Competence-Based Song Recommendation

Lidan Shou Kuang Mao Xinyuan Luo Ke Chen Gang Chen Tianlei Hu
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
{should, mbill, wisp, chenk, cg, htl}@zju.edu.cn

ABSTRACT

Singing is a popular social activity and a good 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 singing competence-based song recommendation framework. 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 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. Finally, we propose a learning-to-rank scheme for recommending songs by singer profile. The experimental study on real singers demonstrates the effectiveness of our approach and its advantages over two baseline methods. To the best of our knowledge, our work is the first to study competence-based song recommendation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.5.5 [**Sound and Music Computing**]: Modeling

Keywords

Song Recommendation, Singing Competence, Learning-to-rank

1. INTRODUCTION

Singing is a popular social activity and a good way of expressing one's feelings. While some people enjoy the experience of conducting a wonderful solo in a karaoke party, many others are upset by their own singing skill due to an unpleasant performance in the past. The truth is, people should blame their own skill of choosing songs rather than singing. It is extremely hard for a girl with a soft voice to sing like Mariah Carey whose songs require loud 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.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

However, song recommendation for singers seemed 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 [7], 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 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[3, 13], we attempt to discover the

voice characteristics of each song, which in turn pose different requirements 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 break the ice by proposing a novel singing competence model which is instantiated as *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[10], voice treatment[21] and vocal training [19]. 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.

The above competence model (singer profile) causes 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 apparently 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 *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 unstudied 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.

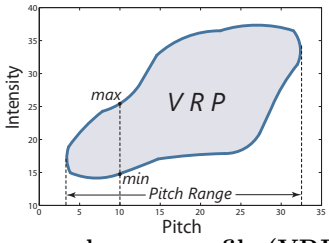


Figure 1: The vocal range profile (VRP) of a singer

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 also present the song profile and describe the method of generating the respective song profile from a database.

(4) The song recommendation is implemented using a multiple criteria learning-to-rank scheme.

(5) Our experiments on a group of users show promising results of the proposed framework.

The rest of our work is organized as follows: Section 2 introduces the related work. Section 3 conducts an overview of the framework. Section 4 and 5 presents the singer/song profile models and the techniques to acquire these profiles. Section 6 describes the learning-to-rank recommendation scheme. The experiments are detailed in Section 7. Finally, Section 8 concludes the paper.

2. 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.

2.1 Vocal Range Profile

As shown in Figure 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. [25] in 1935. Since then, VRP has been widely applied in objective clinical voice diagnosis and singer's vocal training. Many papers [9, 22] 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 [20]. 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.

