

AN EVALUATION TOOL FOR SUBJECTIVE EVALUATION OF AMATEUR VOCAL PERFORMANCES OF “AMAZING GRACE”

Elena Georgieva^{1,2}, Camille Noufi¹, Vidya Rangasayee¹, Blair Kaneshiro¹, Jonathan Berger¹

¹Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, USA

²Music and Audio Research Lab (MARL), New York University, USA

egeorgie@ccrma.stanford.edu

ABSTRACT

In order to study performance characteristics of untrained, amateur singers, we developed an online tool through which coders could evaluate real-world vocal performances of “Amazing Grace” from a Smule dataset. Coders from Stanford University used the online evaluation tool to deliver judgments of age and gender of the performers, as well as skill, likeability, and expressiveness of the vocal performances. Initial results show subjective evaluations of skill, likeability, and expressiveness are highly correlated, and coders rarely gave the highest possible score in any of the three metrics. This online evaluation tool can be used in future computational studies of vocal performance.


1. INTRODUCTION

Singing is accessible to anyone with a voice, and even those who do not self-identify as “singers” have expressed themselves through vocal performance. We present recent work on an online platform for evaluating real-world performances by amateur singers. This tool has the potential to advance the study of natural tendencies of untrained human vocal expression.

We utilize the Digital Archive of Mobile Performances (DAMP), recorded and collated by Smule, Inc.¹ The archive includes vocal recordings from smartphone users using Smule’s karaoke mobile application.² The app allows users to sing along to a karaoke track of their choosing. The Smule app has over 50 million monthly users worldwide, of all ages and levels of vocal training. In the current study, we focused on performances of the song “Amazing Grace” in the Amazing Grace Vocal Performances dataset. The DAMP dataset includes 11,937 performances of this widely familiar American hymn [1].

¹ <https://ccrma.stanford.edu/damp/>

² <https://apps.apple.com/us/app/smule-the-1-singing-app/id509993510>

 © Elena Georgieva^{1,2}, Camille Noufi¹, Vidya Rangasayee¹, Blair Kaneshiro¹, Jonathan Berger¹. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Elena Georgieva^{1,2}, Camille Noufi¹, Vidya Rangasayee¹, Blair Kaneshiro¹, Jonathan Berger¹, “An Evaluation Tool for Subjective Evaluation of Amateur Vocal Performances of “Amazing Grace””, *Extended Abstracts for the Late-Breaking Demo Session of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020.

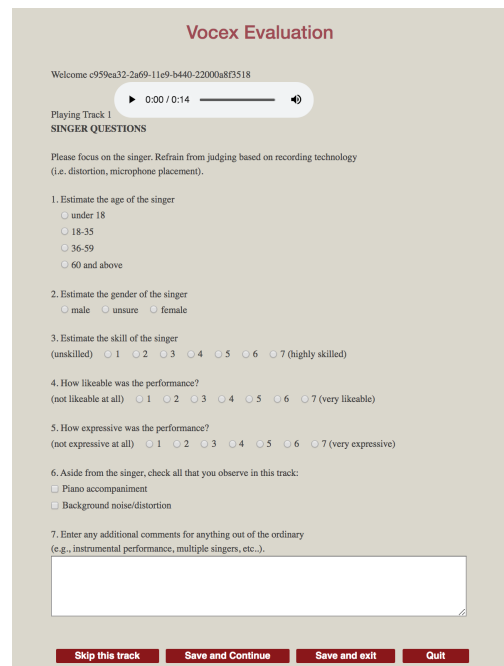


Figure 1. Online vocal expression evaluation interface.

Plenty of researchers study vocal expression in trained, professional-level vocalists [2]. Trained singers are taught a set of principles and techniques in voice lessons and choral singing, for example, how to execute a phrase via volume and breath support or how to fluidly navigate the different vocal registers. Furthermore, many researchers investigate pitch in vocal performance and evaluate the skill or like-ability of vocal performances [3]. Pitch can also be used to characterize vocal performances and singing styles [4]. Pitch is an important objective measure of vocal skill, but is only one factor among many in determining perception of vocal “skill” or “likeability.” [5]. By means of perceptual evaluation, we hope to understand the skill, likeability, and expression of casual singers.

2. METHODS

To gather subjective evaluations of human performances, we created an online evaluation platform and recruited coders to evaluate vocal performances. We were not able to truly crowd-source the evaluations due to the terms of Smule’s Research Data License Agreement. Our plat-

form was hosted on Heroku and data was stored in a Postgres database. The frontend user interface was built using Javascript, supported by a Node.js backend. Each page of our platform included a 14-sec audio recording, followed by six questions with radio buttons, and a space for open-ended comments. Coders were required to be members of the Stanford community, have typical hearing, and be 18 years of age or older and fluent in English. All coders accepted the terms of Smule’s Research Data License Agreement.

