This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Automatic Leaderboard: Evaluation of Singing Quality without a Standard Reference

Chitralekha Gupta, *Student Member, IEEE,* Haizhou Li, *Fellow, IEEE,* and Ye Wang, *Member, IEEE*

***Abstract*—Automatic evaluation of singing quality can be done with the help of a reference singing or the digital sheet music of the song. However, such a standard reference is not always available. In this paper, we propose a framework to rank a large pool of singers according to their singing quality without any standard reference. We define musically motivated absolute measures based on pitch histogram, and relative measures based on inter-singer statistics to evaluate the quality of singing attributes such as intonation, and rhythm. The absolute measures evaluate the goodness of pitch histogram specific to a singer, while the relative measures use the similarity between singers in terms of pitch, rhythm, and timbre as an indicator of singing quality. With the relative measures, we formulate the concept of *veracity* or *truth-finding* for the ranking of singing quality. We successfully validate a self-organizing approach to rank-ordering a large pool of singers. The fusion of absolute and relative measures results in an average Spearman's rank correlation of 0.71 with human judgments in a 10-fold cross-validation experiment, which is close to the inter-judge correlation.**

***Index Terms*—Evaluation of Singing Quality, music-theory motivated measures, inter-singer measures, evaluation by ranking**

I. Introduction

SINGING has always been a popular medium of social recreation. Improving singing abilities is desired by amateur and aspiring singers. Music experts evaluate singing quality with the help of their music knowledge and perceptual appeal. Studies have shown that music experts can evaluate singing quality with high level of consensus when the melody or the song is unknown to them [1], [2]. 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 explore these properties and propose methods to automatically evaluate singing quality without depending on a reference.

Computer-assisted singing learning tools have been reported to be useful for singing lessons [3]–[5]. Recently, karaoke singing apps such as Smule Sing![6], Starmaker[7], and online platforms such as SoundCloud, and Youtube have provided a platform for people to showcase their singing talent, and a convenient way for amateur singers to practice and learn singing. They also provide an online competitive platform for singers to connect with other singers all over the world, and improve their singing skills. Automatic singing evaluation systems on such platforms typically compare a sample singing vocal with a standard reference such as a professional singing vocal [8]–[11] or the song melody notes [12]–[14] to obtain an evaluation score. For example, Perceptual Evaluation of Singing Quality (PESnQ) [8] measures the similarity between a test singing and a reference singing in terms of pitch, rhythm, vibrato, etc. However, such methods are constrained either by the need for a professional grade singer, or the availability of a digital sheet music for every song. The aesthetic perception of singing quality is very subjective and varies between evaluators. As a result even experts often disagree on the perfection of a certain performance [15]. The choice of an *ideal* or *gold-standard* reference singer brings in a bias of subjective choice. Therefore, a reference-independent method of singing quality evaluation is desirable.

Aspiring singers upload cover versions of their favorite songs on these online platforms, that are listened and liked by millions across the globe. However discovering talented singers from such huge datasets is a challenging task [16]. Moreover, often times the cover songs don't follow the original music scores, but rather demonstrate the creativity and singing style of individual singers. In such cases, reference singing or musical score based evaluation method is not an ideal choice.

There have been a few studies on evaluating singing quality without a standard reference. Nakano et al. [2] designed a singing skill evaluation scheme based on pitch interval accuracy and vibrato, which are regarded as the features that function independently from the individual characteristics of singer or melody. They used pitch interval accuracy to measure the consistency of the pitch offset values within a musical semitone grid. 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 folded on to a semitone. If the pitch values have a constant offset from this grid throughout the song, 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 one note throughout the song, the pitch interval accuracy can be unwantedly good because it ignores other aspects of the melody, such as occurrence pattern of notes and their durations.

In [17], we computed the absolute measures (i.e. without a standard reference) to evaluate singing quality by characterizing the shape of the pitch histogram of a singing rendition. We characterized the shape of the peaks in the histogram, the num-

Chitralekha Gupta is a graduate student funded by NUS Graduate School for Integrative Sciences and Engineering (NGS) scholarship, NUS, Singapore. E-mail: chitralekha@u.nus.edu.

Haizhou Li and Ye Wang are with National University of Singapore, Singapore. E-mail: {haizhou.li,dcswangy}@nus.edu.sg

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016, Non-parametric approach to voice morphing.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

ber of peaks, and the concentration of the pitch values around the peaks. This is different from computing the pitch interval accuracy histogram as it leads to a better understanding of the inherent discerning properties of the quality of singing voice without making use of a reference melody. These are the pitch histogram-based *musically-motivated absolute measures*. In this work, we further build upon [17] by introducing a novel autocorrelation-based measure to this group of measures in Section II, while also comparatively and systematically studying all of these measures. However, pitch histogram does not provide the complete picture. For example, the temporal information, such as rhythm, is not captured; furthermore, for an unknown melody, the correctness of pitching the notes is not evaluated. Thus, in this work, we additionally explore relative methods of evaluation, where information about rhythm and note location is retained.

Humans are known to be better at relative judgments, i.e. choosing the best and the worst among a small set of singers [18], [19], than giving an absolute rating. This leads us to the idea of automatically generating a leaderboard of singers, where the singers are rank-ordered according to their singing quality relative to each other. With the immense amount of online uploads on singing platforms, we can now leverage on the comparative statistics between singers as well as music theory to derive such a leaderboard of singers.

We study the research problem of automatic leaderboard generation in the scenarios where a large number of singers perform the singing of the same song. Without the reference singing template or gold-standard, we would like to automatically rank-order the singing vocals by their singing quality. Based on the concept of *veracity*, we believe that good singers sing alike, but bad singers sing very differently to each other. If all singers sing the same song, the good singers would share many characteristics such as the frequently hit notes, the sequence of notes, and the overall consistency in the rhythm of the song. However, different poor singers will deviate from the intended song in different ways. For example, one may be out-of-tune at certain notes, while another may be at some other notes. In this way, the *relative measures based on inter-singer distance* can serve as an indicator of singing quality, that we will discuss in Section III. It is worth noting that excellent singers may stand out of the average, and may differ from other good singers. However, the fundamental quality of the songs, such as pitch, rhythm, and voice timbre should remain consistent. Therefore, the relative measures will provide a broad segregation of singers according to their relative singing quality.

We propose an automatic leaderboard framework that combines the pitch histogram-based measures with the inter-singer distance measures to provide a comprehensive singing quality assessment without relying on a standard reference. We assess the performance of our algorithm by comparing against human judgments.

The automatic leaderboard can be useful as a screening tool for singing competitions, and karaoke applications, where there is a need for large-scale screening of singers. In the context of singing pedagogy, a detailed feedback to a learner about their performance with respect to the individual underlying perceptual parameters such as pitch, rhythm, and timbre, is important. Although humans are known to provide consistent overall judgments, they are not good at objectively judging the quality of individual underlying parameters. We will show that the proposed singing quality evaluation scheme outperforms human judges in this regard. In this paper, we make the following major contributions,

- We introduce novel inter-singer relative measures, based on the concept of veracity algorithm, that rank-orders large number of singing renditions without relying on a reference singing
- We further the study of [17] by introducing a novel autocorrelation-based measure to the group of musically motivated measures and systematically discuss their properties.
- We propose a combination of absolute and relative measures to characterize the inherent properties of singing quality
- We show that our algorithms assess different aspects of singing quality independently, that outperform humans.

This paper is organized as follows. In Section II, we discuss various musically-motivated absolute measures, in Section III, we discuss our idea and approach for inter-singer relative measure computation. We discuss the ranking strategy and fusion methods in Section IV. Data preparation is discussed in Section V, and the experiments and conclusions are discussed in Sections VI and VII, respectively.

II. Musically-Motivated Measures

A subjective assessment study conducted by Nakano et al. [1] found that human judges could evaluate singers with high level of consistency even when the songs are unknown to the judges. This finding suggests that singing quality judgment depends more on common, objective features rather than subjective preference. Moreover, experts make their judgment neither relying on their memory of the song, nor a reference melody. This encourages us to explore methods to quantify singing quality in a reference-independent way.