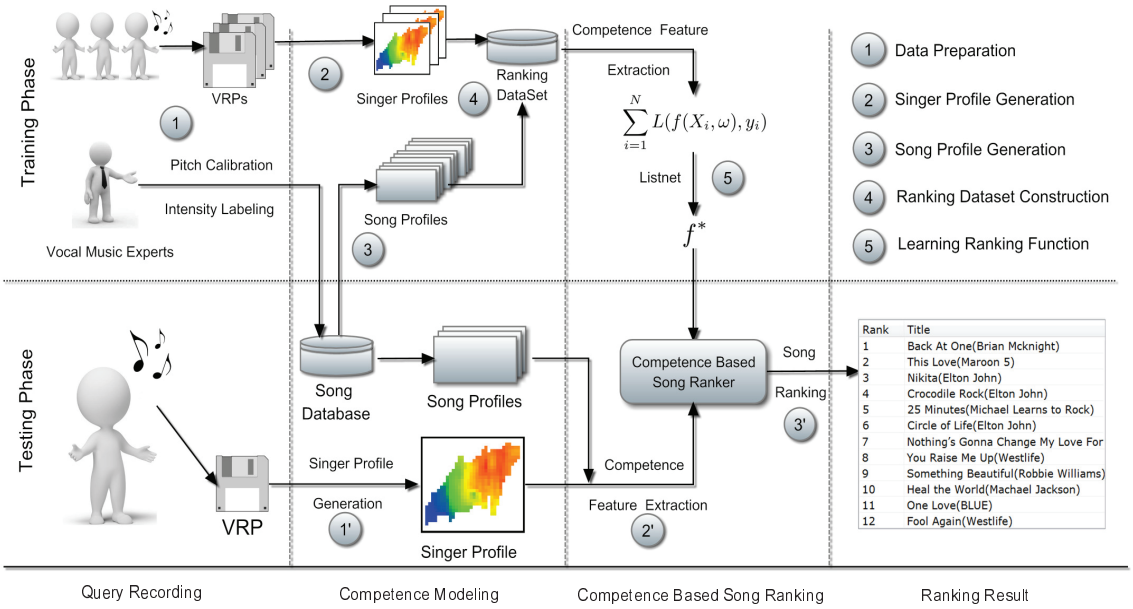


Figure 2: Overview of the competence-based song recommendation framework

2.2 Voice Quality

The technique of *objective voice quality measuring* 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[24] is defined upon the spectrum analyze of voice samples. This measure differs a lot between trained and untrained singers. The other similar examples include tilt[6], and ltasSlope[17]. The last two are meant to discover the singer's singing talent[24]. The above mentioned measures reveal many characteristics of the voice. However, these measures cannot adequately solve our problem, which requires 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. [18] 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 [16] 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 [15], 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 from this work. But our underlying problem and expert knowledge of singing voice quality is totally different from the previous study.

2.3 Song Recommendation

Traditional song/music recommendation focuses on recommending songs by user's listening interests. The earlier studies such as [11] 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[8] 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[23] which requires singing to be performed in the first place. In our preliminary study[14], we demonstrated a system for karaoke song recommendation. This work significantly extends it by formulating the scientific problem of competence-based song recommendation and proposing a novel solution.

3. OVERVIEW OF CBSR FRAMEWORK

As Figure 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.
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.

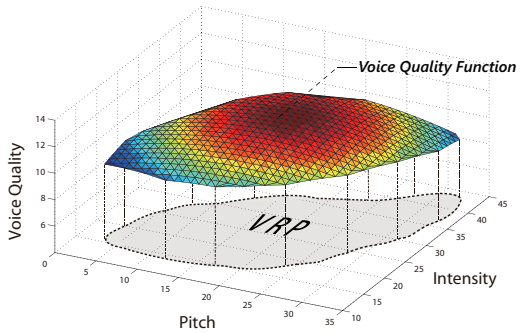


Figure 3: A Singer Profile. Colors on the surface indicate the voice quality.

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.

4. SINGER PROFILES

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

4.1 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 2.1, 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 numerical 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.

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 2-tuples, written as* $<vp, \psi(vp)>$, *where* $vp \in VRP$ *is a vocal point that the singer can voice.*

Figure 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 2-tuples.

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

4.2 Singer Profile Generation

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

4.2.1 Step 1: VRP Generation

Before the VRP recording, the singer has to take some "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 pronounce 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 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 cut all voice files into *voice pieces* with 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 invariable 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.

4.2.2 Step 2: Voice Quality Computation

As mentioned before, there exists no report 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 6 which aims 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).

Supervised Learning

In order to train the quality evaluation function, a number of vocal music experts are requested to annotate the quality

Table 1: Features for voice quality evaluation

Feature Category	Feature Names
Pitch Features	*medianPitch*, *meanPitch*,*sdPitch*,*minPitch*, *maxPitch*,*nPulses*,*meanPeriod*,*sdPeriod*
Frequency Perturbations	*jitter_loc_abs*[1], *jitter_loc*[1], *jitter_rap*[1],*jitter_ppq5*[1]
Amplitude Perturbations	*shimmer_loc*[1], *shimmer_loc_dB*[1], *shimmer_apq3*[1],*shimmer_apq5*[1], *shimmer_apq11*[1]
Spectrum Features	*mean_nhr*[6], *mean_hnr*[26], *singing power ratio*[24],*tilt*[6],*ltasSlope*[17]

of voice pieces in each VRP recording using a software tool called *Praat* [4]. Each expert listens to the recorded WAV files and annotates the voice quality of different parts in each file. The possible annotation scores range from 1 to 5 (the lower the better quality). After an entire file becomes annotated, it will be split into voice pieces for training.

The quality evaluation function is trained as follows. First , several acoustic features are extracted for each voice piece. Table 1 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 spectrum 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.

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 in training. 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.

4.3 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. (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 detailed algorithm is given in Algorithm 1.

The output of the above partitioning algorithm will be used to derive some characteristics of a singer profile. These characteristics are important for understanding singer's competence and learning the recommendation function in Section 6.

