ToDo:

* Highlights and Graphical Abstract
* Format in STAR methods style

1. **Reviewer #1: The authors investigate vocal emotion recognition (VER) performance in musicians with different levels of training and activity in different domains (instrumentalists and singers). They presumed to find no influences of either factor, which was confirmed. However, they found that VER correlated with measures of auditory sensitivity. They conclude that VER operates at a perceptual level (auditory sensitivity) and cognitive levels (musicality, musical activity) may play little role in this respect. The findings do contribute to the literature. Whether they really close a gap, as the authors claim, remains to be seen. I think it is a bold statement to say that a gap has been closed. It feels a bit overambitious to me. Nevertheless, an interdisciplinary community could benefit from learning what musicality can mean and what not.  
   Reviewer #2: The authors find that vocal emotion recognition does not vary between singers and instrumental musicians, or between amateurs and professionals. The finding is important because if music training is the causal factor in links between musicianship and vocal emotion recognition, one would expect differences in performance based on the type of musical background, particularly whether it focuses on the voice, and the extent of training/performing.  
   Reviewer #3: Thank you for the opportunity to review this interesting manuscript. Overall, I find it to be a well-written and well-justified study. The design builds on a previously published experiment and aims to test specific hypotheses regarding vocal emotion recognition across different musical subgroups. The study was preregistered with nine hypotheses that are tested for, and the target sample size (N=40 per group) was met or slightly exceeded. Recruitment procedures, inclusion/exclusion criteria, and compensation protocols were implemented as described. In addition, exploratory analyses were conducted, which is acceptable as long as they are clearly acknowledged as such. The study aligns with previous work, that musicians expertise in vocal emotion recognition is explained by auditory sensitivity, rather than musical training. The study extends a previous study comparing professionals and non-musicians, by including a group of amateurs. Comparisons are also made between singers and instrumentlists.**

*Response:* We thank all three reviewers for their effort and their careful evaluation of this manuscript. Below, we will respond to each point in detail.

1. **Reviewer #1: Humans operate at different, but highly intertwined levels, perceptual, cognitive, social. There is constant interaction between those. The authors conducted a study to suggest that VER is likely "programmed" to some degree in the nervous system. If it could be modified by learning, music is less a candidate than it might seem. This is a fair conclusion from the literature, to which the current study attempts to contribute. Nevertheless, participants respond in full consciousness, thus using their cognitive abilities to allow for inferences about their perceptual skills. So, in other words, there is no direct, albeit converging evidence to sustain the hypothesis of enhanced auditory sensitivity per se to explain individual differences in VER, irrespective of musical training and modality of training (singing vs. playing instruments). Please explain, why musicianship should play a role in such short utterances. I am aware that the "springboard" seems to be initial positive findings, but does that warrant continued investigations? Why should vocalists have an advantage? But the test amounts to show they do not have one. Why it is important to confirm this?** **Any indication in the literature expect the opposite? In other words is this an original hypothesis or an exploration?**

*Response:* You raise several valid questions here. First, we fully agree that emotional inferences like the ones we target here are the results of intertwined perceptual, cognitive and social processes, which all contribute to vocal perception performance. Based on behavioral data alone, they can hardly be disentangled. We do find a consistent link between auditory sensitivity and individual differences in VER, but we never claimed that this is a purely perceptual process. Quite the contrary, based on previous insights from an EEG-study (Nussbaum 2023), we consider it very plausible that cognitive abilities, especially how the auditory information is used to make a conscious decision plays a role. Nevertheless, this does not seem to be affected by the type or amount of musical training.

Why musicianship should play a role in such short utterances? ->

* This has been shown consistently in many studies

Why should vocalists have an advantage, But the test amounts to show they do not have one. Why it is important to confirm this? Any indication in the literature expect the opposite? In other words is this an original hypothesis or an exploration?

