SIP (2018), vol. 7, e10, page 1 of 11 © The Authors, 2018.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2018.10

ORIGINAL PAPER

A technical framework for automatic perceptual evaluation of singing quality

CHITRALEKHA GUPTA[1,2], HAIZHOU LI[3] AND YE WANG[1,2]

Human experts evaluate singing quality based on many perceptual parameters such as intonation, rhythm, and vibrato, with reference to music theory. We proposed previously the Perceptual Evaluation of Singing Quality (PESnQ) framework that incorporated acoustic features related to these perceptual parameters in combination with the cognitive modeling concept of the telecommunication standard Perceptual Evaluation of Speech Quality to evaluate singing quality. In this study, we present further the study of the PESnQ framework to approximate the human judgments. First, we find that a linear combination of the individual perceptual parameter human scores can predict their overall singing quality judgment. This provides us with a human parametric judgment equation. Next, the prediction of the individual perceptual parameter scores from the PESnQ acoustic features show a high correlation with the respective human scores, which means more meaningful feedback to learners. Finally, we compare the performance of early fusion and late fusion of the acoustic features in predicting the overall human scores. We find that the late fusion method is superior to that of the early fusion method. This work underlines the importance of modeling human perception in automatic singing quality assessment.

Keywords: Singing Vocal, Perceptual Evaluation of Singing Quality, Automatic Evaluation, Human Perception

Received 6 February 2018; Revised 29 July 2018; Accepted 30 July 2018

I. INTRODUCTION

Singing quality assessment often refers to the degree to which a particular vocal production meets professional standards of excellence. Music experts assess singing on the basis of musical knowledge and perceptual relevance. However, personal biases and preferences are unavoidable in human judgment. Therefore, an automatic objective singing quality assessment framework can be a useful tool in this context, and could have applications such as karaoke singing scoring, and a practice tool for singing learners.

The assessment of how well a singer sings is based on the rules of singing that are derived from music theory. But the singing quality should also be pleasing to human listeners. Several studies have defined parameters that build a link between singing characteristics and human perceptual relevance [1–3]. Chuan et al. [3] defined six perceptual parameters that characterize the main properties of singing for assessing non-trained singers. These parameters were: *Intonation accuracy*, described as singing in tune; *Rhythm consistency*, described as singing with appropriate tempo speed; *Appropriate vibrato*, described as regular and smooth undulation of frequency of the tone; *Timbre brightness*, described as the brilliance of tone, a sensation of brightness of the spectrum; *Dynamic Range*, described as the pitch range that the subject is able to sing freely throughout, without an inappropriate change in voice quality or any external effort; and *Vocal Clarity*, described as vocal vibrations of a clear, well-produced tone.

For automatic singing quality assessment, it is important to identify the vocal attributes that relate to these human perceptual parameters and objectively define singing excellence. Thus, these perceptual parameters must be expressed in terms of the signal acoustics of singing. Various singing evaluation studies have incorporated different objective acoustic cues to represent the above-mentioned perceptual parameters. For example *Intonation Accuracy* is typically measured by comparing the pitch of a test singer with that of a reference singer or to the Musical Instrument Digital Interface notes [4, 5]. 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 most studies have only incorporated a set of objective acoustic cues that are relevant to a subset of the perceptual parameters for singing evaluation. For example, patents such as [6, 7] have used only volume and pitch as evaluation features, while scientific studies such as [4] have used volume, pitch, and rhythm features.

[1]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore
[2]Computer Science Department, National University of Singapore, Singapore
[3] Electrical and Computer Engineering Department, National University of Singapore, Singapore

Corresponding author:
Chitralekha Gupta
Email: chitralekha@u.nus.edu

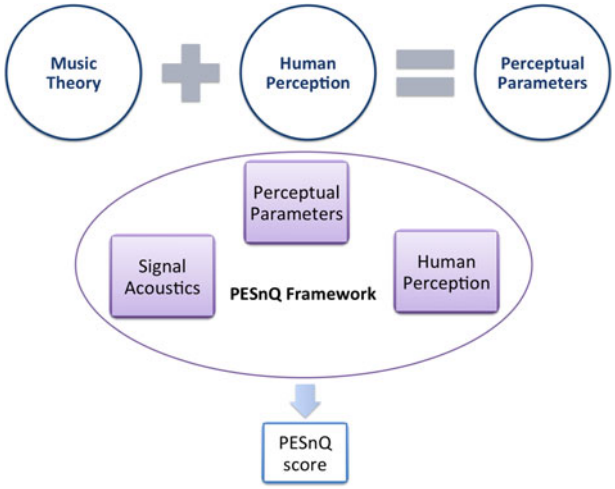

Fig. 1. The concept of the PESnQ framework. The perceptual parameters are motivated by the rules of singing as dictated by music theory and how humans perceive it. Our proposed PESnQ framework comprises elements from signal acoustics, perceptual parameters, and human perception to obtain a perceptually-valid score for singing quality called the PESnQ score.

Our previous work [8] attempts to provide a unified evaluation system that applies music theory and human perception to find the appropriate acoustic features for all the perceptually relevant parameters to obtain the overall evaluation score called Perceptual Evaluation of Singing Quality (PESnQ) score. Figure 1 summarizes the idea of PESnQ. We found that the acoustic features that are characterized by the cognitive modeling concept of the telecommunication standard PESQ [9], allow us to derive a perceptually valid singing quality score that correlates well with the music-expert judgments. This method of singing quality assessment also showed an improvement in performance by ∼96% over the baseline distance-based methods. However, there is a need for further analysis and exploration of different strategies for singing quality assessment within the framework of PESnQ. We believe that the evaluation framework should emulate human perception of judgment which will lead to an improvement in the performance of automatic singing quality scoring.

