Perceptual Evaluation of Singing Quality

Chitralekha Gupta* Haizhou Li† and Ye Wang‡
*‡School of Computing, *NUS Graduate School for Integrative Sciences and Engineering,
†Electrical and Computer Engineering, National University of Singapore, Singapore
*chitralekha@u.nus.edu †haizhou.li@nus.edu.sg ‡wangye@comp.nus.edu.sg

***Abstract*—A perceptually valid automatic singing evaluation score could serve as a complement to singing lessons, and make singing training more reachable to the masses. In this study, we adopt the idea behind PESQ (Perceptual Evaluation of Speech Quality) scoring metrics, and propose various perceptually relevant features to evaluate singing quality. We correlate the obtained singing quality score, which we term as Perceptual Evaluation of Singing Quality (PESnQ) score, with that given by music-expert human judges, and compare the results with the known baseline systems. It is shown that the proposed PESnQ has a correlation of 0.59 with human ratings, which is an improvement of $\sim 96\%$ over baseline systems.**

I. Introduction

Singing is a popular medium of entertainment and pleasure, and a desirable skill to develop. But singing pedagogy remains heavily dependent on human music experts, who are a few in number. The evaluation criterion for singing relies on subjective expert judgments, which are not conveniently available to ordinary people who desire to learn singing. Thus, a system for automatic and reliable evaluation of singing could serve as an aid to singing pedagogy, singing contests, and karaoke systems, in turn making singing training more accessible to the masses.

Singing quality assessment often refers to the degree to which a particular vocal production meets professional standards of excellence. For reliable assessment, it is important to identify vocal attributes that relate to human perceptual ratings and objectively define singing excellence.

Past studies have identified several perceptual features pertaining to singing voice that play significant role in subjective evaluation of singing skill. One study described twelve generally accepted criteria used in the evaluation of Western classical singing by expert music teachers [1], which are: *appropriate vibrato, resonance/ring, color/warmth, intensity, dynamic range, efficient breath management, evenness of registration, flexibility, freedom throughout vocal range, intonation accuracy, legato line,* and *diction*. Oates et al. [2] proposed an auditory-perceptual rating scale for operatic singing voice, which consisted of five perceptual parameters - *appropriate vibrato, ring, pitch accuracy, evenness throughout the range,* and *strain*, and these parameters were proven to be unambiguous and covered all aspects of operatic voice.

However, all of the above parameters may not be suitable for evaluating a non-trained or a novice singer. Chuan et al. [3] defined and verified six perceptual parameters that were of most relevance for assessing non-trained singers. These parameters were: *Intonation accuracy*, described as singing in tune, where suitable key transposition is allowed; *Rhythm consistency*, described as singing with appropriate tempo speed, where slight tempo variation is allowed; *Timbre brightness*, described as brilliance of tone, a sensation of brightness of spectrum; *Appropriate vibrato*, described as regular and smooth undulation of frequency of the tone; *Dynamic Range*, described as the pitch range that the subject is able to sing freely throughout, without inappropriate change in voice quality or any external effort; and *Vocal Clarity*, described as vocal vibrations of a clear, well-produced tone. In this work, we explore different features of audio signal that represent these perceptual parameters for singing evaluation, to develop an objective methodology for singing assessment.

The paper is structured as follows. Section II reviews the related previous work and techniques, the challenges in the area, and formulates the problem of perceptual singing assessment. In Section III, we discuss how singing quality is characterized along with our feature design approach. Section IV describes our experiment methodology of subjective and objective evaluation, and Section V discusses our experiment results. Section VI presents the conclusion of this study and suggestions for the future work.

II. Related work

Objective evaluation of singing has been an area of interest in the recent past. There have been multiple attempts to develop automatic singing evaluation algorithms based on pitch, rhythm, expression, and volume related features. But there are some technical challenges in each of these algorithms, that we will try to address in this work. Also, evaluation can be template-based, in which a test sample is compared against a reference sample, or model-based, in which a test sample is compared against a reference model. The singing evaluation literature, as described in this section, considers template-based evaluation, where the template is either MIDI notes or a reference singing. In our study also, we consider template-based evaluation.

Intonation accuracy or pitch accuracy evaluation is the most common method for singing assessment. In one study, Lal [7] proposed a pitch-based similarity measure to compare a test singing clip to the reference singing clip. But reliable and automatic pitch estimation is a challenging task, and errors in pitch estimation can result in incorrect automatic score. In another study, Tsai and Lee [4] proposed an automatic evaluation system for karaoke singing in which they compared MIDI (Musical Instrument Digital Interface) notes of test

singing to that of the intended reference song to compute the pitch accuracy rating. Although MIDI notes approximately represent the sequence of sung notes, they are unable to represent human voice. Apart from steady notes, singing voice comprises of pitch transitions, modulations, and voice timbre, that are not captured by digitally generated MIDI notes. Also, when singing without background accompaniments, the singers tend to sing at a key they are comfortable in, which may or may not be the same as that of the reference song. In such a scenario, singing the correct sequence of notes with a key transposition should not be penalized [3]. The possibility of key transposition has not been considered in [4], because the key of a song is inherently fixed in karaoke singing due to background accompaniments.

Rhythm consistency is another important feature for singing evaluation. Tsai and Lee [4] evaluated rhythm by comparing the note-onset strength of the background accompaniment of karaoke to that of the test singing. But when we consider the case of singing without background accompaniments (or free singing), such methods cannot be directly applied. Like key transposition, in free singing, the singers can have a slight tempo variation from the reference singing, i.e. slightly but uniformly faster or slower rhythm than the reference song. Such tempo variations should not be penalized [3]. Molina et al. [5] and Lin et al. [6] measured rhythm accuracy without penalizing for a rhythm different from the reference. They evaluated rhythm by aligning test pitch contour with the reference pitch contour using Dynamic Time Warping (DTW), and obtained the rhythm score by computing the deviation of the optimal path from a straight line fit in the cost matrix of the DTW between the pitch contours. This line-fit may be different from the ideal 45 degree straight line, in turn compensating for tempo difference. But aligning test and reference singing using pitch contour makes rhythm assessment dependent on pitch correctness. This poses a problem if the test singer sings with inaccurate pitch (off-tune) but maintains a good rhythm. Inaccurate pitch estimation also creates the same problem. This method will give large deviation from the optimal path due to pitch inaccuracies, despite good rhythm.

Expressive elements such as *appropriate vibrato* are considered to be important cues to distinguish between a well-trained singer and a mediocre singer. Nakano et al. [8] computed acoustic features which are independent from specific characteristics of the singer or melody, such as vibrato features like rate and extent of pitch undulations, to evaluate singing in a case that has no reference singing. But while learning to sing a song, one would try to match their singing with a reference singing in every way possible. In such a case, vibrato evaluation in the presence of reference singing is needed. Vibrato detection and evaluation will also be affected by pitch estimation errors.

