addition, each feature vector is annotated with a human relevance judgement. The feature vector and its corresponding relevance annotation are considered as a learning instance in the loss function. Compared to pointwise or pairwise approaches, the listwise approach acquires higher ranking accuracy in the top ranked results according to [15], as the latter minimizes the loss of the ranking list directly.

In our solution, we employ the Listnet as the learning method. It maps each possible list of scores to a probability permutation distribution and uses the *cross entropy* between these probability distributions as the metric. Thus, the loss function is given by

$$L(y^{(i)}, z^{(i)}(f_\omega)) = -\sum_{\forall g \in \ell} P_{y^{(i)}}(g) log(P_{z^{(i)}}(g)) \tag{3}$$

where $z^{(i)} = (f_\omega(x_1^{(i)}), .., f_\omega(x_{n^{(i)}}^{(i)}))$; $f_\omega(\cdot)$ is the ranking function, and $x_j^{(i)}$ is the feature vector extracted from the i-th singer and the j-th song ($1 \leq j \leq n^{(i)}$ where $n^{(i)}$ is the number of songs relevant to the i-th singer); $y^{(i)} = (y_1^{(i)}, \ldots, y_{n^{(i)}}^{(i)}))$ is the corresponding human annotated relevance score vector, where $y_j^{(i)}$ is the score of the j-th song for the i-th singer; ℓ indicates all possible permutations of relevant songs for i-th singer; P is the permutation probability distribution given by

$$P_{z^{(i)}(f_\omega)}(\ell(j_1, j_2, \ldots, j_{n^{(i)}})) = \prod_{t=1}^{n^{(i)}} \frac{exp(f_\omega(x_{j_t}^{(i)}))}{\sum_{k=t}^{n^{(i)}} exp(f_\omega(x_{j_k}^{(i)}))} \tag{4}$$

We use linear neural network as the ranking function f_ω. Parameter ω is calculated using *gradient descent*.

B. Competence Feature Extraction

Now we shall describe the ranking features [i.e., components of $x_j^{(i)}$ in (3)], which are extracted from each $<$ singer profile, song profile $>$ pair. Specifically, these features capture a song's *term* distribution on various characteristic areas of a singer profile. (See Section V-A for definition of $term$.) As discussed in Section IV-C, each singer profile can be partitioned in 2D into three areas known as the *uncontrollable area*, the *challenging area* and the *well-performed area*. In addition, we can define the 2D area outside the VRP as the *silent area*.

Given a $<$ singer profile, song profile $>$ pair, for any area A in the singer profile, suppose $\{term_1, term_2, \ldots, term_n\}$ are the song terms appearing in A, and their $term\_freq$ and $agg\_duration$ in A are denoted by $\{tf_1, tf_2, \ldots, tf_n\}$ and $\{dur_1, dur_2, \ldots, dur_n\}$ respectively, then the features on this area are defined as follows.

1) *Total TF:* This feature is defined as $\sum_{i=1}^{n} tf_i$.
2) *Total TF-IDF:* Analogous to terms in documents, song terms widely available in different song profiles are less important in distinguishing different songs. For those terms with high/low pitch or loud/soft intensity, they are more important in representing the uniqueness of the song. Thus we compute the TF-IDF value of all terms in the song profile database. If we denote the TF-IDF of $term_i$ in the current song by $tfidf_i$, then the Total TF-IDF of area A is defined as $\sum_{i=1}^{n} tfidf_i$.
3) *Total TF-IVQ (Inverse Voice Quality):* The voice quality of different areas are different. If many song terms are located in the uncontrollable or silent areas, it most probably will be a disaster for the singer to sing that song. Thus, we incorporate the voice quality into the feature definition. The voice quality is firstly averaged on the entire area of A and then inverted (as lower value indicates higher quality). Therefore, the Total TF-IVQ is defined as $\sum_{i=1}^{n} tf_i/avq$, where avq is the *average voice quality* in area A.
4) *Total Duration:* Duration is an important factor affecting the singing performance, especially for the challenging area. Singing a term for a long time in challenging or uncontrollable areas is apparently difficult. Thus, we define the Total Duration as $\sum_{i=1}^{n} dur_i$.
5) *Total TF-IDF Duration:* The duration of each term is also affected by the term importance. The effect of the duration of less important terms should be decreased. So we define this feature as $\sum_{i=1}^{n} dur_i \cdot tfidf_i$.

TABLE III
RANKING FEATURES *(C-Area: Challenging Area; W-Area: Well-Performed Area; U-Area: Uncontrollable Area; S-Area: Silent Area)*

Features	*C-area*	*W-area*	*U-area*	*S-area*
Total TF	√	√	√	√
Total TF-IDF	√	√	√	√
Total TF-IVQ	√	√	√	
Total Duration	√	√	√	√
Total TF-IDF Duration	√	√	√	√
Total Duration-IVQ	√	√	√	

6) *Total Duration-IVQ:* The effect of the duration of each term is also affected by the voice quality in the area. Therefore we define the Total Duration-IVQ as $\sum_{i=1}^{n} dur_i/avq$.