In this study, we provide a systematic analysis of PESnQ to further answer the following research questions:

(i) **Study of human perception for singing quality judgment:** In what way do humans combine the assessment of the perceptual parameters to give an overall singing quality score? Do the human scores for the individual perceptual parameters, i.e. intonation accuracy, rhythm consistency, appropriate vibrato etc. combine to predict the overall singing quality score?
(ii) **Prediction of perceptual parameters:** Based on the acoustic features derived in [8], can a machine predict the perceptual parameters individually?
(iii) **Strategies for overall singing quality scoring:** Can human perception inspired strategy for singing quality assessment result in better scoring? We study and compare two different strategies for scoring: early fusion and late fusion of acoustic features.

The paper is structured as follows. We first discuss our proposed human perception driven framework for evaluation in Section II. We then summarize the acoustic features that characterize singing quality [8] in Section III. Section IV describes our experiment methodology to evaluate our proposed framework along with the evaluation results.

II. FRAMEWORK OF EVALUATION

Singing quality evaluation was once considered to be a highly subjective task. But in the recent years, research has convincingly provided perceptual parameters based on which humans evaluate singing quality. These perceptual parameters are defined on the basis of music theory and human perception. For example, the perceptual parameter *intonation accuracy* involves the concepts of musical notes, pitch, and what a music-expert perceives as accurate intonation. Humans refer to intonation accuracy as the correctness of a produced pitch with respect to the intended musical notes or a reference singing pitch. Once the perceptual parameters are defined, the aim of an automatic singing evaluation framework is to objectively characterize these perceptual parameters based on signal acoustics to predict the singing quality score as would be given by music-experts based on their perception.

The framework for automatic singing quality evaluation can be broadly considered from two perspectives: human perception and signal acoustics. Human perception of singing quality is the subjective ground truth, whereas the signal acoustics provide the objective characterization of the perceptual parameters. The aim of the evaluation framework is to build a bridge between the objective characterizations and the subjective judgments. Figure 2 summarizes our valuation framework. In this section we discuss the strategies for building this framework with respect to these two perspectives.

A) Human perception

Our singing quality evaluation framework is motivated by human perceptual relevance. Our goal is to teach a machine to appreciate singing quality in the same way as human music-experts would do. Thus, based on the identified perceptual parameters for singing judgments, we have considered two different aspects of human perceptual judgments:

1) Cognitive modeling: localized versus distributed errors

According to the cognitive modeling literature, *localized errors* dominate the perception of audio quality [10], i.e. a highly concentrated error in time and frequency is found to have a greater subjective impact than a distributed error. The International Telecommunication Union standard for quality assessment of speech in telephone networks, PESQ (Perceptual Evaluation of Speech Quality) [9] applies this theory in measuring the perceived quality of a degraded speech signal with reference to its original version. That is, a higher weighting is used for localized distortions in PESQ

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

score computation. Motivated by this approach, we applied this concept of audio quality perception for singing quality assessment in our previous work [8] to obtain a novel PESQ-like singing quality score.

PESQ combines the frame-level disturbance values of a degraded audio with respect to the original audio by computing the L_6 norm over split-second intervals, i.e. 320 ms window, 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 weight to localized disturbances than distributed disturbances. In our previous work [8], we applied the same idea of L_6 and L_2 norm to the frame disturbances computed from the dynamic time warping (DTW) optimal path deviation from the diagonal, for a test singing with respect to the reference singing. We applied it to different pitch and rhythm acoustic features, as will be discussed in Section III.4. By incorporating the PESQ-like features, we show a significant improvement in performance over the baseline distance-based features.

2) Human parametric judgment equation

In the recent years, research has provided objective auditory-perceptual parameters that explains the physical and acoustic implications of human judgments. One such work [3] investigates the importance of every perceptual parameter and its contribution to the overall assessment of singing clips. They found that *intonation accuracy* is the most important perceptual contributor to human judgment, when assessing untrained singers.

In the cognitive psychology field of judgment and decision making, studies have found that people often construct a mental representation of the object, person, or situation about which they are making a judgment [11]. Before making a judgment, people construct an internally consistent model that explains as much of the judgment-relevant evidence as possible. This model is central to the judgment process, and plays a causal role in determining the ultimate decision and the degree of confidence that is associated with the decision.

According to music psychology studies, song perception-production in humans is a two-stage process. Humans first convert the perceived audio into a mental *symbolic representation*, and then convert it into a vocal-motor representation to produce the same sound [12]. This means that the underlying mental process of perception is in the form of a symbolic representation, similar to the cognitive psychology studies. This leads us to our hypothesis that humans follow a similar two-stage process to judge the overall singing quality. Humans first convert the perceived singing audio into a weighted representation of the identified perceptual parameters as the *symbolic representation*, and this representation is then mapped to the overall singing quality judgment score.

In cognitive algebra, such mental processes that determine human judgments are modeled as equations [11, 13, 14]. For example, Hoffman [13] fitted linear regression models to predict judgments on the basis of five to ten cues presented in the cases to be judged, concluding that the observable judgments were well fit by algebraic equations. This motivates us to explore the possibility of expressing overall singing quality judgment score in the form of a linear parametric equation in terms of the perceptual parameters. In this paper, we will validate this theory and formulate a parametric equation that models the overall human perceptual scoring as a function of a set of human perceptual parameters.

