Assessment and Classification of Singing Quality Based on Audio-Visual Features

Marigona Bokshi, Fei Tao, Carlos Busso and John H.L. Hansen
The University of Texas at Dallas, Richardson TX 75080, USA
marigona90@hotmail.com,fxt120230@utdallas.edu,busso@utdallas.edu,john.hansen@utdallas.edu

***Abstract*—The process of speech production changes between speaking and singing due to excitation, vocal tract articulatory positioning, and cognitive motor planning while singing. Singing does not only deviate from typical spoken speech, but it varies across various styles of singing. This is due to alternative genres of music, singing quality of an individual, as well as different languages and cultures. Because of this variation, it is important to establish a baseline system for differentiating between certain aspects of singing. In this study, we establish a classification system that automatically estimates singing quality of candidates from an American TV singing show based on their singing speech acoustics, lip and eye movements. We employ three classifiers that include: Logistic Regression, Naive Bayes and K-nearest neighbor (k-NN) and compare performance of each using unimodal and multimodal features. We also compare performance based on different modalities (speech, lip, eye structure). The results show that audio content performs the best, with modest gains when lip and eye content are fused. An interesting outcome is that lip and eye content achieve an 82% quality assessment while audio achieves 95%. The ability to assess singing quality from lip and eye content at this level is remarkable.**

***Index Terms*—Multimodal processing, computational paralinguistics, singer quality assessment**

I. Introduction

Singing can be considered an alternative form of speaking, which significantly deviates from typical speech [1]. The singing voice also varies across different singers, genres of music, gender, and experience in voice training. One particularly important variation factor is the quality of singing skill. Singing quality of a person is determined by how well the singer can match a certain pitch, the occurrence of the singer's formant, the singer's long term average spectra, and other traits [2]–[5]. Typically, trained professional music teachers with several years of music experience are involved to determine singing quality of a person. In music auditions, the process of listening to hours of long auditions by candidates can be a tedious and time consuming task. In order to establish both a quantifiable and automated baseline system for modeling singing quality, an initial goal would be to formulate a system to determine if the person's singing quality is either good or bad. In this study, a classification system is formulated to determine singing quality of a person based on audio (acoustic singing speech) and video features (eye and lip structure and movements). We also restrict the analysis to address a specific genre of music, in order to reduce other sources of variability not related with the subjects.

Studies to evaluate the performance of singers have generally focused on the audio structure for singing. In general, a trained singer will need to find the configurations of their vocal tract that produce the exact acoustics dictated by the phonemes being sung. In order to achieve this, the singer will change their oral articulator components such as: lips, tongue, jaw, velum and larynx in order to create an effective resonance balance. Singers also express emotion differently through facial expression such as closing and opening of their eyes during singing. We hypothesize that these cues are important for determining the level of skills of the singers. This study takes up that challenge to evaluate the contributions of facial cues, in addition to acoustic features, in assessing the performance of singers. We present a classification system that automatically determines the singing quality of a person based on their singing speech, combined with lip and eye movements. We believe that by combining the two modalities (audio and video), we can enhance classification and obtain more accurate results than simply using a single modality.

II. Related Work

Studies have investigated the performance of speech based solutions for singing [1], [6], [7]. This study proposes audio-visual models to assess singing quality. Most prior research in singing has focused on music information retrieval. This includes automatic speech recognition of the singing voice and classification based on features that well represent the singing voice quality [8], [9]. The work by Khunarsal et al. [8] focused on automatic speech recognition of singing based on spectrogram pattern matching. Their proposed to use speech processing methods as well as image processing methods in order to recognize the words that are being sung without the help of text or lyric information. The speech signal, in this case, contains music as well as singing, and the music is considered as noise and is attenuated.

