慶應義塾
Keio University

5322 Endo, Fujisawa
Kanagawa 252-0882, Japan
(+81) 80-6551-4063

November 30th, 2022

RE: *Similarities and differences in a global sample of song and speech recordings* (https://doi.org/10.31234/osf.io/jr9x7)

Dear Dr. Chambers,

We appreciate your invitation to revise and resubmit our manuscript based on the constructive comments of the two Reviewers. We are grateful for the chance to use their constructive feedback to update our planned analyses to allow us to validate our predictions by also using other datasets and to address possible issues of representativeness of our recording sample. We have also added the suggested inter-rater reliability analysis and modified the robustness analysis section to further test whether language lineages affect our predicted results. We have appended a version with tracked changes to this response letter for your convenience.

In addition to this major redesign, we have also clarified several limitations of this study pointed out in this review letter, so that the readers can evaluate our study design and decisions on that transparently.

We feel that the review process prompted substantial improvements to our manuscript. We hope you will find the revised manuscript acceptable for in-principle acceptance.

Sincerely,

Yuto Ozaki and Patrick E. Savage
(on behalf of the authors)

Editor summary (Chris Chambers)

Major Revision

I have now obtained two very helpful expert evaluations of your submission. As you will see, the reviews are broadly positive while also raising a number of issues to consider in revision, including clarification of procedural and methodological details, and consideration of additional (or alternative) analyses.

There is one issue raised in the reviews that I anticipated based on my own reading, which is the validity of sampling exclusively from the team of co-authors. One of the reviewers writes: "...it's of course not ideal to only sample data from co-authors, since co-authors may differ from the randomly sampled participants in many different ways even if they are not aware of the hypotheses being tested. Why not just ask the co-authors to acquire some speech and songs in their native language and do a control analysis on those audios? This approach seems particularly plausible since the current manuscript mainly analyzed songs and spoken descriptions, which are not matched in the content." I think this suggestion is worthy of careful consideration.

We appreciate this constructive suggestion. As described in the manuscript, most existing sources of speech and song recordings are not suitable for our analysis, and collecting new high-quality samples using our approach requires substantial time commitment and co-author expertise that cannot be easily added to the current design. However, after running some additional pilot analyses on the most promising existing source of cross-cultural song/speech data, we have added the following additional analyses that we believe is the most realistic way we can address this concern (we note these analyses should also address Reviewer #1's concern about representativeness of song selection):

> ***"2.7.8: Exploring recording representativeness and automated scalability:* Because our opportunistic sample of coauthors and their subjectively selected "traditional" songs are not necessarily representative of other speakers of their languages, we will replicate our analyses with Hilton, Moser et al.'s (2022) existing dataset, focusing on the subset of languages that can be directly compared. This subset of languages will consist of 5 languages (English, Spanish, Mandarin, Kannada, Polish) represented by matched adult-directed song and speech recordings by ~240 participants (cf. Hilton et al. Table 1).**
>
> ***Because our main analysis method requires time-intensive manual or semi-manual annotation involving the recorded individual that will not be feasible to apply to Hilton et al.'s dataset, we will instead rely for our reanalysis of Hilton et al.'s data on purely automated features. We will then re-analyze our own data using these same purely automated features. This will allow us to explore both the scalability of our own time-intensive method using automated methods, and directly compare the results from our own dataset and Hilton et al.'s using identical methods.***
>
> ***Fig. S7 demonstrate this comparison using pilot data for one feature (pitch height) based on a subset of Hilton et al.'s data that we previously manually annotated, allowing us to simultaneously compare differences in our sample vs. Hilton et al.'s***

sample and automated vs. semi-autoated methods. Even though this analysis focuses on a feature expected to be one of the least susceptible to recording noise (pitch height), our pilot analyses found that these were mildly sensitive to background noise, such that purely automated analyses resulted in systematic underestimates of the true effect size as measured by higher-quality semi-automated methods (Fig. S7). While our recording protocol (Appendix 2) ensures minimal background noise, Hilton et al.'s field recordings were made to study infant-directed vocalizations and often contain background noises of crying babies as well as other sounds (e.g., automobile/animal sounds; cf. Fig. S8), which may mask potential differences and make them not necessarily directly comparable with our results. This suggests the value of comparing our results with Hilton et al.'s using both fully-automated and semi-automated extracted features to isolate differences that may be due to sampling and differences that may be due to the use of automated vs. semi-automated methods.