Timbre brightness is defined as the ring or brilliance of a tone [3], which often relates to voice quality. Singing power ratio (SPR), which is the ratio of highest spectral peak between 2 and 4 kHz and the highest spectral peak between 0 and 2 kHz in voiced segments, has been used previously to separate

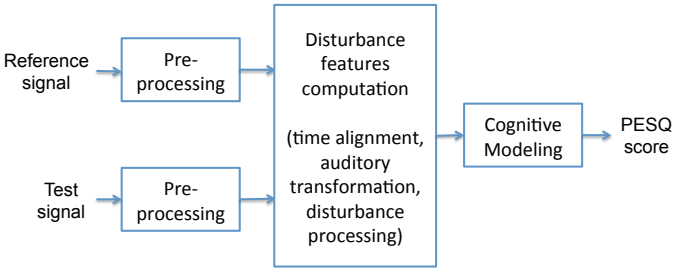


Fig. 1. Overview of PESQ computation

professional singers and non singers [10]. But as pointed out by Tsai and Lee [4], the ringing voice quality which is indicated by high SPR, is typically observed in operatic style of singing. However, operatic style is a specific way of singing that, one may argue, can be unsuitable and undesirable for singing lessons or karaoke performances, especially for beginners. Hence our work here doesn't consider SPR as a parameter for automatic singing evaluation. Prasert et al. [15] developed a more general method to evaluate voice quality in singing based on timbral features, such as Mel Frequency Cepstral Coefficients (MFCC) and Filter Banks (FBANK), and found that MFCCs performed better. We will consider this direction in our study.

As illustrated in [4], most of the singing evaluation studies have been reported in patent documentation that do not discuss the rationale of their evaluation methods, and fail to show results of their qualitative analysis to validate their methods. Comparatively, the number of scientific studies in this area is fewer. Literature suggests that a combination of the various perceptual parameters, as described in [3], would result in the final judgment of a test singing clip. But both, the patents and the scientific studies, have managed to incorporate a set of objective acoustic cues that are relevant to only a subset of the perceptual parameters for singing evaluation. For example, patents such as [21], [22] have used a combination of volume and pitch as evaluation features, while scientific studies such as [4] have used pitch, volume, and rhythm features. We need a unified evaluation system that finds the appropriate weighting of all the perceptually relevant parameters to obtain the final score. The idea of perceptual assessment of singing quality is motivated by the International Telecommunication Union (ITU) standard for quality assessment of speech in telephone networks, PESQ (Perceptual Evaluation of Speech Quality) [12]. PESQ is obtained by comparing the original speech signal with its degraded version (test signal), that is the result of passing the original signal through a communication channel, and predicting the perceived quality of the degraded signal, as shown in Figure 1. We note that objective measurement of signal quality doesn't always correlate with human perception. The ITU benchmark experiments report an average correlation of 0.935 between PESQ scores and human scores, that make PESQ an ideal objective metric. According to cognitive modeling literature, *localised errors* dominate perception of audio quality [11], i.e. a highly concentrated error in time and frequency is found to have a greater sub-

jective impact than a distributed error. This concept has been successfully used in assessing speech signal quality (PESQ), by using a higher weightage for localised distortions in PESQ score computation. Motivated by this approach, we apply this concept of audio quality perception in our work to obtain a novel singing quality assessment method.

III. Singing quality Characterization and Evaluation

In this study, we aim to develop a holistic scoring framework for automatic singing evaluation based on perceptually relevant features, as recommended by music educators. We explore various features, such as pitch, rhythm, vibrato, timbre, volume, and pitch dynamic range, that are perceptually relevant to singing, and try to overcome their technical challenges, to develop a measure for evaluating singing skill of a test singer as compared to an ideal reference singing of a song. We introduce methods to overcome the challenges of key-transposition, and rhythm variation, as well as incorporate other perceptual evaluation parameters such as vibrato, voice quality, pitch dynamics, and volume. We adopt the audio quality perception theory in singing quality assessment by giving high weightage to localised distortions, thus obtaining a novel score for singing quality assessment. We have termed it as Perceptual Evaluation of Singing Quality (PESnQ) score. We compare our results with the known baseline methods for singing evaluation.

In this section, we elaborate on singing quality *characterization* and *evaluation* methods. Singing quality can be characterised by the perceptual parameters identified by human experts, while evaluation is the distance between the target and the reference singing characteristics. Here we describe the acoustic features that determine singing quality, relate these features to the perceptual parameters used for singing assessment, as well as describe the distance parameters defined and used for evaluation. The following sub-sections A-F contain both characterization and evaluation methods, while sub-section G studies the cognitive modeling technique for singing quality evaluation.

A. Intonation accuracy

Pitch is a major auditory attribute of music tones. Pitch of a musical note is defined as the fundamental frequency $F0$ of the periodic waveform. Intonation accuracy or "singing in tune" is directly related to the correctness of the pitch produced as with reference to reference singing. For developing an automatic system that evaluates pitch accuracy, estimation of reliable pitch contours becomes very important. Pitch estimation is an active research area, and various algorithms have been developed for pitch estimation in monophonic speech signals, such as ACF (autocorrelation function)[13], YIN [18], etc. But these methods need adaptations and post-processing to accurately detect pitch in singing waveforms. Babacan et al. [14] compared the different pitch detection algorithms for monophonic singing, and found that parameter settings specific to singing, such as increasing the $F0$ search range to account for wide vocal range of singing, as well as applying post-processing to pitch estimates lead to better pitch estimates. They also found that the autocorrelation-based PRAAT [16] pitch estimator gives best voicing boundaries even without post-processing, while the source-filter model-based STRAIGHT [17] pitch estimator is the most robust algorithm in noisy conditions. The modified autocorrelation-based estimator YIN [18] achieves the best accuracy of pitch detection but it requires a number of post-processing steps depending on the properties of the music type being analysed, as described in [19].

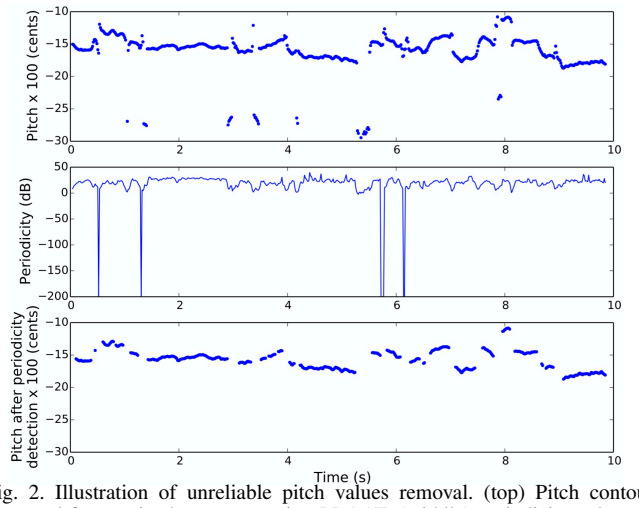


Fig. 2. Illustration of unreliable pitch values removal. (top) Pitch contour extracted from voiced segments using PRAAT, (middle) periodicity values in dB, (bottom) pitch contour, after removal of low periodicity pitch values

