Empirically weighing the importance of decision factors when selecting music to sing

Conference Paper · September 2018

CITATIONS
0

READS
54

4 authors, including:

Karim M. Ibrahim
Nile University
5 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Chitralekha Gupta
National University of Singapore
14 PUBLICATIONS **33** CITATIONS
SEE PROFILE

EMPIRICALLY WEIGHING THE IMPORTANCE OF DECISION FACTORS WHEN SELECTING MUSIC TO SING

Michael Mustaine[1] **Karim M. Ibrahim**[1]
Chitralekha Gupta[1,2] **Ye Wang**[1]
[1] School of Computing, National University of Singapore, Singapore
[2] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore
baronemda@gmail.com, wangye@comp.nus.edu.sg

ABSTRACT

Although music cognition and music information retrieval have many common areas of research interest, relatively little work utilizes a combination of signal- and human-centric approaches when assessing complex cognitive phenomena. This work explores the importance of four cognitive decision-making factors (familiarity, genre preference, ease of vocal reproducibility, and overall preference) influence in the perception of "singability", how attractive a song is to sing. In Experiment One, we develop a model to validate and empirically determine to what degree these factors are important when evaluating its singability. Results indicate that evaluations of how these four factors impact singability strongly correlate with pairwise evaluations ($\rho = 0.692, p < 0.0001$), supporting the notion that singability is a measurable cognitive process. Experiment Two examines the degree to which timbral and rhythmic features contribute to singability. Regression and random forest analysis find that some selected features are more significant than others. We discuss the method we use to empirically assess the complex decisions, and provide a preliminary exploration regarding what acoustic features may motivate these choices.

1. INTRODUCTION

A fundamental task of MIR is to develop of acoustic feature extractors that capture unique characteristics from a recorded piece of sound. However, some acoustic features may not be wholly represented in the acoustic signal, and MIR has been criticized for failing to model analysis based on psychological research [3]. For example, "danceability" - the perceptual experience of grooviness [23, 48] - is a feature available in signal processing packages [9, 10], and open-access APIs [1] using a combination of beat salience and consistency [36]. However, based soley on these acoustic properties the most danceable song would be closer to a steady, metronomic pulse, which clearly does not capture the perceptual nuances of what makes music danceable [14]. The inclusion of psychological acoustic features using signal-only analysis is surprising, given that music is a dynamic system influenced by cognitive [20], cultural, market, and political forces [8]. Despite this knowledge, research is relatively sparse as to how, or to what degree, specific acoustic features influence musical preference. Part of the scarcity may be due to the relative difficulty in quantifying the influence of important psychological features empirically. This work examines the extent to which a cognitive psychology, signal processing, machine learning, and economic decision-making can be used to investigate a previously unexplored psychological perception of "singability": the degree to which a song is attractive to sing. To our knowledge, no empirical study has been conducted which explores whether a feature such as singability can be extracted from a piece of music.

Determining a complex psychological process and decision making strategy like singability is a difficult task. To start, it is intuitively difficult to quantify such a subjective multiple criterion choice in a controlled, scientific manner. Because singability will likely not contain a universally agreed upon set of factors, the major challenge is defining a method that can quantify how - and to what degree - these factors should be incorporated into a model for evaluation. We first introduce some background on closely related concepts to our interpretation of singability from psychological experiments and MIR applications.

1.1 Related Work

Perhaps the most historically relevant psychological research relating to singing preference was initially proposed by Berlyne [7]. Berlyne suggests that music exhibits an inverted-U-shaped relationship for preference, influenced by novelty, complexity, and tone. This model has been replicated independently from a variety of perspectives including personality and preference research [29], and flow

[1] https://developer.spotify.com/web-api/get-audio-features/

states [12].

Prior research on singability has focused on music recommendation systems for digital karaoke applications [25, 26]; they used a competence-based evaluation, and recommended music using an individual's singing proficiency. These systems define singing preference solely on whether one can recreate the original performance [16] and fails to consider other aspects of preference such as familiarity on preference; if the ability to recreate an original version of a song is the sole criteria for determining singable tracks, a naïve extension to improve performance would be to recommend songs based on demographic features such as age, sex, and height through automated assessment of singing voice [46].