Subjective assessment studies suggest that the most important properties for singing quality evaluation are pitch and rhythm [20]–[22]. Pitch is an auditory sensation in which a listener assigns musical tones to relative positions on a musical scale based primarily on their perception of the frequency of vibration [23]. Pitch is characterized by the fundamental frequency $F0$ and its movements between high and low values. Musical notes are the musical symbols that indicate the pitch values, as well as the location and duration of pitch, i.e. the timing information or the rhythm of singing. In karaoke singing, visual cues to the lyric lines to be sung are provided that helps the singer to have more control over the rhythm of the song. Therefore, in the context of karaoke singing, rhythm is not expected to be a major contributor to singing quality assessment. Pitch, however, can be perceived and computed. Therefore, we will focus on the characterization of singing pitch in this section.

A. Pitch Histogram

Pitch histograms are global statistical representations of the pitch content of a musical piece [17], [24]. They represent

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the distribution of pitch values in a sung rendition. A pitch histogram is computed as the count of the pitch values folded on to the 12 semitones in an octave. All pitch values in this study are calculated in the unit of cents (one semitone being 100 cents on equi-tempered octave),

$$f_{\text{cent}} = 1200 \times \log_2 \frac{f_{\text{Hz}}}{440} \quad (1)$$

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

In this work, we use the pitch estimates from the autocorrelation-based pitch estimator PRAAT[25], [26]. Babacan et al. [27] have shown that PRAAT gives the best voicing boundaries for singing voice with the least number of post-processing steps or adaptations, when compared to other pitch estimators such as source-filter model based STRAIGHT[28] and modified autocorrelation-based YIN [29]. We apply one generic post-processing step to remove the frames with low periodicity, as described in detail in [8].

To compute the pitch histogram, it is necessary to remove the key of the song. Previously, Nichols et al. [16] computed the tuning frequency to induce a grid of *correct* pitch frequencies based on an equi-tempered 12 semitone scale, and then computed the histogram of the differences of the pitch values from the nearest correct frequencies. This resulted in a histogram of values within one semitone. Nakano et al. [2] used a filterbank method to obtain the correct frequencies grid, but then computed the one semitone histogram in the same way as in [16]. However, determining the tuning frequency is a challenging task [30], [31]. In this work, we first convert the pitch values to an equi-tempered scale (cents), and then instead of computing the tuning frequency, we subtract the median from the pitch values. Since median does not represent the tuning frequency of a singer, the pitch histogram obtained this way may show some shift across singers. However, it does not affect the strength of the peaks and valleys in the histogram. Also, as the data used in this study is taken from karaoke where the singers sang along with the background track of the song, so the key is supposed to remain same across singers.

We subtract the median of pitch values in a singing rendition, and transpose all pitch values to a single octave, i.e. within -600 to +600 cents. Then we compute pitch histogram H by placing the pitch values into their corresponding bins [32]:

$$H_k = \sum_{n=1}^{N} m_k \quad (2)$$

where H_k is the k^{th} bin count, N is the number of pitch values, m_k = 1 if $c_k \leq P(n) \leq c_{k+1}$ and $m_k = 0$ otherwise, where $P(n)$ is the n^{th} pitch value in an array of pitch values and (c_k, c_{k+1}) are the bounds on k^{th} bin. To obtain a fine histogram representation, we divided each semitone into 10 bins. Thus we have 12 semitones x 10 bins each = 120 bins in total, each representing 10 cents.

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 duration. These dominant notes 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, in the pitch histogram of a good singing vocal of a song, these dominant notes should appear as the peaks, while the transition semitones appear in the valley regions.

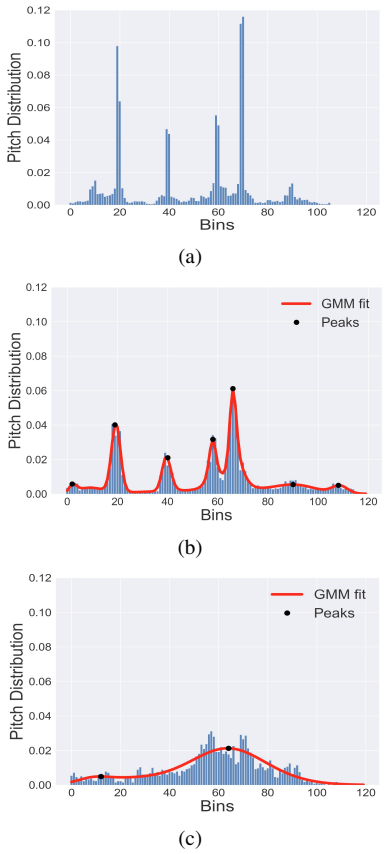


Fig. 1. Normalized pitch histogram for (a) MIDI, and GMM-fit (red line) and detected peaks (black dots) on normalized pitch histogram for (b) good singing (c) poor singing of the song "I have a dream" by ABBA. (1 bin=10 cents)

Figure 1 shows (a) the pitch histogram of MIDI (Musical Instrument Digital Interface) signal, (b) a good singing vocal, and (c) a poor singing vocal, all performing the same song, *I have a dream* by ABBA. The area of histogram is normalized to 1. The MIDI version contains the notes of the original composition, therefore represents the canonical pitch histogram of the song. It is apparent that the good singer histogram should be close to the MIDI histogram. They have four sharp peaks showing that those pitch values are frequently and consistently hit, more than the rest of the pitch values. Since generally a song consists of only a set of dominant notes, so the sharp, narrow, and well-defined spikes of the good singer's pitch histogram indicate that the notes of the song are being hit repeatedly and consistently. On the other hand, the poor singer has a dispersed distribution of pitch values, that reflect that the singer is unable to hit the dominant notes of the song consistently.

Statistical measures kurtosis and skew [2], [16] were used to measure the sharpness of the pitch histogram. These are overall statistical indicators that don't care much about the actual shape of the histogram, which could be informative about the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

singing quality. Therefore, in this work, we characterize the musical properties of singing quality with the 12 semitones pitch histogram. We believe that the shape of this histogram, for example, the number of peaks, the height and spread of the peaks, and the intervals between the peaks contain vital information about the goodness of the sung melody. Although we cannot directly determine the correctness of the notes being sung when the notes of the song are not available, we can measure the consistency of the pitch values being hit, which is an indicator of the singing quality.

In the following sub-sections, we systematically discuss the group of musically-motivated pitch histogram-based measures from a musical perspective. In [17], we formulated several pitch histogram based measures for singing quality evaluation without a reference, that are briefly discussed in sub-sections II-B, II-C1, and II-D. In this paper, a new measure, called the autocorrelation energy ratio measure (Section II-C2), is introduced.

B. From the perspective of overall pitch distribution

This is a group of global statistical measures that computes the deviation of the pitch distribution from a normal distribution. As seen in Figure 1, the pitch histogram of good singers show multiple sharp peaks, while that of poor singers show a dispersed distribution of pitch values. Therefore, we hypothesize that the histogram of a poor singer will be closer to a normal distribution, than that of a good singer [17].

*1) **Kurtosis**:* Kurtosis is a statistical measure (fourth standardized moment) of whether the data is heavy-tailed or light-tailed relative to a normal distribution, defined as

$$\kappa = E\left[\left(\frac{\vec{x}-\mu}{\sigma}\right)^4\right] \tag{3}$$

where $\vec{x}$ is the data vector, which in our case is the pitch values over time, μ is the mean and σ is the standard deviation of $\vec{x}$. A good singer's pitch histogram is expected to have several sharp spikes, as in Figure 1, and thus away from a normal distribution. So a good singer would have a higher kurtosis value than a poor singer.

*2) **Skew**:* Skew is a measure of the asymmetry of a distribution with respect to the mean, defined as

$$\gamma = E\left[\left(\frac{\vec{x}-\mu}{\sigma}\right)^3\right] \tag{4}$$

where $\vec{x}$ is the data vector, μ is the mean and σ is the standard deviation of $\vec{x}$.

The pitch histogram of a good singer has peaks around the notes of the song, whereas that of a poor singer is expected to be more dispersed and spread out symmetrically. So, the pitch histogram of a poor singer is expected to be closer to a normal distribution (see Fig. 1), or more symmetric.

C. From the perspective of pitch concentration

The previous group of measures considered the overall distribution of the pitch values with respect to a normal distribution. However, they do not care about whether the singing vocal hits the musical notes. Next, we would like to quantify the precision with which the notes are being hit.