Algorithm 1: *Singer Profile Partitioning*

```
Input: θ as the threshold;
Input: SP = {t_1, t_2, ..., t_n} as the singer profile ;
Input: PR as the pitch range of SP;
foreach t_i in SP do
    if t_i.vq<θ then
        t_i.vp is marked vp_good
    else
        mark t_i.vp as vp_bad
foreach pt in PR do
    T is the set of t_i whose pitch is pt
    vp_min = arg min_{t∈T} t.vp_good.intensity
    vp_max = arg max_{t∈T} t.vp_good.intensity
    foreach t_i ∈ T do
        if t_i.vp.intensity>vp_min.intensity and
        t_i.vp.intensity<vp_max.intensity then
            t_i.vp is marked as vp_good
        else
            t_i.vp is marked as vp_bad
```

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, $\beta = 4$. Figure 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.

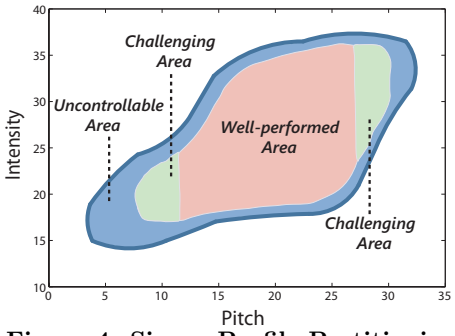


Figure 4: Singer Profile Partitioning

5. 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.

5.1 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 derives to matching the singer profile to the set of terms.

5.2 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 <$pitch_1$, $duration_1$>, <$pitch_2$, $duration_2$>, <$pitch_3$, $duration_3$>, <$pitch_4$, $duration_4$>, ...}, its respective piecewise intensity sequence is {<$intensity_1$, num_1>, <$intensity_2$, num_2>, ..., <$intensity_n$, 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.

6. COMPETENCE-BASED SONG RANKING

We apply the *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.

6.1 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 with 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 [5], 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); $y^{(i)} = (y_1^{(i)}, ..., y_{n^{(i)}}^{(i)}))$ 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

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

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

6.2 Competence Feature Extraction

Now we shall describe the ranking features (i.e. components of $x_j^{(i)}$ in Equation 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 5.1 for definition of *term*.) As discussed in Section 4.3, 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 $\{term_1, term_2, ..., term_n\}$ are the song terms appearing in A, and their $term\_freq$ and $agg\_duration$ in A are denoted by $\{tf_1, tf_2, ..., tf_n\}$ and $\{dur_1, dur_2, ..., dur_n\}$ 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 is really 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 $\sum_{i=1}^{n} dur_i \cdot tfidf_i$.
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 2 shows all the defined 22 features for each area.

Table 2: 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	√	√	√	

7. 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.

7.1 The Datasets

7.1.1 Singer Profile Dataset

For VRP recording, we recruit 55 volunteers including 35 males (mean age=28) and 20 females (mean age=23), 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 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 (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 the recording files of all 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.

7.1.2 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.

7.1.3 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 male ranking dataset, we divided the 100 male's midi songs into 5 subsets. The songs in each subset cover different pitch range and intensities to avoid data skew. We divide the 35 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 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*. As a result, a total number of 700 singing performances will be scored.

The female ranking dataset is built in a similar manner. We divided the 20 female subjects and 100 songs into five

subsets. In the end, 400 performances will be conducted and scored.

Our datasets are relatively small-scale due to resource constraints. However, we have observed sufficient varieties 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.

7.2 Baseline Methods

We compare CBSR scheme against two baseline methods:

Single pitch ranking method (SP) SP ranking method is a simplified version of our scheme. In this method, we regard each 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. The same 22 features will be extracted from the terms appearing in these areas, and then fed into the Listnet for training.

Pitch boundary ranking method (PB) PB ranking method is the most intuitive way of singing song recommendation – the one that we challenge in Section 1. This method only uses singer's pitch range of good quality corresponding to the well-performed area in VRP. The notes whose pitches are within the well-performed area is the determinant for the song ranking. 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 (same area as SP). These features are *Total TF*, *Total TF-IDF*, *Total Duration* and *Total TF-IDF Duration*.