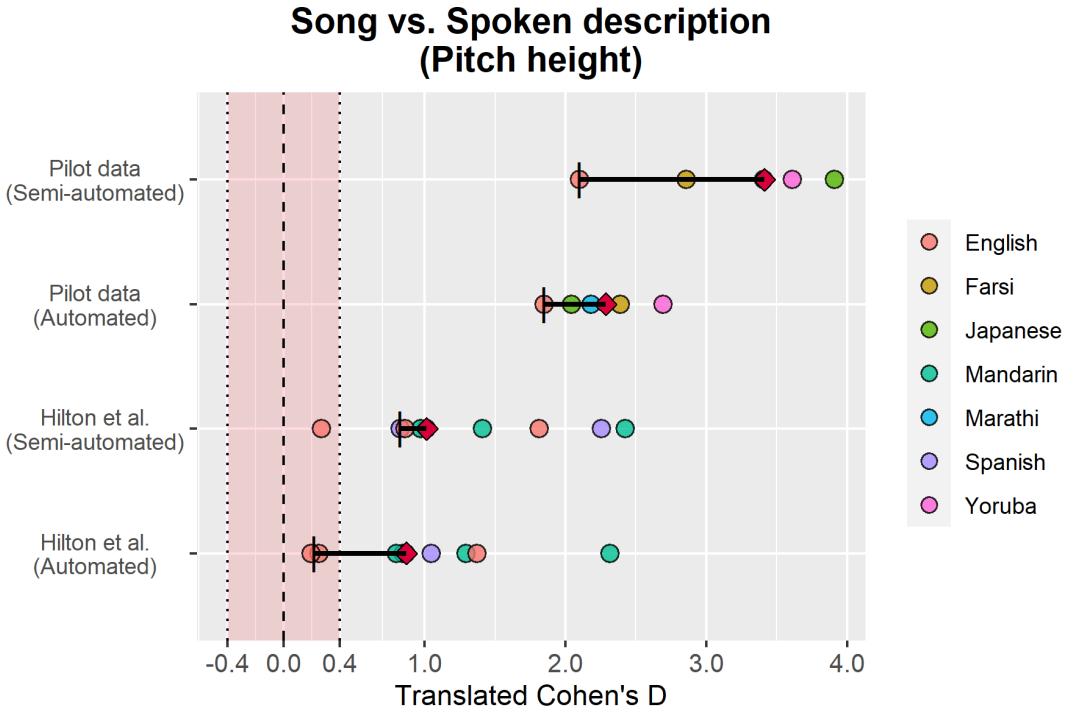

Figure S7. Pilot analysis of a subset of Hilton et al.'s (2022) data (pairs of adult-directed singing/speaking recordings from n=9 participants speaking English, Spanish, or Mandarin) focusing on pitch height. Ozaki et al., (2022) previously analyzed this subset for preliminary analyses using the same method described in S2.1 to avoid contamination by various noises included in audio (vocalization by babies, car noises, etc.), which allows us to explore issues such as whether such extraneous noises are likely to be a concern in our planned fully automated analysis of Hilton et al.'s full dataset (cf. Fig. S8). Although all four conditions demonstrate the predicted trend of song being consistently higher than speech, the effect size varies depending on the dataset and analysis method used (see Section 2.7.8 for discussion).