In our work, we use the pitch estimates from PRAAT, with one generic post-processing step to remove unreliable pitch values. We first use the pitch estimates to determine the voicing boundaries, compute the pitch estimates over all the voiced frames, and then remove the frames with low periodicity, which is determined by harmonic-to-noise ratio (HNR). HNR, also computed in PRAAT, represents the degree of acoustic periodicity expressed in dB. For example, if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is $10\log_{10} 99/1 = 19.95\,\mathrm{dB}$. In determining the valid pitch frames, we remove the ones with $< 98\%$ of energy in periodic part, i.e. $HNR < 10\log_{10} 98/2 \approx 16.9\,\mathrm{dB}$. This threshold is set empirically. By choosing only the voiced segments and removing the frames with low periodicity, spurious $F0$ values are avoided and only reliable pitch values are used. Figure 2 shows an example of pitch contour before and after periodicity-based pitch clean-up. All pitch values in this study are calculated in the unit of cents (one semitone being 100 cents on equi-tempered octave),

$$f_{\mathrm{cent}} = 1200 \times \log_2 \frac{f_{\mathrm{Hz}}}{440}, \tag{1}$$

where 440 Hz (pitch-standard musical note A4) is considered as the base frequency. For singing quality evaluation in terms of pitch accuracy, we first time-align the reference and test singing by using the alignment from DTW between their MFCC vectors. This compensates for any tempo differences or tempo errors between reference and test. Then, we compute

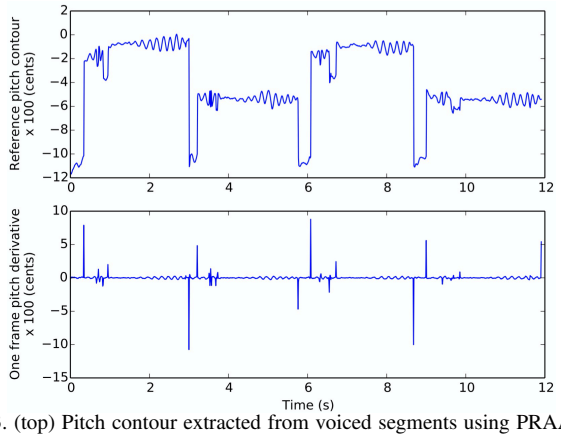


Fig. 3. (top) Pitch contour extracted from voiced segments using PRAAT and after removal of low periodicity frames, and (bottom) corresponding pitch contour derivative with one frame shift.

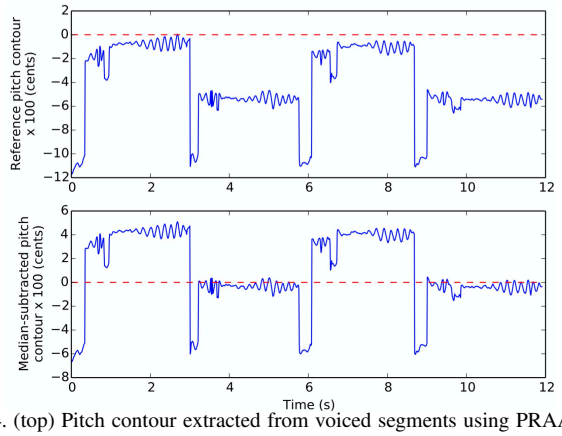


Fig. 4. (top) Pitch contour extracted from voiced segments using PRAAT and after removal of low periodicity frames, and (bottom) corresponding median-subtracted pitch contour.

the DTW distance between the pitch contours of the reference and test singing (termed as *pitch_dist*) for evaluation, which would be an indicator of *intonation accuracy*, as previously used in [5], [4], [7]. But this distance between pitch contours will penalise key transposition, although key transposition is allowed in case of singing without background accompaniments [3]. Hence we use two different methods to make the distance measure insensitive to key transposition:

1) Pitch Derivative: Derivatives of pitch contours of both the reference and the test singing make the resultant contours independent of key shifts. The derivative also emphasises on the transitions between notes, in terms of the magnitude as well as the duration of the change. Note transition expressions, such as glissando, are considered to be a significant indicator of good singing, that get captured by this feature. For a pitch vector $\mathbf{p_a} = \begin{bmatrix} p_1 & p_2 & \dots & p_N \end{bmatrix}^T$, where N is the number of frames, one frame derivative $\mathbf{\Delta p}$ is computed as

$$\mathbf{\Delta p} = \mathbf{p_a} - \mathbf{p_b}, \tag{2}$$

where $\mathbf{p_b}$ is the pitch vector shifted by one frame. Figure 3 shows an example of a pitch contour and its derivative.

2) Median-subtracted Pitch: Subtracting the median of the pitch values of an audio segment is another way to make the pitch contour independent of key-transposition. Here, median is preferred over mean because averaging over all pitch values might get affected by infrequent outlier pitch values, which is avoided by the median. The median-subtracted pitch for a pitch vector $\mathbf{p}$, is computed as

$$\mathbf{p_{medsub}} = \mathbf{p} - \text{median}\{\mathbf{p}\}. \tag{3}$$

Figure 4 shows an example of a pitch contour and its median-subtracted version.

We apply the cognitive modeling theory to these frame-level modified pitch vectors (pitch-derivative and median-subtracted pitch) to obtain the pitch evaluation between reference and test singing (Section III-G).

B. Rhythm Consistency

Rhythm is defined as the regular repeated pattern in music, that relates to the timing of the notes sung, and is often referred to as tempo. Rhythm consistency refers to the similarity of tempo between reference and test singing. As mentioned earlier, a slight variation in tempo is allowed, i.e. uniformly faster or slower tempo compared to the reference. Molina et al. [5] proposed DTW as a procedure for automatic rhythm assessment, and accounted for rhythm variation. They computed the DTW between the reference and the test pitch contours, and analyzed the shape of the optimal path in the cost matrix of DTW. A 45° straight line would represent a perfect rhythmic performance with respect to reference melody, while straight line with an angle $\neq 45^\circ$ would represent good rhythmic performance in a different tempo. So they fit a straight line on the optimal path in the cost matrix of the DTW, and computed the root mean square error of this straight line fit from the optimal path (termed as *molina_rhythm_pitch_dist*),

$$\epsilon = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\epsilon_k^2}, \tag{4}$$

where ϵ_k is the error in linear fit at frame k, and N is the total number of frames.

But aligning test and reference singing using pitch contour makes rhythm assessment dependent on pitch correctness. So if the singer maintains a good rhythm but sings with inaccurate pitch, this algorithm will give a poor score for rhythm. Thus this method works well only when the test singing pitch is same as that of reference singing, even if words are spoken incorrectly. But this method will give a large deviation from the optimal path if the pitch is inaccurate, despite good rhythm.

We propose a modified version of Molina's rhythm deviation measure. Instead of using pitch contour, we use 13 MFCC feature vectors to compute DTW between reference and test singing. MFCCs capture the short-term power spectrum of the audio signal that represents the shape of the vocal tract and thus the phonemes uttered. So when we compute DTW between MFCC vectors, we assume that the sequences of phonemes and words are uttered correctly, thus making this measure independent of off-tune pitch. So we obtain a modified Molina's rhythm deviation measure (termed as *molina_rhythm_mfcc_dist*) that measures the root mean square error (Equation 4) of the linear fit of the optimal path of DTW matrix computed using MFCC vectors.