The above six features are defined in all four areas, except the two voice quality-related ones (Total TF-IVQ and Total Duration-IVQ) for the silent area. These two are undefined as their voice quality is unavailable. Table III shows all the defined 22 features for each area.

VII. REDUCED SINGER PROFILE

Because singer profile models all the vocal points one can produce, it requires each subject to sing an average of 23 pitches. For an untrained singer, recording each pitch is expensive and time-consuming. The singing teacher has to sing the pitch for many times in order to help the subject find the right pitch to sing. In this section, we introduce a reduced singer profile to model user's vocal competence. This model requires lesser human recordings while not losing much recommendation accuracy. The reduced singer profile is a simplified version of singer profile by ignoring less important singer profile areas. This section presents our method for constructing the reduced singer profile.

A. Singer Profile Area Importance

In the Listnet ranking, the recommendation result is affected by the features (see Table III) derived from the notes distribution over different singer profile areas. We analyse the importance of a singer profile area by studying the importance of the features in a singer profile area.

First of all, we will introduce how to determine the importance of a feature. Suppose $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$ is a feature vector extracted from a user's singer profile and a song's song profile where $\mathbf{X_i}$ is a variable represents a competence feature. $\mathbf{Y}$ is a variable represents the rating indicating whether the song is fit for the user. We estimate a competence feature's importance by measuring the correlation between the feature $\mathbf{X_i}$ and the rating $\mathbf{Y}$. We use *information gain*[34] to measure the correlation between the two variables. We use information entropy to measure the uncertainty of a random variable. The information gain measures the decrease of a variable's entropy knowing the value of another variable. The entropy of $\mathbf{X_i}$ is defined as

$$H(\mathbf{X_i}) = -\sum_{k} P(X_i^k) log_2(P(X_i^k)) \tag{5}$$

where X_i^k is a value of $\mathbf{X_i}$, $P(X_i^k)$ is a prior probability for the value of $\mathbf{X_i}$. If we observe the value of variable Y, the entropy of $\mathbf{X_i}$ is defined as

$$H(\mathbf{X_i}|\mathbf{Y}) = -\sum_{j} P(Y^j) \sum_{k} P(X_i^k|Y^j) \log_2(P(X_i^k|Y^j)) \tag{6}$$

where $P(X_i^k|Y^j)$ is the posterior probability of $\mathbf{X_i}$ given the value $\mathbf{Y}$. The decrease of $\mathbf{X_i}$'s entropy knowing the value of $\mathbf{Y}$ is defined as information gain.

$$IG(\mathbf{X_i}|\mathbf{Y}) = H(\mathbf{X_i}) - H(\mathbf{X_i}|\mathbf{Y}) \tag{7}$$

Therefore, if $IG(\mathbf{X_i}|\mathbf{Y})$ is bigger than $IG(\mathbf{X_j}|\mathbf{Y})$, this means $\mathbf{X_i}$ is more correlated with $\mathbf{Y}$ than $\mathbf{X_j}$. It reflects feature $\mathbf{X_i}$ is more important than feature $\mathbf{X_j}$.

Now comes the estimation of the importance of a singer profile area. Because each singer profile area has six features, each feature has a correlation value with the rating. We measure the importance of a singer profile area by averaging the correlation values of the six features in that singer profile area. The bigger the mean correlation value is, the more important the singer profile area will be in recommendation.

Knowing the importance of the singer profile area in recommendation, we can simplify the singer profile model by only keeping the most important part. The analysis result in the experiment part (Section VIII-D3) shows the uncontrollable area is the most important within the three singer profile area. We will model the reduced singer profile only on the uncontrollable area.

B. Reduced Singer Profile Modeling

In this section, we define the reduced singer profile. Reduced singer profile ignores the controllable area in a singer profile. The reduction is based on the fact that the voice quality function is convex. This is determined by the physical structure of our vocal cords. People can vocalise continuous pitches within a fixed range of intensity. The voice quality in each singer profile area is similar. First of all, we define the reduced VRP.

Definition 6: Reduced VRP: Given a VRP, its reduced VRP is made up by the pitches whose vocal points are all located in the uncontrollable area.

Practically, these pitches are located in the leftmost and the rightmost part of the VRP area. In the reduced VRP recording, we only need to traverse the pitches in one's singing limitation. Because the number of pitches to be recorded varies from people, we set to record the η leftmost and τ rightmost pitches(semitones) empirically so as to cover the uncontrollable area of majority users according to the analysis of the current VRPs. Here $\eta = 6$ and $\tau = 3$

Now we can define the reduced version of singer profile.