2. SINGABILITY FACTORS

We examine singability using a synthesis of multiple criterion decision making processes, acoustic feature extraction, and machine learning founded on a theoretical background of music cognition. Based on the general research discussed above, we expand the interpretation of singability from other research [25, 26] to include more factors than just the ability to reproduce the original rendition of a track. For the purpose of this work, singability is defined as a psychological process which includes how attractive a song is to sing without concern of social consternation for being unable to produce the original vocalizations. Based on this refined definition, we consider four factors which could impact singability and include: i) familiarity, ii) genre, iii) preference to listen (listenability), and iv) producibility.

To maintain a realistic scope for exploratory research, we did not include an exhaustive list of potential singability factors. These factors were selected due to their relative presence in the psychological literature. We also were interested in selecting features that would be less demanding to ask crowdsourced workers; other features we did not explore, such as the importance of lyrics or social factors, could be analysed using methodology specific to their disciplines should compelling evidence for singabiliity be found. Next, we highlight research specific to these factors, then describe a method to quantify the prioritization of them when making a complex, multiple criterion decision.

2.1 Familiarity

Familiarity has important influences on preference formation. The mere exposure effect, a foundational psychological process [28, 49], demonstrates that increased exposure to essentially anything increases your preference for it, even when unaware of it's inclusion in your immediate environment [24]. In [32], the mere exposure effect was also found to impact music preference; multiple repetitions of unfamiliar music [28], and random tone sequences [47] increased preferences for them. A possible reason for why familiarity increases preference is because it improves ease of processing [30], impacting the complexity component of Berlyne's optimal complexity model described in Section 1.1.The relationship between familiarity through mere exposure appears to occur early in cognitive processing - Korsakoff amnesics demonstrate increased liking to musical stimuli through increased exposure [19].

However, it is important to consider that increased familiarity does not increase preference in all cases; most people do not actively listen to extremely familiar songs such as Twinkle, Twinkle, Little Star. This still makes sense when considering Berlyne's optimal complexity model (Section 1.1) - extremely familiar music is too simple or not novel enough to engage. Therefore, it is hypothesized that although familiar music is important for singability, music that is too familiar will not be preferred.

2.2 Genre

Genre preference describes a specific aspect of the mere exposure effect through common acoustic features which are hallmark in the genres you typically listen to. For example, Rap music has a high degree of speech, and Metal music generally is high tempo, and with negative valence [4]. This form of familiarity is more active and personal, aligning more closely to the role that individual preference plays in exposure. Neurological evidence for an active mere exposure effect through genre has been demonstrated in brain imaging studies. Using electroencephalography, Mismatch Negativity Responses (MMNs; a spike in brainwave polarization when expectations are violated) can be elicited with tone sequences in the first few trials regardless of formal musical training [38]. In a subsequent study, authors of [39] found that MMN responses, were stronger when genre conventions were defied in a participants preferred musical style. In a study containing 17 million users from over 30 countries, users download tracks of secondary genres acoustic features similar to those of their most preferred genre [4]. For example, users who had clear preferences for Rap music preferentially downloaded tracks from other genres that contained more speech sounds. We therefore hypothesize that genre plays an important role in the selection of a preferred song to sing.

2.3 Listenability

The definition of listenability used for this work refers to how attractive a song is to listen to. Although it may be appealing to suggest that songs that are listenable are by extension singable, they must be considered mutually exclusive. Rap or Metal music for example may fit this category as the vocalizations required are not conducive for singing, but are still highly popular and can be very listenable. Furthermore, listenability is distinct from familiarity, but can be influenced by it. As suggested in Section 1.1, nursery rhymes are highly familiar, but are likely not considered highly listenable or singable by most. Highly listenable songs may also not be familiar because older tracks are played significantly less than newly released songs. Listenability may be best differentiated from familiarity in

that it can be an immediate process, requiring only a single exposure in order to be evaluated as attractive. Cognitive processing of various complex musical features such as genre [21] can happen at millisecond timescales. Listenability is considered an important factor for singablility because it increases the likelihood that a song will be selected to or attended by users (thus directly influence the likelihood it will be sung in the first place) and because they are more salient in memory.

2.4 Producibility

Music cognition research has examined the distinction of singing quality; the perceptual or acoustic features that make trained singers sound better than amateurs. Quality of singing voice has been assessed with respect to full upper resonance in a singer's formant range (known as the singer's formant, a prominent spectral envelope of 3kHz) as of singing voice quality [5]. Professional singers have higher formant intensity than untrained voices; relative amplitudes of singer's formants grew as vocal intensity increased and diminished as pitch rose [35], trained voices have more energy in the formant range but not for all pitches, and males in general have higher formant intensity than females [35]. The singer's formant appears to be a particularly important property for classical operatic singers to project above the orchestra [37].