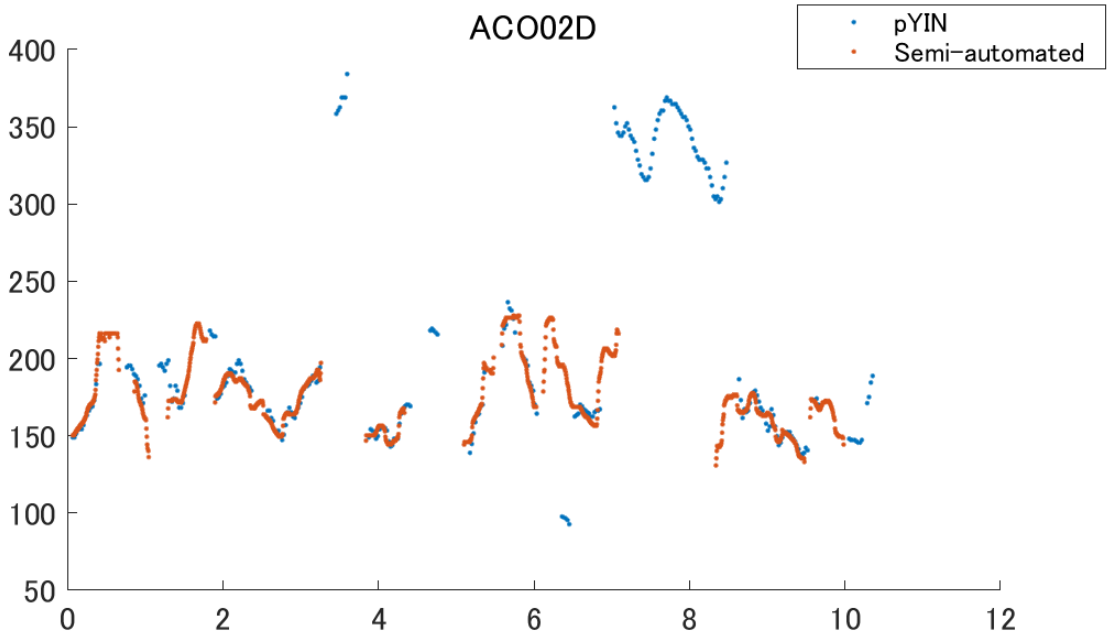

Figure S8. An example of fully-automated vs. semi-automated f0 extraction underlying the analyses in Fig. S7 for one of the field recordings from Hilton et al.'s dataset. AC002D = adult-directed speech [D] from individual #02 from the Spanish-speaking Afro-Colombian [ACO] sample). While the extracted f0 values are generally similar, the fully automated pYIN method sometimes has large leaps, particularly when there are external noises and the main recorded individual stops vocalizing to breathe (here the high-pitched blue contours at around 3.5 and 8 seconds correspond to the vocalizations of a nearby child while the recorded adult male takes a breath)."

Concerning the appropriateness of including audio contributors who are not aware of the hypotheses of the study as co-authors, given the unusual nature of this work, I am happy that they continue as outlined provided they are fully informed at Stage 2 after results are known, and that they have the opportunity to contribute to the interpretation of the results in light of the predictions.

Thanks for your understanding. In any case, all coauthors have now been informed of the hypotheses and provided the opportunity to read and edit the manuscript. (For the record, only 12 of our 80 coauthors [indicated with "*" in the previous author list] had not at the time of original submission.) We have removed those "*" from the author list and modified the text in the footnote to Appendix 2 as follows:

"Because the number of collaborators who could not meet the revised timeline was small enough not to affect our planned power analyses or robustness analyses, we shared the manuscript with all authors and will incorporate those who had not yet made their recordings in the robustness analyses, along with the other authors who made their recordings after knowing the hypotheses."

Overall, based on the reviews and my own reading, I believe your submission is a strong contender for eventual Stage 1 acceptance, and I would like to invite you to submit a revision and point-by-point response to the reviewers. I will likely send your revised manuscript back to at least one of the reviewers for another look.

We are pleased by this encouraging evaluation, and hope our revised manuscript and responses below have addressed all remaining concerns.

Reviewer #1 (Bob Slevc):

This RR describes a plan, along with some pilot data, to collect samples of song and speech (and also of recited lyrics and instrumental music, for exploratory purposes) from a wide variety of language communities, and then analyze these on various dimensions to assess similarities and differences between speech and song.

I really like this overall idea and I think this is a great use of a registered report format. I was initially slightly disappointed that only a few potential dimensions of similarity/differences are included as confirmatory hypotheses, but I think the most valuable part of this project may really be the creation of a rich dataset for use in exploratory work. That is, while I appreciate the (obviously significant) work involved in developing and proposing the confirmatory hypothesis tests, I also want to highlight the usefulness of this undertaking for hypothesis generation.