Coders had to pass a pre-screening stage to qualify to evaluate vocal performances. The purpose of the pre-screening was to guarantee that coders understood the evaluation task and had reasonable ability to listen critically to vocal performances. Twenty-one coders passed the screening and proceeded to evaluate vocal performances of “Amazing Grace”.

To evaluate a single performance, coders rated five aspects of a singer’s voice while listening to the recording: estimates of age (under 18, 18–35, 36–59, 60 and above) and gender (male, unsure, female), skill level (seven options ranging from unskilled to highly skilled), how likeable (seven options ranging from not like-able at all to very like-able), and how expressive the performance was (ranging from not expressive at all to very expressive). A question asking if the coder observed background noise and a space to include additional comments concluded the evaluation of a single performance (Figure 1).

3. RESULTS

We collected 1,598 evaluations covering 1,296 distinct tracks. 881 of these tracks were listened to and evaluated once, 149 twice, and 27 three times. Initial results indicate that many of the singers are judged to be in the 18-35 age group, and the majority of “Amazing Grace” performers are judged to be female.

On average, coders gave performers a skill score of 3.35 (where 1 is “unskilled” and 7 is “highly skilled”). Coders gave performers an average likability score of 3.63, and the expressiveness average was 3.36. Listeners seemed to find vocal performances more like-able than skilled or expressive, on average. These results are skewed to smaller values, as coders rarely gave a rating of 7. Vocal performers were deemed “highly skilled” only 18 times of the 1,600 evaluations, “very like-able” 23 times, and “very expressive” 22 times.

All categories of ratings are positively correlated (all $r \geq 0.72$, Figure 2). Most notably, the perceived skill of a performer and expressiveness of their performance have a correlation of 0.83. Meanwhile, the correlation coefficient of perceived skill and like-ability was 0.73, and the coefficient of like-ability and expressiveness was 0.72.

4. CONCLUSIONS AND FUTURE WORK

In the future, we aim to have each track evaluated three times by three different coders. This will allow us to investigate how much agreement exists between perceptual

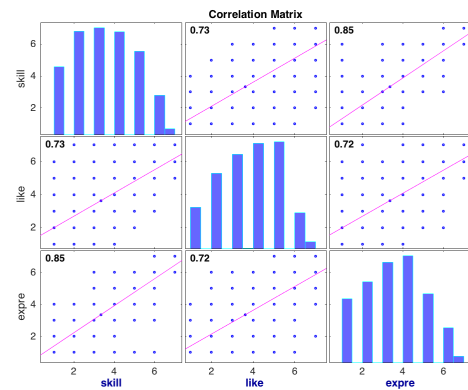


Figure 2. Ratings of skill, likeability, and expression are positively correlated.

evaluations, and what types of performance characteristics produce consistent or contradictory codings. We are also interested in investigating the relationship between traditional, objective measures of a “skilled” voice and the subjective measures we are collecting in the evaluation. We will use pitch track deviation as our initial objective measure, and potentially use a supervised classifier to further investigate the importance of subjective versus objective characteristics in the perception of amateur vocal performances [6].

5. ACKNOWLEDGMENTS

Thanks to Perry Cook for data pre-processing and advice.

6. REFERENCES

- [1] C. Noufi, V. Rangasayee, S. Ciresi, J. Berger, and B. Kaneshiro, “A model-driven exploration of accent within the amateur singing voice,” in *Machine Learning for Music Discovery Workshop, ICML*, 2019.
- [2] K. Scherer, “The expression of emotion in the singing voice: Acoustic patterns in vocal performance,” *The Journal of the Acoustical Society of America*, 2017.
- [3] B. Bozkurt, O. Baysal, and D. Yüret, “A dataset and baseline system for singing voice assessment,” in *Proc. of the 13th International Symposium on CMMR, Matosinhos, Portugal*, 2017.
- [4] M. Panteli, R. Bittner, J. Bello, and S. Dixon, “Towards the characterization of singing styles in world music,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.
- [5] J. Smith, “Correlation analyses of encoded music performance,” *Doctoral Dissertation, Stanford University, U.S.A.*, 2013.
- [6] J. Böhm, F. Eyben, M. Schmitt, H. Kosch, and B. Schuller, “Seeking the superstar: Automatic assessment of perceived singing quality,” in *International Joint Conference on Neural Networks (IJCNN)*, 2017.