Table 1. List of perceptual parameters

Perceptual parameters	Symbol
Intonation accuracy (pitch)	*P*
Rhythm consistency	*R*
Appropriate vibrato	*Vib*
Volume	*Vol*
Voice quality	*VQual*
Pronunciation	*Pronun*
Pith dynamic range	*PDR*

We obtain human judgments for overall singing quality as well as for the seven perceptual parameters relevant to singing quality, that are summarized in Table 1 and discussed in detail in Section III. Our hypothesis is that the overall singing quality score can be approximated by a linear combination of the perceptual parameter scores, that will look like this:

$$\begin{aligned} score = {} & C_1 \times P + C_2 \times R + C_3 \times Vib + C_4 \times Vol \\ & + C_5 \times VQual + C_6 \times Pronun \\ & + C_7 \times PDR. \end{aligned} \tag{1}$$

We train the linear regression model given in equation (1) to obtain the weights C_i of each of these parameters as discussed further in the experiment in Section IV.A.

B) Signal acoustics

Literature shows that human judgment of overall singing quality is based on a set of perceptual parameters such as intonation accuracy, rhythm consistency, etc. [3]. These perceptual parameters can be represented by singing signal acoustic features. However, no studies have elaborated on how these perceptual parameters or the signal acoustics map to the overall singing quality judgment score.

Psychology studies show that the human speech/singing perception-production model converts the perceived audio into some form of symbolic representation, and then converts it into vocal-motor representation to produce or mimic the sound [12]. This indicates that the human perceptual model converts an audio signal into a parametric representation before making a judgment. Based on this theory, we explore methods of obtaining the singing quality judgment from the signal acoustics.

The identified perceptual parameters for singing quality assessment can be objectively represented through signal characterizations, as will be discussed in Section III. We explore two methods of mapping these objective features to

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

the human judgments: the *early fusion method*, where the acoustic features are directly mapped to the overall singing quality judgment; and the *late fusion method*, where the acoustic features are mapped to the perceptual parameters, that are further mapped to the overall judgment.

1) Early fusion

The idea of this method is to combine the acoustic features directly to predict human overall singing quality judgment score. This method is the standard way of computing the overall singing quality judgment score as reported in [4, 5, 8, 15]. In our previous work [8], we generated the overall singing quality judgment score from a linear combination of the cognitive model-based and distance-based acoustic features. We reported and compared the performance of various combinations of these acoustic features. In this work, we use the best performing feature set from our previous work and compare the performance of early and late fusion (discussed in the next sub-section) methods.

2) Late fusion

This method is inspired by the human perception-production model. As discussed in Section II.A.2, we believe that humans first convert the perceived singing audio into a weighted representation of the identified perceptual parameters, which is then mapped to the overall singing quality judgment score. In the same manner, we propose that our machine should first predict the perceptual parameters independently with the help of the acoustic features, and then apply the human parametric judgment equation from Section II.A.2 to fuse these perceptual parameters to give the final overall singing quality judgment score. We believe that this late fusion approach best resembles the process of how humans perceive and judge, and thus will lead to better results.

III. CHARACTERIZATION OF SINGING QUALITY

In this section, we further the discussion in [8] on the acoustic features that characterize singing quality, and relate these features to the perceptual parameters used for singing assessment. The acoustic features are derived from the musical components pertaining to the perceptual parameters. For a more elaborate discussion, please refer to our previous work [8].

3) Intonation accuracy

Pitch of a musical note is an acoustic feature defined as the fundamental frequency $F0$ of the periodic waveform. Intonation accuracy is directly related to the correctness of the pitch produced with respect to a reference pitch. As discussed in our previous work [8], we computed the perceptual feature *pitch_dist* as the DTW distance between the pitch contours of the reference and test singing as an indicator of *intonation accuracy*. This measure has been previously used in [4, 5, 15]. But this distance between pitch contours will penalize key transposition, although key transposition is allowed in case of singing without background accompaniments [3]. Hence we use the acoustic features pitch-derivative ($\Delta\mathbf{p}$) and median-subtracted pitch contours ($\mathbf{p}_{\mathbf{medsub}}$) to make the distance measure insensitive to key transposition. Pitch-derivative also emphasizes the transition period between notes, that represent ornaments such as glissando, that are indicators of singing quality. We then applied the cognitive modeling theory to these frame-level modified pitch vectors (pitch-derivative and median-subtracted pitch) to obtain the PESQ-like perceptual features for intonation accuracy (*pitch_der_L6_L2*, *pitch_der_L2*, *pitch_med_L6_L2*, *pitch_med_L2*) (Section II.A.1).

All pitch values in this study are calculated in the unit of cents (one semitone being 100 cents on equi-tempered octave), where 440 Hz (pitch-standard musical note A4) is considered as the base frequency.

4) Rhythm consistency

Rhythm consistency refers to the similarity of tempo between reference and test singing. A slight variation in tempo should be allowed [3], i.e. uniformly faster or slower tempo than the reference.

Molina et al. [15] and Lin et al. [20] proposed DTW between the reference and the test pitch contours as a procedure for automatic rhythm assessment, and accounted for rhythm variation. They 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 a straight line with an angle $\neq 45°$ 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 the root mean square error of this straight line fit from the optimal path (termed as *molina_rhythm_pitch_dist*) is the perceptual feature that represents rhythm error.

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 an inaccurate pitch, this algorithm will give a poor score for rhythm. Thus this method works well only when the test singer sings in correct pitch, but will give a large deviation from the optimal path if the pitch is inaccurate, despite good rhythm.