We agree with this thoughtful evaluation, which well reflects our own feelings after the process of limiting potential features to just 6 for hypothesis testing. We have added the following text to the end of the abstract to highlight the value of the exploratory part of the study:

> ***"…and provide rich cross-cultural data to generate new hypotheses and inform future analyses of other factors (e.g., functional context, sex, age, musical/linguistic experience) that may shape global musical and linguistic diversity."***

Here, of course, I will offer a few suggestions and comments that might be worth considering before undertaking this large project. In no particular order:

1. I'd like to know more about how songs were chosen. I'm pretty happy with the conditions -- singing/recitation/description/instrumental seem likely to give good samples across the speech-song spectrum -- but I do wonder how much the data from a given language will depend on the specific song that was chosen. I think the prompt in the recording protocol is good, but... do we know, for example, if songs that are "one of the oldest/ most "traditional" (loosely defined) / most familiar to your cultural background" are also representative? (Like, I might worry that the oldest ones are actually a bit unusual such that they stand out in some way.) Relatedly, it might be good to constrain (or at least assess) the type of song. I could imagine systematic differences between, say, lullabies and celebratory music and maybe some of those differences might also show up in their similarity to speech. I don't have reason to think this is likely to be a systematic problem, but I feel like some discussion and/or some plan to assess the "representativeness" of the songs that were chosen might be important.

We thank the reviewer for this important point. We believe that our newly added analysis described above ("2.7.8: Exploring recording representativeness…") will address this issue of song representativeness as well as of participant representativeness. We have also made the following additions to section 2.7.6 ("Exploratory analyses" and the end of the abstract:

> ***"2.7.6: Other factors: In future studies, we also aim to investigate additional factors that may shape global diversity in music/language beyond those we can currently analyze. Such factors include things such as:***
> ***-functional context (e.g., different musical genres, different speaking contexts)***
> ***-musical/linguistic experience (e.g., musical training, mono/multilingualism)***
> ***-neurobiological differences (e.g., comparing participants with/without aphasia or amusia)"***
>
> **Abstract:** ***"…provide rich cross-cultural data to generate new hypotheses and inform future analyses of other factors (e.g., functional context, sex, age, musical/linguistic experience) that may shape global musical and linguistic diversity."***

2. I was also wondering how these 23 language families were chosen. The answer seems to be mostly opportunistically, which I think is fine given that there's a plan to assess bias in the languages sampled.

Correct, as described in Section 2.2.1 ("Language sample / Inclusion criteria": "Coauthors were chosen by opportunistic sampling…"). To help give a better sense of this process, we have added the following new Appendix (#3) reproducing the recruitment email, which also gives you a sense of how we aimed to correct imbalances in global representation:

> ***"Appendix 3: Open call for collaboration to the International Council for Traditional Music (ICTM) email list. Adapted versions of this email were also used later in tandem with in-person recruitment at the conferences described in the main text). Note that in later meetings we decided to relax the restriction of one collaborator per language, in part due to difficulties of defining the boundaries separating languages and the desire to maximize inclusion.***
>
> ***From: Patrick Savage <psavage@sfc.keio.ac.jp>***
> ***Subject: Call for collaboration on global speech-song comparison***
> ***Date: July 15, 2022 9:49:57 JST***
> ***To: "ictm-l@ictmusic.org" <ictm-l@ictmusic.org>***
>
> ***Dear ICTM-L members,***
>
> ***I am emailing to inquire if any of you are interested in collaborating on a project comparing speech and song in diverse languages around the world to determine what, if any, cross-culturally consistent relationships exist.***

I mentioned this project briefly back in January in response to the discussion about Don Niles' post to this list entitled "What is song?". Since then, we have recruited several dozen collaborators speaking diverse languages (see attached rough map), but would like to open up the call to recruit more. As you can see from the map, our current recruitment is quite unbalanced, particularly lacking speakers of indigenous languages of the Americas, Oceania, and Southeast Asia. We hope you can help us correct that!

