Roles of bilingualism and musicianship in resisting semantic or prosodic interference while recognizing emotion in sentences

Cassandra Neumann, Anastasia Sares, Erica Chelini, & Mickael Deroche

A Thesis

in the Department

of Psychology

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts (Psychology) at

Concordia University

Montréal, Québec, Canada

September 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Cassandra Neumann**

Entitled: **Roles of bilingualism and musicianship in resisting semantic or prosodic interference while recognizing emotion in sentences**

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Psychology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____Chair
Dr. Kristen Dunfield

_____Examiner
Dr. Krista Byers-Heinlein

_____Examiner
Dr. Virginia Penhune

_____Supervisor
Dr. Mickael Deroche

Approved by

_____
Dr. Andrew Ryder, Chair of Department

_____
Dr. Pascale Sicotte, Dean, Faculty of Arts and Science

Date: September 19th, 2022

ABSTRACT

Roles of bilingualism and musicianship in resisting semantic or prosodic interference while recognizing emotion in sentences

Cassandra Neumann, B.A.

To infer emotions in speech, listeners can use the way people speak (prosody) or what people say (semantics). We hypothesized that bilinguals and musicians would rely more on prosody than on semantics. In two online experiments, we collected data on 1041 young adults, who listened to sentences with either matching or mismatching semantic and prosodic cues to emotions. Participants then identified the emotion enacted by the speaker's prosody (ignoring semantics; Experiment 1) or in the semantics (ignoring prosody; Experiment 2). In both experiments, performance suffered when prosody and semantics conflicted. Musicians were better at resisting the interference among bilinguals, but not among monolinguals. Thus, the musician advantage may not be due to a difference in weighting prosody over semantics, rather an overall better ability to inhibit irrelevant information. As for the hypothesized bilingual advantage, the findings warn that musicianship is critical to control for.

Key words: bilingualism, musicianship, prosody, semantics, vocal emotion recognition

Acknowledgements

I would like to thank my research supervisor, Dr. Mickael Deroche, for all his guidance throughout these last two years. I would like to also thank the members of my defense committee, Dr. Krista Byers-Heinlein and Dr. Virginia Penhune, for their helpful contributions and constructive feedback. Special thanks go to my colleagues in the Hearing and Cognition Lab, in whom I have found great friendship. I would like to thank my partner, my family, and my friends, for their continuous patience and support.

I also would like to thank all participants on Prolific who gave their time to complete this study. Additionally, I would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada's (NSERC) Discovery Grant awarded to M.D. (ref: DGECR-2020-00106), NSERC's Canada Graduate Scholarship awarded to C.N., les Fonds de recherche du Québec – Nature et technologies (FRQNT) Scholarship awarded to C.N. (#301964), and the Center for Research on Brain, Language, and Music (CRBLM) Scholarship awarded to C.N. The CRBLM is funded by the Government of Quebec via the Fonds de Recherche Nature et Technologies and Société et Culture.

Contribution of Authors

All authors conceptualized the study. C.N. and A.S. created the stimuli and M.D. processed the stimuli. A.S. coded the online experiment. C.N. and E.C. collected, cleaned, and organized the online data. C.N. analyzed the data and wrote the manuscript that was then edited by A.S., M.D., and E.C. All authors reviewed the final manuscript and approved the contents.

Table of Contents

List of Tables

Page

List of Figures

Page

List of Appendices

Roles of bilingualism and musicianship in resisting semantic or prosodic interference while recognizing emotion in sentences

The ability to attribute mental states, including emotions, to oneself and to others results in better social development, communication skills, and empathy (Baron-Cohen et al., 1985). Early in life, infants are highly attracted to social stimuli such as faces and voices, especially those of their mothers (Grossmann, 2010). Infants have also shown the ability to detect the prosodic properties of vocal emotions early on. Mastropieri and Turkewitz (2009) found that newborn infants are more likely to open their eyes when presented with happy vocal stimuli compared to other emotions, a response that intriguingly was only found for their native language, suggesting that affective skills may depend on language mastery. Thus, it is early in life that children begin to recognize prosody (Friend, 2001; Mastropieri & Turkewitz, 2009). However, it is not until age 5 that children begin to consistently label a speaker's emotional state using their tone of voice (Aguert et al., 2010, 2013; Sauter et al., 2013).

