Overall summary:

The paper proposes to examine how vocal aesthetic preferences vary across specific speaking and singing styles. The main novelty of the current proposal is the inclusion of music or song aesthetics as the introduction describes a large literature related to vocal attractiveness in speech. The proposal will examine how much agreement there is among raters in their liking of primarily sung materials, with a specific examination of whether agreement about particular performers differs depending on the genre or style. Interesting metrics from the visual literature, MM1, will be used to assess agreement alongside more traditional metrics of inter-rater reliability (ICC). The proposed study is novel and has the potential to contribute meaningfully to the field of empirical aesthetics, but there are several items that should be addressed before moving to IPA. First, the introduction is a bit hard to follow in that the reader could arrive at multiple different main study questions from the literature review. That is, since the novelty of the current proposal appears to be from a lack of studies related to song, I assumed that the main comparison would be speech vs. song, but that is not the case. The introduction should be streamlined and written to clearly set up the main goals and gaps in the literature that the proposed work is designed to address (e.g., why the test-retest reliability? Why these styles? Why these acoustic features?). The second issue is that the proposed study has a lot of additional metrics (all the perceptual factors) that the participant pool will need to complete that appear to be ancillary to the main question about liking. Instead this study design feels more like a validation of the stimulus set and that the liking question was tacked on as an afterthought. That is a completely normal approach, but for a pre-registration, I feel the study design should be more clean and deliberate so as to answer a specific question without the unintended influence of multiple additional factors built into the study design for exploratory analyses. Finally, I believe the power analysis could be more carefully down with reference to previously published work.

General comments:

Throughout the manuscript there is a differentiation between perceptual features and acoustic features, but this is not defined. For instance, breathiness, tempo, and timbre are described as perceptual features, but they can be just as easily characterized as acoustic features extracted from onset rates, spectral cues, harmonicity from Praat or MIR toolbox-like metrics. Can the authors please define their meaning at the outset so the reader can assess what factors (human vs. algorithm based metrics, musical vs. speech-based metrics, etc) they are differentiating when they describe the contribution of these features.

Abstract:

What is the relevance of vocal attractiveness to the bigger picture of sociobiological signals?

The “type” of vocalization could be interpreted in many ways, but I think you mainly mean speaking vs. singing, although I do see your different types of vocalizations, too. Perhaps your main proposal sentence could read something like: “For instance: why do we like some voices more than others? Does our liking of voices differ depending on whether a person is speaking or singing? Do some voices sound better in some contexts like singing a pop song but not singing an operatic aria?

What is the difference between perceptual and acoustic features? Do you mean features based on participant’s ratings of acoustic characteristics? Do you mean preferences here not perceptual features?

Introduction:

I appreciate the different functional roles, esp. described for ID speech, but it would be good to characterize the role of AD speech as well – I see that you write “Beyond supporting interpersonal communication and conveying semantic information” which I assume is about AD speech, but it would be nice to flesh that out in a sentence or two and then continue with your shift in narrative toward it being a sociobiological signal. This section could do a little more for the reader by summarizing the similarities and differences between music and language.

There needs to be a better segue to the proposed work that motivates why the question is important to study and relevant to human communication. I believe it is important, but more of a motivation is needed to help understand your hypotheses and the context for your proposed study. For instance, why does it matter if vocal attractiveness differs for the same speaker speaking vs. singing? Or is the question more about what features predict attractiveness? Or whether vocal attractiveness is even important for some modalities vs. others?

Vocal attractiveness and sexually dimorphic vocal features: In regards to harmonic-to-noise ratio (called harmonicity in Praat), the cited work is relevant and makes sense, but it a little older. Especially given that there is a strong trend, especially in female voices, to use vocal fry which would have, I assume, a lot more noise in the signal than harmonic information. This was originally associated with California valley girl speaking style (i.e., cool!) but has now pervaded all of north America and likely beyond. It’s work having a sentence about this in this section to update the literature review.

Inter-rater agreement is low (how low?) can you contextualize this for the reader?

RE: Visual data using MM1 – I am having a hard time understanding this section: “They argue that the behavioral relevance of naturally occurring types of stimuli such as landscapes and human faces results in information processing, and hence aesthetic experience, that is highly conserved across individuals.”

What is meant by behavioural relevance?

So is the use of MM1 about being integrative or about comparing with vision or about assessing behavioural relevance? Clarity here will help guide what hypotheses should be or what hypotheses are expected by the reader.

Questions & Hypotheses (Table 1)

I think the main questions are interesting, but it feels like some of the simple effects are missing. For instance, since there are 2 modalities (speech vs. song) and within each modality 2 to 3 sub-types (song: opera, pop, lullaby; speech: AD vs ID), I imagine you would want to examine whether the same speakers are preferred across sung and spoken stimuli and then drill down into whether that interacted with “styles.” However, I do see that the spoken and sung styles are really not comparable between modalities. For instance, song as ID and AD registers as well and that would be a better comparison of differences between speech and song. Or speech has casual conversational styles, conversing with strangers, delivering a speech, or acting on stage, which could be more comparable to the categories chosen for song. So if this is the rationale for not looking at speech and song separately because the subgroups are 1) not balanced and 2) not directly comparable to one another, then that is fine, but I think it’s worth noting this rationale somewhere so the reader understands the rationale for the study design.

Hypotheses 1.1.1 only includes song, but the explanation for it includes speech but does not include a directional hypothesis. Speech should be included in the analyses and some hypotheses should be made either about speech sounds alone or in comparison of speech to song, otherwise I am not sure why the speech stimuli are included.

1.1.2 – in the table these are referred to as rankings, but they are ratings. Describing them as rankings made me think that participants might be doing a ranking task instead of the researchers using the average ratings to effectively rank the performers based on participants ratings.