Collaborators will be expected to make short (~30 second) audio recordings of themselves in four ways:

1) singing a traditional song in their native language

2) reciting the lyrics of this song in spoken form

3) describing the meaning of the song in their native language

4) performing an instrumental version of the song's melody on an instrument of their choice (negotiable)

They will also provide written transcriptions of these recordings, segmented into acoustic units (e.g., syllables, notes) and English translations. Later, they will check/correct versions of these recordings created by others with click sounds added to the start of each acoustic unit. Finally, they will help us interpret the results of acoustic comparisons of these recordings/annotations. Our pilot studies suggest that this should all take 2-4 hours for one set of 4 recordings.

Collaborators will be coauthors on the resulting publication, and will also be paid a small honorarium (pending the results of funding applications). In principle, all audio recordings will be published using a CC BY-NC non-commercial open access license, but exceptions can be discussed on a case-by-case basis (e.g., if this conflicts with taboos or policies regarding indigenous data sovereignty).

We seek collaborators aged 18 and over who are native speakers of diverse languages, but we are open to collaborators who are non-native speakers in cases of endangered/threatened languages where there are few native speaker researchers available. During this first stage, we only plan to recruit one collaborator per language, on a first-come first-served basis in principle (in future stages we will recruit multiple speakers per language).

More details and caveats (e.g., how to interpret "traditional" or "song") can be found in a draft protocol here:
https://docs.google.com/document/d/1qICFXwew7OEj06dkSoR59TlF7HCmVGcudkenMwHRemM/edit

We actually are not quite ready to begin the formal recording/analysis process yet as we are still working out some methodological and conceptual issues (for which we would also welcome your contributions). The reason I am putting out this call now is that I will be presenting at ICTM in Lisbon next week and I know many of you will also be there, so I wanted to use this chance to reach out in case any of you want to meet and discuss in person in Lisbon.

I'll be mentioning more details about this project briefly during a joint ICTM presentation on "Building Sustainable Global Collaborative Networks" at 9am on July 26th (Session VIA01), and would be delighted to meet anyone interested in collaboration following this session or at any other time during the week of the conference.

Please email me (mentioning your native language[s]) if you're interested in collaborating or in meeting in Lisbon to discuss possibilities!

Cheers,
Pat

Dr. Patrick Savage (he/him)
Associate Professor
Faculty of Environment and Information Studies
Keio University SFC (Shonan Fujisawa Campus)
http://compmusic.info"

I wonder if another way to deal with this issue (rather than just sampling a subset of samples that are more balanced) might be a random-effects meta-analysis model, where languages can be nested into language families. I am no expert on meta-analysis, but I believe this is one way that meta-analyses deal with non-independent effects without having to exclude large swaths of data.

We have incorporated this helpful suggestion by employing multi-level meta-analysis models (meta-analytic linear mixed effects models) so that random effects can be separately modeled for each language family by nesting data of each language to the language family as recommended. We updated section "2.6.2 Potential dependency caused by language family lineage" as follows:

"2.6.2 Potential dependency caused by language family lineage
Another potential bias in our design is the unbalanced sample of languages due to our opportunistic sampling design. Related languages are more likely to share linguistic features due to common descent, and sometimes these features can co-evolve following

lineage-specific processes so that the dependencies between the features are observable only in some families but absent in others (Dunn et al., 2011) . Thus, it is possible that our sample of speakers/singers may not represent independent data points. While our study includes a much more diverse global sample of languages/songs than most previous studies, like them our sample is still biased towards Indo-European and other larger languages families, which might bias our analyses. To determine whether the choice of language varieties affects our confirmatory analyses, we will re-run the same confirmatory analyses using multi-level meta-analysis models (linear mixed-effects models; Sera et al., 2019) with each recording set nested in the language family. We will perform model comparison using the Akaike Information Criterion (AIC; Bozdogan, 1987) for the original random-effects model and the multi-level model. The model having the lower AIC explains the data better in terms of the maximum likelihood estimation and the number of parameters (Watanabe, 2018), although critical assessment of information criteria and model selection methods in light of domain knowledge is also important (Dell et al., 2000). If the choice of model technique qualitatively changes the results of our confirmatory hypothesis testing, we will conclude that our results depend on the assumption of the language dependency."