Definition 7: Reduced Singer Profile A reduced singer profile is a set of 2-tuples, written as $< vp, \psi(vp) >$, where $vp \in ReducedVRP$ is a vocal point that the singer can voice.

C. Reduced Singer Profile Generation

Comparing with the singer profile generation, the difference of the reduced singer profile generation lies in the VRP

recording. We perform a reduced VRP recording strategy to reduce the number of pitches to record. The generation process can be divided into reduced VRP recording and model generation.

1) Reduced VRP Recording: In VRP recording, each pitch is an indivisible recording task. Once the singing teacher gives a pitch, the subject will voice the pitch from the softest to the loudest. For reduced VRP recording, we only need to record pitches near one's singing limitation. The biggest challenge is how to find subjects' pitch limitation in the low and high register. We will describe the VRP recording strategy for the low register in detail, the high register will be similar.

There is a straightforward way of recording, we called it *Naive search Recording Strategy* (NRS). Firstly, we find an empirical pitch that most people's lowest pitches will not higher than it. The subject will sing each pitch down until reaching the lowest pitch. However, people's lowest pitches vary from each other. By applying this strategy, some subjects with very low pitch boundaries in their low register sing too many pitches for generating the uncontrollable area, while some subjects' recording pitches are not enough.

Here we introduce a *Binary search Recording Strategy* (BRS) which firstly locates the singing pitch boundary of the subjects and then records η and τ pitches approaching the two pitch boundaries respectively. Suppose P is a list stores all the pitches in music score with increasing order in frequency. We set an empirical pitch range from $P[min]$ to $P[max]$ which most people's lowest pitches will locate in this pitch range.

The lowest pitch localization is described as follows. Begin from the pitch range $P[min]$ to $P[max]$, the subject is asked to sing the pitch in the middle of the range. If the subject can sing the middle one, check the pitch range from $P[min]$ to the middle one using the same strategy. If the subject cannot sing the middle one, check the pitch range from the middle one to $P[max]$ using the same strategy. Iterating until range's begin index and end index satisfy $begin+1 = end$. $P[end]$ is subject's lowest pitch. After finding one's lowest pitch, the subject is asked to sing η pitches higher than the lowest pitch if they were not sung during the localization process.

The recording process for high register is similar. The above recording process can effectively reduce the number of pitches to sing. Many pitches that belong to the uncontrollable area will also be sung in the lowest/highest pitch localization.

2) Model Generation: The reduced singer profile generation is quite similar to the singer profile generation. Firstly, the recording samples will be cut into voice pieces and generate the reduced VRP. Then we use the same voice quality evaluation function learned in Section IV-B2 to calculate the voice quality of each vocal point. Because there is only uncontrollable area in the reduced singer profile, there is no need for singer profile partition.

VIII. Experiments

In this section, we report the experiment setup and results. We first introduce the datasets being used in the experiments. Then we describe the baseline methods which we compare with. We also introduce the metrics which guide the evaluation of the results. Finally, the experimental results are presented and analyzed.

A. The Datasets

1) Singer Profile Dataset: For VRP recording, we recruited 90 volunteers including 45 males (mean age = 25) and 45 females (mean age = 21), with ages varying from 18 to 54. Each singer's VRP is recorded using Audition V3.0. We choose Rode M3 as the recording microphone and M-AUDIO MobilePre USB as the audio card. Before recording, each singer is requested to climb the music scale to "warm-up" their voice. During the recording, a vocal music teacher helps the singers locate their pitch and guide the singer to adjust the singing intensity.

In order to build training dataset for the quality evaluation function, three experienced singing teachers (with 20+ years' experience) are invited to evaluate the voice quality of the recording and annotate different parts of the WAV files using Praat. We provide part recording files of the subjects (20 females and 35 males) to the teachers for voice quality annotation. These files are then split into 6498 female and 17144 male voice pieces with human annotated voice qualities as the training data for *two* quality evaluation functions, one for women and the other for men.

2) Song Profile Dataset: We have collected 200 songs (100 for male, 100 for female) as the training dataset. All singing melodies are calibrated according to their original music scores, and the singing intensity values are annotated by the singing teachers.

3) Ranking Dataset: In order to train the Listnet for song recommendation, we need a ranking dataset which contains manually annotated relevance scores for each $<$ singer profile, song profile $>$ pair.

For building the ranking dataset, we divided the 100 male and female midi songs into 5 subsets respectively. The songs in each subset cover different pitch range and intensities to avoid data skew. We divide the 45 male subjects into 5 groups for 5-fold cross validation, and ensure that their singer profiles are as equally distributed as possible. Each singer is asked to sing some part of the 20 songs in one of the 5 subsets, in front of the 3 singing teachers. Subsequently, the singer teachers choose 1 out of 5 relevance labels, namely *challenging*, *normal*, *easy*, *difficult*, *nightmare*. A total number of 900 singing performances will be scored for male and female respectively.