In [8], we proposed a modified version of Molina's rhythm deviation measure. Instead of assuming that the test singer sings in correct pitch, we assumed that the test singer sings with the correct sequence of words but probably incorrect pitch. Therefore, instead of using pitch contour, we use Mel-frequency cepstral coefficients (MFCC) feature vectors to compute the DTW between the reference and test singing vocals. MFCCs capture the short-term power spectrum of the audio signal that represents the shape of the vocal tract and thus the phonemes and the words 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

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

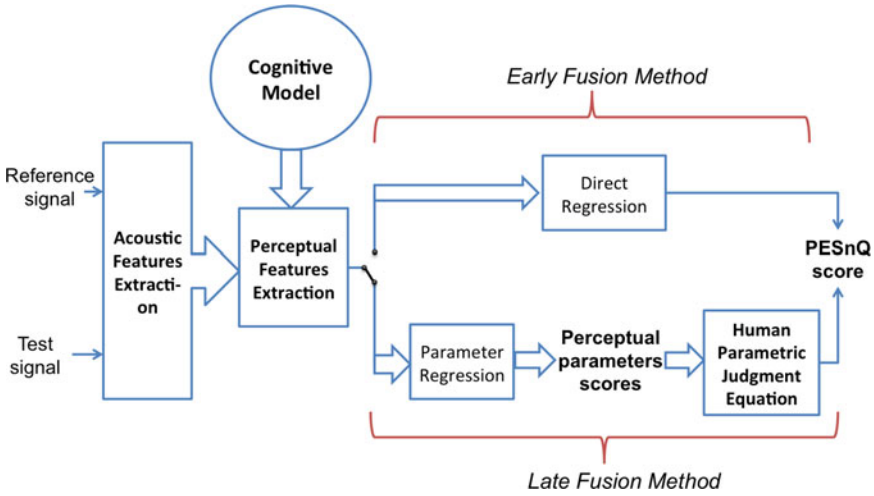

Fig. 2. The diagram of PESnQ scoring with different approaches: early fusion, and late fusion.

(termed as *molina_rhythm_mfcc_dist*) that measures the root mean square error of the linear fit of the optimal path of DTW matrix computed using MFCC vectors.

In [8], we also introduced another way to compute rhythm deviation, while accounting for allowable rhythm variations. We computed the MFCC vectors over a constant window size but at a different frame rate for reference and test singing, such that the number of frames in the reference and the test are equal. This way we compensated for consistent rhythm difference between the reference and test singing, and thus the number of MFCC vectors in reference and test are equal. Then we apply cognitive modeling theory to the DTW between these frame-equalized MFCC feature vectors of reference and test singing to obtain the perceptual feature for rhythm evaluation (*rhythm_L6_L2*, *rhythm_L2*) (see Section II.A.1).

5) Voice quality and pronunciation

Perception of timbre is related to the voice quality and is objectively represented by spectral envelope of the sound, which is captured well by MFCC vectors, as illustrated in [16]. MFCCs also represent phonetic quality, which relates to pronunciation. Thus, in [8], we computed the distance between reference and test singing timbre (termed as *timbral_dist*) by computing the DTW distance between their MFCC vectors, as a measure of voice quality and pronunciation.

6) Appropriate vibrato

Vibrato is the rapid periodic undulations in pitch on a steady note while singing. Presence of vibrato is considered to be an indicator of the quality of singing, thus we computed a measure for the same in our previous work [8]. We computed the vibrato likeliness, rate, and extent features for every valid reference vibrato section and the corresponding test pitch sections, and the Euclidean distance of these features between the reference and the test was defined as the vibrato perceptual feature (termed as *vib_segment_dist*) for evaluation.

7) Volume

It is expected that the volume variations across time for different singers performing the same song will show a similar pattern, thus is used as a common acoustic cue in existing systems [4, 6, 7]. In [8], we implemented the perceptual feature for volume as the DTW distance of the short-term log energy between the reference and the test (termed as *volume_dist*) for evaluation.

8) Pitch dynamic range

The pitch range that a subject is able to sing freely throughout is a good indicator of the quality of singing [3]. Thus in [8], we computed the absolute difference between the highest and the lowest pitch values in an audio segment as a measure for pitch dynamic range, and the distance of this measure between the reference and test singing (termed as *pitch_dynamic_dist*) was the perceptual feature of pitch dynamic range for evaluation.

Table 2 summarizes the acoustic and the perceptual features extracted corresponding to the perceptual parameters.

IV. EXPERIMENTS

For evaluating our hypothesis of a perception-driven singing evaluation framework, we conducted three experiments: (a) Study of human perception for singing quality judgment, (b) Prediction of Perceptual Parameters, (c) Strategies for scoring. In this section, we briefly discuss our dataset, and the three sets of experiments along with their results.[1]

Dataset

The dataset and the subjective evaluation used for the experiments is the same as in our previous work [8]. The audio dataset consisted of 20 recordings, each sung by a different singer with singing abilities ranging from poor to professional. Ten singers sang the song 'I have a dream' by ABBA

[1]The code base for feature extraction and the dataset can be found here: https://github.com/chitralekha18/PESnQ_APSIPA2017.git

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

Table 2. Acoustic features, Perceptual features, and their description corresponding to the human perceptual parameters for singing quality evaluation