Prosody refers to the elements of speech that lack any semantic content. It communicates a speaker's intent, attitude, or affect with the use of acoustic variables such as pitch, intensity, and duration of speech segments (Botinis et al., 2001; Cutler et al., 1997; Lehiste, 1970). For example, anger is typically characterized by high pitch (often with a descending contour), high intensity levels, and a rapid and variable speech rate (Preti et al., 2016). To isolate the role of prosody, many studies have intentionally used semantically neutral sentences. In daily conversation, however, the emotional prosody of speech can sometimes conflict with the semantic context. In such cases, the understanding of emotional prosody is vital for understanding the true message of speech. For example, the utterance "What a great day" has positive or happy semantic content. However, if said in a sarcastic tone of voice, it would

indicate the speaker's discontent. Thus far, the literature shows that when presented with incongruent semantic and prosodic cues to emotions, 4-year-old children will make judgements about a speaker's emotions based on semantic cues, and it is not until 10 years of age that children shift towards using prosodic cues in such situations (Friend, 2000; Friend & Bryant, 2017; Morton & Trehub, 2001). This is surprising given that prosody can be recognized very early in life, before children learn to speak and understand the semantic context of speech (Friend, 2001). This raises the question of how the progressive mastery in a language along with general maturation effects eventually offset the balance between semantics and prosody.

While children are developing the ability to detect emotions in speech prosody, many are also being exposed to a second language (Grosjean, 2010). Being bilingual or multilingual comes with several advantages. In addition to being able to communicate with more people, research has shown that being bilingual may contribute to better cognitive control and metalinguistic awareness, as bilinguals must learn when to use each language depending on the context (Adesope et al., 2010; Bialystok & Craik, 2010; Christoffels et al., 2013; Kroll & Bialystok, 2013; Yow & Markman, 2015, 2016). Thus, bilinguals are constantly making linguistic decisions, and may be better at handling conflicting demands. However, the literature on bilingualism has not always shown an advantage (Paap, 2019; Paap & Greenberg, 2013). While bilinguals are sometimes better at certain cognitive tasks, research has also shown deficits in bilinguals' linguistic abilities, and this remains true across the lifespan (Bailey et al., 2020; Bialystok & Craik, 2010; see meta-analysis by Donnelly et al., 2019). Little is known about the developmental course of paralinguistic cues, such as prosody, in bilinguals. It is well established that both languages known by a bilingual are activated in all contexts, requiring selection mechanisms to attend to the appropriate language in a given listening situation (Bialystok, 2017).

Thus, bilinguals are constantly juggling the semantics of their competing languages. To reduce mental load, bilingual children may progressively learn to use a cue that is more consistent across individuals and languages: prosody.

Emotional prosody is a universal cue recognized across various cultures and languages (Paulmann & Uskul, 2014; Scherer et al., 2001). In a study where 4 year old children were presented with sentences with conflicting prosodic and semantic cues, bilinguals showed an earlier ability to use prosodic cues than monolinguals (Yow & Markman, 2011). These results may be rooted in bilinguals' enhanced executive functioning, specifically in inhibitory control (Bialystok, 1999; Costa et al., 2008; Kovács & Mehler, 2009). Alternatively, bilingual children may simply demonstrate a prosodic bias, as seen in Champoux-Larsson and Dylman's (2018) study. When asked to identify the emotion in the content (i.e., semantics) of the words while ignoring the prosody, bilingual children made *more* mistakes than monolingual children. When asked to identify the emotion in the prosody while ignoring the content, bilingual children made *fewer* mistakes than monolingual children and this difference increased both with age and increased bilingual experience. Thus, bilingual 6–9-year-olds demonstrated a *prosodic bias* whereby they used prosodic cues to detect vocal emotions, even when prosody was the distracting cue. This prosodic bias may continue into adulthood, but only under some conditions (Champoux-Larsson & Dylman, 2021). However, one factor not previously considered was the effect of musical training.