What is the duration of time between test and re-test? I don’t see this in the experimental procedure. What have other studies done? 2 weeks? 2 months? Back-to-back days does not seem long enough to truly examine consistency.

1.1.3 Should include what perceptual and acoustic features you’ll use to predict liking ratings and include what other factors you’ll include, etc.

1.2.1 How did you calculate f? In g-power they have you determine f directly from eta-squared or from variances. Please justify, using the previous papers cited, why you expect an effect of this size and how that relates to the previously found effect sizes (eta-squared). For 1B, what estimates does g-power (z tests section) give you for finding a high correlation between your to dependent correlation coefficients?

1.2.2 I am curious why the Friedman test is warranted? Is it predicted that the residuals will not be normally distributed in the ANOVA? It seems like the mean rating averaged by performer should be fine in an ANOVA unless there are different number of trials per performer? The ANOVA would be a X (Number of performers) by style (n=5) repeated measures analysis as there will be 5 columns per performer, correct?

1.2.3 I know that these are reported in the spirit of explorational analyses, but if they are reported here then they are pre-registered and so they ought to have more detail, I think. If they are truly explorational then perhaps they can be left off of the pre-registration? Given that the authors have extensive experience doing this sort of analysis based off of previous cited work, it seems as though the authors could do some clear pre-registration of a simplified version of their LMEs so that at least some part of this analysis could be evaluated at Stage 1. As such, a power analysis of some sort is warranted here to ensure that the authors have enough power to do even their exploratory analyses.

Also, since the paper is about whether some performers consistently are liked more than others, then perhaps performer should be included as a variable in the LME? Perhaps, if you have enough participants, including speech and song in one model and coding for it would be informative. This would likely require a lot of interaction terms to determine if a feature was useful for song but not speech, for example, so it can get unwieldy (and not converge) quickly, but it would be a stronger way to illustrate that a features usefulness changed depending on the style or modality than two separate models.

Participants. “students and retired subjects” do you think you should restrict to one age group or include some sort of coarse age grouping in your study? It seems very likely that there will be generational differences in aesthetic preferences for voices.

Stimuli. How will you account for F0 differences that were requested for singers of pop and lullaby in your 3rd hypothesized LME? Same for loudness. You will need to be careful about this if concluding that F0 or loudness predicted liking, for example.

“This leads to 110 performances (by 22 singers, each performing three styles of singing and two styles of speaking).” I see why you decided to have the same melody for all singers being judged (apples to apples), but, in order to make strong claims about style and features of that style, it seems like you’d want to include at least one other melody. At the very least, this might make the task a little more enjoyable (and reliable) for participants, given that 22 singers is quite a big number to move through. I know you want variability in performers so keeping the number at 22 is understandable given your hypotheses, but it is worth considering making the study a bit longer for generalizations’ sake.

Acoustic and Perceptual analyses. I see here how you have grouped acoustic and perceptual. Basically, perceptual are provided by participants and acoustics are not. But it’s hard to say that the perceptual features are not the same as or highly correlated with acoustic features. For instance, energy – as an acoustic measure, should be highly correlated with perceptual loudness or F0 calculated by Praat should be highly correlated with participants high-low ratings. How will you enter variables into your LME models? What if there is significant correlation among predictors – will you drop the least correlated? Will you compare models with all correlated acoustic features and all correlated perceptual features? Will you compare models that have a mix of perceptual and acoustic features but only those that are not highly correlated with one another? You have a lot of great variables here, but they are largely overlapping so it makes it hard to understand if the story you’re trying to weave in this case is about the failure of music information retrieval techniques to pick up on the features that real human listeners use, or something else. And the selection of these variables should be justified (for instance why perceptual and acoustic measure of pitch?)

Liking ratings are part of this very large set of perceptual ratings. Is there a reason that each person needs to rate these perceptual features? If your study is truly interested in liking ratings, then it feels like these should be two separate studies or that liking should be asked first so that the large list of features participants need to rate does not bias their liking rating. Further, I wonder how having these perceptual features drives liking ratings on subsequent trials. For instance, a rater might intuitively think that diction is crucial to a good performance and then decides to apply that to all sung stimuli, but would not have considered diction had they not been explicitly asked about it during the study. I worry that all of these perceptual ratings will alter participants liking ratings. It also seems to me that these perceptual ratings do not need to be obtained on a per person basis (that is, do you plan to use perceptual ratings from a given person to predict liking or an average from all participants?).

Procedure, cont’d

Blocks grouped by style – while I agree this is a valid approach, I wonder if the blocking itself might alter the predicted results. Specifically, blocking by style may encourage raters to adopt a set of features for that specific style or genre, whereas varying the style and completely randomizing blocks may encourage people to attend to performer-specific features that are aesthetically pleasing. If you decide to pare down the task to only liking, it could be interesting to run the study with both randomized and blocked presentation of trials perhaps for different melodies (CH vs. NN) or between subjects to examine this methodologically important difference.

Data analyses. This MM1 metric seems interesting and reminds me a lot of jack-knifing techniques for understanding the contribution of that particular rater (or item) to the mean. However, each participant will have a single R value, correct? If so, then I am not sure I understand the next sentence, pasted for clarity below. Wouldn’t the mean of the z-score be 0 and then transforming them back to and r-value would leave them unchanged? I could be missing a step here! I am glad you’re including the ICC as this seems to be a pretty standard metric for assessing agreement across raters.

“The across-observer average MM1 score is computed by 1) transforming individual r-values to z values, 2) computing a mean, and 3) transforming that score back to an r-value for easier interpretability.”

At the end of a proposal I would expect some sort of impact statement about what the predicted results would contribute to the field or what follow-up studies it would spur. I am not sure if this is a typical section for PCI-RR, but it seems important to close out the proposal with this information.