3. It took me a little while to understand how "sameness" would be tested here. On my first read, I was pretty confused (I think in part this is because the manuscript says, when describing the SESOI, "The basic idea is to analyze whether the feature differs between song and speech," which... sounds like a difference test not a 'sameness' test.) The logic is explained much more clearly a bit later, at the bottom of p. 10, but I'd suggest making the logic more clear upfront.

We have removed this confusing sentence and edited the surrounding text to make our logic clearer early on, as follows:

"We test two types of hypotheses, corresponding to the hypothesis of difference and the hypothesis of similarity, respectively. Formally, one type of null hypothesis is whether the effect size of the difference between song and speech for a given feature is null. This hypothesis will be applied to the prediction of the statistical difference. Another type of null hypothesis is whether the effect size of the feature exceeds the smallest effect size of interest (SESOI) (Lakens, 2017). This hypothesis will be applied to the prediction of statistical similarity. In this study, we particularly rely on the SESOI of 0.4 suggested by the review of psychological research (Brysbaert, 2019). There are various ways to quantify the statistical difference or similarity (e.g. Kullbak-Leibler divergence, Jensen-Shannon divergence, Earth mover's distance, energy distance, Ln norm, Kolmogorov-Smirnov statistic). Here we focus on effect sizes to facilitate interpretation of the magnitudes of differences."

More importantly, I'd like to see a bit more justification about why the SESOI is what it is here. I recognize that d=.4 is a "normal" effect size in behavioral research, but this isn't really a standard behavioral study and so... is that effect size a normal/small one for acoustic differences? This isn't just a point of curiosity because the extent to which lack of differences for Hypotheses 4-6 can be taken to support cross-cultural regularities really depends on what counts as similar within those domains. I don't know the best way to choose or justify a SESOI here, but perhaps one simple approach might be to give the reader an intuitive sense by offering examples of the largest

differences (in timbral brightness, pitch interval size, and pitch declination) that would be considered the same under this SESOI (and, conversely, examples of the smallest differences that would be considered *not* the same). Maybe this SESOI could also be justified based on the variability within languages, on the idea that the level of variability that exists within songs/sentences in a language isn't likely to be meaningfully different between languages or something like that. (Obviously that would be hugely impractical to do for all these languages, but maybe just using one or two as a way to estimate between-song/speech variability?) I suppose yet another alternative might be to rely on Bayes Factors, but I'm not sure that actually solves the issue because it would still require justifying the choice of priors which might be even less straightforward than justifying a SESOI.

We thank the reviewers for this important question. We clarify this limitation and add artificially manipulated audio files (summarized in Table S1) to give examples of what such SESOI's would sound like in Sections 1.2 and S6 as follows:

> ***"Our choice of an SESOI of d = 0.4 based on Brysbaert's (2019) recommendation after reviewing psychological studies is admittedly somewhat arbitrary. Future studies might be able to choose a different SESOI on a more principled basis based on the data and analyses we provide here, and the value of our database for such hypothesis generation and exploration is an important benefit beyond the specific confirmatory analyses proposed. However, we currently are faced with a chicken-and-egg problem in that it is difficult to justify an a priori SESOI for analysis until we have undertaken the analysis. The same argument may hold for Bayesian approaches (e.g. highest density regions, region of practical equivalence, model selection based on Bayes factors) independent of the choice of prior distributions. We thus chose to rely on Brysbaert's recommended SESOI of d = 0.4 (and its equivalent relative effect of p_{re} = 0.61) in the absence of better alternatives.***
>
> ***Visual and aural inspection of the distribution of pilot data (Figs. 5 and S5; audio recordings can be heard at https://osf.io/mzxc8/) also suggest that it is a reasonable (albeit arbitrary) threshold given the variance observed across a range of different features and languages. To enable the reader/listener to assess what an SESOI might sound like, we have created versions of the pilot data artificially raising/lowering the temporal rate and pitch height of sung/spoken examples so one can hear what our proposed SESOI would sound like for a range of languages and features (Table S?; audio files also at https://osf.io/mzxc8/.***
>
> …
>
> ***S6 Manipulation of features to demonstrate our designated SESOI (Cohen's D = 0.4).***
> ***Following Brysbaert's (2019) recommendation, we use the relative effect corresponding to 0.4 of Cohen's D as the SESOI for our hypothesis testing. Although the choice of 0.4 of Cohen's D is somewhat arbitrary, we empirically measured how much such differences correspond to the physical attribute of audio using our pilot data focusing on pitch height and temporal rate. For each pair of singing and spoken description recording, we first measured the relative effect (3rd column: Relative effect (pre)). Then, we manipulated the corresponding feature of the song to result in a***