Musical training has been hypothesized to be beneficial for the encoding of speech and more globally for processing language (Patel, 2011; Shook et al., 2013; Tierney et al., 2013; Tierney & Kraus, 2013). This is not surprising given that music and language share many common features (Besson et al., 2011; Hausen et al., 2013; Peretz et al., 2015), including of the

communication of emotions (Paquette et al., 2018). Emotions can be recognized in music as their acoustic properties are similar to emotions depicted in speech (Deroche, et al., 2019a; Juslin & Laukka, 2003). Research has also shown that musicians are better than non-musicians at detecting pitch fluctuations in both music and language (Sares et al., 2018; Schön et al., 2004), and there is some mixed evidence to suggest that they may also be better at recognizing the emotional prosody in speech (Lima & Castro, 2011; Trimmer & Cuddy, 2008). The fact that these findings hold for both speech and music is critical: it suggests that this musician advantage effect (for the "musicality" of speech) may be robust to linguistic or semantic influences in the speech materials. However, in real life, people convey emotions partly through prosody and partly through their choice of words (i.e., semantics), the latter being arguably more straightforward. To date, it is unknown whether adult musicians and non-musicians differ in their use of prosody versus semantics for emotion processing, but there could well be a musician advantage in parsing these cues when they conflict.

Some studies have looked at the individual effects of bilingualism and musicianship on vocal emotion recognition abilities in a single study. Bialystok & DePape (2009) used an auditory Stroop task, where listeners were instructed to attend to prosody or to semantics of single words (and not sentences) with an emotional meaning. They found that musicians (monolingual) responded more quickly than bilinguals and monolinguals (both non-musicians) in the prosody task, but there was no group difference in the semantics task. Similarly, Graham & Lakshmanan (2018) largely replicated the same design but only included a prosody task. They found that musicians (monolingual) had reduced reaction times on incongruent trials and smaller cognitive costs compared to the bilinguals (non-musicians and non-tone second language) but did not differ from monolinguals (non-musicians) or tone language bilinguals.

The current study aims to examine whether bilinguals and musicians weight cues to emotions differently than monolinguals and non-musicians when making judgements about vocal emotions in sentences. More specifically, in situations where semantic and prosodic cues to emotions conflict, we hypothesized that bilingual young adults would either demonstrate a prosodic bias, as seen in children, whereby they would outperform monolinguals when asked to use prosody and ignore semantic cues, but conversely their performance would suffer when asked to use semantic cues and ignore prosodic cues. Alternatively, they may be more resistant to conflicting cues in general (i.e., due to better cognitive control and response inhibition) and would therefore, outperform monolinguals on both tasks. For musicians, a prosodic bias received mixed evidence so we hypothesized that musicians would outperform non-musicians both when asked to use prosodic cues and when asked to use semantic cues to emotions, thus demonstrating a musician advantage regardless of the cue/task in question. To test this, we designed two separate studies to mirror each other, with participants either attending to prosody (Experiment 1) or semantics (Experiment 2) to report the emotion contained in sentences—a sort of emotional Stroop task. Note that these experiments were run between subjects to avoid 1) participants switching between the two tasks possibly changing listening strategies and 2) to optimize the number of trials per task (enhancing the granularity of the dependent variable), while avoiding repetition of the stimuli.