Our datasets are relatively small-scale due to resource constraints. However, we have observed sufficient variations among the singers and songs. Although adding new subjects and data will for sure improve the work, we believe that research on the current datasets can already lead to interesting findings.

B. Baseline Methods

We compare CBSR against two baseline methods.

1) Pitch Boundary Ranking Method (PB): PB ranking method is the most intuitive way of singing song recommendation–the one that we challenge in Section I. This method only uses singer's pitch range of good quality corresponding to the well-performed area in VRP. In this method, we regard each

TABLE IV
PEARSON CORRELATION

Dataset	Without PCA		With PCA	
	Mean	STD	Mean	STD
Male	0.7281	0.0654	0.729	0.0634
Female	0.5565	0.0643	0.5068	0.0717
Hybrid	0.7115	0.0373	0.7083	0.0432

vocal point to be a single dimensional point on the pitch-axis. This is equivalent to projecting the VRP onto the pitch-axis. The voice quality of each 1D vocal point is defined as the average of those 2D points on the same pitch. As a result, we can split the 1D pitch range to obtain controllable/uncontrollable areas, challenging area, and well-performed area. We also use the Listnet to train a ranking function. The ranking features are defined for notes within or outside the well-performed area on 1D pitch range. These features are *Total TF*, *Total TF-IDF*, *Total Duration* and *Total TF-IDF Duration*.

2) CBSR Using Reduced Singer Profile (CBSR-Reduced): CBSR-reduced ranking method uses reduced singer profile to model singer's vocal competence. This method only uses the uncontrollable area and silent area to do the recommendation. We use Listnet to train the ranking function. We use the 10 ranking features defined upon uncontrollable and silent area in CBSR as the features for CBSR-reduced.

C. Evaluation Metric

For the quality evaluation function, we use the Pearson Correlation Coefficient (ρ) as the metric measuring the distance between the human annotated voice quality score and the predicted voice quality. This metric evaluates the linear dependence between two variables.

For the competence-based song recommendation, we adopt the Normalized Discounted Cumulative Gain (NDCG) [33] as our metric for the ranking result. NDCG is for measuring the ranking accuracy which has more than two relevance levels.

D. Experimental Results

We first report the results of the voice quality computation. Next, we compare the ranking accuracy of our CBSR framework against the two baseline methods. Finally, the real recommendation results for singers are demonstrated.

1) Results of Voice Quality Computation: Note that voice quality is computed by learning the quality evaluation function. We learn the linear regression model on male-only (35 men), female-only (20 women) and hybrid (55 people) datasets. Each dataset is randomly split into 5 parts, and then go through 5-fold cross validation. In each trial, four folds are used for training and one remaining fold for testing. We apply *principle component analysis* (PCA) to conduct feature selection before learning and testing. The Pearson correlations of the predicted voice quality and human-annotated voice quality are illustrated in Table IV. The Mean and STD are the average and the standard deviation of the Pearson correlation value calculated from the five trials.

The above result shows large correlation of the predicted voice quality and human annotated voice quality. The male dataset achieves 0.7281 and the hybrid one gives 0.7115. However, the correlation value of Female is lower (0.5565). This is most probably due to the shortage of the female training data. The second finding is that PCA does not improve the voice quality prediction.

2) Singer Profile Demonstration: After learning the quality evaluation function, we are able to generate the singer profile for each subject. Fig. 5 demonstrates six subjects' singer profiles (3 male and 3 female), with the color of each vocal point showing its voice quality. These singer profiles clearly illustrate the different vocal competences of the subjects.

The profiles demonstrate strong correlation between pitch and intensity. With the increase of the pitch, the intensity also becomes higher. The only exception is Fig. 5(f) where the intensity does not increase by pitch in the right part of singer profile. This is because the subject changes from the modal register to the falsetto register (false voice). As an untrained singer, she cannot produce very loud voices in false voice. Fig. 5(a) shows a bass who can perform the low pitch with a rich voice.

The voice quality of these profiles indicate that lower pitch or intensity are more likely to be of bad quality, while high intensity may lead to better quality. This is because in VRP recording, many subjects tend to produce soft voice, no matter whether the voice quality is good or not. When they produce louder voice, some of the subjects are likely to stop voicing when reaching their uncontrollable areas.

Fig. 5 also show clear indication of areas. The dark green and blue pixels indicate the uncontrollable area, while the light green to the yellow ones indicate the challenging area for the singer. The different areas show obvious aggregation of vocal points with similar colors, thus confirming the effectiveness of our singer profile partitioning method.

3) Area Importance Analysis: Table V demonstrates the importance of the three singer profile areas and the silent area with the human rating. From the mean correlation of each area, we can see that the features defined on the silent area which is the area outside one's singer profile acquire the highest correlation with the relevance rating. This finding reveals that the number of notes in silent area is an effective indicator for competence based relevance judgement. This is because if there are many notes in the silent area (area outside subject's VRP), the song will be hard for the singer to sing. For the three singer profile areas, we find the uncontrollable area is more important than well-performed area and challenging area. This is because songs with many notes located in the uncontrollable area will also be hard to perform well. The above finding provides some evidence for why we define the reduced singer profile using the uncontrollable area.

