Automatic Evaluation of Singing Quality without a Reference

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—**Automatic singing quality evaluation methods currently rely on reference singing vocals or score information for comparison. However singers may deviate from the reference singing vocal to personalize the singing that still sounds good. In this work, we present pitch histogram-based methods to automatically evaluate singing quality without any reference singing or score information. We validate the methods with the help of human ratings, and compare with the baseline methods of singing evaluation without a reference. We obtain an average Spearman's rank correlation of 0.716 with human judgments.**

I. Introduction

Singing quality is often judged with respect to professional standards of singing. Music experts make this judgment based on their music knowledge and perceptual appeal. Automatic singing evaluation systems, such as karaoke systems, compare a sample singing vocal with a reference such as a professional singing vocal[1], [2], [12] or the song melody notes [3], [4], [10] to obtain an evaluation score. However such methods of singing quality evaluation are constrained by the need for a professional grade reference singer of the song, or availability of the musical notes of the song.

Recently, online platforms such as Smule Sing![1], Starmaker[2], SoundCloud[3], and Youtube have become popular means to showcase singing talent. Amateur and promising singers upload cover versions of their favorite songs, that are listened and liked by millions across the globe. However discovering talented singers from such huge datasets is a challenging task [6]. Moreover, cover songs are often not intended to be exactly like the original song, rather they display the creativity of a performer's singing style, typically allowing them to recreate the song according to their own taste. Therefore reference-based methods for singing evaluation are not ideal in such cases as the singers may deviate from the original but are pleasing to hear.

Studies have shown that music experts can evaluate singing quality with high consensus when the melody or the song is unknown to them [13], [5]. This suggests that there are inherent properties of singing quality that are independent of a reference singer or melody, which help the music-experts to judge singing quality without a reference.

In this work, we aim to evaluate and rank singers without relying on a reference singer or a melody, with the help of music theory and statistical analysis. we propose new features based on pitch histogram, and study and analyse their behavior in assessing singing quality without reference. We obtain a ranking of all the singers singing a particular song and validate our results with human judgment. This paper is organized as follows. In Section II, we summarize the related work, in Section III, we propose and discuss measures to characterize singing quality based on pitch histogram. Data preparation, including a scalable method to obtain subjective ground-truths, is discussed in Section IV. Finally, we analyze and test the validity of our proposed measures in Sections V, and VI.

II. Related Work

Singing quality has been attributed to several perceptual parameters by the music experts, such as intonation, rhythm, vibrato, timbre, pitch dynamic range etc. [7], [8], [9]. Several studies have shown that out of all of these parameters, intonation accuracy is the highest contributing factor for the overall singing quality rating [9], [1]. Thus in this study we focus on intonation for evaluation.

In reference-based singing evaluation studies, intonation accuracy has been objectively related to the correctness of the pitch (i. e. the fundamental frequency of a periodic waveform) produced with respect to a reference pitch [1], [3], [10], [11], [12]. However, in the absence of a reference, it is a challenge to assess the correctness of pitch.

Only a few studies have attempted to evaluate singing quality without a reference. Nakano et al. [5] used pitch interval accuracy and vibrato-related features to evaluate singing without reference, showing 83.5% accuracy in binary classification of singing quality. For computing the pitch interval accuracy, the fundamental frequency trajectory is fitted to a semitone (100 cents) width grid (corresponding to equal temperament in the Western Music Tradition), i.e. all the pitch values are wrapped on to a semitone. If the pitch values have a constant offset from this semitone grid throughout the song sequence, then the singing was considered to be of good quality. Although pitch interval accuracy is a fair indicator, it ignores other properties of a song. For example, if a singer sings only a single note throughout the song, pitch interval accuracy will classify it as good singing. Therefore it overlooks features such

[1] https://www.smule.com/
[2] https://www.starmakerstudios.com/
[3] https://soundcloud.com/

as occurrence of several notes in a song and different notes being sustained for different durations.

A pitch histogram wrapped on to a 12 semitones (1200 cents) grid preserves the information about the number of frequently hit notes in a song (as discussed in the next section in detail). Furthermore, sharp peaks in the pitch histogram capture note sustenance and thus indicate consistency in hitting the notes independent of any reference. To measure the sharpness of the peaks, Nichols et al. [6] computed kurtosis and skew of the pitch histogram. These are overall statistical indicators but they do not capture the actual shape of the histogram. Therefore there is a need to characterize the finer details of the pitch histogram, that will allow a better understanding of the quality of singing without a reference.