7.3 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 two variables X and Y, ρ is calculated as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}. \quad (5)$$

For the competence-based song recommendation, we adopt the Normalized Discounted Cumulative Gain(NDCG) [12] and Mean Average Precision(MAP) [2] as our metrics for the ranking result. NDCG is for measuring the ranking accuracy which has more than two relevance levels and is calculated as

$$NDCG(n) = \frac{1}{Z_n} \sum_{j=1}^{n} \frac{2^{r(j)} - 1}{log_2(j+1)} \quad (6)$$

where j is the predicted rank position , $r(j)$ is the rating value of j-th document in the ground-truth rank list, Z_n is the normalization factor which is the discounted cumulative gain in the $n-th$ position of the ground-truth rank list.

MAP measures the average precision of all queries in the test set where each query's precision is the average precision computed at the point of each of the relevant documents in the predict rank list. MAP is suitable for evaluation of two level relevances. It can be computed as

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{n=1}^{N} (P@n * rel(n))}{R_q} \quad (7)$$

where Q is the test query set, R_q is q's relevant document, $rel(n)$ is a binary function on the relevance of a given rank, N is the number of retrieved documents. For MAP, we consider the top two of the five human relevance annotation levels as relevant and the remaining threes as irrelevant.

7.4 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.

7.4.1 Results of Voice Quality Computation

Remember 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 3. The Mean and STD are the average and the standard deviation of the Pearson correlation value calculated from the five trials.

Table 3: 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

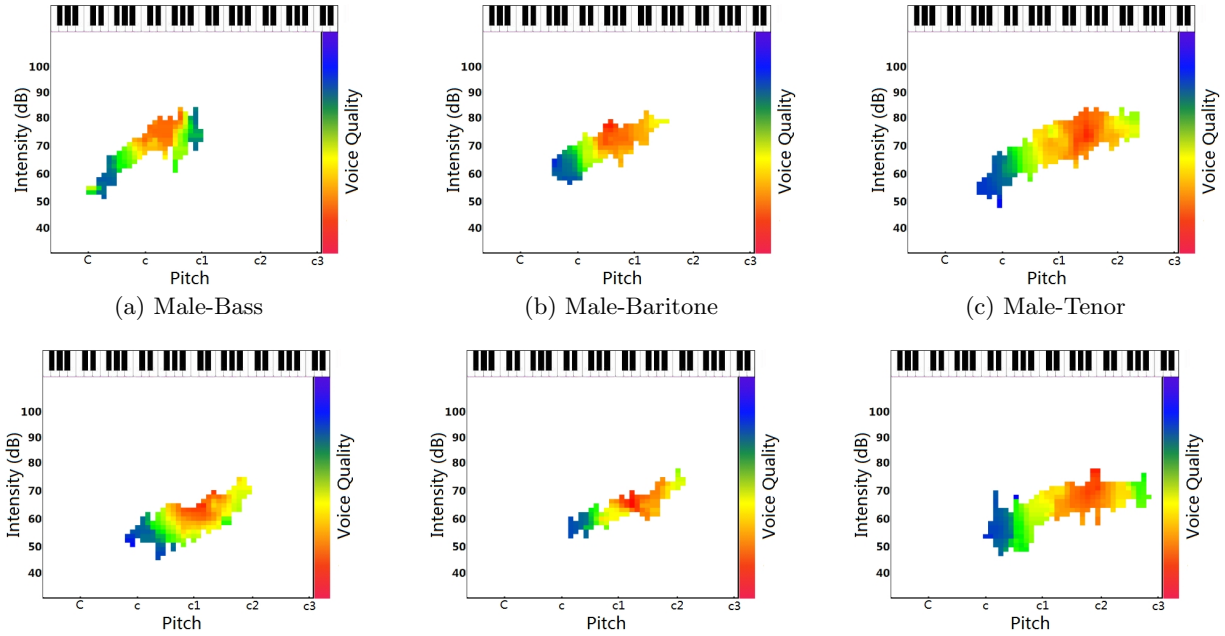
The above result shows large correlation of the predicted voice quality and human annotated voice quality. The male dataset achieves 0.7281 and the hybrid one gives 0.7115. However, the correlation value of Female is lower (0.5565). This is most probably due to the shortage of the female train data.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

7.4.2 Singer Profile Demonstration

After learning the quality evaluation function, we are able to generate the singer profile for each subject. Figure 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 Figure 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. Figure 5(a) shows a bass who can perform the low pitch with a rich voice.