4) Binary Search Recording Strategy: Because we have all the subjects' complete VRP data, we can simulate the reduced VRP recording process using binary search recording strategy and count the number of each subject's pitches (semitones) to record.

According to our data, Table VI shows the range for most subjects' singing pitch limitations. For example, male's lowest singing pitch will be located in the range from 73.4 Hz to 155.5 Hz.

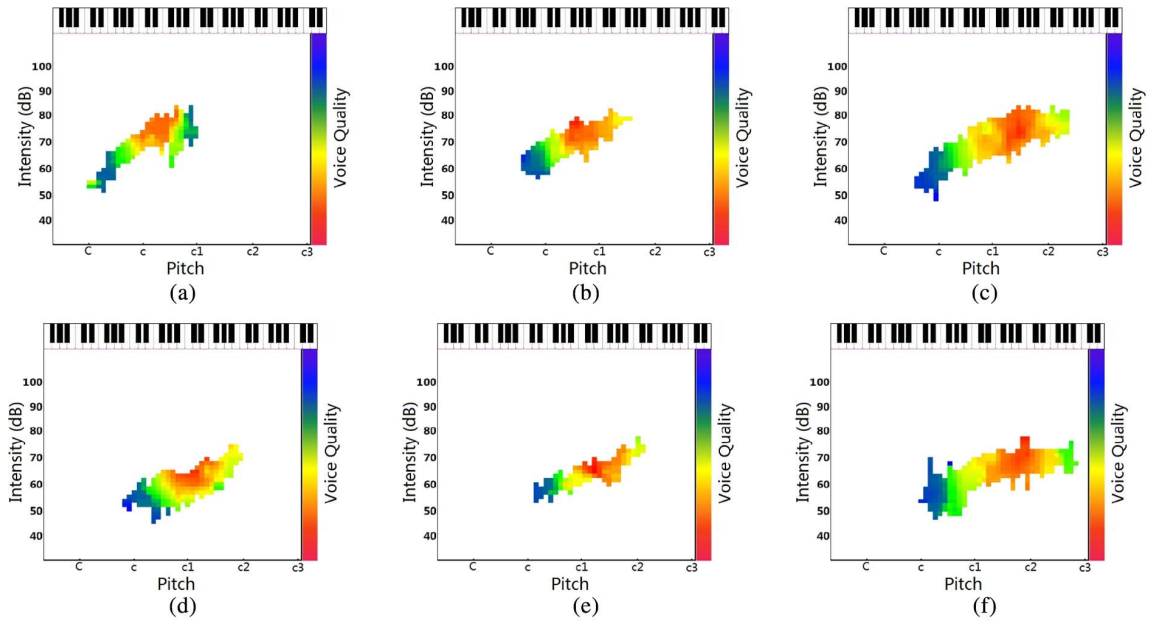


Fig. 5. Singer profiles of subjects. (a) Male-bass. (b) Male-baritone. (c) Male-tenor. (d) Female-bass. (e) Female-baritone. (f) Female-tenor.

TABLE V
SINGER PROFILE AREA IMPORTANCE

Features	*S-area*	*U-area*	*C-area*	*W-area*
Total TF	0.132	0.0298	0.0195	0.0183
Total TF-IDF	0.103	0.0334	0.0216	0.0198
Total TF-IVQ		0.0338	0.0166	0.0172
Total Duration	0.131	0.0267	0.0148	0.0177
Total TF-IDF Duration	0.103	0.0334	0.0216	0.0198
Total Duration-IVQ		0.0304	0.0181	0.0166
Mean Correlation	**0.117**	**0.0313**	**0.0187**	**0.0182**

TABLE VI
PITCH LIMITATION

Gender	**Low Register**		**High Register**	
	Min(Hz)	**Max(Hz)**	**Min(Hz)**	**Max(Hz)**
Male	73.4	155.5	261.6	587.4
Female	123.4	261.6	392.0	988.2

TABLE VII
NUMBER OF PITCHES TO RECORD

Gender	**Low Register**			**High Register**		
	BRS	**NRS**	**Locate**	**BRS**	**NRS**	**Locate**
Male	7.02	8.33	3.71	4.18	7.64	3.82
Female	7.05	8.10	3.70	4.35	10.10	4.00

Then we apply the binary search recording strategy during the reduced VRP recording for each subject. We compare the binary search recording strategy (BSR) with the naive search recording strategy (NRS) described in Section VII-C on the number of recording pitches. Table VII is the mean number of pitches requires to record for each subject. For the 6 pitch recording tasks in the low register, BRS requires to record a mean of 7.02 pitches and 7.05 pitches for male and female respectively. NRS requires to record a mean of 8.3 and 8.1 pitches for male and female respectively. For the 3 pitch recording tasks in the high register, the advantage is more obvious. BRS requires to record 3.8 and 4 pitches while NRS requires 7.6 and 10.1 pitches for male and female respectively. The column "Locate" in Table VII represents the mean number of pitch needed to sing for locating the pitch boundary. The results proves the effective of the binary search recording strategy in reducing the workload of the reduced VRP recording.