We would essentially want to measure the concentration of the pitch values in the pitch histogram. Multiple sharp peaks in the histogram indicate precision in hitting the notes. Moreover, the intervals between these peaks contain information about the relative location of these notes in the song indicating the musical scale in which the song was sung.

*1) **Gaussian mixture model-fit (GMM-fit)**:* To capture the fine details of the histogram, we fit a mixture of Gaussian distributions to model the pitch histogram. Figure 1(b) and (c) show the GMM-fit for a good and a poor singer respectively. 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.

To characterize the peaks in the histogram, we detect the local maximas in the GMM-fit [33]. Figure 1(b) and (c) show the detected local maximas. A point is considered to be a peak candidate if it has the maximal value, while being preceded and succeeded by a lower value [33]. Empirically, a peak candidate is considered to be the actual local maxima if it is the highest peak within at least $\pm$50 cents. We characterize singing quality on the basis of the detected peaks in the following two ways.

Firstly, we measure the spread around the peaks, that indicates the consistency of hitting the same notes, that we call the Peak Bandwidth (ρ_b), defined as:

$$\rho_b = \frac{1}{N^2}\sum_{i=1}^{N} w_i^2 \tag{5}$$

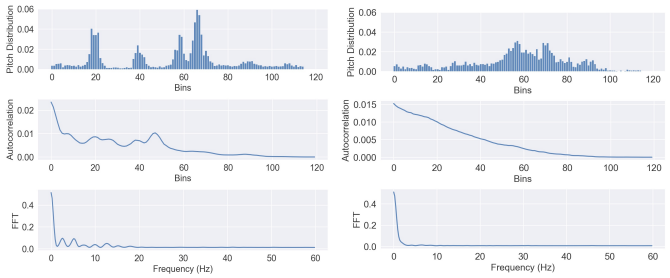
where w_i is the 3 dB half power down width of the i^{th} detected peak. Since a pop song is expected to have more than one or two significant peaks, we additionally penalize if there are only a small number of peaks, by dividing by the number of peaks N. Therefore, the peak bandwidth measure averaged over the number of peaks becomes inversely proportional to N^2.

Secondly, we measure the percentage of pitch values around the peaks, called Peak Concentration (ρ_c) measure, defined as:

$$\rho_c = \frac{\sum_{j=1}^{N}\sum_{i=bin_j-\Delta}^{bin_j+\Delta} A_i}{\sum_{k=1}^{M} A_k} \tag{6}$$

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, each bin represents 10 cents. Human perception is known to be sensitive to pitch changes, but the smallest perceptible change has been debated upon. Scientists agree that average adults are able to recognize pitch differences of as small as 25 cents reliably [34]. Thus, in equation 6, Δ is the number of bins on either sides of the peak to be considered for measuring peak concentration. It represents the allowable range of pitch change without being perceived as out-of-tune. We empirically consider the Δ values of $\pm$5 and $\pm$2 bins, i.e. $\pm$50 cents and $\pm$20 cents respectively, which along with the center bin (10 cents), are a total of 110 cents and 50 cents, respectively. We term these measures as $\rho_{c_{110}}$ and $\rho_{c_{50}}$ respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing



(a) (b)

Fig. 2. The normalized pitch histogram (1 bin = 10 cents) (top), autocorrelation of the histogram (middle), and the magnitude of the Fourier transform of the autocorrelation (bottom) for (a) good singing (b) poor singing.

*2) **Autocorrelation**:* Singers are supposed to sing mostly around the 12 semitones. The minimum interval is one semitone, and the intervals between the musical notes should be one or multiples of a semitone, that can be observed if we perform autocorrelation on the pitch histogram. If a good singer hits the correct notes all the time, we expect to see sharp peaks at multiples of semitones in the Fourier transform of the autocorrelation of the pitch histograms. This is evident from Figure 2 (bottom tier), where the magnitude spectrum of the autocorrelation of a good singing pitch histogram has energy in the higher frequencies representing the interval pattern of the dominant notes in the pitch histogram. On the other hand, the energy is concentrated only in the zero frequency component in the magnitude spectrum of the autocorrelation of the poor singing pitch histogram.

We compute the autocorrelation energy ratio measure or α as the ratio of the energy in the higher frequencies to the total energy in the Fourier transform of the autocorrelation of the pitch histogram,

$$\alpha = \frac{\sum_{f=4Hz} |Y(f)|^2}{\sum_{f=0Hz} |Y(f)|^2} \tag{7}$$

where,

$$Y(f) = F\left(\sum_{n=1}^{120} y(n) y^*(n-l) \right) \tag{8}$$

i.e. the Fourier transform of the autocorrelation of the histogram $y(n)$ where n is the bin number, and total number of bins is 120, and l is the lag. The assumption of sampling frequency for Fourier transform is 120 corresponding to the 120 bins in the pitch histogram and the corresponding autocorrelation. The lower cut-off frequency of 4 Hz in the numerator of equation 7 corresponds to the assumption that at least 4 dominant notes are expected in a good singing rendition, i.e. 4 cycles per second.

D. Clustering based on musical notes

As discussed earlier, a song typically consists of a set of dominant musical notes. Although the melody of the song is unknown, we can imagine that the pitch values, when sung correctly, will be clustered around these dominant notes. Therefore, they serve as a natural reference for evaluation. We explore two ways of measuring this clustering behavior.

*1) **k-Means Clustering**:* Tightly grouped clusters of pitch values across the histogram indicate that most of the pitch values are close to the cluster centers which means that the same notes are hit consistently. Keeping this idea in mind, we apply k-Means clustering to the pitch values, where $k = 12$ for the 12 semitones in an octave.

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. 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:

$$\zeta = \frac{1}{L} \sum_{i=1}^{k} d_i^2 \tag{9}$$

where L is the total number of 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

$$d_i^2 = \sum_{j=1}^{L_i} \left(p_{ij} - c_i \right)^2 \tag{10}$$

where p_{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, L_i is the number of pitch values in i^{th} cluster, and i ranges from $1, 2, ..., k$ number of clusters.

The difference between this measure and the ρ_b measure is that ρ_b is a function of the number of the dominant peaks, whereas in ζ, the number of clusters are fixed to 12 corresponding to all the possible semitones in an octave. Thus, they are different in capturing the influence of the dominant notes on the evaluation measure.

*2) **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. Equations 9 and 10 hold true for this method too, the only difference is that the cluster boundaries are fixed in binning or β method at 100 cents.

In summary, we have eight musically-motivated absolute measures for evaluating singing quality without a reference (Table I): κ, γ, ρ_b, $\rho_{c_{110}}$, $\rho_{c_{50}}$, ζ, β, and α.

III. Inter-Singer measures

For the first time, we propose an approach for evaluating singing quality without a reference by leveraging on the general behaviour of the singing vocals of the same song by a large number of singers. This novel approach uses inter-singer statistics to rank-order the singers in a self-organizing way.

The problem of discovering good singers from a large pool of singers is similar to that of finding *true* facts from a large amount of conflicting information provided by various websites [35]–[37]. The truth-finder algorithm utilizes the relationships between websites and their information. A website is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if the same information is provided by many trustworthy websites. The premise of the truth-finder algorithm is the heuristic that there is only one true version of a fact, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
LIST OF MUSICALLY-MOTIVATED ABSOLUTE AND INTER-SINGER RELATIVE MEASURES

Measure Group	Sub-group based on	Measure names
Musically-motivated absolute measures	Overall pitch distribution	Kurtosis (κ), Skew (γ)
	Pitch concentration	Peak bandwidth (ρ_b), Peak concentration ($\rho_{c_{110}}$, $\rho_{c_{50}}$), Autocorrelation energy ratio (α)
	Clustering	k-Means (ζ), Binning (β)
Inter-singer distance based relative measures	Pitch	$pitch\_med\_dist$, $pitch\_med\_L2$, $pitch\_med\_L6\_L2$, $pitchhist12DDistance$, $pitchhist120DDistance$, $pitchhistKLD12$, $pitchhistKLD120$
	Rhythm	$molina\_rhythm\_mfcc\_dist$, $rhythm\_L2$, $rhythm\_L6\_L2$
	Timbre	$timbral\_dist$

the true fact should appear in the same or similar way on different websites. Moreover, the false *facts* will be different and dissimilar between websites, because there can be many different ways of falsifying. Our hypothesis about singing quality follows the same heuristics [35]. We believe that a song can be sung correctly by many people in one consistent way, but incorrectly in many different, dissimilar ways. So, the goodness of a perceptual parameter of a singer is proportional to his/her similarity with other singers with respect to that parameter.