* Suggestions of causality several times (but absence of evidence)
* ToDo
* This is an original hypothesis (and a preregistered one)

1. **The psychoacoustic transformation of stimuli materials is an interesting feature of the study. It suggests that no sensitivity can compensate for a reduction of stimulus information that leaves the participants merely with timbre differences. However, some emotions seem to be affected more or less strongly by morph type (interaction), which seems to make the picture more complicated. Indeed, the selection of emotions (p.10) appears not intuitive. The rationale to use happiness and pleasure (alongside fear and sadness) as distinct categories should be explained. Is there a circumplex-idea in the background? Could it be confirmed that the four emotions occupy four quadrants? Otherwise, there could be a confound arising from similarity and differences of underlying dimensions of the four emotions. - Please explain.**

*Response:* Yes, indeed, a reduction of stimulus information to timbre information does have an impact on the recognition of all four emotions (although to different degrees) across all participants. However, participants are able to use timbre information to some degree, since emotion recognition is still above chance in this condition in all emotions (cf. Figure 3). And indeed, emotions are affected to different degrees by this manipulation, as indicated by the Morph Type x Emotion interaction. We have found this interaction consistently across a number of studies, also with partly different stimuli (e.g. Nussbaum 2022, SCAN; Nussbaum 2023, Brain Sciences and Nussbaum 2024, BJOP). In short, we always see a similar pattern: in general, pitch (F0) information plays a more dominant role, relative to timbre. However, in emotions with high intensity (especially happiness), the difference is much larger. In emotions with lesser intensity (like pleasure and sadness) the contribution of timbre and F0 information is more balanced. We have not discussed this in full detail here, because it was not the focus of the present study, but you find a detailed discussion here: <https://doi.org/10.1093/scan/nsac033>. To make this transparent to the readers now, we added the following: *“Note that the effects of our F0 and timbre manipulation on emotion recognition also provided a complete replication of the pattern observed in professional musicians and non-musicians in our previous study (Nussbaum et al., 2024). For an even more detailed reflection on the roles of F0 and timbre for emotion recognition (irrespective of musicality), please also refer to Nussbaum et al. (2022).” (page 28)*

Concerning the choice of emotions: Indeed, the rationale behind the choice was to balance positive and negative emotions and their emotional intensity. When using basic emotions only, happiness as the only positive emotion is usually the “oddball”, which can be problematic. Thus, we added a second one, which was pleasure. The stimulus material was validated in a previous rating study to ensure that the four emotions occupy the four quadrants of valence/intensity. We added this information in the methods section: “*A prior validation study with 20 raters confirmed that the two positive and two negative emotions had different degrees of emotional intensity (happiness, M=3.40±0.06, vs. pleasure, M=2.88±0.07, t(19)=9.57, p < 0.001, and fear, M=3.01±0.06, vs. sadness, M=2.78±0.07, t(19)=6.58, p < 0.001, on a rating scale from 1 to 4].* ” (page 10/11)

1. **I wonder whether VER could be "boiled down" to the ability to recognize and differentiate pitch contours vocal utterances, with anything going up indicating arousal, and downward indicating relaxation to guess the correct emotion on the basis of contour alone seems rather plausible. So, in our daily interactions, we use this skill myriads of times. Why should even 10.000 hours of music training in 10 years (to reach expertise) change anything in this basic skill? So, why are the findings not trivial after all? Please explain, especially in the light of testing a Null hypothesis.**

*Response:* Our findings in Figure 3 show that emotional inferences are possible based on timbre information alone (although to a lesser degree than based on pitch information). Further, emotion perception performance is best, when both pitch + timbre information is available (in the Full) condition. Thus, while pitch contour undoubtedly plays a dominant role, it is not the only acoustic parameter that signals emotional information and can be picked up by listeners. Further, while we show that pitch contour plays a big role in general, a conclusion like “upward = higher arousal” and “downward = less arousal” seems oversimplified. It could also be something like “high pitch variation = higher arousal” and “low pitch variation = lower arousal”. However, the specific way in which emotion is inferred based on pitch cues was not the focus of the present study, so we refrain from any speculation about this matter. We hold, however, the viewpoint that VER can not simply be boiled down to the perception of pitch contour direction.