Nakano et al. [10] focused on classification of singing speech based on pitch interval accuracy and vibrato features. Automatic classification was achieved by combining pitch with vibrato features in order to classify the singing quality of a person on a binary groups as either good or bad. This is achieved without any help from score information of the sung melody and achieves a classification accuracy of 83.5%. The work of Dalla Bella et al. [11] mainly focused on acoustic analysis of sung performance between occasional and professional singers. Various measures of pitch and time accuracy were computed. The acoustical analysis of sung performance

978-1-5386-0462-5/17/$31.00 ©2017 IEEE.

were targeted for vowels. Other studies have also analyzed pitch accuracy during singing performance [12], [13].

Brown et al. [14] explored the perceptual differences between professionally trained and untrained singers based on features such as singer's formant, percent-jitter, percent-shimmer, and fundamental frequency. In that study, perceptual classification was performed between professional and untrained singer's singing. They also analyzed the differences between singing and speaking. All subjects were asked to sing an excerpt from "America, the beautiful" and their performance was compared across gender, for different feature parameters. The study showed that classification of singing between the two groups reached a classification accuracy of up to 87%, whereas the comparison of singing and speaking achieves a classification of 57%. Omori et al. [15] proposed the *singing power ratio* (SPR) for assessment of singing voice quality, which estimates the ratio of the harmonics peaks observed in frequency bands 2- 4 kHz and 0-2 kHz.

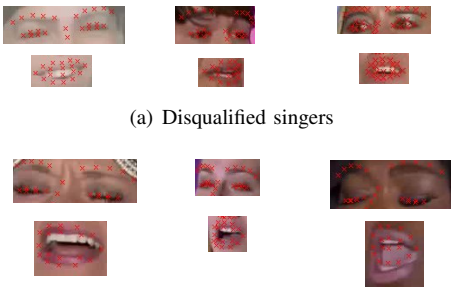
A number of other studies consider the singer's formant as a characteristic of trained singers. Singer's formant is explained as an insertion of an extra formant between the third and fourth formants, and is seen as a peak in the spectrum of a sung vowel [4]. Barrichelo et al. [2] explains the singer's formant as a clustering that occurs when the third, fourth and fifth formants are close in frequency and the peak appears in the vicinity of 3 kHz in all vowel spectra sung by male singers and by altos. This phenomenon happens only during singing, and is mostly useful for operatic singers that need to be heard over an orchestra. The physical configuration of the vocal tract in order to reach this optimal frequency is altered so that the singer's larynx is lowered, creating a longer vocal tract length and therefore new resonant structure.

Although previous research shows results on classification tasks for determining singing quality, none of the previous studies focus on combining other modalities in determining the singing quality of a person. In this current study, we consider both audio and visual features (eyes and lips) in order to perform classification of singing quality of trained and untrained singers of an American TV singing show.

III. Motivation and Resources

A. Motivation

This study performs classification of singing quality based on audio-visual features (speech, lip motion and eye movements). The speech signal is quite reliable in representing the acoustic differences between professional singers and untrained singers. We previously noted that trained singers sometimes obtain what is called a singer's formant, which is a quality that amplifies their voice in settings such as orchestra music. We believe that such qualities and other acoustic features such as fundamental frequency, and formant frequencies, can help establish a reliable classification system to distinguish singing quality. However, including another visual modality can help improve classification. For this study, we chose to incorporate features based on lip and eye movements of trained and untrained singers. Figure 1 shows examples showing the orofacial and eye area of qualified and nonqualified singers.



(a) Disqualified singers

(b) Qualified singers

Fig. 1. Face Features of (a) three nonqualified singers, and (b) three qualified singers

In general, a trained singer will need to find effective articulatory configurations of the vocal tract that produce the exact acoustics dictated by the phonemes being sung. These singers therefore configure their articulators such as oral cavity, jaw and lips in order to produce phonemes at a certain predetermined pitch. Skilled singers produce the right articulatory structure to achieve their goals (e.g., hyperarticulation around the orofacial area). Also, different singers move their eyes in order to express emotion while singing. These observations motivate us to explore lip and eye features as complementary modalities to classify signing quality. This section describes the database use for the study.

http://ieeexplore.ieee.org/document/8304579/

B. Database

One of the contributions of this study is to find accessible recordings to create audiovisual models to assess the quality of singers. Current databases include only speech. Given the goals of our project, we create our own corpus by using audio-visual data from videos downloaded from a video-sharing website. The videos correspond to 96 auditions for an American TV talent singing show. Most candidates that we selected sang pop genre, helping us to specialize our corpus to only one music style, reducing ambiguity and variability across different singing styles. The participants are not professional singers at the time of the audition, although some of them have become well-known professional singers.