relative effect equal to 0.61 (corresponding to 0.4 of Cohen's D) and 0.5 (corresponding to no difference, 0.0 of Cohen's D). Specifically, we shifted down the entire f0 for pitch height and slowed down the playback speed for temporal rate. The 4th and 5th columns show actual scale factors identified at each recording and feature. For example, the first row indicates the f0 of the sung version needed to be shifted 730 cents downward to manipulate the difference in this feature between singing and spoken description to be as small as our proposed SESOI of Cohen's D = .4. Similarly, the sixth row indicates the IOIs of singing needed to be multiplied by 0.472 (i.e., each sung note sped up to be 47.2% as short as the original duration) to make no difference against the spoken description recording, meaning the playback speed of singing should be over 2x faster than the the original recording. Although there are only 5 recording pairs and this measurement does not directly provide the justification for using 0.4 of Cohen's D, we can see how the current SESOI threshold corresponds to the physical attribute of audio by comparing the 4th and 5th columns (106 cents for pitch height and factor of 0.091 for temporal rate in average), which to we authors seems reasonabl borderlines for listeners to notice the change in audio content. The corresponding audio examples are available in our OSF repository (https://osf.io/mzxc8/files/osfstorage/638491c81daa6b1394759086).

Table S1. Overview of our pilot recordings with key features (pitch height [f0] and temporal rate [1/IOI]) manipulated to demonstrate what real examples of song and speech might sound like if they the differences were non-existent ("equivalence") or negligible (as small as our chosen SESOI [Smallest Effect Size Of Interest]).

Vocalizer	Feature	Relative effect (p_{re})	Manipulation to demonstrate SESOI (p_{re} = 0.611)	Manipulation to demonstrate equivalence (p_{re} = 0.5)
D. Sadaphal (Marathi)	f_0	0.992	-730 cents (i.e., pitch is transposed down such that sung pitch is more than half an octave lower than the original)	-860 cents
Nweke (Yoruba)	f_0	0.995	-930 cents	-1030 cents
McBride (English)	f_0	0.931	-650 cents	-770 cents
Hadavi (Farsi)	f_0	0.978	-430 cents	-480 cents
Ozaki (Japanese)	f_0	0.997	-1300 cents	-1430 cents
D. Sadaphal (Marathi)	IOI	0.931	x 0.544 (i.e., playback speed is increased by almost 2x such that the duration of each sung note is only 54.4% as fast as the original)	x 0.472

Nweke (Yoruba)	**IOI**	**0.831**	**x 0.622**	**x 0.499**
McBride (English)	**IOI**	**0.836**	**x 0.530**	**x 0.415**
Hadavi (Farsi)	**IOI**	**0.932**	**x 0.396**	**x 0.324**
Ozaki (Japanese)	**IOI**	**0.939**	**x 0.393**	**x 0.320**

4. I was also a little confused about the segmentation. The introduction notes that each recording will be manually annotated by the coauthor who recorded it, but later on the methods section notes that "Those annotations will be created by the first author (Ozaki) because the time required to train and ask each collaborator to create these annotations would not allow us to recruit enough collaborators for a well-powered analysis." (I think maybe this is referring to the pilot data vs. the proposed data, but I am not totally sure.)