Why should even 10.000 hours of music training in 10 years (to reach expertise) change anything in this basic skill? -> In an extensive meta-analysis, Juslin and Laukka (2003) showed that emotions expressed by voices and by music share a similar acoustic code. Based on this, early theoretical frameworks like the OPERA hypothesis (Patel 2011) proposed a causal effect of musical training on non-musical skills, if five specific conditions are met: Overlap (between the neural networks involved), (2) Precision (i.e. high auditory-motor demands), (3) Emotion (i.e. musical activity has to be considered as rewarding), (4) Repetition, and (5) Attention. Based neuroscientific insights, can be argued that singing may share a larger overlap and requires higher precision in the use of the vocal apparatus than other forms of musical engagement, which – based on this framework – could causally affect non-musical skills like vocal emotion perception. Against this backdrop, the proposition of a null hypothesis is not trivial at all. In fact, it may seem counterintuitive. However, we pursued this approach, because for vocal emotion perception (and other non-musical skills), the OPERA hypothesis has not stood up to rigorous empirical examination (Schellenberg & Lima 2024). Nevertheless, suggestions of causal effects of musical training on non-musical skills in the absence of convincing empirical evidence are still prominent in the literature, which is why studies like ours are important.

1. **Reviewer #2: In general, the writing is very clear. My one main suggestion is to emphasize the big picture more. If individual differences in vocal emotion recognition are caused by an environmental factor (i.e., music training/performing in this instance), one would expect such differences to depend on the quality and quantity of the environmental intervention, yet none was found.**

*Response:* Yes, one of the key main points of our story is that there is no evidence for a causal effect of music training on vocal emotion recognition, and this is why we assumed that there should not be an effect of the environmental intervention (in quality and quantity), which is what we found. We adjusted several part in the manuscript, to make this clearer:

1. **I also thought that the ms could be streamlined a bit more by not repeating details that were included in the authors' earlier publication (Nussbaum et al., 2024), but rather referring the reader to the previous article. The authors do this to some extent already. I'm wondering if they could do it more without making reading the earlier paper a prerequisite for understanding the present submission.**

*Response:* We went through the methods section again and checked whether we can reduce redundancy by referring to the previous publication. However, in our view, it’s very important that reading the earlier paper is not the prerequisite for understanding the current one and we feel that a certain level of detail is crucial for that, especially regarding the acoustic manipulations, although they take up a lot of “space”. If the editor or the reviewers have specific ideas for aspects we could leave out without compromising comprehensibility, we are very open to it. So far, we have only shortened a few details:

“*~~After substantial preprocessing (e.g. manual mapping of time- and frequency anchors in each stimulus),~~”* (page 11)

“*~~As browser, we recommended Google Chrome, and excluded Safari for technical reasons.~~”* (page 12)

1. **Reviewer #3: Major issues: Q1. Transparency and preregistration  
   The study claims to follow open science principles through preregistration. However, the preregistration was submitted on May 3, 2025, more than two years after the original study (Nussbaum, 2024), when some data had already been collected and analyzed. While the current study includes new data collection, the timing of the preregistration, shortly before manuscript submission, raises concerns about whether it qualifies as a true preregistration. Most importantly, this makes it unclear whether the hypotheses were formulated a priori or post hoc. Although I don’t consider this as an attempt to circumvent p-hacking, I believe these issues should be addressed explicitly if the benefits of preregistration and open science are to be fully realized.  
   This is my major concern and I would like the authors to elaborate on this, in the manuscript primarily, but of course also in their response.**

*Response:* There seems to be a confusion. The preregistration was submitted/published on June 9, 2023 (link: <https://doi.org/10.17605/OSF.IO/76PV5>). This was \*before\* the very first datapoint of the new dataset was collected (data collection started on June 12). Of course, it was \*after\* the data of the previous study (Nussbaum 2024) were collected, analyzed, and (almost) published, but this was made fully transparent and explicit in the preregistration throughout (e.g. the description and the explanation of preexisting data). We checked, but we could not find the date May 3 2025 anywhere? There was something like an sever update/migration of OSF earlier this year, maybe that when some metadata was changed? However, our preregistration was published in June, 2023 and we can also provide some (automatic) e-mail correspondence/notification with a time stamp on request.