III. Singing Quality Characterization Without a Reference

Pitch is an important perceptual parameter for singing quality evaluation. All pitch values in this study are calculated in the unit of cents (one semitone being 100 cents on equi-tempered octave),

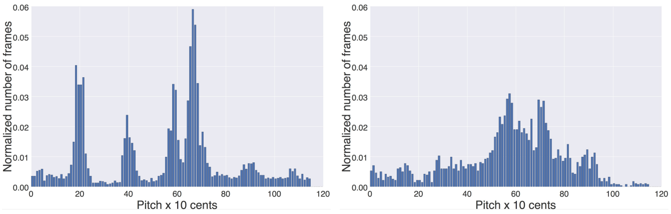
$$f_{\text{cent}} = 1200 \times \log_2 \frac{f_{\text{Hz}}}{440}, \tag{1}$$

where 440 Hz (pitch-standard musical note A4) is considered as the base frequency.

Pitch histograms are global statistical representations of the pitch content of a musical piece [14]. They represent the distribution of pitch values in a sung rendition. Features calculated from them have been used for genre classification, similarity retrieval, as well as singing evaluation [14], [6]. A pitch histogram is computed as the count of the pitch values folded on to the 12 semitones in an octave. To obtain a finer representation, we further divided each semitone into 10 bins. Thus we have 12 semitones x 10 bins each = 120 bins in total, each representing 10 cents.

In order to evaluate a singer without a reference, we must rely on the inherent discerning qualities of singing that distinguish good singing quality from poor singing quality. Since pitch plays a primary role in singing judgment, we focus on the properties of the pitch in a song for this work. The melody of a song typically consists of a set of dominant musical notes (or pitch values). These are the notes that are hit frequently in the song and sometimes are sustained for long durations. These dominant notes of the song are a subset of the 12 semitones present in an octave. The other semitones may also be sung during the transitions between the dominant notes, but are comparatively less frequent and not sustained for long durations. Thus on plotting a pitch histogram of a good singing vocal of a song, wrapped on to an octave, these dominant notes appear as the peaks, while the transition semitones appear in the valley regions.

Figure 1 shows an example each of a good singing vocal and a poor singing vocal pitch histograms, both singing the same song. The histogram heights are normalized to sum to 1. Note that the good singer histogram has sharp peaks showing that the singer frequently and consistently hits certain pitch values more than the rest of the pitch values. Since generally a song consists of only a set of dominant notes, the "spikiness" of the pitch histogram of the good singer indicates that the notes of the song (wrapped onto an octave) are being hit repeatedly and consistently. On the other hand, the poor singer has a dispersed distribution of pitch values indicating that the singer is unable to hit the dominant notes of the song consistently.



(a) (b)

Fig. 1. Normalized Pitch Histogram for (a) good singing (b) poor singing.

We formulate and analyze the following seven statistical measures for singing quality evaluation when the song or melody is unknown to characterize these discerning properties of the pitch histogram. In our experiments, we test the reliability of each of these measures with the help of subjective evaluation.

A. Kurtosis

Kurtosis is a statistical measure used in the literature for quantifying the quality of singing without a reference [6]. Kurtosis is the fourth standardized moment, defined as

$$Kurt = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \tag{2}$$

where X is the data vector, μ is the mean and σ is the standard deviation of X.

Kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution. A good singer's pitch histogram is expected to have several dominant spikes, as in Figure 1, and thus away from a normal distribution. So a good singer would show a higher kurtosis value than a poor singer.

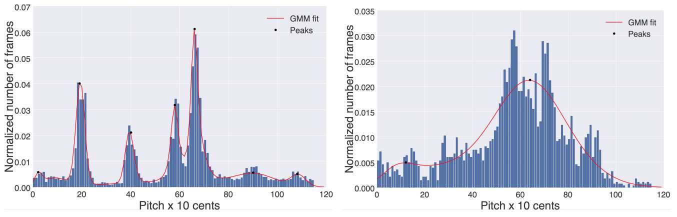
B. Skew

Skew is another measure used in the literature for singing quality assessment [6]. It is a measure of the asymmetry of a distribution with respect to the mean, defined as

$$Skew = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{3}$$

where X is the data vector, μ is the mean and σ is the standard deviation of X.

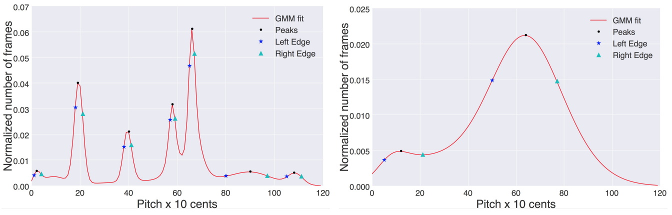
The pitch histogram of a good singer has peaks around the note locations of the song, whereas the histogram of a poor singer is expected to be more dispersed and spread out evenly. So the pitch histogram of a good singer is expected to be more asymmetric with respect to the mean than that of a poor singer.