We also introduce another way to compute rhythm deviation, while accounting for allowable rhythm variations. We compute 13 MFCC vectors over a 32 ms long window for every 16 ms of the reference singing, and then compute the corresponding frame rate for the test singing such that the number of frames in reference and test are the same. This way we compensate for constant rhythm difference between reference and test singing, and thus the number of MFCC vectors in reference and test are equal. Then we apply cognitive modeling theory to these frame-equalized MFCC feature vectors to obtain the rhythm evaluation between reference and test singing (see Section III-G).

C. Voice Quality and Pronunciation

Timbre is related to the voice quality and describes the perceived quality of a tone produced by the singer. Perception of timbre is physically represented by spectral envelope of the sound, which, as mentioned earlier, is captured well by MFCC vectors, as illustrated in [15]. MFCCs also represent phonetic quality, which relates to pronunciation. Thus, we compute the distance between reference and test singing timbre (termed as *timbral_dist*) by computing the DTW distance between their 13 MFCC vectors. This measure represents two parameters - voice quality and pronunciation.

D. Appropriate Vibrato

Vibrato is the rapid periodic undulations in pitch on a steady note while singing. Studies have found that vibrato oscillations are within 5-8 Hz, and their extent is between 30-150 cents [20]. Vibrato is considered to be a fair indicator of the quality of singing, hence we would like to evaluate it. For a fully automated evaluation system, the idea is to first detect the vibrato sections in the reference, then find the corresponding time-aligned pitch segments in the test, and finally compute measures to compare the reference and test vibrato segments. Another way could be to compare vibrato-specific feature vectors of every frame from test and reference. However, the frames in test that correspond to those in reference that do not contain vibrato, should not be given a high score, as we are not giving marks for "absence of vibrato". Thus detection of vibrato sections as the first step is necessary.

Nakano et al. [8] applied short-term Fourier transform to the first order differential of $F0$ and defined vibrato likeliness $P_v(t)$ as the product of power $\Psi_v(t)$ and sharpness $S_v(t)$ as:

$$\Psi_v(t) = \int_{F_L}^{F_H} \hat{X}(f,t)df, \quad S_v(t) = \int_{F_L}^{F_H} \left| \frac{\partial \hat{X}(f,t)}{\partial f} \right| df, \quad (5)$$

$$P_v(t) = \Psi_v(t) S_v(t), \quad (6)$$

where (F_L,F_H) is the range of vibrato rate set as 5 and 8 Hz respectively, and $\hat{X}(f,t)$ is the power spectrum $X(f,t)$ normalized over f:

$$\hat{X}(f,t) = \frac{X(f,t)}{\int X(f,t)df}. \quad (7)$$

If the value of vibrato likeliness is greater than a threshold, the section is detected as *vibrato section*. However, the problem with this measure of vibrato likeliness is that the obtained likeliness values are not normalized, which makes it difficult to set a singer-independent threshold for vibrato detection. In this study, we have modified the vibrato likeliness measure as the ratio of energy in the power spectrum of $F0$, $X(f,t)$, between 5 to 8 Hz (F_L,F_H) to the total energy in the spectrum (Equation 8). A similar feature was used by Amir et al. [9].

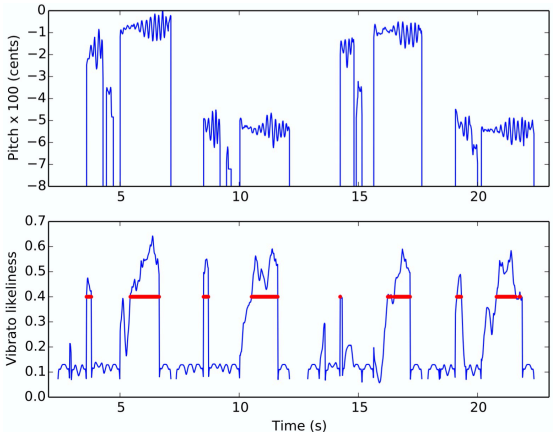


Fig. 5. (top) Pitch contour extracted from reference singing, and (bottom) modified vibrato likeliness $P_{v_{\text{mod}}}(t)$, vibrato sections marked in red.

$$P_{v_{\text{mod}}}(t) = \frac{\int_{F_L}^{F_H} X(f,t)df}{\int X(f,t)df}. \quad (8)$$

This measure gives a normalized score between 0 and 1, unlike the score obtained by Nakano et al. Also it is a good indicator of concentration of energy in the vibrato oscillation frequency range. We compute this modified vibrato likeliness score over every 512 ms frame (i.e. 32 samples, similar to [8]) of the reference singing segment, and empirically set a threshold of 0.4 to detect the valid vibrato segments in the reference singing as shown in Figure 5.

We map the time stamps of the detected vibrato segments in reference to that of the aligned test pitch contour to obtain potential vibrato segments in the test. For these segments, we compute three vibrato-related features - modified vibrato likeliness ($P_{v_{\text{mod}}}(t)$ from Equation 8), extent, and rate. The extent and rate features are the ones defined by Nakano et al.:

$$\frac{1}{\text{rate}} = \frac{1}{N}\sum_{n=1}^{N} R_n \qquad \text{extent} = \frac{1}{2N}\sum_{n=1}^{N} E_n, \quad (9)$$

where R_n (in sec) is the time period of n^{th} oscillation, computed as the difference between alternate zero-crossing time instants, and E_n (in cents) is the difference between the maximum and the minimum pitch value in the n^{th} oscillation. As a post-processing step, we discard any detected reference vibrato section from vibrato evaluation that does not have at least one whole oscillation present. Thus we have modified vibrato likeliness, rate, and extent features for every valid reference vibrato section and corresponding test pitch section. We compute the Euclidean distance of these features between the reference and the test to obtain the vibrato distance score (termed as *vib_segment_dist*) for evaluation.

E. Volume

Dynamics of volume reflect the relative loudness or softness of different parts of the song. It is expected that there will be a similar pattern of volume variations across time when different singers perform the same song [4]. Apart from Tsai and Lee's work, various singing evaluation patents have incorporated volume as an acoustic cue in their systems [21], [22]. In this study, we implement the volume feature used by Tsai and Lee's system, i.e. short-term log energy over 30 ms window, and then compute the DTW distance of this feature between the reference and the test (termed as *volume_dist*) for evaluation.

F. Pitch Dynamic Range

The pitch range that a subject is able to sing freely throughout is a good indicator of quality of singing [3]. Thus we compute the absolute difference between the highest and the lowest pitch values in an audio segment as a feature for pitch dynamic range. The distance of this feature between reference and test singing (termed as *pitch_dynamic_dist*) is an indicator of the similarity of the test singing range to the expected singing range, and is used for singing quality evaluation.