The next question is how to measure the similarity between singers. Let's first define a feature that represents a perceptual parameter of singing quality, say pitch contour. Suppose that all singers are singing the same song, we compare this feature of a singer with every other singer by a distance metric. According to our hypothesis, a good singer will be similar to the other good singers, therefore they will be close to each other, whereas a poor singer will be far from everyone. Figure 3 is a radial visualization of the Euclidean distance between the pitch contours of 100 singers, where the centre represents the singer of interest, and the radial distance of each dot from the centre represents his/her distance with one of the other 99 singers. The dots are placed at different polar angles in the plot. The polar angles are not part of the similarity metric. They are just for the purpose of visualization. It is evident that the best singers (top-ranked) are similar to other singers, therefore they are cluttered around the center, whereas the poorest singer is distant from everybody else. The observation validates our hypothesis that good singers are similar, and poor quality singers are dissimilar. This also points us to a method of ranking singers by their similarity with the peer singers.

In the following sub-sections, we discuss our metrics to measure the inter-singer distance, as summarized in Table I. These metrics measure the distance in terms of the perceptual parameters pitch, rhythm, and timbre. We also discuss singer characterization methods using these distance metrics.

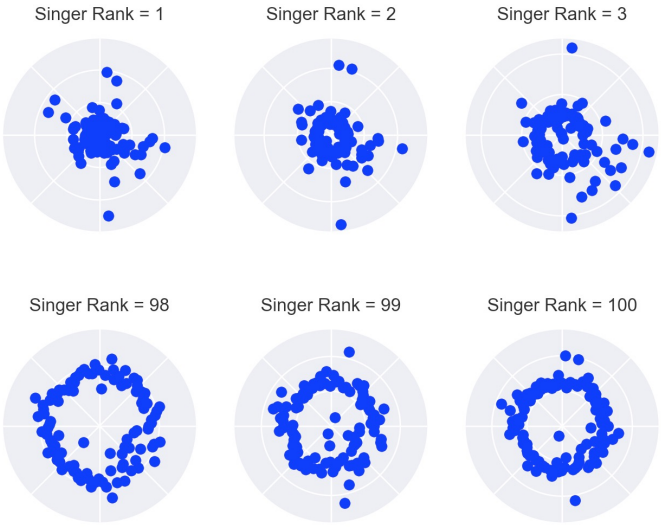


Fig. 3. Visualization of the pitch-based relative measure distance metric *pitch_med_dist* between each singer and the remaining 99 singers, for the best 3 (top row) and the worst 3 (bottom row) singers among 100 singers singing the song "Let it go".

A. Musically-Motivated Inter-Singer Distance Metrics

We now discuss how to measure inter-singer similarity by examining their pitch, rhythm and timbre in the singing.

1) ***Pitch-based Relative Distance****:*
Intonation or pitch accuracy is directly related to the correctness of the pitch produced with respect to a reference singing [8], [20], [22], [22]. In this work, we apply them to compare one singer with another, instead of a reference. The distance metrics used are the dynamic time warping (DTW) distance between the two median-subtracted pitch contours (*pitch_med_dist*), the Perceptual Evaluation of Speech Quality (PESQ)-based [38] cognitive modeling theory [39]-inspired pitch disturbance measures *pitch_med_L6_L2* and *pitch_med_L2* [8].

Additionally, in this work, we compute pitch histogram-based relative distance metrics. As seen in Figure 1, there is a clear distinction between the pitch distribution of a good and a poor singer. We compute the Kullback-Liebler (KL) Divergence between the normalized pitch histograms to measure the distance between the histograms of singers. Moreover, as the pitch histogram is computed after subtracting the median of the pitch values, not the actual tuning frequency in which the song is sung, the pitch histograms may be shifted by a few bins across singers. To account for this shift, we also compute DTW-based distance of the 12-bin and 120-bin histograms between singers as relative measures (*pitchhist12KLdist, pitchhist120KLdist, pitchhist12Ddist, pitchhist120Ddist*).

2) ***Rhythm-based Relative Distance****:*
Rhythm or tempo is defined as the regular repeated pattern in music, that relates to the timing of the notes sung. In karaoke singing, rhythm is determined by the pace of the background music and the lyrics cue on the screen. Therefore rhythm inconsistencies in karaoke singing only occurs when the singer is unfamiliar with the melody and/or the lyrics of the song.

Mel-frequency cepstral coefficients (MFCC) capture the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

short-term power spectrum that represents the shape of the vocal tract and thus the phonemes uttered. So, if the words are uttered at the same pace by two singers, then their rhythm is consistent. Thus, we compute the DTW alignment between two singer utterances with respect to their MFCC vectors. In this work, we use the three best performing rhythm measures from [8] to compute inter-singer rhythm distance: a modified version of Molina et al.'s [14] rhythm deviation measure (termed as *molina_rhythm_mfcc_dist*) that computes the root mean square error of the linear fit of the optimal path of DTW matrix computed using MFCC vectors, PESQ-based *rhythm_L6_L2*, and *rhythm_L2*.

3) ***Timbre-based Relative Distance****:*
Perception of timbre often relates to the voice quality [12], [20]. Timbre is physically represented by the spectral envelope of the sound, which is captured well by MFCC vectors, as shown in [40]. We compute the *timbral_dist* as the DTW distance between the MFCC vectors between the renditions of two singers.

B. Singer Characterization using Inter-Singer Distance

According to our theory, the distance as defined in Section III-A between a singer and others is indicative of the singer's singing quality. We now explore three methods to characterize a singer based on these inter-singer distance metrics, that we call relative scoring methods that give rise to the relative measures. We will refer to Figure 4 in the section, that demonstrates the relative measure computation from the *pitch_median_dist* distance metric with the three methods for the best and the worst singer out of 100 singers of a song.

1) ***Method 1: Affinity by Headcount*** $s_h(i)$*:*
We can set a constant threshold D_T on the distance value across all singer clusters and count the number of singers within the set threshold as the relative measure or score. If a large number of singers are similar to that singer, then the number of dots within the threshold circle will be high, as can be seen in Figure 4(a). If $dist_{i,j}$ is the distance between the i^{th} and j^{th} singers, the singer i's relative measure $s_h(i)$ by this headcount method is

$$s_h(i) = |dist_{i,j} < D_T : \forall j \in Q, j \neq i| \tag{11}$$

where, Q is the set of singers, and $|.|$ is the count of the number of points satisfying the expression within.

2) ***Method 2: Affinity by*** k^{th} ***Nearest Distance*** $s_k(i)$*:*
We can fix the number of singers k as the threshold, and consider the distance of the k^{th} nearest singer as the relative measure, as seen in Figure 4(b), where we set $k = 10$. If this distance is small, the singer is likely to be good. Singer i's relative measure ($s_k(i)$) of method 2 can be described as follows,

$$s_k(i) = dist_{i,j=k}; k \neq i \tag{12}$$

3) ***Method 3: Affinity by Median Distance*** $s_m(i)$*:*
The median of the distances of a singer from all other singers can be assigned as the relative measure, which represents his/her overall distance from the rest of the singers (Figure 4(c)). The median is taken instead of the mean to avoid outliers. If this distance is small for a singer, the singer is likely to be good. The singer i's relative measure by this method is