(a)

(b)

Fig. 2. GMM-fit and detected peaks on the Normalized Pitch Histogram for (a) good singing (b) poor singing (the y-axis scales are different for better visibility).



(a)

(b)

Fig. 3. Edges of the detected peaks in the GMM-fit on the Normalized Pitch Histogram for (a) good singing (b) poor singing (the y-axis scales are different for better visibility).

C. GMM-fit

Both kurtosis and skew measures consider the overall distribution of the pitch values with respect to a normal distribution. However they do not consider the shape of the histogram in detail, i.e. the shape (spread and height) of the peaks, and the number of peaks in the histogram. These details provide more insights about the quality of singing. For example, a low spread around the peaks (that is, sharp peaks) indicates that the note was consistently hit by the singer, and vice versa. Similarly if a large percentage of pitch values are in the peaks, then it indicates more consistently sung notes, thus good singing. Moreover, typically a Western popular song (not Rap or Metal) is expected to have more than two or three dominant notes. So singers showing very few peaks in their pitch histogram would indicate poor singing.

To capture these fine details of the histogram, we wanted to model the histogram with something more than just a normal distribution. We fitted a mixture of Gaussian distributions to model the pitch histogram. A Gaussian Mixture Model (GMM) should be able to fit a histogram with several dominant peaks, as well as a dispersed histogram, thus providing a less noisy approximate representation of the histogram. After experimenting with different numbers of mixtures, we found that a high number of mixtures are required for fitting the histogram of good singers as they have many concentrated sharp peaks. Therefore, empirically we set the number of mixtures as 150. Figure 2 shows the GMM-fit for the good and the poor singer.

The idea is to design a measure that characterizes the shape of the histogram, i.e. the peaks and the valleys, that captures the inherent discerning characteristics of singing quality. To characterize the peaks in the histogram, we first detect the local maximas in the GMM-fit. A point is considered to be a peak candidate if it has the maximal value, while being preceded and succeeded by a lower value [15]. Empirically, a peak candidate is considered to be the actual local maxima if it is the highest peak within at least ±50 cents. Figure 2 shows the detected local maximas.

We characterize singing quality on the basis of the detected peaks in the two following ways.

1) Peak-Bandwidth Measure: The spread around the peaks indicates the consistency of hitting the same notes. The smaller the spread, the higher the consistency, and therefore better the singer. This hypothesis has been previously explored on a semitone grid by [5] and [6], but we apply it on the GMM-fit of the 12 semitones pitch histogram.

To compute the spread around a peak or the peak bandwidth (BW), we consider the half power down point or 3dB bandwidth (i.e. peak amplitude/$\sqrt{2}$ as the left and the right edges around the peak), as shown in Figure 3. Half of the extent between the right and the left edges is termed as the peak standard deviation σ. Therefore a measure of poor singing quality is directly proportional to σ^2 , i.e. larger the spread poorer is the singing quality. Moreover, since a pop song is expected to have more than one or two significant peaks in the pitch histogram, the measure of poor singing quality should be inversely proportional to the number of peaks detected N. Thus we define the average peak-BW measure as:

$$PeakBW = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_i^2}{N} \tag{4}$$

where σ_i^2 is the variance of the i^{th} detected peak.

Note that this measure has an inverse relation with the singing quality, that is, lower the value of $PeakBW$, better is the singing quality.

2) Peak-Concentration Measure: The percentage of pitch values at and around the peaks measures the concentration of pitch values in the peaks. That is, it indicates the amount of time a singer spends on singing the intended notes compared to the non-intended notes. If this percentage is high, it means that most of the pitch values are concentrated around the peaks, indicating that the singer hits the same notes consistently and does not spend time singing the other notes. This measure takes the height of the peaks into consideration, which is also an indicator of the duration of the sustained long notes of the song. We define peak-concentration measure as

$$PeakConc = \frac{\sum_{j=1}^{N}\sum_{i=bin_j-\Delta}^{bin_j+\Delta} A_i}{\sum_{k=1}^{M} A_k} \tag{5}$$

where N is the number of peaks, bin_j is the bin number of the j^{th} peak, A_i is the histogram value of the i^{th} bin, and M is the total number of bins, i.e. 120 here. Δ is the number of bins on either sides of the peak to be considered for measuring peak concentration. We have considered ±5 and ±2 bins, i.e. a total of 110 cents and 50 cents respectively around a peak. We term these measures as $PeakConc_{110}$ and $PeakConc_{50}$ respectively.