Methods

Participants

A total of 1086 participants across two experiments were recruited through Prolific, an online recruitment platform. To ensure that participants had sufficient knowledge of English to complete the experiment, recruitment was open only to specific English-dominant countries (Australia, Ireland, New Zealand, United Kingdom, and United States). Four separate batches were collected for each experiment: bilingual musicians, bilingual non-musicians, monolingual musicians, and monolingual non-musicians. The batches were based on filters for bilingualism and musicianship available in Prolific. Forty-five participants either had technical difficulties (e.g., downloading the materials or browser issues) or did not complete the experiment and were thus excluded from the analyses. None of the participants had concerns about their hearing but two participants (0.19%) reported having mental health issues (still included). The final sample included 526 participants ($N_{Females}$ = 271, N_{Males} = 253, and $N_{Prefer\ not\ to\ say}$ = 2) in Experiment 1 and 515 participants ($N_{Females}$ = 298 and N_{Males} = 217) in Experiment 2, all between the ages of 18-41 years old (M = 24.80, SD = 5.50).

Within the experiment, participants were asked about their language and musical background, and based on these answers (*not* the original Prolific filters), they were divided into four groups: bilingual musicians ($N_{Experiment\ 1}$ = 177, $N_{Experiment\ 2}$ = 171) bilingual non-musicians ($N_{Experiment\ 1}$ = 114, $N_{Experiment\ 2}$ = 101), monolingual musicians ($N_{Experiment\ 1}$ = 138, $N_{Experiment\ 2}$ = 144), and monolingual non-musicians ($N_{Experiment\ 1}$ = 97, $N_{Experiment\ 2}$ = 99). Participants were asked "How many languages do you know in total?" and then asked to give the name of language. Following this, for each language reported, participants were asked "At what age did you begin learning this language?", "How proficient are you in this language?", and "In the past

year, how much have you used this language in daily life? 0 = Never, 10 = Exclusively." The same questions were then asked for instruments played. The group classification was intentionally simple: monolinguals were participants who reported knowing only one language, English, while bilinguals reported knowing two or more languages. Similarly, non-musicians were participants who did not play any musical instrument, while musicians reported playing one or more instruments. Note that this is not to deny the considerable variability within these groups. There is a notorious heterogeneity among bilinguals (e.g., de Bruin, 2019; Luk, 2015) and among musicians (Daly & Hall, 2018), so we recorded further information about their age of acquisition, proficiency, and use of each of their language or musical instrument. Some participants' expertise in either their language or musical abilities bilingualism or musicianship could be questioned. For example, it is known that early-trained musicians (before age 7) have behavioral benefits in auditory tasks (Bailey et al., 2020) and changes in cortical and sub-cortical brain networks compared to late-trained musicians (those beginning after age 7; Penhune, 2019; Shenker et al., 2022; Vaquero et al., 2020). In our sample, of our 315 musicians, only 124 reported learning their first instrument before age 7. On this basis, one might be tempted to narrow down our musician group definition (and the same holds for bilingualism), but this is a slippery slope: given the dangers of dichotomizing such continuous variables (MacCallum et al., 2002), we did not reclassify participants using arbitrary cut-offs from these metrics. We did, however, explore bilingualism and musicianship as continuous variables in regression approaches (for example with proficiency scores; see Appendix D), and the findings were in line with our current categorical approach.

Protocol

Once recruited through Prolific, participants were redirected to the experimental interface hosted on Pavlovia (an online platform for behavioural experiments), that was designed using the PsychoPy software (Peirce et al., 2019). All participants provided informed consent in accordance with the Institutional Review Board at Concordia University (ref: 30013650) and were compensated £3.90 for their online participation.

After providing informed consent, participants were provided with written instructions on how to complete the task. Participants were asked to adjust the volume of their device to a comfortable level before beginning a practice block. The practice block consisted of 16 trials, half of which were congruent (matching semantics and prosody), and the other half were incongruent (differing semantics and prosody). Participants were asked to attend to the prosody of each sentence in Experiment 1 or the semantics in Experiment 2. After the presentation of each sentence, the participants were asked to click on the word of the emotion that was expressed out of four possible options: angry, calm, happy, or sad. To pass the practice block, the participants had to obtain a minimum of 75% correct (12 out 16 trials correct). If this was not obtained, participants continued repeating the practice block until 75% was attained. Feedback on performance was provided for practice trials but not for test trials. After completing the practice, participants moved on to the test phase.