1. **Q2. Coding of PROMS Responses  
   The manuscript states:   
   "In alignment with the approach by Nussbaum et al. (2024), we recoded responses in the PROMS from 0 to 1 in 0.25 steps starting with the “definitely” correct option down two the “definitely” incorrect option (thus, “don’t know” was always coded with 0.5) and subtracted 0.5 from the final measure"  
   Should "down two" be "down to"? Is "definitely correct" coded as 0 or as 1? Why is 0.5 added and then subtracted for the “don't know option”. Please explain more clearly.**

*Response:* Good point – “definitely correct” was coded as 1 and “definitely incorrect” was coded as 0. We adjusted this in the manuscript. Thus, each response gets a number between 1 and 0. We then subtracted 0.5. This was mainly done to facilitate interpretability, because this way a positive score indicates more positive/correct ratings, a negative score indicates more incorrect ratings and a score of zero indicates responses at chance level. But of course, the subtraction of a constant does not affect the statistical analyses of group differences. We rephrased this paragraph to make it clearer:

*“In alignment with the approach by Nussbaum et al. (2024), we first recoded responses in the PROMS from 1 to 0 in 0.25 steps starting with the “definitely” correct option down two the “definitely” incorrect option (thus, “don’t know” was always coded with 0.5). For the final measure, we then subtracted 0.5, because this way, a positive score indicates that participants were more correct/confident, a negative score indicates more incorrect/uncertain ratings, and a score of zero indicates responses at chance level. For statistical analyses, we used the averaged performance across trials for each subtest.” (page 15)*

1. **Q3. Justification of Priors  
   In Bayesian analysis, the choice of priors is critical. While “default setting of priors” are convenient and often appropriate, they may not reflect the best assumptions for a specific research question. The manuscript does not justify the choice of priors, which is particularly relevant given that Bayes factor outcomes are sensitive to prior assumptions. Without such justification, it is unclear whether the default priors aligns with the expected effect size or the theoretical framework of the study (which anticipates a medium effect, as mentioned in the power analysis). Please provide a rationale for the chosen priors.**

*Response:* For our Bayesian analysis, we oriented ourselves at Neves et al. 2025 (<https://doi.org/10.1016/j.cognition.2025.106102>) who tested a null hypotheses for a related research questions and with similar sample sizes. For their analysis, they used default priors and it is also argued in Ly et al 2016 (<http://dx.doi.org/10.1016/j.jmp.2015.06.004>) that default priors are appropriate for the testing of null hypotheses (as the anticipated effect is, in fact, a nulleffect and not a medium-sized one). We have added an explanatory part to the methods sections:

“*These analyses were conducted in JASP Version 0.19.3 (JASP Team, 2025) using default priors, which have been considered appropriate for testing null hypotheses based on similar sample sizes (Ly et al., 2016; Neves et al., 2025). Further, we ensured that our Bayesian inference did not depend critically on the choice of priors by running robustness checks (see data analysis files on OSF).”* (page 15)

1. **Q4. Sensitivity Analysis  
   The manuscript does not report whether results are robust to different prior choices. This is a common but important omission in Bayesian reporting. I recommend a sensitivity analysis to demonstrate the reliability of the findings across varying prior assumptions.**

*Response:* In fact, we did include a robustness check in our Bayesian analysis, which we uploaded on OSF but did not mention that in the manuscript. We fixed that now (cf. our response to the previous point).