D. k-Means Clustering

The density of pitch values across the histogram bins is an indicator of how well the pitch values are clustered together. Tightly grouped clusters indicate that most of the pitch values are close to the cluster centers which means the same notes are hit consistently. Keeping this idea in mind, we apply k-Means clustering to the pitch values such that 12 clusters are formed. We chose k=12 for the 12 semitones in an octave. k-Means clustering algorithm optimizes the cluster centroids and boundaries by minimizing the sample distances within the clusters, while maximizing the distances between the clusters [17], [16].

Figure 4 shows the 12 cluster centroids for the two types of singing quality, good and poor. If a centroid is located close to a peak in the histogram, it implies that a large number of samples (or pitch values) have a small distance from the centroid. Moreover, when two centroids are closely spaced, the average distance of the samples from the centroid in each of those clusters will be less. We can see that the centroids around the highest peaks of the good singer's histogram are closely spaced, implying smaller sample distances. Therefore, whether the pitch values are tightly or loosely clustered can be represented by the average distance of each pitch value to its corresponding cluster centroid. So this distance is inversely proportional to the singing quality, i.e. smaller the distance, better the singing quality. We define the average cluster distance as

$$kMeans = \frac{1}{N}\sum_{i=1}^{k} d_i^2 \qquad (6)$$

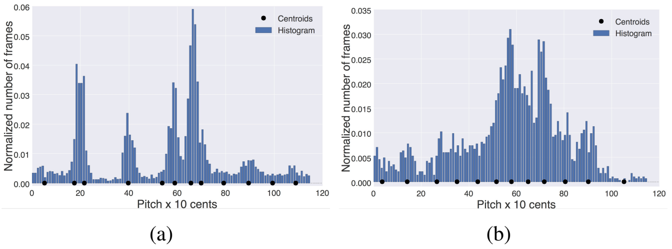
where N is the total number of pitch values (or frames with valid pitch values), and d_i is the total distance of the pitch values from the centroid in i^{th} cluster, defined as

$$d_i^2 = \sum_{j=1}^{N_i} \left(x_{ij} - c_i\right)^2 \qquad (7)$$

where x_{ij} is the j^{th} pitch value in i^{th} cluster, c_i is the i^{th} cluster centroid obtained from the k-Means algorithm, N_i is the number of pitch values in i^{th} cluster, and i ranges from $1, 2, ..., k$ number of clusters.

E. Binning

Another way to measure the clustering of the pitch values is by simply dividing the 1200 cents (or 120 pitch bins) into 12 equi-spaced semitone bins, and computing the average distance of each pitch value to its corresponding bin centroid. The bin centroid is the average of the pitch values present in that bin. This method is simpler in computation than the k-means clustering method. Equations 6 and 7 hold true for this method too, the only difference being that the cluster boundaries are fixed in binning method at 100 cents (or 10 pitch bins).



(a) (b)

Fig. 4. Centroids of the 12 k-means clusters along with the Normalized Pitch Histogram for (a) good singing (b) poor singing (the y-axis scales are different for better visibility).

Thus we have seven objective statistical measures for evaluating singing quality without a reference: $Kurt$, $Skew$, $PeakBW$, $PeakConc_{110}$, $PeakConc_{50}$, $kMeans$, and $Binning$, out of which $Kurt$ and $Skew$ are baseline measures.

In the next sections, we evaluate the performance of each of these objective measures for singing quality evaluation without a reference. We first discuss our dataset (Section IV-A), then we describe our method to obtain subjective ground-truth annotations on a large scale using a crowd-sourcing platform (Section IV-B). Next, we discuss the performance of our proposed measures in evaluating the songs in our dataset without a reference, and validate them against the subjective ground-truths (Sections V and VI).

IV. Data Preparation

A. Audio Dataset

Our dataset consists of 5 popular Western songs. All the songs are rich in steady notes and rhythm, as summarized in Table I. The dataset consists of a mix of songs with long and sustained as well as short duration notes. They also have a range of different tempos (beats per minute).

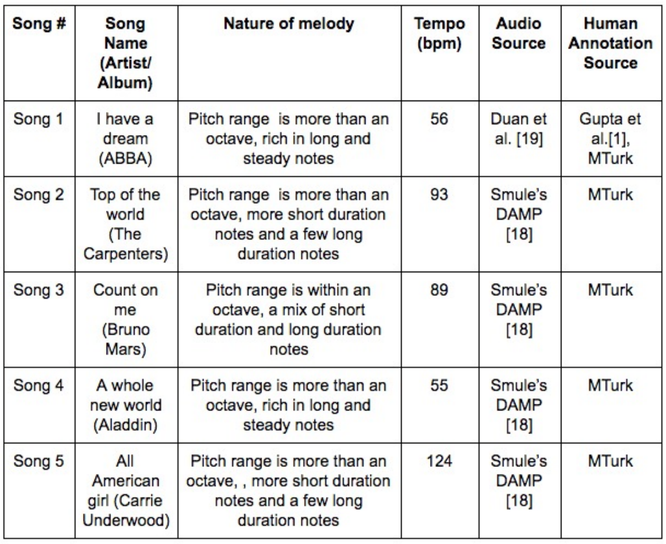
10 different singers sang each song. The singers were subjectively chosen such that the entire spectrum from poor to good singing quality is represented. The singers for 4 out of the 5 songs were taken from Smule's DAMP dataset [18]. And singers for one song "I have a dream" were taken from the singing corpus used in [1]. The reason for this distribution is because the song from [1] is annotated by professional musicians for singing quality, that we wanted to use for validation purposes, as described in the next sub-section.