We also count the mean number of pitches for getting each complete singer profile. The number of pitches is 23.04 and 23.35 for male and female respectively. For reduced VRP recording using BRS, we only need to record a mean of 11.2 and 11.4 pitces for each male and female's VRP respectively. This shows that the reduced VRP recording reduces 50% of the recording task comparing with the original VRP recording.

5) Ranking Accuracy: To study the ranking accuracy, we divide the male and female ranking datasets into five subsets for cross validation. In each trial, four subsets are used for training, and one for testing. The NDCG@n results reported are all averaged from the 5-fold cross validation.

Fig. 6 shows the ranking accuracy measured from NDCG@n on male and female ranking datasets. Apparently, CBSR outperforms the two baseline methods. CBSR outperforms PB by an average of 37% and 22% on male and female dataset respectively. This indicates the effect of the uncontrollable area and the voice quality which PB ignores in recommendation. CBSR-reduced which uses reduced singer profile is only 6% and 5% worse than CBSR for male and female respectively, while 50% of the recording task is saved. This shows the reduced singer profile is effective in modeling singer's vocal competence.

Fig. 7 is the correlations between the Listnet loss function and the measure of NDCG during CBSR's learning process on male and female ranking dataset. We can see the learning process converge after about 250 iterations. For CBSR's converge behavior

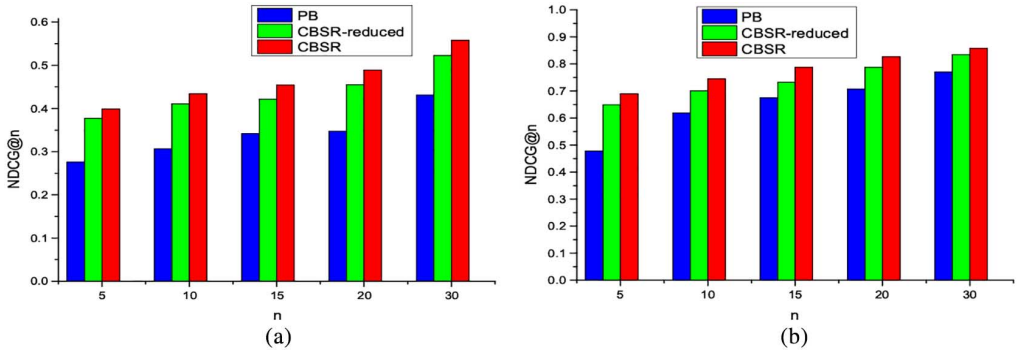


Fig. 6. Ranking accuracy in NDCG@n on male and female ranking datasets. (a) Male. (b) Female.

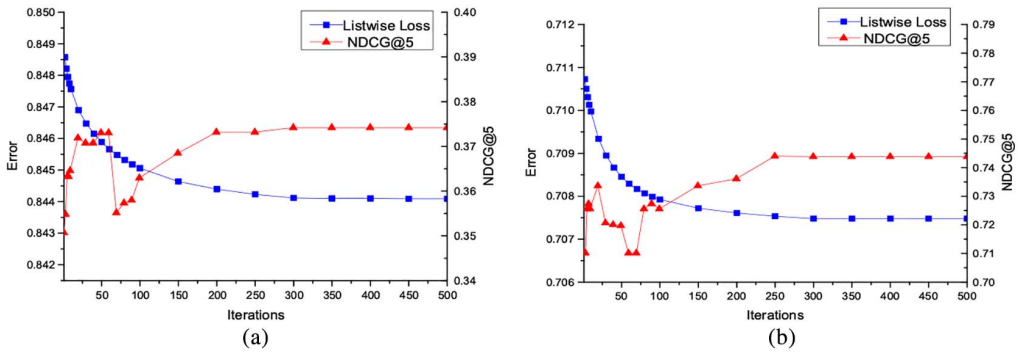


Fig. 7. Coverage behavior of the listnet on male and female ranking datasets.(a) Male (b) Female.

on both male and female dataset, we find the NDCG@5 first increases in the first 50 iterations and then decrease until 70 iterations. After that, NDCG@5 increases until the listwise loss of Listnet reaches its limit.