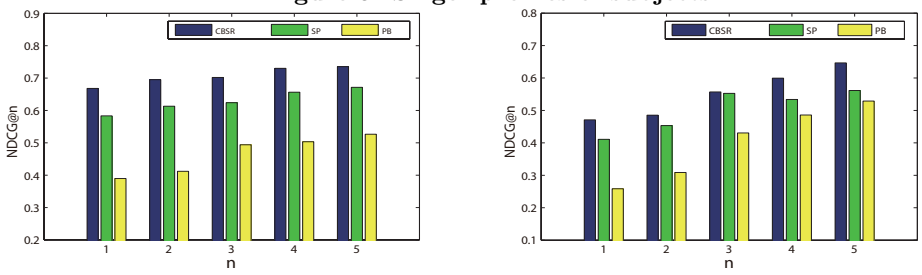
The voice quality of these profiles indicate that lower pitch or intensity are more likely to be of bad quality, while high intensity may lead to better quality. This is because in VRP recording, many subjects tend to produce soft voice, no matter whether the voice quality is good or not. When asked to



(a) Male-Bass (b) Male-Baritone (c) Male-Tenor

(d) Female-Bass (e) Female-Baritone (f) Female-Tenor

Figure 5: Singer profiles of subjects



(a) Male (b) Female

Figure 6: Ranking accuracy in NDCG@n on male and female ranking datasets

produce louder voice, some of the subjects are likely to stop voicing when reaching their uncontrollable area.

Figure 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.

7.4.3 *Ranking Accuracy*

To study the ranking accuracy, we divide the male and female ranking datasets into five subsets for cross validation. In each trial, three subsets are used for training, one for validation, and one for testing. The validation dataset is used to tune the number of iteration in Listnet training. For feature selection, we also perform PCA to reduce the feature dimensionality to 9. The MAP and NDCG@n results reported are all averaged from the 5-fold cross validation.

Figure 6 shows the NDCG values of top 5 ranked songs on male and female ranking datasets. Apparently, our CBSR outperforms the two baseline methods. Our CBSR scheme outperforms SP (which does not consider intensity) by an average of 9%. This is a clear indication of the importance of the intensity dimension when considering song recommendation. Baseline method SP also outperforms PB, due to the effect of considering the challenging area and the voice quality. These results illustrate the superiority of CBSR scheme.

Table 5: Ranking accuracy in terms of MAP

Algorithm	CBSR	SP	PB
Male	0.682	0.521	0.503
Female	0.494	0.461	0.439

Table 5 shows the results of the MAP value for CBSR, SP, and PB on the male and female ranking datasets. The ranking accuracy of PB measured from all ranking result is satisfactory, this is because PB has a good ability of identifying "difficult" songs which contain significant number of notes outside the singer's pitch range. However, as for recommending top ranked songs that are "challenging but manageable" for the singer, then the CBSR scheme performs much better.

7.4.4 *Query Demonstration*

In this part we demonstrate the concrete recommendation results of two tenor singers using CBSR and PB. Their singer profiles are already depicted in Figure 5(c) and Figure 5(f). Their *highest voicing pitch* are f2 and #a2 for the male and female respectively. The *major singing intensity levels* are {2,3,4} and {1,2,3} for the male and female, where 1, 2, 3, 4 represent the intensity *whisper*, *soft*, *normal* and *loud* respectively. Table 4 demonstrates the top five ranking results of CBSR and PB. The Pitch/Intensity column indicates the highest pitch and major singing intensity levels of the song. The results show all songs recommended by CBSR are chal-

Table 4: Comparison of CBSR and PB Recommendation Results

Query	Top-5 CBSR Query Results			Top-5 PB Query Results		
	Song Name	Rank Value	Pitch/Intensity	Song Name	Rank Value	Pitch/Intensity
Male-Tenor	Never say goodbye	0.921	#c2/{2,3,4}	Apologize	0.986	#a1/{2,3}
	You and I	0.91	d2/{2,3,4}	My love	0.983	a1/{2,3}
	Guilty	0.909	#d2/{2,3,4}	My december	0.976	c1/{1,2,3}
	Tripping	0.905	#d2/{2,3}	I'll be there for you	0.971	d2/{2,3}
	Careless whisper	0.874	e2/{2}	As long as you love me	0.97	#g1/{3}
Female-Tenor	Memory	0.943	#f2/{1,2,3}	Listen	0.973	#f2/{2,3,4}
	I will always love you	0.938	#f2/{2,3}	Stay	0.972	d2/{2,3}
	Bleeding love	0.938	a2/{2,3}	Heartbeats	0.965	c2/{1,2}
	Time to say goodbye	0.921	a2/{2,3}	My heart will go on	0.953	#d2/{2,3,4}
	Hero	0.906	e2/{2,3}	Hero	0.952	e2/{2,3}