1. **Q5. Rationale for Emotional Categories  
   The rationale for selecting the emotional categories could be expanded. While the inclusion of two positive (Pleasure, Happiness) and two negative (Fear, Sadness) emotions provides a balanced valence structure, it is unclear whether the authors considered the temporal and psychological distinctions. For example, Pleasure is often immediate and biologically driven, whereas Happiness is more enduring and psychologically constructed. Similarly, Fear is typically short-lived and reactive, while Sadness may be more prolonged. Where these distinctions considered in the design? Could comparisons between short-term and long-term, or comparisons between positive and negative grounded emotions offer additional insights? And if so, as an exploratory analysis.**

*Response:* Concerning our rationale for selecting the emotional categories, please refer to our response to point 3. Regarding the temporal domain, our voice morphing approach ensures that the timing is kept constant across all emotions by holding the time anchors constant. However, that does not imply that the emotional information does unfold similarly over time. It is possible, for example, that the emotional information in happiness available earlier than in e.g. in sadness. We discussed this possibility in a previous EEG study, because we observed somewhat emotion-specific timing in the electrophysiological responses. This was not the focus of the present study, but we mentioned this work in the discussion for the interested reader (page 28).

1. **Q6. Inclusion of Big5 Personality Traits  
   In table 4, the BIG5 personality traits are reported, although they are not linked to any of the preregistered hypotheses. Was analysis conducted for exploratory purposes? If so, please clarify and consider presenting these results in a separate exploratory section to avoid confusion.**

*Response:* The purpose of Table 1 and Table 4 is to provide insights about the comparability of the two groups. This is a cross-sectional design, so there is always a risk of undetected confounds, but at least we wanted to make sure that the groups would not differ with regard to “common candidates” (i.g. socioeconomic status, personality, or affective state), or if so, be transparent about it. In line with our reporting style in our previous study (Nussbaum 2024), we put this at the beginning of the results section to give the readers a good overview over the sample characteristics before we move on to the hypotheses. But we are open to relocate or reframe this sample overview, if this is strongly preferred by the editor or the reviewer. In any case, we added a line:

*“First, we checked for comparability of groups on a number of demographic and psychological aspects.” (page 16)*

*“Again, we first checked for comparability of groups.”* (page 24)

1. **Q7. Terminology in Hypotheses H5-H7  
   Some terminology used in hypotheses H5-H7 is not introduced in the methods section. See “General-ME”, “Perception Subscale”, and “selfrated singing abilities”. For many readers, these terms only become clear when reviewing the results, particularly Table 4. Please define these terms earlier in the manuscript.**

*Response:* We tried to improve readability by slightly rephrasing these hypotheses directly:

*“H5: Averaged-VER and Full-VER are correlated with the general musical sophistication (General-ME), measured by the Gold-MSI.*

*H6: Averaged-VER and Full-VER are correlated with the Perception Subscale of the Gold-MSI*

*H7: Averaged-VER and Full-VER are correlated with self-rated singing abilities Subscale of the Gold-MSI.” (page 20/21)*

1. **Q8. Hypothesis H9   
   I did not understand the rationale for hyptothesis H9. Please explain.**

*Response:* We tried to outline the rationale for Hypothesis 9 at the beginning of section 4 (Part III). In principle, we assumed that amateur and professional musicians would perform equal in our emotion recognition task. Thus, we predict a similar null finding as for the contrast between singers and instrumentalists. The rationale behind this being that the link between musicality and vocal emotion perception is not based on the quality or quantity of formal musical training. However, back then, we also considered a number of studies that also found that amateurs were actually better than professionals in a number of non-musical abilities, and we therefore phrased this hypothesis more carefully and including this possibility as well. Now, there is new evidence that challenges the notion that amateurs have superior cognitive abilities than professionals ([Talamini et al. preprint](https://www.researchgate.net/profile/Francesca-Talamini-2/publication/394797163_Musical_Expertise_and_Cognitive_Abilities_No_Advantage_for_Professionals_over_Amateurs/links/68a717ad6327cf7b63d88525/Musical-Expertise-and-Cognitive-Abilities-No-Advantage-for-Professionals-over-Amateurs.pdf)). However, we obviously cannot change the hypothesis (nor their wording) post-hoc to align it with new evidence.

Please note that our hypothesis targets F0 and the Full morphs condition only, because we assumed that performance would be equal for Timbre anyway. However, we admit that we missed making this explicit in our preregistration.