IX. Conclusion and Future Work

In this paper, we study the novel competence-based song recommendation problem. We modeled singer's vocal competence as singer profile which takes voice pitch, intensity, and quality into account. We proposed a supervised learning method to train voice quality evaluation function, so that voice quality could be computed at query time. A reduced version of singer profile is also proposed to reduced the recording task in competence modeling. We also proposed a song model, which enabled matching with the singers. The proposed models allowed us to build a learning-to-rank scheme for song recommendation relying on human-annotated ranking datasets. The experiments demonstrated the effectiveness of our approach and its advantages compared to two baseline methods.

For future work, we plan to study the differences of singer profile before and after vocal training. For trained singer, the controllable area will expand while the uncontrollable area will shrink. By analyzing the singer profile, we can recommend songs that the subjects can perform well after vocal training.

Acknowledgment

This research was carried out at the NUS-ZJU SeSaMe Centre.

References

[1] S. K. Wolf, D. Stanley, and W. J. Sette, "Quantitative studies on the singing voice," *J. Acoust. Soc. Amer.*, vol. 6, no. 4, pp. 255–266, 1935.
[2] A. M. Sulter, H. K. Schutte, and D. G. Miller, "Differences in phonetogram features between male, and female subjects with, and without vocal training," *J. Voice*, vol. 9, no. 4, pp. 363–377, 1995.
[3] Y. Zhu and M. S. Kankanhalli, "Precise pitch profile feature extraction from musical audio for key detection," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 575–584, Jun. 2006.
[4] L. Heylen, F. L. Wuyts, F. Mertens, M. D. Bodt, and P. H. V. d. Heyning, "Normative voice range profiles of male, and female professional voice users," *J. Voice*, vol. 16, no. 1, pp. 1–17, 2002.
[5] H. K. Schutte and W. Seidner, "Standardizing voice area measurement/phonetography," *Folia Phoniatr (Basel)*, vol. 35, no. 6, pp. 286–288, 1983.
[6] C. Watts, K. Barnes-Burroughs, J. Estis, and D. Blanton, "The singing power ratio as an objective measure of singing voice quality in untrained talented, and nontalented singers," *J. Voice*, vol. 20, no. 1, pp. 82–88, 2006.
[7] J. Shen, J. Shepherd, and A. H. H. Ngu, "Towards effective content-based music retrieval with multiple acoustic feature combination," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1179–1189, Dec. 2006.
[8] G. Peeters, IRCAM, Paris, France, "A large set of audio features for sound description," Tech. Rep., 2004.
[9] J. Peter and H. Pabon, "Objective acoustic voice-quality parameters in the computer phonetogram," *J. Voice*, vol. 5, no. 3, pp. 203–216, 1991.
[10] Y. Yu, R. Zimmermann, Y. Wang, and V. Oria, "Scalable content-based music retrieval using chord progression histogram, and tree-structure LSH," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1969–1981, Dec. 2013.
[11] L. Zhang, M. Song, N. Li, J. Bu, and C. Chen, "Feature selection for fast speech emotion recognition," in *Proc. ACM Multimedia*, 2009, pp. 753–756.
[12] Y. Maryn, P. Corthals, P. V. Cauwenberge, N. Roy, and M. D. Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech, and sustained vowels," *J. Voice*, vol. 24, no. 5, pp. 410–426, 2010.
[13] K. Hoashi, K. Matsumoto, and N. Inoue, "Personalization of user profiles for content-based music retrieval based on relevance feedback," in *Proc. ACM Multimedia*, 2003, pp. 110–119.