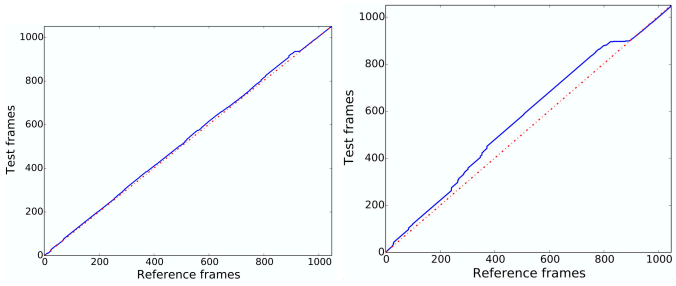
G. Cognitive Modeling for Evaluation

PESQ standard [12] incorporates the audio perception theory that a localized error in time has a larger subjective impact than a distributed error [11]. PESQ combines the frame-level disturbance values of an audio file by computing the L_6 norm over split-second intervals, i.e. over 20 frames (320 ms) window (with 50% overlap and no window function), and the L_2 norm over all these split-second disturbance values over the length of the speech file. The value of p in L_p norm is higher for averaging over split-second intervals, to give more weightage to localized disturbances. L_p norm is computed as:

$$L_p \text{ norm} = \left(\frac{1}{\mathrm{N}} \sum_{\mathrm{m}=1}^{\mathrm{N}} \text{disturbance}[\mathrm{m}]^{\mathrm{p}} \right)^{\frac{1}{\mathrm{p}}}. \qquad (10)$$

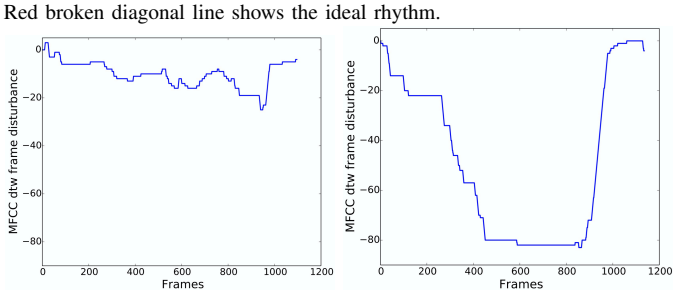
where N is the total number of disturbance values, over index m. Similarly in singing, errors are time-localized; for example, only certain notes may become off-tune or only certain sections may be sung with bad rhythm. Therefore, in this study we explore the possibility of applying the same cognitive modeling concept as in PESQ, for singing quality evaluation.

We first compute the frame-level disturbance values of the following singing features: pitch derivative $\mathbf{\Delta p}$, median-subtracted pitch $\mathbf{p_{medsub}}$, and frame-equalized MFCC feature vectors for rhythm. That is, we compute the optimal path in the cost matrix of DTW between the respective feature vectors of reference and test. If the pitch or rhythm in test singing matches with that of the reference, it would give a 45° optimal path in the corresponding DTW cost matrix. Figure 6 illustrates the optimal path of singing with good and poor rhythm accuracy. Deviation of the best alignment path from the diagonal represents error in that characteristic (pitch or rhythm). We compute the number of frames that deviate from the ideal diagonal path for every frame, called *rhythm frame disturbance* R_d, *pitch derivative frame disturbance* $P_{d_{\mathbf{\Delta p}}}$, and *median-subtracted pitch frame disturbance* $P_{d_{\mathbf{p_{medsub}}}}$. Larger deviations indicate poor intonation/rhythm accuracy. Figure 7 shows an example of the frame disturbance of good and poor rhythm.



(a) (b)

Fig. 6. Optimal path in DTW cost matrix for (a) good rhythm (b) poor rhythm. Red broken diagonal line shows the ideal rhythm.



(a) (b)

Fig. 7. Frame disturbance for (a) good rhythm (b) poor rhythm.

Next, we compute the L_6 norm over split-second intervals and L_2 norm over all these split-second disturbance values over the length of the sung file for all of the frame-level disturbance values mentioned above - $P_{d_{\mathbf{\Delta p}}}$ (termed as *pitch_der_L6_L2*), $P_{d_{\mathbf{p_{medsub}}}}$ (termed as *pitch_med_L6_L2*), and R_d (termed as *rhythm_L6_L2*). And for performance comparison, we also compute the L_2 norm of all the disturbance values over the entire file, to observe the effectiveness of the cognitive modeling method for singing evaluation. To summarize, we compute three kinds of evaluation features: L_2 norm, L_6+L_2 norm (PESQ-based), and DTW distance (feature groups: L2, L6+L2, and dist respectively). The summary of evaluation features is listed in Table III.

We define *Perceptual Evaluation of Singing Quality* (PESnQ) as the score generated from a system comprising of a combination of PESQ-based, L_2 norm, and DTW distance-based features. In the following sections, we will explore different combinations of these features to build various singing evaluation systems and investigate the factors that can impact their performance, such as type and definition of features, the PESQ-based perceptual distance features, and their combinations.

IV. Experiments

A. Data

To test our methodology for singing evaluation, we chose two popular English songs - "I have a dream" by ABBA

($\sim$2 min), and "Edelweiss" ($\sim$1 min) from the movie "The Sound of Music". These songs have steady rhythm throughout the song, and are rich in long steady notes and vibrato. We needed monophonic sung recordings of these songs from singers with a range of singing ability - poor to professional. Duan et al. [23] recorded these two songs from 20 singers, but the range of singing ability in that dataset was limited to mediocre to good level, where the singers had some exposure to vocal training or were talented to sing in tune and rhythm. However, to cover the entire spectrum of singing ability, we needed samples from the two extremes - poor singers and professionally trained excellent singers. So we first obtained the dataset from Duan et al. and then recruited a few subjects to fill the gap at the two ends of the spectrum. We recruited two professionally trained singers and five students from NUS with no past experience in singing. These subjects were a mix of native and non-native English speakers, but all were proficient in speaking in English, similar to [23]. To be consistent with the previous dataset [23], we followed their procedure for collecting audio data from the new recruits. Subjects were asked to familiarize themselves with the two songs beforehand. Audio data was collected in a sound-proof audio recording studio at 16-bit and 16 kHz. A metronome was fed to the subject via headphone to serve as a guide for singing: "I have a dream" at 56 bpm, and "Edelweiss" at 32 bpm. These settings were same as that in [23]. Except for the metronome beats, no other accompaniment was provided. Lyrics for the songs were provided for the subject's reference while recording.

From the previous and the newly collected dataset, we selected 20 recordings for singing quality evaluation. Each was sung by a different singer with singing abilities ranging from poor to professional. Ten singers sang the song "I have a dream", and the other ten sang "Edelweiss". We obtained subjective evaluation ratings from music experts for these 20 recordings and ensured that this dataset is well representative of the singing skill spectrum (see Section IV-B). We also obtained objective evaluation scores for these recordings using our features and methods, and the known baseline methods.

