An Evaluation Tool for Subjective Evaluation of Amateur Vocal Performances of “Amazing Grace”

Elena Georgieva, Camille Noufi, Vidya Rangasayee, Blair Kaneshiro, Jonathan Berger

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.**
- This online evaluation tool can be used in future computational studies of vocal performance.

Dataset

- Digital Archive of Mobile Performances (DAMP)[1], recorded and collated by Smule, Inc.
- Vocal recordings from smartphone users using Smule’s karaoke mobile application[2].
- DAMP includes 11,937 performances of “Amazing Grace,” a widely familiar American Hymn.

Methods

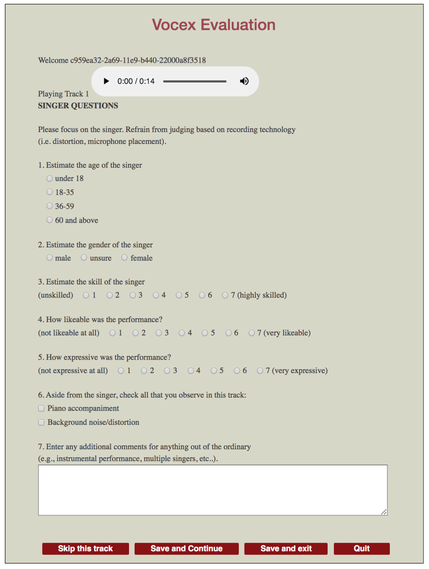

Figure 1. Online vocal expression evaluation interface.

- Goal: to gather subjective evaluations of human performances, specifically performances by casual singers.
- Evaluation platform was hosted on Heroku, data stored in Postgres database.
- Coders were members of the Stanford community and had to pass a pre-screening stage to qualify to evaluate vocal performances.
- Each audio recording: 14 second in length
- Evaluation form was 6 questions with radio buttons and a space for open-ended comments.
- 2 demographic questions, 3 subjective questions

Evaluate the **skill** of the singer.
How **likeable** was the performance?
How **expressive** was the performance?

Results

Number of evaluations collected: 1,598
Number of distinct tracks evaluated: 1,296
21 coders evaluated performances

- On average, listeners seemed to find vocal performances more likeable than skilled or expressive (average score 3.63 vs. 3.35 and 3.36, respectively).
- Results skewed to lower values—coders rarely gave a rating of 7 for any of the questions.
- All categories are positively correlated ($r \geq 0.72$)
- Most notably, the perceived skill of a performer and the expressiveness of their performance have a correlation of 0.83

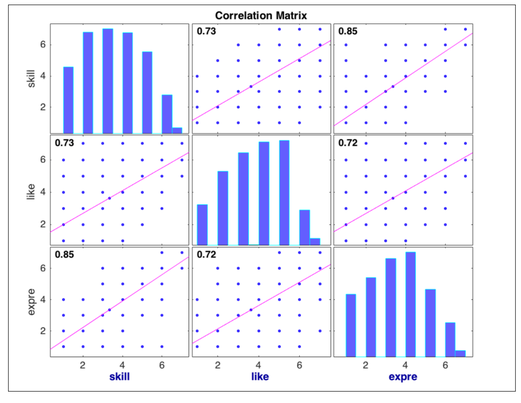

Figure 2. Ratings of skill, likeability, and expression are positively correlated.

References

[1] https://ccrma.stanford.edu/damp **[2]** https://apps.apple.com/us/app/smule- the-1-singing-app/id509993510 **[3]** C. Noufi et al., “A model-driven exploration of accent within the amateur singing voice,” ML4MD, ICML, 2019. **[4]** K. Scherer, “The expression of emotion in the singing voice: Acoustic patterns in vocal performance,” JASA, 2017. **[5]** B. Bozkurt et al., “A dataset and baseline system for singing voice assessment,” CMMR, 2017. **[6]** M. Panteli, et al., “Towards the characterization of singing styles in world music,” IEEE ICASP, 2017. **[7]** J. Smith, “Correlation analyses of encoded music performance,” Doctoral Dissertation, Stanford University, 2013. **[8]** J. Böhm, et al., “Seeking the superstar: Automatic assessment of perceived singing quality,” IJCNN, 2017.

Contact: egeorgie@ccrma.stanford.edu