We selected video compilations of qualified and nonqualified candidates from different cities and states within The United States. The ground truth information of singing quality information for each candidate is provided by the judges as they select or disqualify the candidate from the audition. Qualified singer are participants that were positively evaluated by the judges in their auditions and were invited to participate in the show ("good" singers). Participants that were not invited to the show are considered as nonqualified singers ("bad" singers). This dataset provides a great opportunity to create audiovisual models to assess the quality of the singers, where reliable ground truth is provided by three judges who are music experts with experience to assess singing quality.

We manually annotate parts of the video showing frontal faces of the singers. These videos allow us to use automatic

TABLE I
NUMBER AND GENDER DISTRIBUTION OF SINGERS.

	Male	Female	Total
Qualified	25	30	55
Nonqualified	21	20	41

algorithm to extract facial features. These segments range in duration from 5 to 15 seconds. Each video is a compilation of only one category of singers: qualified or disqualified. Both categories have a mixture of male and female singers, making it possible to consider gender differences. We collect approximately one hour of singing data for each category: qualified singing candidates and nonqualified singing candidates. Table I shows the number of singers and the gender distribution.

IV. METHODOLOGY

We are interested in classifying qualified versus nonqualified singers. First, we extract the audio and video signals from the segments, which we individually pre-process. Then, we extract acoustic (Section IV-A) and facial features (Section IV-B), which are used for classification. We explore various machine learning algorithms (Section IV-C) using unimodal and multimodal features.

A. Acoustic Features

The audio was extracted from the videos. The audio is originally sampled at 44 kHz, but we down-sample all audio files to 8 kHz. This sampling frequency preserve most of the spectral frequencies associated with speech from the audio signal. We derive Mel-frequency cepstral coefficients (MFCCs) from the audio. We use 12 MFCCs in addition to their delta and delta-delta coefficients (first- and second-order frame-to-frame difference). This approach incorporate the temporal dynamic, as well as static spectral content, creating a 36 dimensional feature vector representing the speech signals. MFCCs are the most commonly used features in speech recognition and speaker identification systems, and have been shown to represent the envelope of the shape of the vocal tract well. We believe these features are suitable for our current classification system as well.

B. Facial Features

Figure 1 shows examples of lip and eye regions for qualified and nonqualified singers. We observe that the mouth and eye areas are important cues, so we focus our analysis on these features. We automatically extract visual features from the videos. We use IntraFace [16] to extract the coordinates in pixels of facial features, including the location of the lips and eyes. InterFace is a robust toolkit developed by the Carnegie Mellon University that detects and tracks 66 facial features landmarks in a video. It works well under different head poses. From them, we use the x and y coordinates of 17 facial landmarks that represent the lip locations and 10 facial landmarks that represent the locations around each eye. Figure 2 shows these facial landmarks of a qualified singer. These coordinates are normalized to account for changes in the zoom of the camera resulting in faces with different sizes. Then, we use these features as vertices to calculate the lip and eye areas on a frame-by-frame basis (function "polyarea" in Matlab).

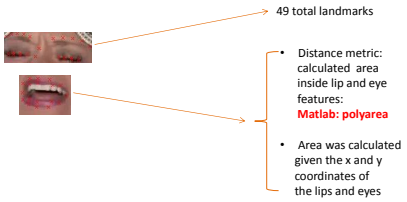


Fig. 2. Face features of a qualified singer extracted with IntraFace

C. Classification

For classification, we use Weka [17] – a data mining software in Java developed by the University of Waikato. We extract acoustic and facial features for the frames recorded by each speaker, and all their frames are labeled as either "good", for qualified singers, or "bad", for non-qualified singers. Since we only have a limited number of recordings, we use a 10-fold cross validation technique to train and test the system. Cross validation works by splitting the entire data into n folds (in our case n=10) and create n separate experiments. In each experiment, we use one fold for testing and remaining data for training. We then collect all scores from all experiments to determine final decision accuracy of the system. Cross-validation is an effective technique to use when the amount of training data is limited, and to avoid randomly picking a test set that may not be representative of the singers/subjects.