[14] D. Goldberg, D. A. Nichols, B. M. Oki, and D. B. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.
[15] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. ICML*, 2007, pp. 129–136.
[16] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Inf. Retrieval*. Reading, MA, USA: Addison-Wesley, 1999.
[17] L. Zhang, Y. Xia, K. Mao, and Z. Shan, "An effective video summarization framework toward handheld devices," *IEEE Trans. Ind. Electron.*, vol. 62, no. 2, pp. 1309–1316, Feb. 2015.
[18] D. Deliyski, "Acoustic model, and evaluation of pathological voice production," in *Proc. Eurospeech*, 1993, pp. 1969–1972.
[19] E. Yumoto, W. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Amer.*, vol. 71, no. 6, pp. 1544–1550, 1982.
[20] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: Musical information retrieval in an audio database," in *Proc. ACM Multimedia*, 1995, pp. 231–236.
[21] K. Mao, X. Luo, K. Chen, G. Chen, and L. Shou, "myDJ: Recommending karaoke songs from one's own voice," in *Proc. SIGIR*, 2012, p. 1009.
[22] L. Shou, K. Mao, X. Luo, K. Chen, G. Chen, and T. Hu, "Competence-based song recommendation," in *Proc. SIGIR*, 2013, pp. 423–432.
[23] K. Mao, J. Fan, L. Shou, G. Chen, and M. S. Kankanhalli, "Song recommendation for social singing community," in *ACM Multimedia*, 2014, pp. 127–136.
[24] P. Boersma and D. Weenink, Univ. of Amsterdam. Amsterdam, The Netherlands, "Voice," Oct. 2006 [Online]. Available: http://www.fon.hum.uva.nl/praat/manual/Voice.html
[25] L. Zhang, Y. Gao, Y. Xia, Q. Dai, and X. Li, "A fine-grained image categorization system by cellet-encoded spatial pyramid modeling," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 564–571, Jan. 2015.
[26] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Breaking the glass ceiling object recognition and segmentation," in *Proc. ISMIR*, 2012, pp. 379–384.
[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer, " ver. 5.3.06, Accessed: May 1, 2007.
[28] J. P. Pabon and R. Plomp, "Automatic phonetogram recording supplemented with acoustical voice-quality parameters," *J. Speech Hearing Res.*, vol. 31, no. 4, pp. 710–722, 1988.
[29] L. G. Heylen, F. L. Wuyts, F. W. Mertens, and J. E. Pattyn, "Phonetography in voice diagnoses," *Acta Oto-Rhino-Laryngologica*, vol. 50, no. 4, pp. 299–308, 1996.
[30] R. Speyer, G. H. Wieneke, I. v. Wijck-Warnaar, and P. H. Dejonckere, "Efficacy of voice therapy assessed with the voice range profile (phonetogram)," *J. Voice*, vol. 17, no. 4, pp. 544–559, 2003.
[31] B. Schneider, M. Zumtobel, W. Prettenhofer, B. Aichstill, and W. Jocher, "Normative voice range profiles in vocally trained, and untrained children aged between 7, and 10 years," *J. Voice*, vol. 24, no. 2, pp. 153–160, 2010.
[32] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1233–1243, May 2012.
[33] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proc. SIGIR*, 2000, pp. 41–48.
[34] J. Ross Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

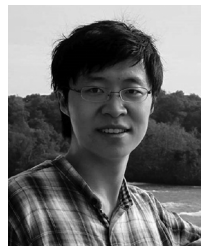
Kuang Mao is currently working toward the Ph.D. degree at the College of Computer Science, Zhejiang University, Hangzhou, China.

From 2013 to 2014, he was a Research Intern with the SESAME Group at the National University of Singapore, Singapore. His research area includes recommendation systems, singing song recommendation, graph ranking algorithms, and probabilistic modeling.

Lidan Shou received the Ph.D. degree in computer science from the National University of Singapore, Singapore.

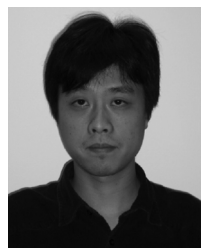
He is currently a Professor with the College of Computer Science, Zhejiang University, Hangzhou, China. Prior to joining the faculty, he had worked in the software industry for over two years. His research interests include spatial database, data access methods, visual and multimedia databases, and web data mining.

Dr. Shou is a member of the ACM.



Ju Fan received the B.Eng. degree in computer science from the Beijing University of Technology, Beijing, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Haidian, China, in 2012.

He is currently a Research Fellow with the School of Computing, National University of Singapore, Singapore. His research interest includes crowdsourcing-powered data analytics, spatial-textual data processing, and database usability.



Gang Chen received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China.

He is a Professor with the College of Computer Science and the Director of the Database Lab, Zhejiang University, Hangzhou, China. He has successfully led investigations in research projects that aim at building China's indigenous database management systems. His research interests range from relational database systems to large-scale data management technologies supporting massive Internet users.

Dr. Chen is a member of the ACM and a senior member of China Computer Federation.

Mohan S. Kankanhalli (M'92–SM'09–F'14) received the B.Tech. degree from IIT Kharagpur, Kharagpur, India, and the M.S. and Ph.D. degrees from the Rensselaer Polytechnic Institute, Troy, NY, USA.

He first joined the Institute of Systems Science, National University of Singapore (NUS), Singapore, in 1998 as a Researcher. He then became a Faculty Member of the Department of Electrical Engineering, Indian Institute of Science, Bangalore, India. He was the Vice Dean of Academic Affairs and Graduate Studies at the School of Computing, NUS, from 2008 to 2010, and Vice Dean of Research from 2001 to 2007. He is currently a Professor with the Department of Computer Science, NUS. He is also the Associate Provost for Graduate Education at the NUS. His current research interests include multimedia systems (content processing and retrieval) and multimedia security (surveillance and privacy).

Dr. Kankanhalli is actively involved in organizing of many major conferences in the area of multimedia. He is on the editorial boards of several journals including the *ACM Transactions on Multimedia Computing, Communications, and Applications*, the *Springer Multimedia Systems Journal*, the *Pattern Recognition Journal*, and the *Multimedia Tools and Applications Journal*. He has been recently awarded a large grant by Singapore's National Research Foundation to set up the Centre for Sensor-Enhanced Social Media, Singapore.