Although measures regarding whether an individual has vocal training can be assessed through the singer's formant, producibility is not contingent on these features. For example, untrained singers with self-expressed singing talent have identical pitch matching accuracies when compared to trained singers [45]. Producibility based on vocal features which indicate professional training may also not be appropriate because the correlation between genre preference and training does not align with what is popularly sung; individuals with more musical training show increased preferences for "serious" genres such as Classical and Jazz, but not other genres such as Pop [17].

3. EXPERIMENT ONE: VALIDATING SINGABILITY

To our knowledge, there is no prior work that examines whether what people think makes a song singable correlates with what they actually select in natural settings. For example, [41] instructed professional musicians to evaluate recordings of top-three placing performances from piano competitions under three conditions, recordings with: i) video only, ii) audio only, or iii) audio and video. Participants accurately ranked the video-only condition more consistently with who won the competition than in any other condition; the audio-only condition was the least consistent. This work establishes that it is possible that our impressions of what features are important in our musical preferences may not be internally consistent.

We combine a series of psychological analysis methods to establish whether singability can be consistently assessed among individuals using a set of 50 popular song excerpts. To establish a bottom-up ground truth, a forced alternative choice (FAC) experiment is conducted with pairs of songs; a complex decision-making model known as Analytic Hierarchical Process (AHP) [33] is used to determine top-down impressions. We then rank songs based on their assessed singability using both methods (FAC and AHP) to determine whether there is consistency between what we think is singable, and what our decisions end up inevitably being. An additional benefit of using AHP is that it can weigh the degree to which each of the four features described above contributes to an individuals choice to sing a song. Because AHP is less commonly used, we briefly describe AHP and how it is conducted prior to reporting experimental structure.

3.1 Analytic Hierarchical Process

AHP is an technique to quantify how, and to what degree, subjective criteria influence a complex decision making task. The validity of the AHP has been examined extensively [44], and has been used within government, business, and healthcare [42]. Figure 1 illustrates the final importance values for each factor and are now described. Determining singability using AHP involves breaking down the decision problem into a set of global priorities (green boxes). Global priorities are a set of general factors that are suspected to influence the decision-making process. After global priorities are determined, levels within each priority (local priorities; blue boxes) are established. Once priorities have been established, the importance of each factor can be systematically evaluated to determine their contribution to the final decision. Decision makers weigh the importance of each of these priorities using multiple pairwise comparisons, and require the decision maker to evaluate every priority relative to another. For instance, a worker is asked "how important was it that the vocals were easy to reproduce, as opposed to moderately difficult". Because more than one worker answered the same question multiple times, we take the average importance value from all comparisons as the final importance value. Priorities are calculated by dividing the importance of the first comparison over the other. A pairwise comparison matrix is generated after all evaluations are made by multiplying the entries of each row and taking the nth root of the product. The roots are then summed and normalized to produce an eigenvector representing the priority importance. [2]

3.2 Methods

The dataset contains excerpts of 50 songs (ten songs from five genres) from the top 50 Billboard chart songs between the years of 2011-2015. Selected songs had equal numbers of male and female singers (five per sex per genre). In order to reduce high degrees of familiarity, songs from the bottom of the list were selected. 15-seconds of audio was extracted from each artist's official YouTube channel. Audio was extracted from the video as mp3 files.

[2] For in-depth example, see: `http://rad.ihu.edu.gr/fileadmin/labsfiles/decision_support_systems/lessons/ahp/AHP_Lesson_1.pdf`

A two-part online survey was crowdsourced using Amazon's Mechanical Turk. [3] Although some research suggests that the quality of crowdsourced data is more diverse and at times better than data collected in traditional laboratory settings [6], additional metrics which validate or refine analysis highlighted in Section 1.1 should be considered. The first part of the experiment consisted of a series of FACs. Workers were instructed to listen to excerpts of two songs. They were asked to determine which song was more singable, listenable, and whether either of the songs were familiar. Workers repeated this paradigm for five pairs of songs in total.