lenging but manageable. In contrast, those recommended by PB contain either easy-to-sing songs such as "My December" and "Heartbeats" or very tough songs in intensity such as "My heart will go on" and "Listen", because it ignores the singing intensity and challenging but manageable singing area of the singer.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we broke the ice to study the 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. 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 extend the competence model to adopt more singer characteristics, such as timbre. Another direction is to simplify the competence acquisition.

Acknowledgments

The work is supported by the National Science Foundation of China (GrantNo. 61170034).

9. REFERENCES

[1] http://www.fon.hum.uva.nl/praat/manual/Voice.html.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: Breaking the glass ceiling. In *ISMIR*, pages 379–384, 2012.

[4] P. Boersma and D. Weenink. *Praat: doing phonetics by computer (Version 5.3.06) [Computer program]*, Retrieved May 1, 200. from http://www.praat.org/.

[5] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007.

[6] D. Deliyski. Acoustic model and evaluation of pathological voice production. In *Proceedings of Eurospeech*, pages 1969–1972, 1993.

[7] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by humming: Musical information retrieval in an audio database. In *ACM Multimedia*, pages 231–236, 1995.

[8] D. Goldberg, D. A. Nichols, B. M. Oki, and D. B. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[9] L. Heylen, F. Wuyts, F. Mertens, M. D. Bodt, and P. V. de Heyning. Normative voice range profiles of male and female professional voice users. *Journal of voice*, 16:1–17, 2002.

[10] L. G. Heylen, F. L. Wuyts, F. W. Mertens, and J. E. Pattyn. Phonetography in voice diagnoses. *Acta Oto-Rhino-Laryngologica*, 50:299–308, 1996.

[11] K. Hoashi, K. Matsumoto, and N. Inoue. Personalization of user profiles for content-based music retrieval based on relevance feedback. In *ACM Multimedia*, pages 110–119, 2003.

[12] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, pages 41–48, 2000.

[13] H. Kirchhoff, S. Dixon, and A. Klapuri. Multi-template shift-variant non-negative matrix deconvolution for semi-automatic music transcription. In *ISMIR*, pages 415–420, 2012.

[14] K. Mao, X. Luo, K. Chen, G. Chen, and L. Shou. mydj: recommending karaoke songs from one's own voice. In *SIGIR*, page 1009, 2012.

[15] 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. *Journal of voice*, 24:410–426, 2010.

[16] J. P. Pabon and R. Plomp. Automatic phonetogram recording supplemented with acoustical voice-quality parameters. *Journal of Speech and Hearing Research*, 31:710–722, 1988.

[17] G. Peeters. A large set of audio features for sound description. Technical report, IRCAM, 2004.

[18] J. Peter and H. Pabon. Objective acoustic voice-quality parameters in the computer phonetogram. *Journal of Voice*, 5:203–216, 1991.

[19] 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. *Journal of voice*, 24:153–160, 2010.

[20] H. Schutte and W. Seidner. Recommendation by the union of european phoniatricians (uep): Standardizing voice area measurement/phonetography. *Folia Phoniatr (Basel)*, 35:286–288, 1983.

[21] R. Speyer, G. H. Wieneke, I. van Wijck-Warnaar, and P. H. Dejonckere. Efficacy of voice therapy assessed with the voice range profile (phonetogram). *Journal of Voice*, 17:544–559, 2003.

[22] A. M. Sulter, H. K. Schutte, and D. G. Miller. Differences in phonetogram features between male and female subjects with and without vocal training. *Journal of voice*, 9:363–377, 1995.

[23] W.-H. Tsai and H.-C. Lee. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:1233–1243, 2012.

[24] 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. *Journal of voice*, 20:82–88, 2006.

[25] S. K. Wolf, D. Stanley, and W. J. Sette. Quantitative studies on the singing voice. *The Journal of the Acoustical Society of America*, 6:255–266, 1935.

[26] E. Yumoto, W. Gould, and T. Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71:1544–1550, 1982.