We slightly adjusted the paragraph before the hypothesis, to make the rationale more transparent:

“*Mostly, we predicted that amateurs and professional musicians would be comparable regarding vocal emotion recognition. However, as previous evidence reviewed above showed that amateurs can differ from professionals in cognitive abilities which could be linked to emotional sensitivity, we also considered the option that amateurs could be more proficient at making emotional inferences than professionals, reflected in our H9 below*.” (page 22)

1. **Q9. Definition of Singers (and Instrumentalists)  
   On page 9, singers are defined as individuals who “must not currently be instrumentalists in an orchestra or a band". Did you control for participants’ musical history, such as previous experience playing an instrument? If so, how?**

*Response:* This is an important point and one that we picked up for critical reflection in our discussion. While it may seem trivial at first glance, making a clear-cut distinction between singers and instrumentalists is actually very complicated on a practical level to recruit mutually exclusive groups. Thus, a few of our singers did report playing an instrument in the past (details are all on OSF) and many instrumentalists also sing (although not on a regular basis or in an ensemble). One possibility would have been to exclude these individuals but at the potential cost of statistical power. Instead, we opted for transparent description of our samples and an honest discussion of this limitation on page 31.

1. **Q10. Stimulus Validation  
   Were the stimuli validated? Figure 5 in Nussbaum (2024) shows confusion rates for different emotions, but perhaps there were specific stimuli that were consistently misclassified? Are misclassifications a result of poor interpretation or poor stimuli representation? For example, in Figure 3, is the poor performance in Timbre condition a result of poor stimuli quality or poor perception? Please elaborate.**

*Response:* Yes, stimuli were validated, for details please refer to our response to point 3.

We also calculated confusion rates for the present sample, which can be found in the supplemental materials on OSF (and references in the manuscript).However, we did not check whether specific individual stimuli were consistently misclassified.We acknowledge that this could be an insightful analysis, but since this was not part of our preregistered analysis plan and due to concerns about statistical power, we did not perform it with the present dataset. However, the data is openly available now, leaving the option for future research.

Please note that confusion matrices looks very similar between all groups (singers vs. instrumentalists; and amateurs vs. professionals vs. non-musicians). Thus, while we cannot exclude the possibility that single stimuli may have been misclassified, we would argue that this would not invalidate the conclusions of the study, because this would probably have affected all groups similarly.

We can only speculate whether misclassifications are the result of poor interpretation or poor representation, because with the present paradigm, we study both in conjunction. Given that vocal emotion classification is highly comparable across groups, we would rather interpret this pattern as a result of the stimulus material and conclude that timbre is less emotionally informative relative to the pitch contour or the combined information. Please note that we can only make inference about the relative importance of cues (i.e. pitch vs timbre). Further, we also refrained from interpreting absolute differences between emotions (i.e. whether happiness is overall better recognized than pleasure or fear), because this depends highly on the stimulus material and response biases of participants, which is why did not consider this informative.

1. **Q11. Conclusion and causality  
   The conclusion that “the ability to recognize emotions in voices is not primarily determined by the type or amount of musical activity, but rather by individual differences in auditory sensitivity” is based on correlations in table 3 and table S10. First, are both tables including control for musical education? Or is it only controlled for in table S10. Please explain the differences. Second, “determined by” implies a causal relationship. Since this is not an experimental setup where individuals’ auditory sensitivity is manipulated, the causal claim has to be based upon theoretical assumptions (for example by motivating that no other potential causes occur).**

*Response:* Any implication of causality was unintentional and therefore we agree that “determined” was a poor choice of wording. Please refer to our response to point 18, where we list several changes we applied to the manuscript in order to avoid causal language.

In section 3.2. (Data analysis), we report that we controlled alle correlations for formal musical education as well (according to our preregistered plan), but this made no difference for the observed patterns. For simplicity, we kept the uncontrolled correlations in the manuscript (S3) and the ones controlled for musical education (S10) in the supplemental materials on OSF.