B. Collecting Subjective Ground-truths

To validate objective measures for singing evaluation, we need subjective ratings as ground-truth. Reliable subjective ratings for singing quality can be provided by trained or professional music experts. However, obtaining such ratings at a large scale is a challenging task. Music experts may not be easily available, and the process of obtaining these ratings from them is time consuming, and expensive.

In our previous work [1], we recruited 5 professional musicians to provide singing quality ratings for 10 singers singing the song "I have a dream". The ratings were on a likert scale of 5, for overall singing quality. These judges were

TABLE I
SUMMARY OF THE SONGS USED IN THE DATASET.



Song #	Song Name (Artist/ Album)	Nature of melody	Tempo (bpm)	Audio Source	Human Annotation Source
Song 1	I have a dream (ABBA)	Pitch range is more than an octave, rich in long and steady notes	56	Duan et al. [19]	Gupta et al.[1], MTurk
Song 2	Top of the world (The Carpenters)	Pitch range is more than an octave, more short duration notes and a few long duration notes	93	Smule's DAMP [18]	MTurk
Song 3	Count on me (Bruno Mars)	Pitch range is within an octave, a mix of short duration and long duration notes	89	Smule's DAMP [18]	MTurk
Song 4	A whole new world (Aladdin)	Pitch range is more than an octave, rich in long and steady notes	55	Smule's DAMP [18]	MTurk
Song 5	All American girl (Carrie Underwood)	Pitch range is more than an octave, , more short duration notes and a few long duration notes	124	Smule's DAMP [18]	MTurk

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 and/or music teachers. The subjective ratings obtained from them showed high inter-judge correlation (0.82), and can be considered as the ideal ground-truth. However the process of recruiting these judges is expensive and time consuming.

To collect reliable human judgments for singing quality in a scalable way that is also cost effective, we propose a method to leverage on crowd-sourcing platforms such as Amazon mechanical turk (MTurk). To the best of our knowledge, crowd-sourcing platforms have not been used for singing quality judgments before. We would like to study how to obtain reliable singing quality judgment data from MTurk. A method of proving reliability of the MTurk data is to observe the correlation between the MTurk data and that from a laboratory-controlled experiment [20].

Best-Worst Scaling

Due to their music training, professional musicians are able to rate singing quality on a scale of 5 reliably and consistently. However on crowd-sourcing platforms, we cannot be very sure of their absolute ratings. Absolute rating methods on a likert scale are known to have problems [21], [22]. They are supposed to be less useful, because the judges are not forced to discriminate between items, so they are likely to give similar ratings for multiple items. Moreover, the scale is arbitrary, i.e. each rating on the scale is not precisely defined. Nevertheless, the ratings that we obtained in a lab-controlled environment showed good distribution over the full range of ratings, and high inter-judge correlation because we ensured that the judges were professional musicians. But it is not possible to ensure this on crowd-sourcing platforms.

The Best-Worst Scaling (BWS) is a popular method to obtain human judgments for rank ordering a set of items according to their preference. The judges in this case are asked to select the best and the worst option from a given small set of items. This is repeated over all the combinations of item sets. At the end of this exercise, the items can be rank ordered according to the aggregate BWS scores of each item, given by

$$BWS_{score} = \frac{N_{best} - N_{worst}}{N} \tag{8}$$

where N_{best} and N_{worst} are the number of times the item is marked as best and worst respectively, and N is the total number of times the item appears.

The BWS scores will tell us the order and the strength of importance of all items. A positive BWS score for an item means it is chosen as the most appealing more often than the least appealing, and vice versa. A zero score means it is chosen as the most and least appealing an equal number of times or it has never been chosen as the most and least appealing [22].

This method overcomes the problems of the absolute rating methods, because people in general are good at picking the extremes, but their preferences for anything in between might be fuzzy and inaccurate [22], as discussed above. Thus we use this method to obtain singing judgments from MTurk users.