$$s_m(i) = median(dist_{i,j}); \forall j \in Q, j \neq i \tag{13}$$

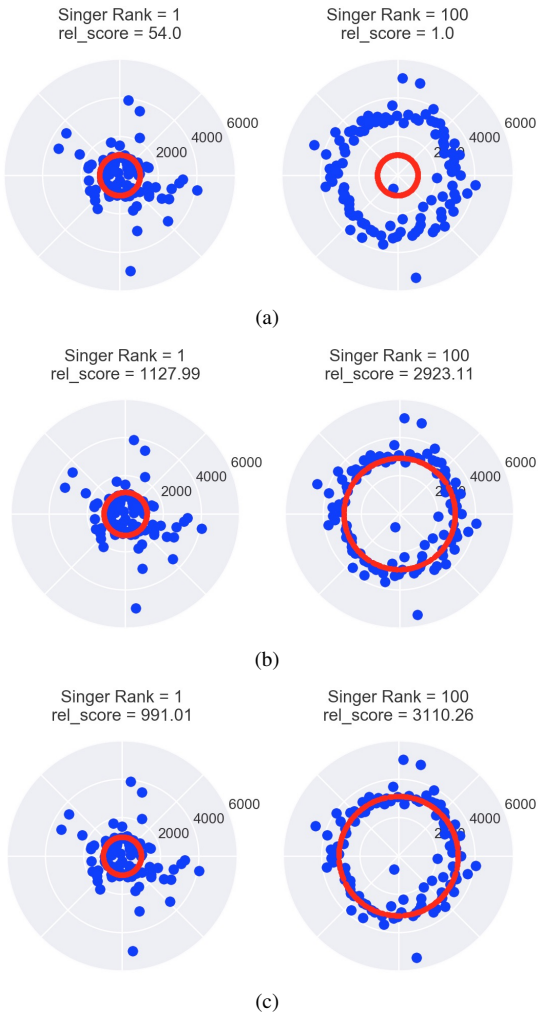


Fig. 4. Demonstration of relative scoring methods from the *pitch_med_dist* measure for the best (Rank 1) and the worst (Rank 100) singer of an example song (Song 1, snippet 1), along with the respective relative measure values or scores using: (a) Method 1: Affinity by Headcount (b) Method 2: Affinity by k^{th} Nearest Distance, $k = 10$ (c) Method 3: Affinity by Median Distance. The red circle in (a) and (b) are the thresholds, while for (c) it is the median value.

IV. Ranking Strategy, and Fusion Methods

In this section, we discuss our singer ranking strategy and score/measure and system fusion methods.

A. Strategy for Ranking

The primary objective of a leaderboard is to inform where a singer ranks with respect to the singer's contemporaries. As the best-worst scaling (BWS) theory [41] says, humans are known to be able to choose the best and the worst in a small set of choices, which over many such sets results in rank-ordering of the choices. However, when humans are asked to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

numerically rate singers on a scale of say 1 to 5, they do not reveal discriminatory results. Therefore, it makes sense to study how the absolute and relative measures reflect the ranking, and design our algorithm towards a better prediction of the overall rank-order of the singers.

Given a set of measure values or scores $S = S_1, S_2, ..., S_T$, where S_i represents a score of the i^{th} singer, and, T is the total number of singers of a song, the singers can be rank-ordered as:

$$\Gamma = (S_1, S_2, ..., S_T) \tag{14}$$

where,

$$S_1 \leq S_2 \leq ... \leq S_T \tag{15}$$

It is worth noting that all absolute and relative measures are song independent. We further normalize the measures by the length of the song so that they are independent of song duration.

B. Strategies for Score Fusion

Each of the absolute and relative measures can provide a rank-ordering of the singers. To arrive at an overall ranking of the singers, we hope to find ways to combine or fuse them together for a final decision. One way to compute an overall ranking is by computing an average of the ranks (AR) of all the measures. This method of score fusion does not need any statistical model training, but gives equal importance to all the measures. Considering that some measures are more effective than others, we also use a linear regression (LR) model that gives different weights to the measures. Owing to the success of neural networks and the possibility of a non-linear relation between the measures and the overall rank, we also explore neural network models to predict the overall ranking from the absolute and the relative measures. One of the neural network models (NN-1) consist of no hidden layers, but a non-linear sigmoid activation function. The other neural network model (NN-2) consist of one hidden layer with 5 nodes, with sigmoid activation function for both the input and the hidden layers. The models are summarized in Table II.

We also investigate the performance of the fusion of the two scoring systems, i.e. fusion of the 8 absolute measures system and the 11 relative measures system. One method to combine them is *early-fusion* where we incorporate all the scores from the evaluation measures to get a 19 dimensional score vector for each snippet. Another method of combining the measures is *late-fusion*, where we compute the average of the ranks predicted independently from the absolute and the relative scoring systems.

V. Data Preparation

To evaluate singing quality without a reference, we conducted experiments using the musically-motivated absolute measures, the inter-singer distance based relative measures, and the combinations of these measures. In this section, we discuss the singing voice dataset and the subjective ground-truths used for these experiments.

TABLE II
SUMMARY OF THE FUSION MODELS. ($\mathbf{r}_i$ = RANK-ORDERING OF SINGERS ACCORDING TO i^{th} MEASURE; N = # OF MEASURES; $\mathbf{x}$ = MEASURE VECTOR; $\mathbf{w}^i$ = WEIGHT VECTOR OF i^{th} LAYER; $\mathbf{b}$ = BIAS; $S(.)$ = SIGMOID ACTIVATION FUNCTION; y = PREDICTED SCORE); AR: AVERAGE RANK, LR: LINEAR REGRESSION.

#	Model	Description	Equation
1	AR	Equally weighted sum of individual measure ranks	$y = \frac{1}{N}\sum_{i=1}^{N} \mathbf{r}_i$
2	LR	Weighted sum of measures	$y = \mathbf{b} + \mathbf{w}^T\mathbf{x}$
3	NN-1	MLP with sigmoid activation, no hidden layer	$y = S(\mathbf{b} + \mathbf{w}^T\mathbf{x})$
4	NN-2	MLP with sigmoid activation, 1 hidden layer with 5 nodes	$y = S(\mathbf{b^{(2)}} + \mathbf{w^{(2)}}S(\mathbf{b^{(1)}} + \mathbf{w^{(1)}}^T\mathbf{x}))$

TABLE III
SUMMARY OF THE SINGING VOICE DATASET. NOTES CAN BE OF SHORT, LONG OR MIXED DURATIONS; BPM = BEATS PER MINUTE

#	Song Name	Nature of Melody		Tempo (bpm)
		Pitch Range	Note duration	
1	Let it go (Frozen)	more than an octave	mix	68
2	Cups (Pitch Perfect)	within an octave	short	130
3	When I was you man (Bruno Mars)	more than an octave	mix	73
4	Stay (Rihanna)	within an octave	mix	112

A. Singing Voice Dataset

In the automatic leaderboard experiments, we assume that all singers sing the same song. We construct a database that consists of 4 popular Western songs each sung by 100 unique singers (50 male, 50 female) extracted from Smule's DAMP dataset [42]. DAMP dataset consists of 35k solo-singing recordings without any background accompaniments. The selection of songs was based on the available number of unique singers in the dataset, and equal distribution between males and females, to avoid gender bias. Our selected subset of songs were the most popular four songs in the DAMP dataset with more than 100 unique singers singing them. All the songs are rich in steady notes and rhythm, as summarized in Table III. The dataset consists of a mix of songs with long and sustained as well as short duration notes with a range of different tempi in terms of beats per minute (bpm).

We divide every song into 4 snippets, where each snippet is of approximately 20 seconds in duration. Such short duration clips are recommended for the relative measure computation as shorter duration segments are less prone to misalignments during DTW [8], [43], [44].

B. Subjective Ground-Truth

We need subjective ratings as ground-truth to validate the objective measures for singing evaluation. We can obtain consistent ratings from professionally trained music experts. However, obtaining such ratings at a large scale may not be always possible, as it can be time consuming, and expensive. We have shown in [17] that crowd sourcing platforms, such as Amazon mechanical turk (MTurk), is effective to obtain reliable human judgments of singing vocals. We showed that the ratings provided by MTurk users correlated well with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the ratings obtained from professional musicians in a lab-controlled experiment [17]. The Pearson's correlation between lab-controlled music-expert ratings and filtered MTurk ratings for various parameters are as follows: overall singing quality: 0.91, pitch: 0.93, rhythm: 0.93, and voice quality: 0.65. We continue to use MTurk to derive the subjective ground-truth in this paper.

While it is possible that professional musicians rate singing quality at an absolute scale of 5 consistently, we cannot be sure about the ratings through crowd sourcing. Also, absolute ratings are known to not discriminate between items, and each rating on the scale is not precisely defined [18], [41]. Therefore, we used a method of relative rating called *best-worst scaling* (BWS) which can handle a long list of options and always generates discriminating results as the respondents are asked to choose the best and worst option in a choice set [18], [19]. At the end of this exercise, the items can be rank-ordered according to the aggregate BWS scores of each item, given by

$$B = \frac{n_{best} - n_{worst}}{n} \tag{16}$$

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 Spearman's rank correlation between the MTurk experiment and the lab-controlled experiment reported in [17] was 0.859.