1. **Reviewer #1: The causal language should be reconsidered, even though it is taken on board from the literature, in part. This is a correlational study.**

*Response:* Since our study specifically challenges the claim that musical training /activity is causally involved in the link between musicality and vocal emotion perception, any impressions of implied causality by the choice of wording is unintentional. We scanned me manuscript and slightly adjusted the wording in the following parts:

* *“auditory skills promote vocal emotion recognition” -*> “*auditory skills are linked to vocal emotion recognition*” (page 4)
* “*Overall, the few data that are available do not provide clearcut, let alone causal evidence for a specific benefit in vocal emotion recognition by singing over playing an instrument.*” (page 6)
* “*emotion recognition is not related to formal training, but rather to natural differences in auditory sensitivity*” (page 8)
* *“these results suggest that the link between musicality and vocal emotion recognition is driven by individual differences in auditory sensitivity” -> we changed “driven by” to “associated with”* (page 27)
* “*This adds a new perspective to the accumulating evidence that the link between musicality and vocal emotion recognition is predominantly driven by individual differences in natural auditory sensitivity.” -> changed “driven by” to “associated with*”(page 32)

1. **The number of hypotheses seems inappropriate to me, at least. As stated above, commenting on a correlation table is sufficient. The hypotheses 3-7 are not at the same level of the surrounding ones. That should be reflected in the writing.**

*Response:* ToDo

1. **Reviewer #2: The ms is written in APA style but some of the references diverge from APA (e.g., title case for article titles, abbreviated journal names).**

*Response:* Thank you for pointing that out. We scanned through our references and formatted them in APA-style.

1. **São Luis Castro should be abbreviated Castro, S. L. in the reference list.**

*Response:* Thanks for noticing. We fixed this.

1. **p. 30 "...future research should incorporate..." Change "incorporate" to "consider"**

*Response:* Changed as suggested.

1. **Why/how were the four subtests from the PROMS chosen? These aren't the same as those that are included in the mini-PROMS.**

*Response:* This is true. We chose these specific subtests because we considered them most informative for our research. Because we are contrasting pitch (F0) and timbre in the voice, we included the subtests that we assumed to capture similar aspects in the musical domain (melody, pitch and timbre) and then added on rhythm task to capture sensitivity to temporal information as well. We included only four subtests to keep overall duration of the study within reasonable limits. We added a small clarification to the manuscript: “[...] comprised of the four subtests „Melody“, „Pitch“, „Timbre”, and „Rhythm“, which we considered most informative for the present research question.” (page 13)

1. **Baldé et al. (2025) now has volume, issue, and page numbers. It came out in the June issue of JEPHPP.**

*Response:* We added the missing information.

1. **Reviewer #3: Minor issues: Q12. Clarity in Reporting Simple Main Effects. The reporting of simple main effects following significant interactions is exemplary in format, however, there are some points that require clarification. See for example sentence: "…all pairwise comparisons |ts(77)| = 2.57, ps = .012, ds = 0.28 [0.06 0.49], except for Fear vs. Sadness (|t(87)| = 1.13, p = .261)." This begins with comparisons between F0 and Timbre (conditions), but ends with comparisons Fear vs Sadness (emotions). Please clarify and make sure all relevant comparisons are reported.**

*Response:* In this particular paragraph, we calculated the performance difference between the F0 and the timbre conditions and compared this difference between emotions. We carefully checked the sentence, and we think that this is described here correctly. However, we made this more explicit by adding a little specification and changing eg. “*M = 0.34*” to “*MF0-Timbre = 0.34*”. (page 18)

1. **Q11 Epsilon HF: in table 2, explain epsilon{HF}, for example in a footnote. (Huynh-Feldt I assume).**

*Response:* We added the following note: “Note. **εHF =** Huynh–Feldt (HF) epsilon correction factor in case of violation of the sphericity assumption.” (page 18)

1. **Q12. Missing Legend in Table 3: Table 3 lacks a legend. Please add one to clarify the meaning of the columns and values.**

*Response:* Good point. We added a table caption “Spearman correlations between the PROMS and vocal emotion recognition performance” and more information in the legend: “VERAvg: VER performance averaged across all trials, Full-Morphs: VER in the Full Morph condition only, F0-Morphs: VER in the F0 Morph condition only,, Timbre-Morphs: VER in the Timbre Morph condition only, PROMSAvg: music perception performance averaged across all four subtests of the PROMS (Pitch, Melody, Timbre, and Rhythm).” (page 21)

1. **Q13. Terminology Consistency: The manuscript uses both “Vocal emotion recognition” and “vocal emotion perception”. Please ensure consistent terminology throughout the manuscript or clarify distinctions if several terms are intentionally used.**

*Response:* We have changed “vocal emotion perception” to “vocal emotion recognition” throughout.