Perceptual parameters	Acoustic features	Perceptual features	Description
A) Intonation accuracy	Pitch	pitch_dist	DTW distance between pitch contours
	Pitch derivative Ap	pitch_der_L2	L2-norm of frame disturbances of DTW between pitch derivative contours
		pitch_der_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between pitch derivative contours
		pitch_der_dist	DTW distance between pitch derivative contours
	Median subtracted pitch Pmedsub	pitch_med_L2	L2-norm of frame disturbances of DTW between median subtracted pitch contour
		pitch_med_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between median subtracted pitch contour
		pitch_med_dist	DTW distance between median-subtracted pitch contours
B) Rhythm consistency	Pitch	molina_rhythm_pitch_dist	Rhythm distance computed by the method in [6]
	MFCC	rhythm_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between MFCC vectors
		rhythm_L2	L2-norm of frame disturbances of DTW between MFCC vectors
		molina_rhythm_mfcc_dist	Modified version of the method in [6] by computing rhythm distance using mfcc vectors instead of pitch
C) Voice quality & pronunciation	MFCC	timbral_dist	DTW distance between MFCC features
D) Appropriate vibrato	Pitch	vib_segment_dist	DTW distance between vibrato features of only the valid vibrato segments
E) Volume	Energy	volume_dist	DTW between log energy contours
F) Pitch dynamic range	Pitch	pitch_dynamic_dist	Comparison of the difference between max and min pitch values

(∼2 min), and the other ten sang 'Edelweiss' from the movie 'The Sound of Music'. These songs were chosen as they have a steady rhythm, and are rich in long steady notes and vibrato.

Subjective Evaluation

We developed a website[2] to collect the subjective ratings for this dataset, where the task was to listen to the audio recordings and evaluate the singers' singing quality, compared with a professionally trained reference singer. The reference singing of both the songs was from one professional singer from the dataset of [17], different from our test singing evaluation dataset of 20 singers. Five professional musicians trained in vocal and/or musical instruments in different genres of music such as jazz, contemporary, and Chinese orchestra, were the human judges for this task.

The judges were asked to provide an overall singing quality score between 1 and 5 to each of the 20 audio recordings compared with the corresponding reference singing of the song. The judges were also asked to rate the recordings based on each of the perceptual parameters: 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.

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. Also, the inter-judge correlation of all the questions showed a correlation of higher than 0.60. Thus these parameters are judged by music experts consistently.

[2] https://slions.smcnus.org/welcome.php

We computed the average of the overall singing quality score as well as each of the perceptual parameter scores given to each of the 20 singers over the five human judges as the subjective evaluation ground-truths. We also observed from the overall average singing quality scores that our dataset is well representative of the entire singing skill spectrum (see [8] for details).

Pre-processing

As a pre-processing step, we first semi-automatically split every audio recording into shorter segments of approximately 20 sec duration, as discussed in our previous work [8]. This is done by aligning MFCC feature vectors of a test audio to the reference by DTW such that manually marked segment boundaries of the reference can be automatically aligned to the test audio files. We need these short audio segments because alignment errors propagation is expected to be less in short duration segments compared with relatively longer segments. From here on, each of the features is computed for each of these short segments and are called utterances. The subjective evaluation for a test audio recording is assumed to hold for every such utterance of that recording. We have 80 such utterances for the song 'I have a dream', and 40 utterances for the song 'Edelweiss', in total 120 test singing utterances.

A) Study of human perception for singing quality judgment

As discussed in Section II.A.2, based on psychology studies, we hypothesize that humans follow a two-stage process to judge the overall singing quality, i.e. we first convert

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

Table 3. Pearson's correlation between individual perceptual parameters human scores. (*vq*: voice quality, *pronun*: pronunciation, *pdr*: pitch dynamic range)

	Intonation	Rhythm	Vibrato	Volume	vq	pronun.	pdr	Overall
Intonation	1.0	0.919	0.910	0.786	0.890	0.793	0.965	0.961
Rhythm		1.0	0.863	0.847	0.860	0.758	0.947	0.923
Vibrato			1.0	0.684	0.923	0.839	0.871	0.960
Volume				1.0	0.716	0.677	0.856	0.731
vq					1.0	0.822	0.889	0.945
pronun.						1.0	0.783	0.848
pdr							1.0	0.945
Overall								1.0

the perceived singing audio into a weighted representation of the identified perceptual parameters, and then map this representation to the overall singing quality judgment score.

To evaluate this theory, we first observe the effect of the individual perceptual parameters on the overall singing quality scores in our dataset, and then train a regression model to estimate the weights of the parametric equation that approximates the overall score (equation (1)). This will give us insight about the important perceptual parameters that affect the human judgment. We also verify if the trends observed are comparable with that reported in the literature. This is an important validation step because these human ratings would serve as the references for the rest of our experiments. Table 3 shows the matrix of Pearson's correlation between the individual perceptual parameters. The last column of Table 3 shows the Pearson's correlation between the average of the overall ratings by five judges and the individual perceptual parameter scores.

Results and Interpretation

Table 3 shows that intonation accuracy is the highest contributing factor to the overall performance rating. Chuan et al. [3] also report this same observation. Moreover, we find that appropriate vibrato also has a strong correlation with the overall score. Chuan et al. [3] did not consider vibrato as an evaluation criterion in their experiments because it rarely appeared in their dataset that consisted of only untrained singers. However, our dataset consisted of a mix of trained and amateur singers, and literature suggests that appropriate vibrato is an important cue to distinguish between them [1, 2, 18]. This is also reflected from the strong correlation between intonation and vibrato (0.910), meaning that good intonation is likely to have good vibrato. Thus the high contribution of vibrato in determining singing quality is reasonable. The other parameters also show high correlation with the overall score, except the volume parameter. We believe that the reason for relatively weaker correlation for volume is lack of clear definition. As seen in Section I, volume is not considered as subjective assessment criteria in the music literature [2, 3]. However, the volume is one of the key features in the objective evaluation literature [4, 6] because this measure is easy to compute and pattern-match with a reference template, though difficult to rate subjectively. This explains the relatively weak correlation of volume with the overall rating.