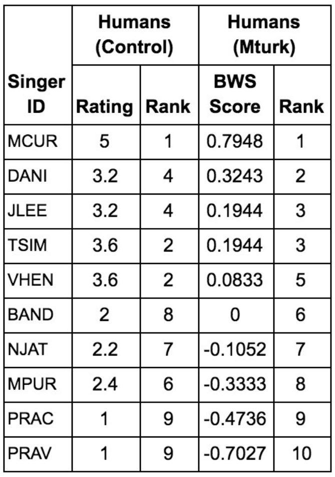
MTurk Data Reliability Test

We conducted an MTurk experiment where we prepared sets of three different singers singing the same song. We asked listeners to choose the best and the worst tracks in each of the sets based on singing quality. There are ${}^{10}C_3$ number of ways to choose 3 singing tracks from 10 singers of a song, i.e. 120 sets. This experiment was conducted separately for each of the 5 songs of Table I. Therefore there were in total $120 \times 5 = 600$ sets.

We applied filters to the MTurk users by asking for their experience in music and asked them to annotate musical notes. We accepted their attempt only if they had some sort of formal training in music, and could write the musical notations successfully. For example, a user whose attempt was accepted had mentioned in his/her music skill description, “I am a classical voice teacher, and double bass teacher. I also play the piano and sing in public quite often.” We made exceptions for the music notations if they mentioned that they were trained in music, but haven’t learnt the Western music notation style. We also applied a filter on the time spent in performing the task to remove the less serious attempts where they may not have spent time listening to the tracks. Empirically we set the time threshold as 50 seconds, i.e. an attempt is accepted only if it took more than 50 seconds to complete. We paid US $0.01 for every user-attempted set that was valid according to our filters.

With the help of the BWS method, we obtained BWS scores and ranks of the 10 singers for each of the 5 songs. We first wanted to verify our hypothesis that the BWS method for singing judgment using MTurk can provide reliable scores and ranks. So we correlated the BWS scores and ranks of the Song

TABLE II
CONTROL VS MTURK: PEARSON'S SCORE CORRELATION IS 0.931; SPEARMAN'S RANK CORRELATION IS 0.859.



Singer ID	Humans (Control)		Humans (Mturk)	
	Rating	Rank	BWS Score	Rank
MCUR	5	1	0.7948	1
DANI	3.2	4	0.3243	2
JLEE	3.2	4	0.1944	3
TSIM	3.6	2	0.1944	3
VHEN	3.6	2	0.0833	5
BAND	2	8	0	6
NJAT	2.2	7	-0.1052	7
MPUR	2.4	6	-0.3333	8
PRAC	1	9	-0.4736	9
PRAV	1	9	-0.7027	10

1 'I have a dream" with that of the lab-controlled music expert judgments of that song obtained from [1]. As shown in Table II, the human ratings from the professional musicians has a high Pearson's correlation of 0.931 with the BWS scores from MTurk users. The corresponding Spearman's rank correlation is 0.859. This high correlation shows that BWS method for singing judgment can provide a scalable and cost effective solution to the problem of obtaining subjective annotations for singing evaluation. Thus we use the BWS scores from the MTurk users for the rest of the songs.

V. EXPERIMENTS

In this section we study the behavior of the proposed measures for evaluating singing quality without a reference. We describe the procedure of objectively evaluating singing quality without a reference in terms of the seven statistical measures discussed in Section III. We obtain singing evaluation scores and ranks of singers from the baseline and the proposed evaluation measures and compare them against the human judgments obtained in Section IV-B.

First, we estimate the pitch contour of the singing vocal. For monophonic singing, the autocorrelation-based PRAAT [24] pitch estimator is reported to give the best voice boundaries with minimal post-processing [23]. We use PRAAT to obtain the pitch estimates (in cents) with one generic post-processing step to remove unreliable pitch values. We remove the frames with low periodicity which is determined by the harmonic-to-noise ratio (HNR), as discussed in [1]. By choosing only the voiced segments and removing the frames with low periodicity, spurious $F0$ (pitch) values are avoided and only reliable pitch values are used.

In this study, we solely focus on the steady notes, and not on the quality of vibrato. Therefore we would like to minimize the effect of vibrato and any other spurious values. Hence we pass the obtained pitch contour through a butterworth low pass filter with a cut-off frequency of 2 Hz set empirically such that the vibrato regions are smoothened, while ensuring that pitch values around note transitions are not lost.

For evaluating singing quality without a reference, the key of the song would be unknown. To avoid any effect of key difference between singers, we subtract the median pitch value from the pitch contour. This step brings down the pitch contour about 0 cents.

We wrap these pitch values on to one octave, i.e. 1200 cents, and compute the 120 bin pitch histogram, as discussed in Section III. We compute all the seven measures of evaluation for every singer of the 5 songs and compare the singer ranks obtained for each song from these measures with the ranks from human judgments. We compare the ranks instead of the scores because the subjective and the objective scores may or may not be linearly related, however their rank correlation represents the strength and direction of their association without assuming linearity.