After completing the FAC section, workers were then instructed to complete an AHP after briefly reflecting on the choices they made when selected between pairs of songs. Different levels of local priorities are established for each global priority. Five levels for genre were selected, Rock, Pop, Alternative, Country, and Rap music were selected; two levels for familiarity (low and high); three levels for producibility (easy, medium, hard) and; three levels for preference to listen (low, medium, high). In order to keep the task as simple as possible for workers, we reduced the number of local priorities to as little as possible - unlike producibility and listenability, only two levels were selected for familiarity because we wanted to know whether any prior knowledge of a piece would influence their choice.

Importance values were calculated by taking the average response for each priority across all respondents. Lastly, we requested workers to report only their sex - we did not collect information regarding worker age, socioeconomic status, or ethnicity. The reasoning for this was two-fold: i) Mechanical Turk demographic variability is in general more diverse than traditional laboratory data collection [6], and; ii) we were interested in establishing the general existence of a psychological perception from a rarified set of possible influencers before examining how dynamic anthropological and sociological factors modulate the preference. A benefit of using AHP is that it is a simple process to add or remove global priorities and replicate the experiment easily with new variables and interactions.

3.3 Analysis

Pairwise comparisons were conducted for all 50 songs (1225 pairs), each job instructed users to evaluate 5 pairs (245 jobs), and each job was assessed 3 times (735 surveys conducted). 88 submissions (11%) were rejected for incorrectly answering a confirmatory test question. [4] A worker was compensated $0.07 USD per job and could perform up to five surveys. 245 unique respondents (44% male, 56% female) completed the survey. On average, workers agreed with each other that one song was more singable than another 77.8% of the time. We examined whether individuals selected a song as more singable based on the sex of the artist. A binomial test indicated that individuals selected same-sex singers slightly more often ($53\%; p < 0.001$), though the difference was marginal.

Once AHP priorities were calculated, songs were segmented into bins for the familiarity (high and low), and listenability (high, medium, and low) categories based on the survey responses. Figure 1 represents the global and local priority values generated through the Mechanical Turk survey. Song rankings for AHP were derived for each song by producing a rank-order based off the product of local priority values for genre, listenability, and familiarity. For example, a Rock song which was in the top 50th percentile for familiarity, and the bottom 33rd percentile for listenability would receive a singability value of $0.227 * 0.613 * 0.299 = 0.0416$. Producibility was not included in the calculation because this feature is relative to an individual's skill at singing and can only be evaluated for each user, as opposed to each song. Ranks for the FAC portion of the experiment were generated by ordering the amount of times any given song within a pairwise comparison was selected by the user as more attractive to sing.

Once ranks were generated for the bottom-up (FAC), and top-down (AHP) processes, we conducted a Spearman-ρ rank correlation. Ranks derived from FAC are highly correlated with ranks derived from the AHP ($r_s = 0.691, p < 0.0001$). 47.61% of the variance in rank could be accounted for across ranked derived from FAC and AHP. Figure 2 plots the ranks derived for each song excerpt. Each song's coordinates represent the FAC derived rank (x-axis) to the AHP derived rank (y-axis). Significant Spearman-ρ correlations were also found comparing Billboard ranks to FAC ($r_s = 0.518, p < 0.001$) and AHP ($r_s = 0.540, p < 0.0001$).

3.4 Discussion

The purpose of experiment one was to derive a method that can determine whether people's heuristic impressions of preference reliably predicts their actual decisions. The highly significant correlation ($p < 0.0001$) and large effect size ($r^2 = 0.4761$), supports the hypothesis that people's top-down assessments of singability are features they actually use when making the decision. This finding is significant because it supports the notion that a less labour intensive process is needed for determining a music-cognitive process; you do not need to conduct a bottom-up comparison for the entire corpus of music to determine general preference. The results suggest that listenability is the most important feature follwed by: familiarity, genre, and producibility. The importance values for most local priorities are generally intuitive; easily produced, familiar music we like to listen to are important factors we use when deciding to sing something. Rock was the most important genre (22.7% importance), followed by Pop (21.3%), Alternative (19.6%), Country (19.3%), and Rap (16.8%). The significant binomial correlation also indicates that user demographic information such as sex should be considered when recommending music to sing. Although the preference for same-sex singers (3%) does not account for a

[3] https://www.mturk.com/

[4] Workers were instructed to select whether an excerpt from Michael Jackson's Billie Jean was more familiar than an unreleased composition from one of the authors