In this work, we conducted a pairwise BWS test on MTurk where a listener was asked to choose the better singer among a pair of singers singing the same song. We presented one excerpt of approximately 20 seconds from every singer of a song (the same 20 seconds for all the singers of a song). There are ${}^{100}C_2$ number of ways to choose 2 singers from 100 singers of a song, i.e. 4,950 Human Intelligence Tasks (HITs) per song. This experiment was conducted separately for each of the 4 songs of Table III. Therefore there were in total $4,950 \times 4 = 19,800$ HITs.

We screened the MTurk users in the same way as we did in [17]. We asked the users for their experience in music and asked them to annotate musical notes as a test. We accepted their attempt only if they had some formal training in music, and could write the musical notations successfully. We also monitored the time spent by the MTurk users in performing the task to remove the less serious attempts in case some may not finish listening to the snippets.

VI. Experiments

In Sections II, and III, we designed various musically-motivated absolute and relative objective measures that, we believe, can assess the inherent properties of singing quality that are independent of a reference. When the absolute and relative measures are appropriately combined, we generate a leaderboard of singers ranked in the order of their singing ability. Figure 5 shows the overview of this framework. Various methods to combine the absolute and relative measures are explored, as discussed in Section IV-B. The rank-order of the individual measures are averaged to obtain an average rank (AR). Moreover, we train the linear regression (LR) model, and the two different neural network models (NN-1, NN-2) (as discussed in Section IV-B) in 10-fold cross-validation. We ensure that in every fold, equal number of singers are present from every song, both in train and test data. The absolute and relative measure values are the inputs to these networks, while the human BWS scores given in equation 16 are the output values to be predicted. The loss function for the neural nets is mean squared error, with adam optimizer. All computations are done using scikit-learn [45].

We conduct several experiments to investigate the role of the absolute and the relative measures individually in predicting the overall human judgment, and the methods of combining these measures. Moreover, we compare the ability of our machine-based measures and humans in predicting the performance of the underlying perceptual parameters.

In this section, we first discuss the baseline system performance from the literature, and the achievable upper limit of performance in the form of the human judges' consistency in evaluating singing quality. Then we describe our experiments and analyse the results.

A. Baseline

As discussed earlier, Nakano et al. [2] and Bohm et al. [15] attempted to evaluate singing quality without a reference. They used the global statistics kurtosis and skew to measure the consistency of pitch values. These are two of our eight absolute measures. Moreover, [15] used the Interspeech ComParE 2013 (Computational Paralinguistics Challenge) feature set as baseline. It comprises of 60 low-level descriptor contours such as loudness, pitch, MFCCs, and their 1st and 2nd order derivatives, in total 6,373 acoustic features per audio segment or snippet [46]. We extract this same set of features using OpenSmile toolbox [47] to create our baseline for comparison. We conducted a 10-fold cross-validation experiment using the snippet 1 from all the songs to train a linear regression model with these features. The Spearman's rank correlation between the human BWS rank and the output of this model is 0.39. This rank correlation value is an assessment of how well the relationship between the two variables can be described using a monotonic function. This implies that with the set of features used in the literature, the machine predicted singing quality ranks has a positive but a low correlation with that given by humans.

B. Performance of Human Judges

In a pilot study [8], we recruited 5 professional musicians to provide singing quality ratings for 10 singers singing a song. These judges were 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 of 0.82. This shows that humans do not always agree with each other, and there is, in general, an upper limit of the achievable performance of any machine-based singing quality evaluation. Thus the goal of our singing evaluation algorithm is to achieve this upper limit of correlation with human judges.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

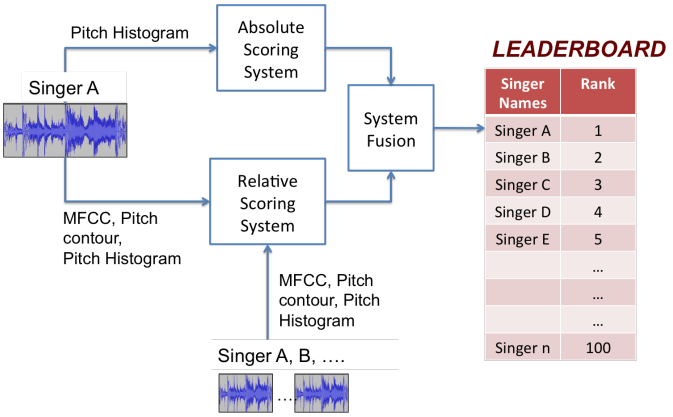


Fig. 5. Overview of the framework for automatic singing quality leaderboard generation, consisting of fusion of musically-motivated absolute scoring system and inter-singer distance based scoring system.

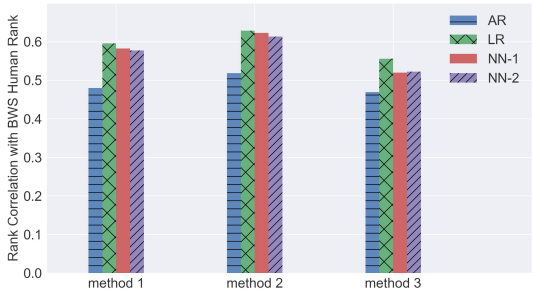


Fig. 6. Spearman's rank correlation performance of the three methods for inter-singer distance measurement (Section III-B): Method 1: Affinity by Headcount; Method 2: Affinity by 10^{th} Nearest Distance; Method 3: Affinity by Median Distance. Models are as listed in Table II.

C. Experiment 1: Comparison of Singer Characterization Methods using Inter-Singer Distance

In this experiment, we perform a preliminary investigation to compare the three singer characterization methods discussed in Section III-B: headcount (method 1), k^{th} nearest distance (method 2), and median distance (method 3). We obtained the relative measures from these methods for each of the 11 inter-singer distance measures. Figure 6 shows the Spearman's rank correlation of the human BWS ranks with ranks from these relative measures used with the six models of Table II, over the snippet 1 of all the 4 songs for the three methods. To observe the best case scenario for method 1 (headcount method), its distance threshold is optimized for each measure for snippet 1. The number of singers threshold for method 2 (k^{th} nearest distance method) is empirically set as 10 singers, assuming that roughly at least ten percent of singers in a large pool of singers would be good. In this way, if the distance of a particular singer from the 10^{th} nearest singer is small, it means that the singer sings very similarly to 10 singers, thus the singer is good.

We observe that the k^{th} nearest distance method (method 2) performs better than the other two methods for all the six models. The result suggests that our assumption that at least ten percent in a pool of singers would be good, serves our purpose. Method 3, i.e. the median of the distances of a particular singer from the rest of the singers assumes that half of the pool of singers would be good singers, which is not a reliable assumption, therefore this method performs the worst.

With the preliminary findings, we decide that the relative measures are computed using the k^{th} nearest distance method (method 2) in the rest of the experiments.

D. Experiment 2: Evaluating the measures individually

We analyze how well can each of the absolute and relative measures individually predict the ranks of the singers. Figure 7 shows the Spearman's rank correlation of each of the 8 absolute and the 11 relative score vectors with the human BWS ranks. We can see that all the derived measures show a positive correlation with humans, although some correlate better than others. The newly introduced autocorrelation energy ratio α measure shows the best correlation among the absolute measures. This suggests that the interval pattern of the dominant notes in the histogram carry important information about singing quality. The $\rho_{c_{50}}$ shows better performance than $\rho_{c_{110}}$, which agrees with the finding in the literature that human ear is sensitive to changes in pitch as small as 25 cents [34].

The relative measures, in general, perform better than the absolute measures, which suggests that the inter-singer comparison method is closer to how humans evaluate singers. The pitch-based relative measures perform better than the rhythm-based relative measures. This is an expected behavior for karaoke singings, where the background music and the lyrical cues help the singers to maintain their timing. Therefore, the rhythm-based measures do not contribute as much in rating the singing quality. Among the relative measures, *pitchhist120DDistance* performs the best, along with the KL-divergence measures, showing that inter-singer pitch histogram similarities is a good indicator of singing quality. The *pitch_med_dist* measure follows closely, indicating that the comparison of the actual sequence of pitch values and the duration of each note give valuable information for assessing singing quality. These aspects are not captured by the pitch histogram-based methods.