B. Subjective Evaluation

We developed a website to collect subjective ratings for this dataset. The task was to listen to the audio recordings and evaluate the singers' singing quality, compared to a professionally trained reference singer (also provided on the website). The reference singing of both the songs were from one professional singer from the dataset of [23], different from our test singing evaluation dataset of 20 singers. Five professional musicians were the human judges to complete this task. These judges have been trained in vocal and/or musical instruments in different genres of music such as jazz, contemporary, and Chinese orchestra, and all of them were stage performers. One of them has also been a music teacher for more than 2 years. The task could be done in multiple sittings, a few recordings each time. Their evaluations were saved in our database, that they could revisit later.

The website had two songs, each with 10 audio tracks, sung by different individuals (as described in Section IV-A). For every track, the corresponding lyrics were displayed on the screen. This is followed by a questionnaire, where the judges were asked to give an overall singing quality score out of 5 to each of these audio recordings compared to the reference singing of the song. The judges were also asked to separately rate each of the recordings based on pitch (*intonation accuracy*), rhythm (*rhythm consistency*), expression/vibrato (*appropriate vibrato*), voice quality (*timbre brightness*), articulation, relative volume, and pitch dynamic range on a likert scale of 1 to 5. Additionally, an optional question was asked to know if the music expert considers any other parameters that the singer could improve upon, apart from the ones already listed.

The average inter-judge (Pearson's) correlation of the overall singing quality question was 0.82, which shows a high agreement of singing quality assessment amongst the music experts. Table I shows the inter-judge correlation of all the questions that used a likert scale. Most of the questions showed correlation of higher than 0.60. Thus these parameters are judged by music experts coherently. However, the questions on pronunciation and volume showed lower inter-judge correlation. Since the lyrics were already provided to the singers, there was little room for mispronouncing words because of unfamiliar lyrics. The only way mispronunciations could have happened was due to mother-tongue influence in non-native English. A possible reason for less agreement on pronunciation ratings is unclear definition of mispronunciation in singing, which leads to influence of other factors on this rating. An example of disagreement was when a singer, whose mother-tongue was English, but who had poor singing skills, was rated poorly for pronunciation by a couple of judges, while the other three judges rated the singer high for the same parameter. So in this case, poor singing seems to influence the perception of pronunciation. We believe that the reason for disagreement in case of the relative volume question is also because of lack of clear definition. As seen in Section I, volume never showed up in the literature on subjective assessment of singing in non-trained singers [3], [2], but volume was one of the key features in the objective evaluation literature [4], [21] because this measure is easy to compute objectively and pattern-match with a reference template, but difficult to rate subjectively. This explains the low agreement on volume parameter.

We computed the average of the overall singing quality score given to each of the 20 singers over the 5 human judges. We found that this data represents the complete singing skill spectrum. Table II shows the number of singers with different overall singing abilities categorized by average human ratings.

C. Objective Evaluation

Here we describe automatic systems built using combinations of the features from Section III. Our automatic singing evaluation framework is the same as that of PESQ (Figure 1).

As a pre-processing step, we first split every audio recording into shorter segments of approximately 20 sec duration. This is done by using DTW to align MFCC feature vectors of the test audio with that of the reference audio that is marked with

TABLE I
INTER-JUDGE CORRELATION FOR THE QUESTIONNAIRE QUESTIONS.

Question	Inter-judge correlation
Overall singing quality	0.82
How would you rate the singer in terms of pitch accuracy?	0.81
How would you rate the singer in terms of rhythm accuracy?	0.75
How would you rate the singer in terms of vibrato/expression quality?	0.65
How would you rate the singer in terms of voice quality?	0.68
How would you rate the singer in terms of pronunciation quality?	0.53
How would you rate the singer in terms of relative volume?	0.46
How would you rate the singer in terms of pitch dynamic range?	0.67

TABLE II
NUMBER OF SINGERS WITH DIFFERENT LEVELS OF OVERALL SINGING ABILITY, CATEGORIZED BASED ON AVERAGE HUMAN RATINGS.

Avg. score range	1.0 - 1.8	1.8 - 2.6	2.6 - 3.4	3.4 - 4.2	4.2 - 5.0
# of singers	5	3	8	2	2

segment boundaries. Rough segment boundaries for test audio file are obtained from this method, and then a quick manual check and correction of these segments is done, if needed. We need these short audio segments because alignment errors propagation is expected to be less in short duration segments compared to relatively longer segments. From here on, each of the features are computed for each of these segments. The subjective evaluation for a test audio recording is assumed to hold for every segment of that recording. We have 80 such segments for the song "I have a dream", and 40 segments for the song "Edelweiss", in total 120 test singing segments.

We then compare each of the corresponding reference and test audio segments in terms of pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range related objective features, as listed in Table III. The methods to compute these features are described in Section III.

To investigate the factors that influence the performance of machine-based singing quality evaluation, we use combinations of the various objective features to design two baseline and 9 test evaluation systems (Table IV). Baselines A and B are the systems purely consisting of features extracted from the singing evaluation literature. Baseline A consists of pitch distance feature [5], [4], [21], [22] and Molina et al.'s pitch-based rhythm feature [5], while Baseline B has an additional volume distance feature [4], [21]. So Baselines A and B are the comparison benchmarks of this study. Also these systems would reveal the impact of the additional volume feature. Systems 1 and 2 are modified-baselines A and B respectively with the difference of the pitch-based rhythm feature [5] being replaced with the MFCC-based modified version (see Section III-B). These systems will provide insight about the definition of the objective feature for rhythm consistency, i.e. if the MFCC-based rhythm feature is better than the pitch-based version. System 4 contains PESQ-based L6+L2 norm features along with distance features but no L2-norm feature, while System 5 is the one with L2-norm features but without L6+L2 features. System 6 contains only the distance features. Systems 4, 5, and 6 should show the impact of the PESQ-

TABLE III
EVALUATION FEATURES GROUPED BASED ON THE SINGING CHARACTERISTICS (OR PERCEPTUAL FEATURES).

Perceptual feature	Feature Name	Description	Feature Group
Intonation accuracy	pitch_dist	DTW distance between pitch contours	dist
	pitch_der_L2	L2-norm of frame disturbances of DTW between pitch derivative contours	L2
	pitch_med_L2	L2-norm of frame disturbances of DTW between median subtracted pitch contour	L2
	pitch_der_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between pitch derivative contours	L6+L2
	pitch_med_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between median subtracted pitch contour	L6+L2
	pitch_der_dist	DTW distance between pitch derivative contours	dist
	pitch_med_dist	DTW distance between median-subtracted pitch contours	dist
Rhythm consistency	rhythm_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between MFCC vectors	L6+L2
	rhythm_L2	L2-norm of frame disturbances of DTW between MFCC vectors	L2
	molina_rhythm_pitch_dist	Rhythm distance computed by the method in [5]	dist
	molina_rhythm_mfcc_dist	Modified version of the method in [5] by computing rhythm distance using mfcc vectors instead of pitch	dist
Voice quality and Pronunciation	timbral_dist	DTW distance between MFCC features	dist
Appropriate Vibrato	vib_segment_dist	DTW distance between vibrato features of only the valid vibrato segments	dist
Volume	volume_dist	DTW between log energy contours	dist
Pitch Dynamic Range	pitch_dynamic_dist	Comparison of the difference between max and min pitch values	dist

based perceptual features, compared to the distance features commonly used in singing evaluation literature. System 3 consists of PESQ-based (L6+L2) features as well as all other distance and L2-norm based features, except for the rhythm distance feature of [5] and its modified version. System 7 adds the MFCC-based modified rhythm distance feature to System 3, while System 8 adds the pitch-based rhythm feature [5] to System 3. System 9 adds both these rhythm distance features to System 3. Comparison of Systems 3, 6, 7, 8, and 9 will provide insight about the interaction between the objective features that they comprise of, in terms of their performance in predicting the overall singing quality rating.