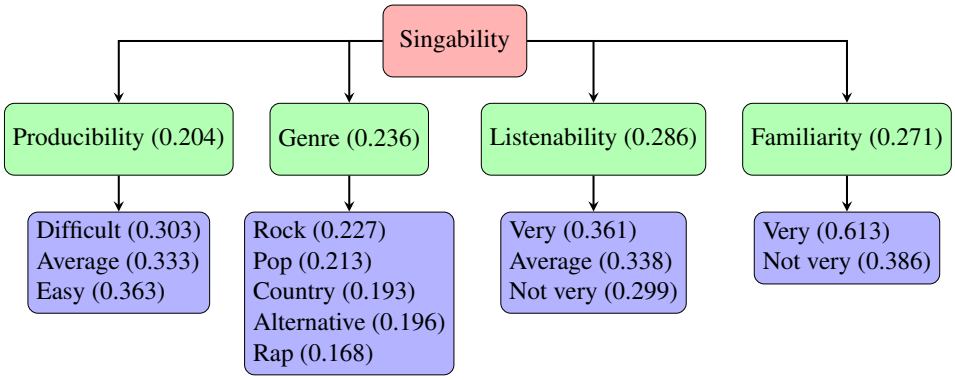

Figure 1. Analytic Hierarchy Process for Singability derived from Mechanical Turk experiment. The most important global priority was familiarity, followed by preference to listen, genre, and producibility.

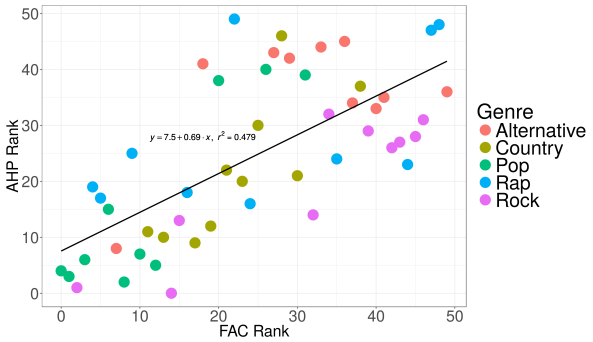

Figure 2. FAC-to-AHP Rank Scatterplot. X-axis represents ranks derived from the AHP analysis for a given track. Y-axis represents ranks deried from FAC analysis. Linear regression line for this data is plotted ($\hat{y} = 0.6922x + 7.8490$)

high degree of difference, recommendation systems based on human behaviours are relatively rare and can improve user satisfaction in generally unexplored ways.

A downside of this current investigation is that the variation of importance of global and local priorities was quite low, ranging between 2-3% across most factors. A more pronounced effect may be achievable using a more controlled, laboratory recruited participant pool. Producibility was also not a factor used to generate AHP ranking. The rationale for this is that there is no clear or simple way to evaluate vocalization difficulty of an excerpt relative to an MTurk worker's actual skill, whereas measures of familiarity and preference have a high degree of comorbidity with qualitative assessments [49].

An important component missing from this analysis is determining whether specific acoustic features influence ranking in meaningful way; is a song that is more singable one that generally has more pronounced vocals, or a faster tempo? Experiment two is a preliminary exploration into assessing whether some acoustic features are more important than others for determining singability based on the ranks generated through the AHP.

4. EXPERIMENT TWO: FEATURE IMPORTANCE EXPLORATION

After establishing that singability is a measurable cognitive process, the natural next step is analysis of acoustic features. Evaluating the importance of acoustic features related to singability may enable us to establish whether, or which, specific auditory signals contribute to this complex decision-making task. Experiment two provides preliminary, exploratory analysis into the importance of a specific set of acoustic features when evaluating singability.

4.1 Methods

Similar to [43], we extract perceptually-relevant features for singability under two categories: timbral and rhythmic. Signal processing is conducted using a combination of LibRosa [27], MIRToolbox [22], and vocal analysis work in [18]. 24 features (4 rhythmic and 20 timbral) in total were assessed. An averaged value for each feature was extracted for each song every 15-seconds. Timbral features include: Vocal-to-Accompaniment Ratio (VAR) [40], High Frequency Energy (HFE) [11], Mel-band Frequency Cepstral Coefficients (MFCC) 1-5 [13], spectral centroid mean and deviation [34], spectral roll off mean and deviation, and root mean squared (RMS) of energy mean and deviation [31]; rhythmic features include: tempo, zero-crossing mean and deviation [15], event density [2], and syllabic rate [18]. These features were selected for exploratory purposes due to their ubiquity in signal processing toolkits and MIR research.