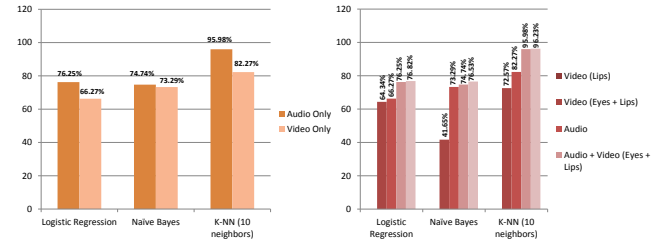
We trained and tested our system using a Logistic Regression linear classifier, the Naïve Bayes non-linear classifier, and a k-NN classifier. The classification is conducted at the frame level. We compare results across each of these classifiers and investigate which classifier is more suitable for classifying our dataset with the highest accuracy. To evaluate the complementary information provided by audio-visual modalities, we implement an early fusion framework, where we concatenate the corresponding features.

V. EXPERIMENTAL EVALUATION

We obtain our results by training three classifiers (Logistic Regression, Naïve Bayes and k-NN) using the methodology described in Section IV. We compare our results based on: unimodal features (audio-only and video-only – Section V-A), and multimodal features (audio + video – Section V-B).

A. Unimodal Features

The first evaluation consists of determining the discriminative power of each modality. Therefore, we train the classifiers using either audio or facial (eyes + lips) features. Figure 3(a) shows the accuracy of the classifiers. Naïve Bayes shows the lowest accuracy for audio features and Logistic Regression shows the lowest accuracy for video features. k-NN shows the highest accuracies for both audio (95.98%) and visual (82.27%) features. We observe that classifiers trained with audio features outperform the ones trained with facial features for the three machine learning algorithms considered in this study. As expected, acoustic features provide the most discriminative information to determine the quality of the singers, reaching 95.98% accuracy. Good singers are able to modify their vocal cavity to reach target articulations (e.g., hitting the right formats). As a result, the spectral component in the



(a) (b)

Fig. 3. (a) Unimodal feature accuracy: audio-only and video-only (lips + eyes), (b) Multimodal and unimodal feature accuracy comparison.

acoustic signal is particularly important. This information is efficiently captured by MFCCs.

It is also remarkable that the accuracy for the k-NN classifier trained with facial features reaches 82.27% accuracy. This result demonstrates the discriminative power of the facial features proposed in this study. The high performance suggests that facial features can be combined with acoustic features to improve the classification performance. The audio-visual fusion is particularly appealing in cases where the audio features cannot be robustly extracted (e.g., presence of noise).

B. Multimodal Features

After evaluating unimodal classifiers, we evaluate the classification improvement when we concatenate the audio and facial features, creating an extended feature vector (e.g., feature level integration). More sophisticated fusion approaches are left as future work. Figure 3(b) shows the results, which indicate a slight improvement in performance for all classifiers. The highest classification accuracy improvement (2% absolute) was achieved with the Naïve Bayes classifier. The improvement in accuracy demonstrates the complementary information between the modalities, supporting our claim that lip and eye features help in singing quality assessment when added to audio features.

Figure 3(b) also shows the performance for face-only features for two conditions: when we only use lip features, and when we use lip and eye features. The accuracy improves by up to 31% for the Naïve Bayes classifier and about 10% for the k-NN classifier. This confirms our hypothesis that eye movements fused with lip movements increase accuracy.