Systems 3, 4, 6-9 consist of combinations of PESQ-based, L2-norm, and DTW distance-based features. Thus the score generated from these systems is termed as the PESnQ score.

We build a Linear Regression (LR) model and a Multi-Layer Perceptron (MLP) model with one hidden layer for each of these systems using Weka [24], in two modes: A) train and test on overall singing quality score averaged over the 5 judges in 10-fold cross validation, and B) Leave-one-judge-out, i.e. train on 4 judges in 10-fold cross validation, and test on 1 judge. The R-squared correlation values (computed in Weka) between the various system outputs and human ratings are shown in Table V.

TABLE IV
THE OBJECTIVE FEATURES THAT DESCRIBE THE VARIOUS SINGING EVALUATION SYSTEMS.

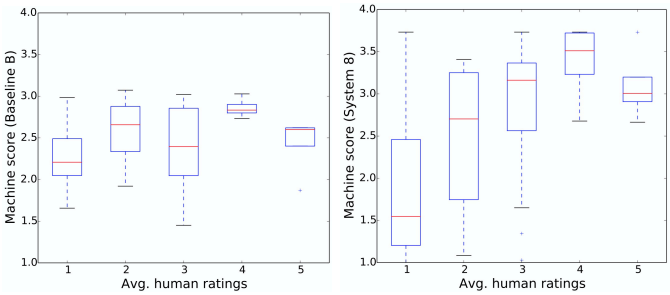
System Name	Description	Feature List
Baseline A	Consists of pitch distance and pitch-based rhythm distance [5] features	pitch_dist, molina_rhythm_pitch_distance
Baseline B	Consists of Baseline A features along with volume-based distance features	Baseline A + volume_dist
System 1	Modified Baseline A - modification: pitch-based rhythm distance feature [5] changed to MFCC-based rhythm distance feature	pitch_dist, molina_rhythm_mfcc_distance
System 2	Modified Baseline B - modification: pitch-based rhythm distance feature [5] changed to MFCC-based rhythm distance feature	Baseline B + volume_dist
System 3	Consists of L2, L6+L2, and dist features, except pitch-based [5] and MFCC-based rhythm distance features	rhythm_L2, pitch_der_L2, pitch_med_L2, rhythm_L6_L2, pitch_der_L6_L2, pitch_med_L6_L2, timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 4	Consists of L6+L2, and dist features, except pitch-based [5] and MFCC-based rhythm distance features	rhythm_L6_L2, pitch_der_L6_L2, pitch_med_L6_L2, timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 5	Consists of L2, and dist features, except pitch-based [5] and MFCC-based rhythm distance features	rhythm_L2, pitch_der_L2, pitch_med_L2, timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 6	Consists of only dist features, except pitch-based [5] and MFCC-based rhythm distance features	timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 7	Consists of union set of features from System 2 and System 3	System 2 + System 3
System 8	Consists of union set of features from Baseline B and System 3	Baseline B + System 3
System 9	Consists of union set of features from Baseline B, System 2 and System 3	Baseline B + System 2 + System 3

V. RESULTS AND DISCUSSION

From the subjective evaluation, we wanted to see if the parameters pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range perceptually combine to predict the overall singing quality score. We trained and tested the two models in the leave-one-judge-out mode (mode B) as mentioned in Section IV-C. We used the average subjective ratings for each of these parameters to predict the average subjective overall singing quality rating. Mode A, i.e. using the average score over all the judges, is not applicable in this case because of overlap between train and test data. The predictions showed the maximum average leave-one-judge-out correlation of 0.87 (Table V). This is the maximum correlation achieved amongst human judges, thus it is also the upper bound of the achievable performance of machine-based singing quality evaluation. We asked an optional question to the human judges to find out if there are other perceptual features that are important to singing quality assessment. Most of the answers were associated with one of these seven parameters, e.g. "key changes in the middle of the song" is indicated by the pitch accuracy parameter, etc. But there were a few comments which were indeed not covered in those seven parameters, such as "inability to sustain long notes". Nonetheless, with the high correlation between parameter-based prediction of overall score and the actual overall score, we can safely consider that the current set of seven perceptual features are

TABLE V
CORRELATION BETWEEN SYSTEM OUTPUT AND HUMAN OVERALL SINGING QUALITY RATINGS.

System configs	Average Overall Score		Leave out judge 1		Leave out judge 2		Leave out judge 3		Leave out judge 4		Leave out judge 5		Avg. leave-one-judge-out	
	LR	MLP	LR	MLP	LR	MLP	LR	MLP	LR	MLP	LR	MLP	LR	MLP
Human judge	-	-	0.93	0.96	0.87	0.87	0.78	0.75	0.95	0.88	0.83	0.83	0.87	0.86
Baseline A	0.30	0.26	0.36	0.39	0.30	0.34	0.45	0.36	0.31	0.36	0.35	0.37	0.35	0.36
Baseline B	**0.30**	0.29	0.36	0.40	0.30	0.40	0.45	0.36	0.31	0.39	0.35	0.36	0.35	**0.38**
System 1	**0.36**	0.27	0.38	0.50	0.35	0.55	0.43	0.40	0.37	0.50	0.40	0.36	0.39	**0.46**
System 2	0.34	0.30	0.38	0.36	0.35	0.39	0.43	0.30	0.37	0.36	0.34	0.32	0.37	0.35
System 3	0.48	0.55	0.61	0.66	0.57	0.56	0.54	0.54	0.61	0.66	0.46	0.51	0.56	0.59
System 4	0.50	0.55	0.61	0.64	0.57	0.56	0.54	0.53	0.60	0.65	0.48	0.48	0.56	0.57
System 5	0.49	0.55	0.61	0.65	0.57	0.56	0.54	0.53	0.61	0.66	0.49	0.50	0.56	0.58
System 6	0.53	0.47	0.58	0.62	0.56	0.68	0.52	0.52	0.58	0.67	0.41	0.43	0.53	0.58
System 7	0.48	0.53	0.61	0.66	0.57	0.67	0.51	0.53	0.61	0.71	0.46	0.54	0.55	0.62
System 8	0.42	**0.59**	0.61	0.68	0.58	0.70	0.54	0.57	0.61	0.73	0.45	0.54	0.56	0.64
System 9	0.43	0.56	0.61	0.69	0.59	0.72	0.53	0.57	0.61	0.74	0.47	0.56	0.56	**0.66**



(a) (b)

Fig. 8. Machine score vs. Average Human Rating for overall singing quality with (a) Baseline B (b) System 8.

good predictors of the overall singing quality. So we designed objective methods to obtain automatic scores for each of these parameters for building an automatic singing quality evaluation system.