1. **Q14. Comparisons between Professionals and Non-Musicians: In Table 4, some variables (Extraversion, Attention to Detail, Social Skills, Rhythm) show differences between professionals and non-musicians. These comparisons are not reported in the table. I suggest either including them where significant or or explicitly stating that such comparisons are reported elsewhere (Nussbaum, 2024?).**

*Response:* Indeed, we did not focus on these comparisons, because they are already reported and discussed in the previous publication. This is mentioned in the manuscript under 4.2 Method: “We focused our analysis on the comparison of amateurs with the other two groups, because the comparison of professional musicians and non-musicians is reported in Nussbaum et al. (2024).” To make this more transparent to the readers, we added a note to Table 4 as well: “*For a detailed description and discussion of the differences between professional musicians and non-musicians, please refer to the previous publication (Nussbaum et al., 2024).*” (page 24)

1. **Reviewer #1: The authors do not mention in the Abstract, which emotions were considered or whether a model representing dimensions of emotion recognition was used. Therefore, it is not obvious what they mean by emotion recognition upfront.**

*Response:* We added this information to the abstract now: *“[…] based on short vocal utterances expressing happiness, pleasure, fear, and sadness.” (p2)*

1. **The authors use causal language at times, which is understandable due to the wide-spread use of it in the literature. However, in their own terms, they should rather consider to talk about associations, even when the literature they cite implies causality in contexts such as correlational studies. So, please avoid causal language.**

*Response:* Please refer to our response to point 18, where we list the detailed changes we made to address this concern.

1. **The findings are relevant as they extend the current knowledge to assert that professional or amateur vocalists or instrumentalists do not differ in their VER performance, neither do personality, or music sophistication play any role. However, this may still not be too surprising in light of previous literature, although this conclusion can now be drawn on a planned study.   
   For example, https://doi.org/10.1037/emo0000770 found that "natural musical abilities" without training are sufficient to excel in VER, limiting the role of musical training, a study that the authors cite themselves. Note that the sample in that study also was substantial (N=169). This leaves the authors' claim of a substantial sample size as a reason to conduct this study with a bit of a sour taste, albeit the sample size was sufficient for their study. I do not really trust the claim as expressed in the first line of 5.3.**

*Response:* ToDo

1. **There is ample research by Schellenberg to suggest that musical talent or aptitude rather than musical practice and expertise predicts a range of aspects of auditory functioning, especially in the linguistic or speech domain** [**https://doi.org/10.1037/xlm0000798**](https://doi.org/10.1037/xlm0000798)**.**

*Response:* ToDo

1. **The authors report a software error (p.13), but offer no account for how that error could have affected the results.**

*Response:* ToDo

1. **I am struggling with the authors' decision to test 9 hypotheses. I believe that hypothesis-driven approaches are mandatory, but there is a line to be drawn between hypotheses based on previous findings in the literature and ad hoc hypotheses (working hypotheses). Therefore, I would rather like to see the basis for H3-H7 more clearly. Otherwise, a correlation table is sufficient with an associated description of implications. Also, considering an arbitrary N/hypothesis ratio, it turns out that 10 individuals on average account for one hypothesis. Given the massive load of data acquired in the 25 minutes of experimenting (fatigue-effects being ignored), I think we are approaching a grey area in experimental economy. The authors should perhaps run tests whether the quality of responses at the beginning and at the end of the 25-minute period is still the same.**

*Response:* ToDo

1. **Table 5 reports an emotion x morph type interaction. Did I miss a discussion of that effect, or follow-up analyses? Please explain.**

*Response:* ToDo