Then we train a linear regression model in 10-fold cross validation using WEKA [19] to estimate the weights C_i of the equation (1) that predicts the overall singing quality score from the individual perceptual parameter scores. We check the correlation of these predictions with the subjective ground-truths. This experiment tests the possibility of defining singing quality judgment score in the form of a parametric equation. The linear regression model describes how humans relate the individual human perceptual parameters with the overall song-level scoring. The 10-fold cross validation resulted in a Pearson's correlation of 0.966 with the human overall judgment, and the linear equation that gives this prediction is:

$$\begin{aligned} score = {} & 0.161 \times P + 0.189 \times R + 0.297 \times Vib - 0.27 \\ & \times Vol + 0.181 \times VQual + 0.115 \times Pronun \\ & + 0.311 \times PDR - 0.061, \end{aligned} \tag{2}$$

where P, R, Vib, Vol, $VQual$, $Pronun$, PDR are the human scores for the perceptual parameters intonation accuracy (pitch), rhythm consistency, appropriate vibrato, volume, voice quality, pronunciation, and pitch dynamic range, respectively. This is the human parametric judgment equation for singing quality evaluation.

The weights of the perceptual parameters in this equation show the contribution of each of the parameters to the overall score. We see that intonation, rhythm, vibrato, voice quality, and pitch dynamic range are high contributors, while volume and pronunciation are low contributors. This trend is consistent with the trend observed in the last column of Table 3.

The strong correlation of the predicted overall score with the human ratings based on this linear parametric equation indicates the possibility that humans evaluate singing quality in a two-stage manner, i.e. first evaluate the individual perceptual parameters and then express the overall judgment as a weighted combination of these perceptual parameters. Therefore to build an automatic system that emulates this two-stage process of singing quality evaluation, we should first predict the perceptual parameters and then use this parametric judgment equation to predict the score.

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

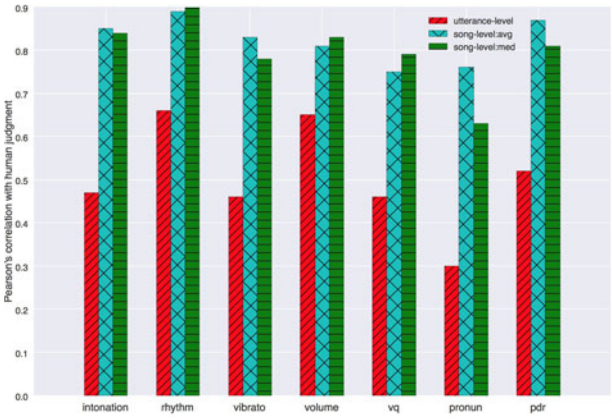

Fig. 3. Performance of the best set of acoustic features from [8] in predicting the individual perceptual parameters when trained separately for each of them, at utterance-level, and song-level (average and median). (*vq*: voice quality, *pronun*: pronunciation, *pdr*: pitch dynamic range).

B) Prediction of perceptual parameters

In this experiment, we would like to observe how well we can predict the individual perceptual parameter scores using the best system that we designed in [8]. This experiment is motivated by our theory that the human evaluation process can be best emulated by predicting the perceptual parameters first and then combining these parameter scores. Providing scores for individual perceptual parameters also means giving more meaningful feedback to a learner, i.e. providing a breakdown of an overall score in terms of musical parameters understood by humans.

In our previous work [8], we tested various combinations of cognitive as well as distance-based acoustic-perceptual features to predict the overall singing quality judgment. We found that the *System 8* that consisted of a combination of PESQ-like, L2-norm, and distance-based pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range features performed the best in predicting the overall singing quality rating, with a correlation of 0.59 with human ratings at utterance-level. This was an improvement of ∼96% over the baseline that only employs distance-based features.

Now we use the features of System 8 to train (10-fold cross validation) linear regression models separately for each of the individual perceptual parameters. Prediction is done at *utterance-level* and *song-level*. Song-level scores are computed in two ways: (a) by computing the *average* over all the utterance scores of a song, and (b) by computing the *median* of all the utterance scores of a song. The median-based song-level scores avoid the impact of outlier utterance-scores. Figure 3 shows the correlation of the predictions of the individual parameters with the average human scores.

Results and interpretation

Figure 3 shows that the song-level predictions correlate with the human scores more than the utterance-level predictions. Intuitively this trend can be explained by the fact that humans give their ratings after listening to the entire audio recording based on their overall impression. Short excerpts of the song may not represent the overall impression of the song, thus resulting in noisy scores. Thus a song-level automatic score for the song should correlate more with the song-level human judgment.

We observe that all the individual perceptual parameters are predicted well at the song-level with a correlation of more than 0.7. Moreover, the average and the median score correlation at song-level evaluation are comparable. This means that the utterance scores did not have many outliers, hence these song-level scores were not affected drastically.

These results show that it is possible to predict the individual perceptual parameter scores reliably and hence provide a more meaningful feedback to a learner in terms of a breakdown of an overall judgment score into perceptually and musically relevant scores. Moreover, having reliable objective measures for evaluation also helps in avoiding subjective biases that can creep into human judgments. For example, if a singer sings with good pronunciation, but has bad intonation, we observe that some judges tend to poorly rate pronunciation in such cases because humans tend to get influenced by the overall impression. Objective evaluation helps in avoiding such biases.