Training and testing the various singing evaluation systems (Table V) on average overall score (Mode A) shows that System 8 performs the best with a correlation of 0.59 with the average human ratings, as compared to 0.30 of Baseline B. This shows that a combination of PESQ-based, L2-norm, and distance-based pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range related objective features, can predict the overall singing rating with an improvement of ∼96% over the current baseline systems (Baseline B) that use only a subset of these features. Figure 8 shows the outputs of these two systems versus the average human ratings. Ideally the machine scores should be directly proportional to the human ratings. This relation is visually much more evident in the plot of System 8, compared to that of Baseline B. This result shows that our designed features that take key transposition and rhythm variations into account have a positive impact on the system performance.

Addition of volume feature in Baseline B shows only a slight improvement over Baseline A. System 1, which is the modified version of Baseline A, performs better than the baselines. This shows that our MFCC-based modified rhythm distance feature performs better than the baseline pitch-based rhythm distance feature [5]. This supports our theory that

mistakes in pitch will degrade the baseline pitch-based rhythm distance feature. In our dataset, the subjects were proficient in English and had rehearsed the songs before recording. Thus, they made few mistakes in the lyrics while singing. However, they were restricted by their singing ability. Thus, the MFCC-based modified version of the baseline rhythm distance feature, which is robust to pitch errors, is more suitable in this case.

An interesting finding is that System 4 shows improvement over System 5. This is also evident when System 4 (PESQ-based and distance features) is compared to System 6 (only distance features). PESQ-based L6+L2 features provide an improvement of 3.7% over only distance features. Although earlier works relied on distance metric alone, our results show that adding features based on cognitive modeling theory improves machine correlation with human perceptual judgment.

The leave-one-judge-out experiments (Mode B) show that the output of our system trained on 4 judges correlates well with the 5th judge consistently. Thus, our system is able to generalize when trained on 4 judges. System 9 shows the best average correlation of 0.66. This is closer to the upper-bound of achievable correlation compared to the baseline system that shows correlation of 0.38. We also notice that the performance of some of our systems is comparable to that of the human judges. For example, System 9 shows correlation of 0.74 for leave-out-judge4 experiment, which is comparable to human judges' leave-out-judge3 correlation values. So, our system is close to reproducing judgments from a human music expert.

VI. Conclusion

We presented a framework for automatic perceptual evaluation of singing quality. From the subjective judgments of music experts, we found that pitch, rhythm, voice quality, vibrato, pronunciation, volume, and pitch dynamic range are the perceptual parameters that can reliably predict the overall singing quality. We designed objective features to automatically evaluate each of these perceptual features, while overcoming the challenges of the well-known baseline features. Our pitch evaluation features avoided penalizing for overall key transposition, and our rhythm evaluation features avoided penalizing for uniform rhythm variation, even when the pitch is off-tune. Also we designed features according to the cognitive modeling theory for audio perception in speech, used in the PESQ standard. We found that this theory could be applicable for singing evaluation also. Based on these features, we compared various systems trained to predict the overall singing quality. The predicted PESnQ score from a system having a combination of PESQ-based, L2-norm, and distance based features (System 8), showed a correlation of 0.59 with human ratings. This is a $\sim$96% improvement over the system built on baseline features. In future, we would explore the possibilities of indicating as feedback, the type and the precise location of the error, so that this framework can be developed into a comprehensive singing training tool.

Acknowledgment

We thank Dania Murad from Sound and Music Computing Lab, National University of Singapore, for her support in data collection and in building the subjective evaluation website.

References

[1] J. Wapnick, and E. Ekholm, "Expert consensus in solo voice performance evaluation," *J. of Voice,* vol. 11(4), pp. 429, 1997.

[2] J. M. Oates, B. Bain, P. Davis, J. Chapman, and D. Kenny,"Development of an auditory-perceptual rating instrument for the operatic singing voice," *J. of Voice,* vol. 20(1), pp. 71-81, 2006.

[3] C. Chuan, L. Ming, L. Jian, and Y. Yonghong,"A study on singing performance evaluation criteria for untrained singers," *9th ICSP,* vol. 20(1), pp. 1475-1478, 2008.

[4] W. H. Tsai, and H. C. Lee,"Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 20(4), pp. 1233-1243, 2012.

[5] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," *IEEE ICASSP,* pp. 744-748, 2013.

[6] C. H. Lin, Y. S. Lee, M. Y. Chen, and J. C. Wang,"Automatic singing evaluating system based on acoustic features and rhythm," *IEEE ICOT,* pp. 165-168, 2014.

[7] P. Lal, "A comparison of singing evaluation algorithms," *Interspeech,* 2006.

[8] T. Nakano, M. Goto, and Y. Hiraga,"An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," *Rn,* vol. 12, pp. 1, 2006.

[9] N. Amir, T. Erlich, N. Grabstein, and J. Fainguelernt, "Automated evaluation of singers' vibrato through time and frequency analysis of the pitch contour using the DSK6713," *16th IEEE Int. Conf. on Digital Signal Processing,* pp. 1-5, 2009.

[10] K. Omori, A. Kacker, L. M. Carroll, W. D. Riley, and S. M. Blaugrund, "Singing power ratio: quantitative evaluation of singing voice quality," *J. of Voice,* vol. 10(3), pp. 228-235, 1996.

[11] M. P. Hollier, M. O. Hawksford, and D. R.. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *IEE Proc. Vision, Image and Signal Processing,* vol. 141(3), pp. 203-208, 1994.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE ICASSP,* vol. 2, pp. 749-752, 2001.

[13] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol. 25(1), pp. 24-33, 1977.

[14] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," *IEEE ICASSP,* pp. 7815-7819, 2013.

[15] P. Prasert, K. Iwano, and S. Furui, "An automatic singing voice evaluation method for voice training systems," 音講論集 春季, pp. 911-912, 2008.

[16] P. P. G. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[17] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA,* pp. 59-64, 2001.

[18] A. De Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The J. of the Acoustical Society of America,* vol. 111(4), pp. 1917-1930, 2002.

[19] B. Bozkurt, "An automatic pitch analysis method for Turkish maqam music," *J. of New Music Research*, vol. 37(1), pp. 1-13, 2008.

[20] J. Sundberg, "The science of the singing voice," *Northern Illinois Univ. Press,* 226p., 1987.

[21] T. Tanaka, "Karaoke Scoring Apparatus Analyzing Singing Voice Relative to Melody Data," *U.S. Patent,* No. 5889224, 1999.

[22] P. C. Chang, "Method and Apparatus for Karaoke Scoring," *U.S. Patent,* No. 7304229, 2007.

[23] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," *IEEE APSIPA,* pp. 1-9, 2013.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11(1), pp. 10-18, 2009.