4.2 Analysis

To determine whether specific features are more common in singable songs, we first conduct multiple linear regression comparing the AHP generated numeric values to the 24 extracted acoustic features. The multiple-comparisons F-Test was marginally significant ($F(23, 36) = 1.779, p = 0.05915, r^2 = 0.2328$), independent regressions yielded significant two features (Deviation of RMS and MFCC 5) and six marginally significant features (Deviations of spectral roll off, MFCCs 1 and 3, and means of RMS, spectral

Feature	t-value	p-value
Deviation of MFCC 5	-2.919	0.00604**
Deviation of RMS	-2.539	0.01558*
Deviation MFCC 1	1.994	0.05372·
Deviation spectral roll off	1.934	0.06099·
Mean of zero crossings	-1.864	0.0702·
Mean spectral centroid	1.738	0.09072·
Deviation MFCC 3	1.713	0.09525·
Mean of RMS	1.703	0.09713·

Table 1. Individual Linear Regression Significance Table. Multiple comparisons F-Test was marginally significant ($F(23, 36) = 1.779, p = 0.05915, r^2 = 0.2328$). $\cdot p < 0.1, * * p < 0.01, *p < 0.05$

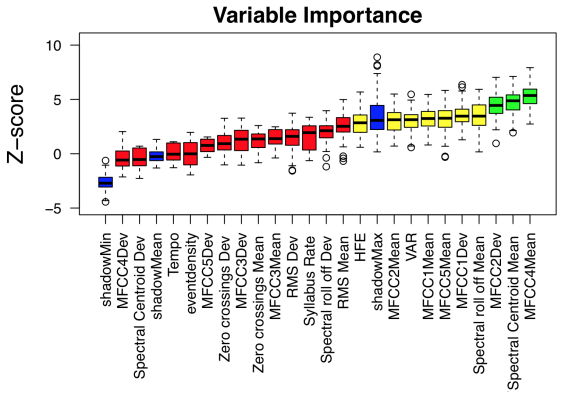

Figure 3. Random forest with regression model. Z-scores represent the relative importance of a feature in the determination of AHP-generated values for songs. Colours represent significance values: significant (green), marginally significant (yellow), not significant (red), and anchor values (blue).

centroid and zero-crossing). Table 1 provides a summary of analysis for all marginally significant features.

A statistical disadvantage of relying on standard linear regression analysis only is that multiple comparisons increasingly introduces type-I error with each added feature. We employ a random forest for regression and compare significant features across both models. An added benefit of using random forest is that it can assess the relative importance of each feature in the evaluation of singability. Three features significantly influenced AHP-generated singability scores (Mean of spectral centroid and MFCC 4, and deviation of MFCC 2), and six marginally influenced AHP-generated singability scores (Syllabic rate, VAR, HFR, mean of RMS, and deviation of MFCC). Figure 3 presents the relative importance of each feature (x-axis) as a Z-score (y-axis). Features that were at least marginally significant across both models included: mean of spectral centroid and RMS, and deviation of MFCC 1. Features that were at least marginally significant in the random forest model that were not significant using independent linear regressions included: mean of MFCC 4, spectral roll off, VAR, HFE, and syllabic rate.

4.3 Discussion

Both sets of analyses suggest that acoustic features may influence perceptions of singability. However, the models disagree on which features are maximally important in this decision. The three significant features that were shared across models (mean of RMS, spectral centroid, and deviation of MFCC 1) suggest that more singable songs are in general louder, brighter, and timbral fluctuations in high frequency energy may be particularly important when selecting music to sing to. Features where there was a disagreement in singability across models include zero-crossings, spectral roll off, VAR, HFE, and syllabic rate. This suggests that types of percussive sounds, pronounced vocals, and higher than average frequency in vocalizations and syllabic rate, may also contribute to evaluations of singability. The marginal significance of the multiple-comparisons F-test indicate that acoustic features may influence judgements of singability, however additional analysis needs to be conducted in order to demonstrate the validity of this assertion (see Section 5). Future work should investigate whether less common features, such as chorusness [1], are more relevant to singability.

Compared to Experiment One, the results from Experiment Two are less intrepretable. It may be that the our corpus size, or that extracting high-level acoustic features from 15-second excerpts is insufficient sampling for this kind of analysis.

5. CONCLUSIONS

The methods utilized in both experiments may be useful for others in the refinement of psychologically-based music features such as danceability, or enable the exploration of other previously unexamined features.