C) Strategies for scoring

In the experiment in Section IV.A, we found that the overall singing quality score can be modeled with a linear parametric equation, or equation 2. We showed that the human scores for the individual perceptual parameters collectively predict the overall scores via this model. This motivates us to apply this model on the machine scoring in Experiment B and obtain the overall singing quality score. This approach of late fusion for automatic overall scoring emulates the two-stage process of singing quality judgment by humans, as discussed in the previous sections. In this experiment, we compare this strategy with the early fusion strategy of scoring.

The acoustic feature set consists of the same features as in Experiment B, i.e. a combination of PESQ-like and distance-based features. We compare the overall song-level singing quality scoring by two methods:

(i) early fusion, i.e. by using the acoustic-perceptual features directly [8], and
(ii) late fusion, i.e. by using the individual parameter predictions from Experiment B and applying equation 2 for overall scoring.

Table 4 shows the Pearson's correlation of the predictions obtained from two methods by 10-fold cross validation, where the ground-truth was the overall singing quality score averaged over all the five judges. All the correlation values are statistically significant with $p < 0.001$.

Table 5 shows a leave-one-judge-out experiment, where the two methods predict the singing quality score of the 5th

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

Table 4. Comparison of pearson's correlation of the human overall judgment with the predicted overall PESnQ score by early and late fusion methods

	Early fusion	Late fusion
Song-level: average	0.725	0.904
Song-level: median	0.747	0.855

Table 5. Comparison of pearson's correlation of predicting the 5th judge in a leave-one-judge-out experiment by early and late fusion methods

	Leave out judge 1	Leave out judge 2	Leave out judge 3	Leave out judge 4	Leave out judge 5	Avg.
Human judges	0.929	0.872	0.780	0.948	0.831	0.872
Early fusion	0.745	0.703	0.696	0.730	0.622	0.699
Late fusion	0.785	0.722	0.726	0.780	0.683	0.739

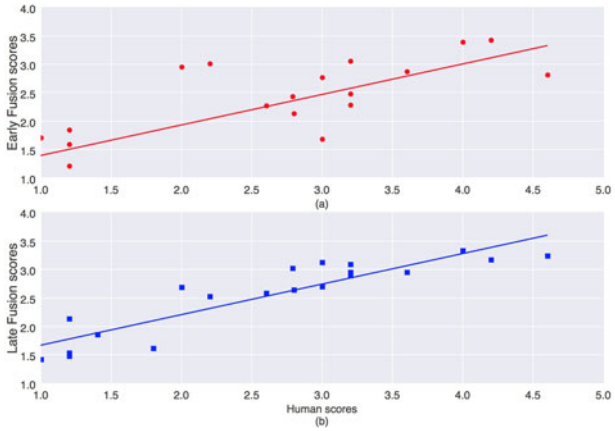

Fig. 4. (a) Early Fusion versus (b) Late Fusion to obtain the PESnQ score. Pearson's correlation of early fusion method is 0.725 and that of late fusion method is 0.904, both with statistical significance of $p < 0.001$.

judge i.e. train on four judges in 10-fold cross validation, and test on one judge.

Results and Interpretation

From Table 4, we find that the overall singing quality scoring from the late fusion method is superior to that from the early fusion method. We obtain a song-level averaged score prediction correlation of 0.904 with that of human scores by late fusion method, compared with 0.725 by early fusion method. This shows that the late fusion method emulates human perception better than the early fusion method. This means that our hypothesis that the late fusion approach most resembles the process of how humans perceive and judge singing quality is valid.

Average and median song-level scores show comparable performance like in Experiment B indicating that there may not be many outliers in the utterance-level scores. Figure 4 demonstrates the correlation of the machine song-level scores from early and late fusion with the human scores. We observe that the correlation in case of late fusion is higher and there is no gross misjudgment by the machine. This is an encouraging result as it highlights the significance of modeling human perception for building a machine that assesses singing quality.

We also verify that the correlation between the machine scores for the individual parameters and the overall singing quality by late fusion method (Table 6) follow a trend similar to that observed in human scores (Table 3). Intonation accuracy shows the highest correlation with the overall singing quality, while volume shows low correlation, same as in human scores. This similarity in trends implies that our designed objective features and perceptually-driven framework for evaluation generates perceptually reliable scores both for overall singing quality as well as for the related musical parameters understood by humans.

In the leave-one-judge-out experiment, we see that the maximum correlation achieved by the human judges to predict all the song ratings of an unseen judge is 0.87, thus indicating an upper bound of the achievable performance of any system that tries to emulate a human judge. The leave-one-judge-out experiment (Table 5) is different from the overall singing quality scoring experiment (Table 4) by the fact that leave-one-judge-out predicts the scores given by a judge for all the songs when trained on four judges, while the other experiment predicts the score of any unseen singing when trained on all the judges. We find that late fusion is better at predicting the unseen judge than the early fusion method, thus showing that the perception-driven framework is able to generalize better. We also notice that the performance of our framework is sometimes comparable with that of the human judges. For example, in the leave-out-judge3 experiment, machine performance is comparable with the human judges' performance in predicting the unseen judge. These results show that our framework of evaluation closely emulates a human music expert.

V. CONCLUSIONS

In this work, we presented a technical framework for the automatic PESnQ. We obtained a human judgment parametric equation that predicts the overall singing quality human score from the individual perceptual parameter human scores. We showed that the best set of acoustic features from our previous work [8] can predict the individual perceptual parameter scores reliably. So then we applied the human judgment parametric equation on the predicted perceptual parameter scores to obtain the prediction of the overall singing quality score. We show that the singing quality score obtained from the late fusion method has a higher correlation of 0.904 with human judgment than that of 0.725 from the early fusion method. Thus, we provided a systematic analysis to show that a framework that emulates the human perceptual process of singing quality assessment results in better scoring.