VI. Conclusions

In this study, we performed classification of singing skill based on audio, lip and eye features. We noted that our results improved (by up to 2% absolute) when we adding lip and eye features, compared to using the audio modality only. We also noticed a significant improvement in performance (up to 31%) when fusing lip and the eye features together. The results support our hypothesis that lip and eye features can help determine singing skill and can be used to supplement audio features as a primary backbone to determine singing quality. We believe that we could obtain better results if we used a database where the singers are always front-facing the camera. One application for this study would be to perform automatic singing skill assessment.

In the future, we could incorporate other features that represent different facial landmarks. One type of features used in this case would be the Gabor filter. Also, it would be possible to choose dynamic fusion techniques instead of directly concatenating the features, and weigh the different features differently (i.e, system fusion versus feature fusion).

References

[1] M. Mehrabani and J. Hansen, "Dimensionality analysis of singing speech based on locality preserving projections." in *14th Annual Conference of the International Speech Communication Association, INTERSPEECH 2013*, Lyon,France, Aug. 2013, pp. 2910–2914.

[2] V. Barrichelo, R. Heuer, C. Dean, and R. Sataloff, "Comparison of singer's formant, speaker's ring, and lta spectrum among classical singers and untrained normal speakers," *Journal of voice*, vol. 15, no. 3, pp. 344–350, 2001.

[3] J. Estis, A. Dean-Claytor, R. Moore, and T. Rowell, "Pitch-matching accuracy in trained singers and untrained individuals: the impact of musical interference and noise," *Journal of Voice*, vol. 25, no. 2, pp. 173–180, 2011.

[4] J. Sundberg, *The acoustics of the singing voice*. Scientific American, 1977.

[5] S. Dalla Bella, "Defining poor-pitch singing," *Music Perception: An Interdisciplinary Journal*, vol. 32, no. 3, pp. 272–282, February 2015.

[6] M. Mehrabani and J. H. Hansen, "Singing speaker clustering based on subspace learning in the gmm mean supervector space," *Speech Communication*, vol. 55, no. 5, pp. 653–666, June 2013.

[7] M. Mehrabani and J. Hansen, "Language identification for singing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4408–4411.

[8] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Singing voice recognition based on matching of spectrogram pattern," in *International Joint Conference on Neural Networks, 2009. IJCNN 2009.* Atlanta,GA,USA: IEEE, June 2009, pp. 1595–1599.

[9] B. Kostek and P. Zwan, "Automatic classification of singing voice quality," in *5th International Conference on Intelligent Systems Design and Applications, 2005. ISDA'05.* Wroclaw,Poland: IEEE, Sept. 2005, pp. 444–449.

[10] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *International Conference on Spoken Language (ICSLP 2006)*, Pittsburgh, PA, USA, September 2006, pp. 1706–1709.

[11] S. Dalla Bella, G. Jean-François, and I. Peretz, "Singing proficiency in the general population," *The journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1182–1189, 2007.

[12] P. Pfordresher, S. Demorest, S. Bella, S. Hutchins, P. Loui, J. Rutkowski, and G. Welch, "Theoretical perspectives on singing accuracy: an introduction to the special issue on singing accuracy (part 1)," *Music Perception: An Interdisciplinary Journal*, vol. 32, no. 3, pp. 227–231, 2015.

[13] S. Demorest, P. Pfordresher, S. Bella, S. Hutchins, P. Loui, J. Rutkowski, and G. Welch, "Methodological perspectives on singing accuracy: an introduction to the special issue on singing accuracy (part 2)," *Music Perception: An Interdisciplinary Journal*, vol. 32, no. 3, pp. 266–271, 2015.

[14] W. Brown, H. Rothman, and C. Sapienza, "Perceptual and acoustic study of professionally trained versus untrained voices," *Journal of Voice*, vol. 14, no. 3, pp. 301–309, 2000.

[15] K. Omori, A. Kacker, L. Carroll, W. Riley, and S. Blaugrund, "Singing power ratio: quantitative evaluation of singing voice quality," *Journal of Voice*, vol. 10, no. 3, pp. 228–235, September 1996.

[16] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Portland, OR: IEEE, June 2013, pp. 532–539.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.