Experiment One establishes a method for measuring complex cognitive decision making processes like singability in an operationalized manner. A major limitation of this operationalization is that it did not consider social and contextual features influencing singing preference. As described in Section 3.2, a benefit of using AHP is that including or removing global priorities is simple; future work should consider the role that other factors (such as social context and song lyrics) may play in the evaluation of singability.

Experiment Two provides a preliminary exploration of the extent acoustic features influence singability scores generated in experiment one. Two statistical models, one simple and the other more complex, were used to determine what features may be contributing most to the evaluation of singability. Significant features in common across the two models suggests that further signal analysis will be important future work.

This exploratory work does not definitively establish singability as a core feature of the music. Rather we suggest that it provides compelling evidence to support a perceptual process of singability, and a refinable methodology to explore or support other properties involving cognition.

6. ACKNOWLEDGEMENTS

This project is funded by Smule Inc. We would also like to thank our reviewers for their insightful comments and feedback.

7. REFERENCES

[1] Hooked: a Game for Discovering what Makes Music Catchy. (Proceedings of the 12th International Society for Music Information Retrieval Conference):245–250, 2013.

[2] Samer A Abdallah and Mark D Plumbley. Probability as metadata: event detection in music using ica as a conditional density model. In *Proc. 4th Int. Symp. Independent Component Analysis and Signal Separation (ICA2003)*, pages 233–238. Citeseer, 2003.

[3] Jean-Julien Aucouturier and Emmanuel Bigand. Seven problems that keep mir from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*, 41(3):483–497, 2013.

[4] Michael D Barone, Jotthi Bansal, and Matthew H Woolhouse. Acoustic features influence musical choices across multiple genres. *Frontiers in psychology*, 8:931, 2017.

[5] Wilmer T Bartholomew. A physical definition of good voice-quality in the male voice. *the Journal of the Acoustical Society of America*, 5(3):224–224, 1934.

[6] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800, 2011.

[7] Daniel E Berlyne. Novelty, complexity, and hedonic value. *Attention, Perception, & Psychophysics*, 8(5):279–286, 1970.

[8] Denise D Bielby and C Lee Harrington. Managing culture matters: Genre, aesthetic elements, and the international market for exported television. *Poetics*, 32(1):73–98, 2004.

[9] Dmitry Bogdanov, Joan Serra, Nicolas Wack, and Perfecto Herrera. From low-level to high-level: Comparative study of music similarity measures. In *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*, pages 453–458. IEEE, 2009.

[10] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. pages 493–498. Citeseer, 2013.

[11] Lauren B Collister and David Huron. Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review*, 3(3):109–122, 2008.

[12] Mihaly Csikszentmihalyi. *Flow and the psychology of discovery and invention*. New York: Harper Collins, 1996.

[13] Jeremiah D Deng, Christian Simmermacher, and Stephen Cranefield. A study on feature analysis for musical instrument classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(2):429–438, 2008.

[14] Anders Friberg, Erwin Schoonderwaldt, Anton Hedblad, Marco Fabiani, and Anders Elowsson. Using perceptually defined music features in music information retrieval. *arXiv preprint arXiv:1403.7923*, 2014.

[15] Fabien Gouyon, François Pachet, Olivier Delerue, et al. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*, 2000.

[16] Chu Guan, Yanjie Fu, Xinjiang Lu, Enhong Chen, Xiaolin Li, and Hui Xiong. Efficient karaoke song recommendation via multiple kernel learning approximation. *Neurocomputing*, 2017.

[17] David J Hargreaves, Chris Comber, and Ann Colley. Effects of age, gender, and training on musical preferences of british secondary school students. *Journal of Research in Music Education*, 43(3):242–250, 1995.

[18] K.M. Ibrahim, D. Grunberg, K. Agres, C. Gupta, and Y. Wang. Intelligibility of sung lyrics: A pilot study. Suzhou, China, 2017.

[19] Marcia K Johnson, Jung K Kim, and Gail Risse. Do alcoholic korsakoff's syndrome patients acquire affective reactions? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(1):22, 1985.

[20] Stefan Koelsch and Walter A Siebel. Towards a neural basis of music perception. *Trends in cognitive sciences*, 9(12):578–584, 2005.

[21] Carol L Krumhansl. Plink:" thin slices" of music. *Music Perception: An Interdisciplinary Journal*, 27(5):337–354, 2010.

[22] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A matlab toolbox for music information retrieval. *Data analysis, machine learning and applications*, pages 261–268, 2008.