Another interesting observation is the high correlation of the *timbral_dist* measure. It indicates that voice quality, represented by the timbral distance, is an important parameter when humans compare singers to assess singing quality. This observation supports the timbre-related perceptual evaluation criteria of human judgment [20], [48], [49] such as *timbre brightness, color/warmth, vocal clarity, strain*. The timbral distance measure captures the overall spectral characteristics, thus represents the timbre-related perceptual criteria.

E. Experiment 3: Absolute Scoring System: The fusion of absolute measures

In this experiment, we evaluate the performance of the combination of musically-motivated pitch histogram-based absolute measures that were introduced in Section II in ranking the singers. Table IV, second column shows the Spearman's rank correlation between the human BWS ranks and the ranks predicted by absolute measures with different fusion models. Our preliminary experiments show that the pitch histogram for the full song provide a better representation than the histogram

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

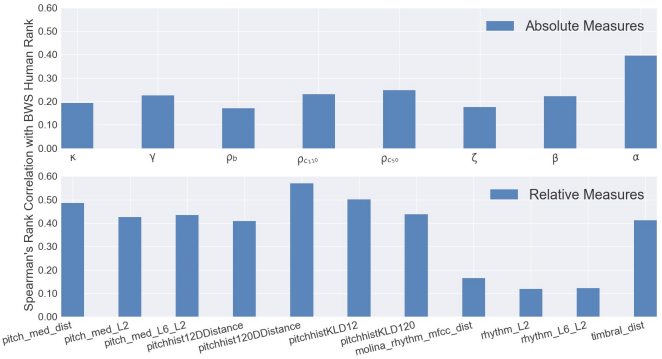


Fig. 7. Spearman's rank correlation of the individual absolute measures (top) and relative measures (bottom) with human BWS ranks.

of a short duration snippet of the song because more data results in better statistics. Therefore, we compute the absolute measures for the full songs (more than 2 minutes duration) in all the experiments.

One hidden layer in the neural network model (NN-2) performs better than the one without a hidden layer (NN-1), as well as the LR model. This indicates that non-linear combination of the measures provides a better prediction of human judgment. Interestingly, the average of ranks (AR) performs comparably with NN-2, suggesting that all measures are informative in making meaningful ranking. It also indicates that although the measures individually may not have performed equally well (Figure 7), each of them captures a different aspect of the pitch histogram quality, therefore, combining them with equal weights results in a comparable performance.

It is important to note that there are specific conditions when the absolute measures fail to perform [17]. By converting a pitch contour into a histogram, information about timing or rhythm is lost. The correctness of the note order also cannot be evaluated through the pitch histogram. Moreover, the relative positions of the peaks in the histogram cannot be modeled without a reference, i.e. incorrect location of peaks goes undetected. For example, 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 absolute measures would not be able to detect it. Pitch histogram also loses the information about localized errors, i.e. errors occuring for a short duration. According to cognitive psychology and PESnQ measures [8], [38], [39], localized errors have greater subjective impact than distributed errors. Therefore, if a singer sings incorrectly for a short duration, and then corrects himself/herself, the absolute measures are unable to capture it.

F. Experiment 4: Relative Scoring System: Evaluating the fusion of relative measures

In this experiment, we investigate the performance of the combination of the inter-singer relative measures computed from method 2 in Section VI-C. Table IV, third column shows the Spearman's rank correlation between the human BWS ranks and the ranks predicted by the relative measures with the different fusion models. We evaluate the four different snippets from each song and average the ranks over these snippets. It is worth noting that, according to the preliminary experiments, we found that the samples of longer duration lead to better statistics, therefore, more accurate scores.

TABLE IV
SUMMARY OF THE PERFORMANCE OF ABSOLUTE AND RELATIVE MEASURES, AND THEIR COMBINATIONS. THE VALUES IN THE TABLE ARE SPEARMAN'S RANK CORRELATION BETWEEN HUMAN BWS RANKS AND THE MACHINE GENERATED RANKS AVERAGED OVER 4 SNIPPETS.(ALL P-VALUES$<$0.05)

Model	Absolute Measures	Relative Measures	Early-fusion	Late-fusion
AR	**0.4796**	**0.6396**	**0.6877**	**0.7059**
LR	0.4205	0.5737	0.6413	0.6426
NN-1	0.3975	0.5799	0.6385	0.6407
NN-2	0.4711	0.6153	0.6636	0.6692

The combinations of the relative measures outperform the combinations of the absolute measures. This is consistent with the observation in Section VI-D where the relative measures individually outperform the absolute measures. Like the absolute measures, average of ranks (AR) performs better than the other score fusion models, indicating that all relative measures are informative in making meaningful ranking.

G. Experiment 5: System Fusion: Combining absolute and relative scoring systems

In this experiment, we investigate the combinations of the 8 absolute and 11 relative measures by early-fusion and late-fusion methods (Section IV-B). The rank correlation between the BWS ranks and the ranks obtained from early-fusion method averaged over four snippets is reported in column 4, Table IV, and that from late-fusion is in column 5.

The results suggest that the late-fusion of the systems show a better correlation with humans than early-fusion. This suggests that the predictions coming separately from the absolute and relative measures provide different and equally important information. Therefore, equal weighting to both shows better correlation with humans. Moreover, a simple rank average shows a better performance than the complex neural network models. It is encouraging to see that the individual measures describe different aspects of singing quality, and correlate with human judgments to a different extent. It is important to note that the process of converting values to ranks is inherently non-linear (see Section IV-A).

H. Experiment 6: Humans versus Machines

We note that human judgment on individual singing quality, for example, between pitch and rhythm, tends to be influenced by their overall judgment of the rendition. In other words, when humans judge a song as having bad pitch, they tend to consider them having bad rhythm as well. We would like to study how the objective evaluation techniques compare with human judgment in terms of independent judgment on perceptual parameters, such as pitch, rhythm, and timbre.

In this experiment, we use the data from our previous work [8], where music experts were asked to rate each singer on a scale of 1 to 5 with respect to the three aspects of perceptual quality, namely pitch, rhythm, and timbre individually. Figure 8(a) shows that human ratings for the three perceptual aspects

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2947737, IEEE/ACM Transactions on Audio, Speech, and Language Processing

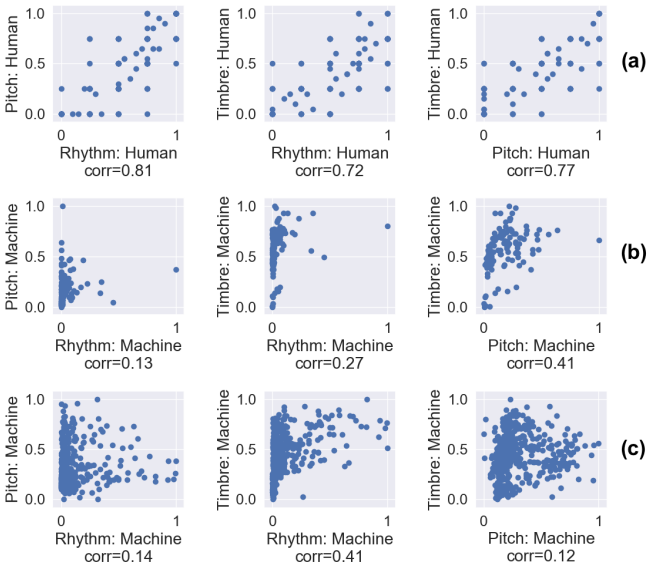


Fig. 8. Humans vs. Machines: Correlation between scores given individually for pitch, rhythm, and timbre by (a) human experts on the data in [8], (b) machine, on the same data as in (a), and (c) machine, on the data used in this work, as in Table III.

are highly correlated amongst each other. On the same data (each song is split into multiple short duration segments for processing), the machine scores for the three aspects are significantly less correlated (Figure 8(b)). We also verified this observation on the data used for the experiments in this current work (Figure 8(c)). It is clear that machines are able to assess the perceptual aspects of the singing rendition more objectively than humans. Such machine assessment can be useful for a learner to understand how to improve upon the individual parameters.

I. Discussion

With both absolute and relative measures, the proposed framework effectively addresses the issue with pitch interval accuracy [2] by looking at both the pitch offset values as well as other aspects of the melody. The absolute measures such as ρ_c, ρ_b, and α, characterize the shape of the pitch histogram of a given song. Furthermore, the relative measures compare a singer with the rest of the singers singing the same song. It is unlikely for all singers in a large dataset to sing one note throughout the song.