VI. RESULTS AND DISCUSSION

Spearman's rank correlation for all the measures and for each song is shown in Figure 5. We see that the proposed new measures in general perform better than the baseline measures $Kurt$, and $Skew$.

The GMM-fit measures consistently perform well, which shows that capturing the shape of the pitch histogram indeed characterizes singing quality. However in case the GMMs do not model the peaks accurately, there is a possibility of error. For example, in case of Song 5, a good singer is ranked very low because in the singer's pitch histogram, there are two peaks lying very close to each other, that the GMM captures as one peak. $kMeans$ and $Binning$ measures perform relatively better in such cases as they do not rely on the exact shape of the histogram, rather they look at the relative density of the pitch clusters. Also the global statistics provided by $Kurt$ and $Skew$ are better indicators in such cases.

Since each of these measures capture different aspects of the pitch histogram, we hypothesize that their combination should perform better. Table III shows Spearman's correlation of the ranks obtained from different combinations of the ranks from different measures. For example, the combination "Baseline" in the second column is the resultant rank when the ranks from the measures $Kurt$ and $Skew$ are added. These resultant ranks are correlated with the BWS ranks of human judges from MTurk. We can see that on an average, the combination of the proposed measures outperforms the baseline combination. Moreover the combination of all the measures outperforms the baseline and the proposed measures combinations. This indicates that all the measures capture different aspects and thus contribute in different ways to improve the performance. On comparing the last two columns, we find that a combination of top two best performing measures shows better correlation than the best measure. This confirms our hypothesis that combination of measures is a better method to evaluate singing quality than relying on a single measure.

Additionally, we compared the objective measure ranks with that provided by the control human judges for Song 1 (Table

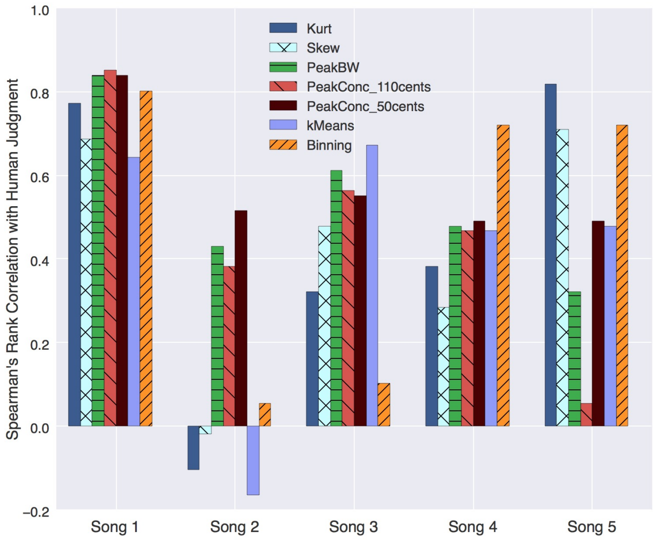


Fig. 5. Spearman's rank correlation of the seven measures with the BWS score ranks (human judgments) for the 5 songs.

TABLE III

RANKS FROM DIFFERENT MEASURES ARE ADDED, AND THE RESULTANT RANK IS CORRELATED (SPEARMAN'S) WITH BWS RANKS. BASELINE: $Kurt$+$Skew$, PROPOSED MEASURES: $PeakBW$ + $PeakConc_{110}$ + $PeakConc_{50}$ + $kMeans$ + $Binning$.

Songs	Baseline	Proposed Measures	All Measures	Best Single Measure	Two Best Performing Measures	
Song 1	0.723	**0.851**	0.799	**0.851**	0.850	*PeakConc$_{110}$, PeakBW*
Song 2	-0.042	0.455	0.503	**0.515**	0.477	*PeakConc$_{50}$, PeakBW*
Song 3	0.403	0.638	0.571	0.673	**0.734**	*kMeans, PeakBW*
Song 4	0.335	0.564	0.539	0.721	**0.729**	*Binning, PeakConc$_{50}$*
Song 5	0.745	0.489	0.636	0.818	**0.827**	*Kurtosis, Binning*
Avg.	0.433	0.599	0.610	0.716	**0.723**	-

III). We observe that the proposed measures outperform the MTurk evaluators in terms of the correlation with the control human judges (i.e. professional musicians). In spite of the various filters and conditions imposed, MTurk user data is still noisy compared to professional musician data. Therefore the fact that proposed measures are closer to the professional musicians than the MTurk users are, is an encouraging result because it confirms that the proposed measures evaluate the singing quality in the way professional musicians do.