[23] Guy Madison. Experiencing groove induced by music: consistency and phenomenology. *Music Perception: An Interdisciplinary Journal*, 24(2):201–208, 2006.

[24] George Mandler, Yoshio Nakamura, and Billie J Van Zandt. Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):646, 1987.

[25] Kuang Mao, Ju Fan, Lidan Shou, Gang Chen, and Mohan Kankanhalli. Song recommendation for social singing community. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 127–136. ACM, 2014.

[26] Kuang Mao, Lidan Shou, Ju Fan, Gang Chen, and Mohan S Kankanhalli. Competence-based song recommendation: Matching songs to ones singing skill. *IEEE Transactions on Multimedia*, 17(3):396–408, 2015.

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[28] Max Meyer. Experimental studies in the psychology of music. *The American Journal of Psychology*, 14(3/4):192–214, 1903.

[29] Adrian C North and David J Hargreaves. Subjective complexity, familiarity, and liking for popular music. *Psychomusicology: A Journal of Research in Music Cognition*, 14(1-2):77, 1995.

[30] Nathan Novemsky, Ravi Dhar, Norbert Schwarz, and Itamar Simonson. Preference fluency in choice. *Journal of Marketing Research*, 44(3):347–356, 2007.

[31] Costas Panagiotakis and Georgios Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on multimedia*, 7(1):155–166, 2005.

[32] Isabelle Peretz, Danielle Gaudreau, and Anne-Marie Bonnel. Exposure effects on music preference and recognition. *Memory & Cognition*, 26(5):884–902, 1998.

[33] Thomas L Saaty. How to make a decision: the analytic hierarchy process. *European journal of operational research*, 48(1):9–26, 1990.

[34] Emery Schubert, Joe Wolfe, and Alex Tarnopolsky. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*, pages 112–116. sn, 2004.

[35] H-J Schultz-Coulon, R-D Battmer, and H Riechers. Der 3-khz-formant–ein mass für die tragfähigkeit der stimme? *Folia Phoniatrica et Logopaedica*, 31(4):302–313, 1979.

[36] Sebastian Streich and Perfecto Herrera. Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization. In *Proceedings of the 118th AES Convention*, 2005.

[37] Johan Sundberg. Level and center frequency of the singer's formant. *Journal of voice*, 15(2):176–186, 2001.

[38] Mari Tervaniemi, Minna Huotilainen, and Elvira Brattico. Melodic multi-feature paradigm reveals auditory profiles in music-sound encoding. *Frontiers in human neuroscience*, 8(July):496, 2014.

[39] Mari Tervaniemi, Lauri Janhunen, Stefanie Kruck, Vesa Putkinen, and Minna Huotilainen. Auditory profiles of classical, jazz, and rock musicians: Genre-specific sensitivity to musical sound features. *Frontiers in psychology*, 6:1900, 2016.

[40] Wei-Ho Tsai, Dwight Rodgers, and Hsin-Min Wang. Blind clustering of popular music recordings based on singer voice characteristics. *Computer Music Journal*, 28(3):68–78, 2004.

[41] Chia-Jung Tsay. Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences*, 110(36):14580–14585, 2013.

[42] Suppawong Tuarob and Conrad S Tucker. Quantifying product favorability and extracting notable product features using large scale social media data. *Journal of Computing and Information Science in Engineering*, 15(3):031003, 2015.

[43] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[44] Omkarprasad S Vaidya and Sushil Kumar. Analytic hierarchy process: An overview of applications. *European Journal of operational research*, 169(1):1–29, 2006.

[45] Christopher Watts, Jessica Murphy, and Kathryn Barnes-Burroughs. Pitch matching accuracy of trained singers, untrained subjects with talented singing voices, and untrained subjects with nontalented singing voices in conditions of varying feedback. *Journal of Voice*, 17(2):185–194, 2003.

[46] Felix Weninger, Martin Wöllmer, and Björn Schuller. Automatic assessment of singer traits in popular music: Gender, age, height and race. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 37–42, 2011.

[47] William R Wilson. Feeling more than we can know: Exposure effects without learning. *Journal of personality and social psychology*, 37(6):811, 1979.

[48] Maria AG Witek, Eric F Clarke, Mikkel Wallentin, Morten L Kringelbach, and Peter Vuust. Syncopation, body-movement and pleasure in groove music. *PloS one*, 9(4):e94446, 2014.

[49] Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1, 1968.