The experiments in this paper show that 100 renditions from different singers constitute a database for a reliable automatic leaderboard ranking. The absolute measures in the framework are independent of the singing corpus size, while the relative measures are scalable to a larger corpus.

The experimental results show that the derived absolute and relative measures are reliable reference-independent indicators of singing quality. We have focused the evaluation on the fundamental quality of the songs, such as pitch, rhythm, and voice timbre. It is noted that another level of singing quality appreciation, such as expressions, flexibility, and agility that project artistry, creativity, and personality, requires further studies. While the automatic leaderboard works well for first stage screening of a large number of singers, the resulting performance-related parameters can be useful for analysis by music experts as well.

One can expect variations in these objective measures across different genres and styles of singing. For example, the criteria of evaluation of a rap singing will be different from that of a jazz singing, or a Chinese opera singing from a Western classical singing. This work explores Western pop, as the first step in the direction of a large-scale reference-independent singing evaluation framework. In the future, other singing styles and music genres need to be investigated.

VII. Conclusions and Future Work

In this work, we successfully introduce a self-organizing method to produce a leaderboard of singers according to their singing quality without relying on a reference singing sample or musical score, by leveraging on musically-motivated absolute measures and *veracity* based inter-singer relative measures. The baseline method (Section VI-A) shows a correlation of 0.39 with human assessment using linear regression, while the linear regression model with our proposed measures shows a correlation of 0.64, and the best performing method shows a correlation of 0.71, which is an improvement of 82.1% over the baseline. This improvement is attributed to:

- the musically-motivated absolute measures, that quantify various singing quality discerning properties of the pitch histogram, and
- the novel *veracity* based musically-informed relative measures that leverages on inter-singer statistics and overcomes the drawbacks of using only absolute measures

We find that the two kinds of measures provide distinct information about singing quality, therefore a combination of them boosts the performance.

We find that the proposed ranking technique provides objective measures for perceptual parameters, such as pitch, rhythm, and timbre independent, that human subjective assessment fails to produce.

Human experts, in general, are more consistent amongst themselves than the machine scores, with a correlation of 0.82 (Section VI-B). Thus, the machine scores remain to be improved. Inclusion of other perceptual parameters such as vibrato and pronunciation can further improve the machine scores. Extension of the proposed methods to music genres other than Western pop also needs to be investigated in future.

References

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

[2] ——, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Ninth International Conference on Spoken Language Processing*, 2006.

[3] D. Hoppe, M. Sadakata, and P. Desain, "Development of real-time visual feedback assistance in singing training: a review," *Journal of computer assisted learning*, vol. 22, no. 4, pp. 308–316, 2006.

[4] G. F. Welch, D. M. Howard, and C. Rush, "Real-time visual feedback in the development of vocal pitch accuracy in singing," *Psychology of Music*, vol. 17, no. 2, pp. 146–157, 1989.

[5] D. M. Howard and G. F. Welch, "Microcomputer-based singing ability assessment and development," *Applied Acoustics*, vol. 27, no. 2, pp. 89–102, 1989.

[6] Smule, "Sing! karaoke app." https://www.smule.com, 2008.

[7] Starmaker, "Starmaker karaoke app." https://www.starmakerstudios.com, 2010.

[8] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proceedings of APSIPA Annual Summit and Conference*, vol. 2017, 2017, pp. 12–15.

[9] P.-C. Chang, "Method and apparatus for karaoke scoring," Dec. 4 2007, uS Patent 7,304,229.

[10] P. Lal, "A comparison of singing evaluation algorithms," in *Ninth International Conference on Spoken Language Processing*, 2006.

[11] C.-H. Lin, Y.-S. Lee, M.-Y. Chen, and J.-C. Wang, "Automatic singing evaluating system based on acoustic features and rhythm," in *Orange Technologies (ICOT), 2014 IEEE International Conference on*. IEEE, 2014, pp. 165–168.

[12] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2012.

[13] T. Tanaka, "Karaoke scoring apparatus analyzing singing voice relative to melody data," Mar. 30 1999, uS Patent 5,889,224.

[14] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 744–748.

[15] J. Böhm, F. Eyben, M. Schmitt, H. Kosch, and B. Schuller, "Seeking the superstar: Automatic assessment of perceived singing quality," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1560–1569.

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

[17] C. Gupta, H. Li, and Y. Wang, "Automatic evaluation of singing quality without a reference," in *Proceedings of APSIPA Annual Summit and Conference*, 2018.

[18] J. Louviere, I. Lings, T. Islam, S. Gudergan, and T. Flynn, "An introduction to the application of (case 1) best–worst scaling in marketing research," *International Journal of Research in Marketing*, vol. 30, no. 3, pp. 292–303, 2013.

[19] B. Çişman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 677–684.

[20] C. Cao, M. Li, J. Liu, and Y. Yan, "A study on singing performance evaluation criteria for untrained singers," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*. IEEE, 2008, pp. 1475–1478.

[21] G. F. Welch, "The assessment of singing," *Psychology of Music*, vol. 22, no. 1, pp. 3–19, 1994.

[22] C. Gupta, H. Li, and Y. Wang, "A technical framework for automatic perceptual evaluation of singing quality," *APSIPA Transactions on Signal and Information Processing*, vol. 7, 2018.

[23] C. J. Plack, A. J. Oxenham, and R. R. Fay, *Pitch: neural coding and perception*. Springer Science & Business Media, 2006, vol. 24.

[24] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," *Journal of New Music Research*, vol. 32, no. 2, pp. 143–152, 2003.

[25] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[26] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 1, pp. 24–33, 1977.

[27] 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," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7815–7819.

[28] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[29] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[30] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," *arXiv preprint arXiv:1706.02921*, 2017.

[31] S. Dieleman, P. Brakel, and B. Schrauwen, "Audio-based music classification with a pretrained convolutional network," in *12th International Society for Music Information Retrieval Conference (ISMIR-2011)*. University of Miami, 2011, pp. 669–674.

[32] G. K. Koduri, J. Serrà Julià, and X. Serra, "Characterization of intonation in carnatic music by parametrizing pitch histograms," in *Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: Proceedings of the 13th International Society for Music Information Retrieval Conference; 2012 Oct 8-12; Porto, Portugal. Porto: FEUP Ediçoes, 2012.* International Society for Music Information Retrieval (ISMIR), 2012.

[33] E. Billauer, "function PeakDet, MATLAB (Converted to python)," https://gist.github.com/endolith/250860, [Online; accessed 20-May-2018].

[34] I. Peretz and K. L. Hyde, "What is specific to music processing? insights from congenital amusia," *Trends in cognitive sciences*, vol. 7, no. 8, pp. 362–367, 2003.

[35] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[36] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 1003–1012.

[37] X. Lin and L. Chen, "Domain-aware multi-truth discovery from conflicting sources," *Proceedings of the VLDB Endowment*, vol. 11, no. 5, pp. 635–647, 2018.

[38] 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," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[39] M. P. Hollier, M. Hawksford, and D. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 141, no. 3, pp. 203–208, 1994.

[40] P. Prasertvithyakarn, "An automatic singing voice evaluation method for voice training," , *2008-03*, pp. 911–912, 2008.

[41] A. Marley, T. N. Flynn, and V. Australia, "Best worst scaling: theory and practice," *International Encyclopedia of the Social & Behavioral Sciences*, vol. 2, no. 2, pp. 548–552, 2015.

[42] S. Inc., "Digital Archive Mobile Performances (DAMP)," https://ccrma.stanford.edu/damp/, [Online; accessed 15-March-2018].

[43] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *Proceedings of the 19th ISMIR Conference*, 2018, pp. 600–607.

[44] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Fifth International Conference on Spoken Language Processing*, 1998.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[46] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[47] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[48] J. Wapnick and E. Ekholm, "Expert consensus in solo voice performance evaluation," *Journal of Voice*, vol. 11, no. 4, pp. 429–436, 1997.

[49] J. M. Oates, B. Bain, P. Davis, J. Chapman, and D. Kenny, "Development of an auditory-perceptual rating instrument for the operatic singing voice," *Journal of Voice*, vol. 20, no. 1, pp. 71–81, 2006.