ACKNOWLEDGEMENTS

We thank NUS ALSET for financially supporting this work: Grant no. R-252-000-696-133. We also thank Dania Murad

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

Table 6. Pearson's correlation between individual perceptual parameter score predictions and overall singing quality (PESnQ) scoring by late fusion method. (*vq*: voice quality, *pronun*: pronunciation, *pdr*: pitch dynamic range)

	Intonation	Rhythm	Vibrato	Volume	vq	pronun	pdr	Overall
Intonation	1.0	0.84	0.801	0.494	0.707	0.647	0.855	0.952
Rhythm		1.0	0.811	0.704	0.824	0.613	0.830	0.912
Vibrato			1.0	0.634	0.831	0.555	0.627	0.856
Volume				1.0	0.788	0.639	0.574	0.537
vq					1.0	0.702	0.700	0.818
pronun						1.0	0.653	0.672
pdr							1.0	0.896
Overall								1.0

from Sound and Music Computing Lab, NUS, for her support in data collection and in building the subjective evaluation website.

ETHICAL STANDARDS

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The authors obtained approval from NUS Institute Review Board (NUS-IRB) for collecting the singing audio data, and ratings from the music-experts. The IRB approval number is Q17130.

REFERENCES

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

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

[3] Chuan, C.; Ming, L.; Jian, L.; Yonghong, Y.: A study on singing performance evaluation criteria for untrained singers, in *9th International Conference on Signal Processing, ICSP 2008*, 2008, Beijing, China, 1475–1478.

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

[5] Lal, P.: A comparison of singing evaluation algorithms. *Interspeech*, (2006).

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

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

[8] Gupta, C.; Li, H.; Wang, Y.: Perceptual Evaluation of Singing Quality, in *Proceedings of APSIPA Annual Summit and Conference*, Kuala Lumpur, Malaysia, 2017.

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

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

[11] Hastie, R.; Pennington, N.: Cognitive approaches to judgment and decision making. *Psychol. Learn. Motiv.*, **32** (1995), 1–31, Academic Press.

[12] Hutchins, S.; Sylvain, M.: The Linked Dual Representation model of vocal perception and production. *Front. Psychol.*, **4** (2013), 825.

[13] Hoffman, P.: The paramorphic representation of clinical judgment. *Psychol. Bull.*, **57** (2) (1960), 116–131.

[14] Stevenson, M.; Busemeyer, J.; Naylor, J.: Judgment and decision-making theory, in Dunnette, M. D.; Hough, L. M.: Eds., Handbook of industrial and organizational psychology, **vol. 1**, 2nd ed.. Palo Alto, CA: Consulting Psychologists Press, 1990, 283–374.

[15] Molina, E.; Barbancho, I.; Gómez, E.; Barbancho, A.M.; Tardón, L.J.: Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. *IEEE ICASSP*, (2013), 744–748.

[16] Prasert, P.; Iwano, K.; Furui, S.: An automatic singing voice evaluation method for voice training systems. 音講論集 春季, (2008), 911–912.

[17] Duan, Z.; Fang, H.; Li, B.; Sim, K.C.; Wang, Y.: The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. *IEEE APSIPA*, (2013), 1–9.

[18] Nakano, T.; Goto, M.; Hiraga, Y.: An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. *Rn*, **12** (2006), 1.

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

[20] Lin, C.H.; Lee, Y.S.; Chen, M.Y.; Wang, J.C.: Automatic singing evaluating system based on acoustic features and rhythm. *IEEE ICOT*, (2014), 165–168.

Chitralekha Gupta received the B.E. degree from Maharaja Sayajirao University (MSU) Baroda, India, and M.Tech. degree from Indian Institute of Technology Bombay (IIT-B), India in 2008 and 2011 respectively. She is currently a Ph.D. candidate at National University of Singapore (NUS). She has previously worked as a software developer at Dell R&D, and as a researcher at Airbus Defense and Space, Bangalore, India.

Her research interests are speech and music signal processing, music information retrieval, and their applications in education and health. She is currently working on automatic evaluation techniques for speech and singing voice signals.

Haizhou Li received the B.Sc, M.Sc, and Ph.D degrees from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990, respectively. He is now a Professor at the

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press

Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore.

Dr. Li is currently the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015–2018). He has served as the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of the Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012 and INTERSPEECH 2014.

Dr. Li was the recipient of National Infocomm Awards 2002, and President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by Nokia Foundation and an IEEE Fellow in 2014.

Ye Wang received the B.Sc from South China University of Technology, China in 1983, M.Sc from Braunschweig University of Technology, Germany in 1993, and Ph.D from Tampere University of Technology, Finland in 2002. He is an Associate Professor in the Computer Science Department at the National University of Singapore (NUS). Before joining NUS, he was a Member of the Technical Staff at Nokia Research Center in Tampere, Finland for 9 years.

His research interests are in the area of Sound and Music Computing (SMC); in particular, Music Information Retrieval (MIR), with an emphasis on applications in health and learning. Two active projects focus on the design and evaluation of systems to support (1) therapeutic gait training using rhythmic auditory stimulation (RAS); and (2) ear training and singing practice for rehabilitation and language learning.

https://doi.org/10.1017/ATSIP.2018.10 Published online by Cambridge University Press