TABLE IV

RANKS FROM DIFFERENT MEASURES ARE ADDED FOR SONG 1, AND THE RESULTANT RANK IS CORRELATED (SPEARMAN'S) WITH CONTROL HUMAN RANKS. MTURK BWS SCORE RANKS ARE ALSO CORRELATED.

Song	Baseline	Proposed Measures	All Measures	Two Best Performing Measures		MTurk BWS
Song 1	0.673	**0.899**	0.834	0.895	*PeakBW, PeakConc$_{110}$*	0.859

A. Drawbacks

Although the measures proposed in this work perform well in general, there are specific conditions when they fail to perform, as discussed below.

Other perceptual parameters

By converting a pitch contour into a histogram, we lose the time sequence. Therefore information about time correctness or rhythm is lost. Moreover, information about vibrato is either completely lost because of the low-pass filtering step, or it appears as a spread around the histogram peaks that degrades the performance. Noticeably, all the measures perform poorly in case of Song 2 (Figure 5). Song 2 was a relatively easy song where almost all the singers could hit the right notes. So the human judges rated them based on other perceptual parameters relevant to singing quality such as voice quality, vibrato, and pronunciation. The pitch histogram is unable to capture these aspects of singing quality evaluation. For example, a singer of Song 2 sings the notes of the song correctly, but cannot keep up with the lyrics. The human judges have ranked her low, however the objective measures fail to capture this aspect and rank her high.

Correctness of notes

None of the measures model the relative positions of the peaks in the histogram. Therefore, incorrect location of peaks go undetected. If a song consists of five notes, and a singer sings five notes precisely but they are not the same notes as that present in the song, then the objective measures would not be able to detect it.

Key shift

If a song consists of a legitimate key shift (less than an octave) in the middle of the song, then the objective measures will fail, because the key change will appear as additional peaks shifted by a constant amount from the original peaks.

Localized errors

Pitch histogram also loses the information about localized error or error that occurs for a short duration. According to cognitive psychology and PESnQ measures [26], [25], [1], localized errors have greater subjective impact than distributed errors. Therefore if a singer sings incorrectly for a short phrase, and then corrects himself/herself, the objective measures are unable to capture it.

VII. CONCLUSIONS

In this work, we presented methods to automatically evaluate singing quality without relying on a reference singing or score information. We proposed pitch histogram based methods such as GMM-fit and clustering that capture the shape and density of the histogram. We confirmed our hypothesis that these finer details of the histogram provide discerning information about singing quality, that the baseline global statistical measures kurtosis and skew do not provide. Combination of measures results in an even better performance showing that each of these measures model different aspects of the histogram.

We also showed that crowd-sourcing platforms can provide a scalable method of obtaining reliable singing quality judgment ground-truths by applying appropriate filters and constraints. We find that our proposed methods perform even better than MTurk users when correlated with the lab-controlled professional musician judgments.

Although the proposed measures provide reliable judgments in general, there is scope for improvement. These measures are unable to capture the other perceptually relevant parameters such as rhythm, vibrato, voice quality and pronunciation. Thus in the future, a complete framework that captures every aspect of singing quality independent of any reference needs to be explored.

References

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

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

[3] 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.

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

[5] 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.

[6] E. Nichols, C. DuHadway, H. Aradhye, and R. Lyon, "Automatically discovering talented musicians with acoustic analysis of youtube videos," *Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE,* pp. 559-565, 2012.

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

[8] 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.

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

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

[11] 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.

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

[13] T. Nakano, M. Goto, and Y. Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," *9th International Conference on Music Perception and Cognition,* 2006.

[14] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," *Journal of New Music Research,* 32(2), pp.143–152, 2003.

[15] E. Billauer, function PeakDet, MATLAB (Converted to python), https://gist.github.com/endolith/250860, Accessed on 20th May 2018.

[16] E. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, 21, pp. 768–769, 1965.

[17] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory,* 28(2):129–37, 1982.

[18] Smule Sing! Karaoke, Digital Archive of Mobile Performances (DAMP) https://ccrma.stanford.edu/damp/. Last Accessed: 1st March 2018.

[19] 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.

[20] B. Naderi, T. Polzehl, I. Wechsung, F. Kster, and S. Mller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," *In Sixteenth Annual Conference of the International Speech Communication Association,* 2015.

[21] A. Marley, T. Flynn, and V. Australia, "Best worst scaling: theory and practice", *Handbook of Choice Modelling,* Edward Elgar Publishing, Leeds (UK), 2012.

[22] J. Louviere, I. Lings, T. Islam, S. Gudergan, and T. Flynn, "An introduction to the application of (case 1) bestworst scaling in marketing research," *International Journal of Research in Marketing,* 30(3), pp.292–303, 2013.

[23] 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.

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

[25] 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.

[26] 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.