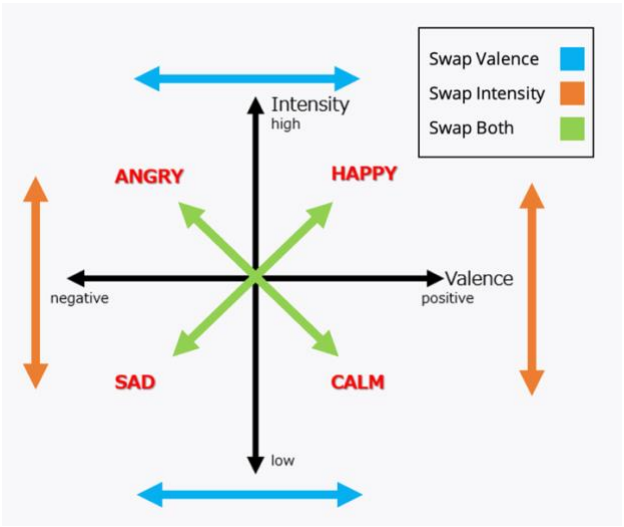
The test phase consisted of 144 trials split across three blocks (48 trials per block). In each block, half of the trials (24) were congruent, and the other half were incongruent. Trials were equally divided into the four emotions: angry, calm, happy, or sad. Participants in the first experiment were asked to attend to the prosody of each sentence and ignore semantics. Participants in the second experiment were asked to attend to semantics and ignore prosody.

Participants were presented with audio recordings of the sentences and asked to choose which emotion was expressed out of the four possible options.

Each of the three blocks differed in the way in which the semantic and prosodic cues to emotions were swapped in the incongruent trials (see Figure 1). In the *swap valence* block, the valence, or positive-negative dimension, of the emotions was swapped (e.g., a semantically angry sentence enacted with a happy prosody). In the *swap intensity* block, the intensity, or high-low energy dimension, of the emotions was swapped (e.g., a semantically happy sentence enacted with a calm prosody). Finally, in the *swap both* block, both the intensity and valence of the emotions were swapped (e.g., a semantically angry sentence enacted with a calm prosody). The order in which these three blocks were presented was counterbalanced across participants.

Figure 1

Three different block types in the test phase



Note. The blue arrows show a swap in valence, the orange arrows show a swap in intensity, and the green arrows show a swap in both intensity and valence.

Finally, the participants were asked about their language and musical background. Specifically, they were asked to name each of their languages and each of their musical

instruments, detailing the age of start, proficiency, and frequency of use or practice. Proficiency and use were rated on a 10-point Likert scale, where 0 was not proficient at all or never used, and 10 was the most proficient or used all the time. As previously mentioned, monolinguals were participants who reported knowing only one language, English, while bilinguals reported knowing two or more languages, regardless of their proficiency, use, or age of acquisition. Similarly, non-musicians were participants who did not play any musical instrument, while musicians reported playing one or more instruments, regardless of their proficiency, use, or age of acquisition. The experiment took on average 25 minutes (SD = 11) to complete. The amount of time taken to complete the experiment did not differ by group (F (3,1033) = 0.86, p = .462, η^2 = 0.002), or by study (F (1,1033) = 0.49, p = .483, η^2 < .001), nor was there an interaction between the two (F (3,1033) = 1.29, p = .277, η^2 = 0.004).

Stimuli

All stimuli were created by the experimenters and produced by four speakers (2 males and 2 females) to generate variability and prevent listeners from learning speaker-specific manners of conveying emotions (either through voice characteristics or speaking style). The list of 144 sentences can be found in Appendix A and contained 36 semantically angry sentences (e.g., “My sister gets on my nerves”), 36 semantically calm sentences (e.g., “Baths are relaxing”), 36 semantically happy sentences (e.g., “Let's go to Disneyland”), and 36 semantically sad sentences (e.g., “His grandmother died”). The speakers read each sentence with the prosody of all four emotions to create congruent and incongruent stimuli, resulting in 576 recordings from each speaker. Thus, there was a total of 2304 stimuli in the full set. Of these, 144 were randomly selected for each participant, with no repetition of sentences.