We have clarified the difference between the segmented text (provided by each coauthor) and the onset/breath annotations (created by Ozaki from the segmented text, then checked by each coauthor) in section 2.1 as follows:

> ***"In order to maximize efficiency and quality in our manual annotations, we adopt the following 3-step process:***
>
> 1) ***Each coauthor sends a text file segmenting their recorded song/speech into acoustic units and breathing breaks (see Appendix 1 for examples).***
> 2) ***The first author (Ozaki) creates detailed millisecond-level annotations of the audio recording files based on these segmented texts. (This is the most time-consuming part of the process).***
> 3) ***Each coauthor then checks Ozaki's annotations (by listening to the recording with "clicks" added to each acoustic unit) and corrects them and/or has Ozaki correct them as needed until the coauthor is satisfied with the accuracy of the annotation."***

Along these lines, I appreciated the assessment of interrater reliability for the pilot data, but I wonder if it might be important to get some measure of reliability from the to-be-analyzed data as well. (Like, the analysis of pilot data helps assess confounds in how people segment their own recordings, but if one author does all recordings, might there be some differences in languages that author knows well vs. does not know or something like that?) I'm sure it would be impractical to have multiple raters for all items, but perhaps some subset (or excerpts of some subset, as in the pilot reliability analyses) could be segmented by two people to assess reliability?

We have incorporated this helpful suggestion by adding a new analysis (section 2.7.7) as follows:

> ***"2.7.7: Reliability of annotation process: Each of Ozaki's annotations will be based on segmented text provided by the coauthor who recorded it, and Ozaki's annotations will***

be checked and corrected by the same coauthor, which should ensure high reliability and validity of the annotations. However, in order to objectively assess reliability, we will repeat the inter-rater reliability analyses shown in Fig. S1 on a subset of the full dataset annotated independently by Savage without access to Ozaki's annotations. Like Fig. S1, these analyses will focus on comparing 10s excerpts of song and spoken descriptions, randomly selected from 10% of all recording sets (i.e., 8 out of the 80 coauthors, assuming no coauthors withdraw). Ozaki's annotations corrected by the original recorder will be used as the "Reference" datapoint as in Fig. S1, and Savage's annotations (also corrected by the original recorder) will correspond to the "Another annotator" datapoints in Fig. S1. Note however that we predict that Savage's corrected annotations will be more analogous to the "Reannotation" data points in Fig. S1, since in a sense our method of involving the original annotator in checking/correcting annotations is analogous to them reannotating themselves in the pilot study."

5. The power analysis seems reasonable enough to me, except doesn't the power here depend on the specific languages the recordings come from? (like, shouldn't it be about the necessary variability of language families or whatever rather than just overall number of recordings?)

Actually, our power analysis does not take into account the specific languages used. We have clarified this as follows in section 2.5:

"While it would be ideal to have models that capture how languages (and other factors such as sex, age, etc.) influence the song-speech difference, we do not have enough empirical data or prior studies to build such models at this moment. Hence we simply treat each recording data without such factors, controlling for language family relationships separately in our robustness analyses. Future studies may be able to better incorporate such factors in a power analysis based on the data our study will provide."

These points aside, I think this is a really interesting proposal that is a great fit for a registered report and I look forward to seeing how it develops!

Thanks for the enthusiasm and constructive suggestions!

Reviewer #2 (Anonymous reviewer):

1A. The scientific validity of the research question(s)

-- The question addressed here is how speech and songs differ in their acoustic properties. The question is certainly interesting.

1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

-- The hypotheses are driven by the pilot analysis and are valid.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

-- The methods are largely valid. However, it's of course not ideal to only sample data from co-authors, since co-authors may differ from the randomly sampled participants in many different ways even if they are not aware of the hypotheses being tested. Why not just ask the co-authors to acquire some speech and songs in their native language and do a control analysis on those audios? This approach seems particularly plausible since the current manuscript mainly analyzed songs and spoken descriptions, which are not matched in the content.

We have incorporated this helpful suggestion, as described above in response to the editor who highlighted this point.

A related issue: I'm not sure if it's valid to include audio contributors who are not even aware the hypotheses of the study as co-authors.

Thank you for raising this important point. Please see the editor's comment above regarding this point and our response.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses

-- Yes but the statistics should be improved in the final version.

We thank the reviewer for this advice. Along with the suggestion by the 1st reviewer, we updated our analysis and its statistical formalization (cf. Sections 1.2, 2.5, and S3). We also did an additional round of editing for grammar and clarity (all changes can be viewed in the tracked change version at our OSF repository: https://osf.io/mzxc8/files/osfstorage/63878e40c86e28261c88f556).

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

-- I don't think there are such issues.

Thank you for the succinct and constructive comments!