We conducted an analysis on the semantics of each sentence using the *word2vec* algorithm (see Appendix B for more details on this analysis). It confirmed that, overall, each set of sentences contained semantic content that reflected the intended emotion. However, this was somewhat difficult to demonstrate and can perhaps be improved with more advanced packages (Raji & de Melo, 2020). Similarly, we conducted an analysis on the prosody of each sentence, demonstrating that emotions were enacted by the four speakers as expected: angry productions were particularly fast and dynamic in their intensity contours, while sad productions were slow and more stationary; happy productions were particularly high in pitch and well intonated, while sad and calm productions were low and more monotonous. In each metric, however, it is clear that speakers had their own style (see Appendix C for more detail) and were only partially consistent with one another in how they conveyed emotions.

Equipment

Given that the present experiments took place online, we did not have rigorous control over the participants equipment and quality of sound. To address this limitation, we asked participants to indicate the audio device they were using (headphones, earbuds, external speakers, or default output from their PC/laptop), and asked them to rate the quality of their audio from 0-10 (where 0 is poor and 10 is excellent). There were no differences between groups in audio quality (F (3, 1037) = 0.22, p = .881, η^2 <.001; M = 6.3, SD = 0.86). There were also no differences between groups in the type of audio device used (χ^2 (9, N = 1041) = 16.52, p = .057), with about 26% of participants listening through headphones, 19% through earbuds, 17% through speakers, and 38% through their default computer output.

Analyses

The measures of performance focused on sensitivity (*d'* values) and reaction times. Participants' responses were collapsed into confusion matrices, which were translated into hits and false alarm rates for each emotion. From these rates, we calculated *d prime* (*d'*) values for each participant, which were then used as the dependent variable in linear mixed effects models to examine the recognition of emotional prosody in Experiment 1 and the recognition of emotional semantics in Experiment 2. There were two between-subject fixed factors *musicianship* and *bilingualism*, where participants were either classified as a musician or non-musician and classified as a bilingual or monolingual, respectively. Finally, there was a within-subject fixed factor of *trial type* (incongruent or congruent condition). These models always contained random intercepts by subject, and random intercepts by emotion. Chi-square tests were conducted, after each fixed term was progressively added to the model to evaluate main effects and interactions. Scores were also analyzed on a trial-by-trial basis (using logistic regressions; see Appendix E) and the findings were consistent with the main analysis.

In the aforementioned analysis, the type of incongruency was ignored (i.e., block type was not considered). However, we designed this experiment such that the emotions portrayed by the semantic and prosodic cues were swapped in a particular fashion in each block: valence-based, intensity-based, or both valence and intensity. To examine this factor, *d'* values by block type were also used as the dependent variable in linear mixed effects models to examine the differences in performance by block type and group allocation averaged across the four emotions. For simplicity (i.e., to avoid complex 4-way interactions), we used the interference effect in *d'* units (congruent-incongruent) as the dependent variable, with *musicianship*, *bilingualism*, and *block type* (swap valence, swap intensity, or swap both) as fixed factors. This

model contained random intercepts by subject. See Appendix F for the results and discussion of the block type results.

Finally, the logarithm of the reaction time was used as the dependent variable in linear mixed effects models to examine how quickly participants responded as a function of *trial type*, *musicianship,* and *bilingualism* as fixed factors. This model again contained random intercepts by subject, and random intercepts by emotion. Each model was run using the *lme4* package in *r* (Bates et al., 2014) and were run separately for Experiment 1 and Experiment 2. The *emmeans* package in *r* (Lenth, 2021) was used for all post-hoc comparisons.

Results

Demographics

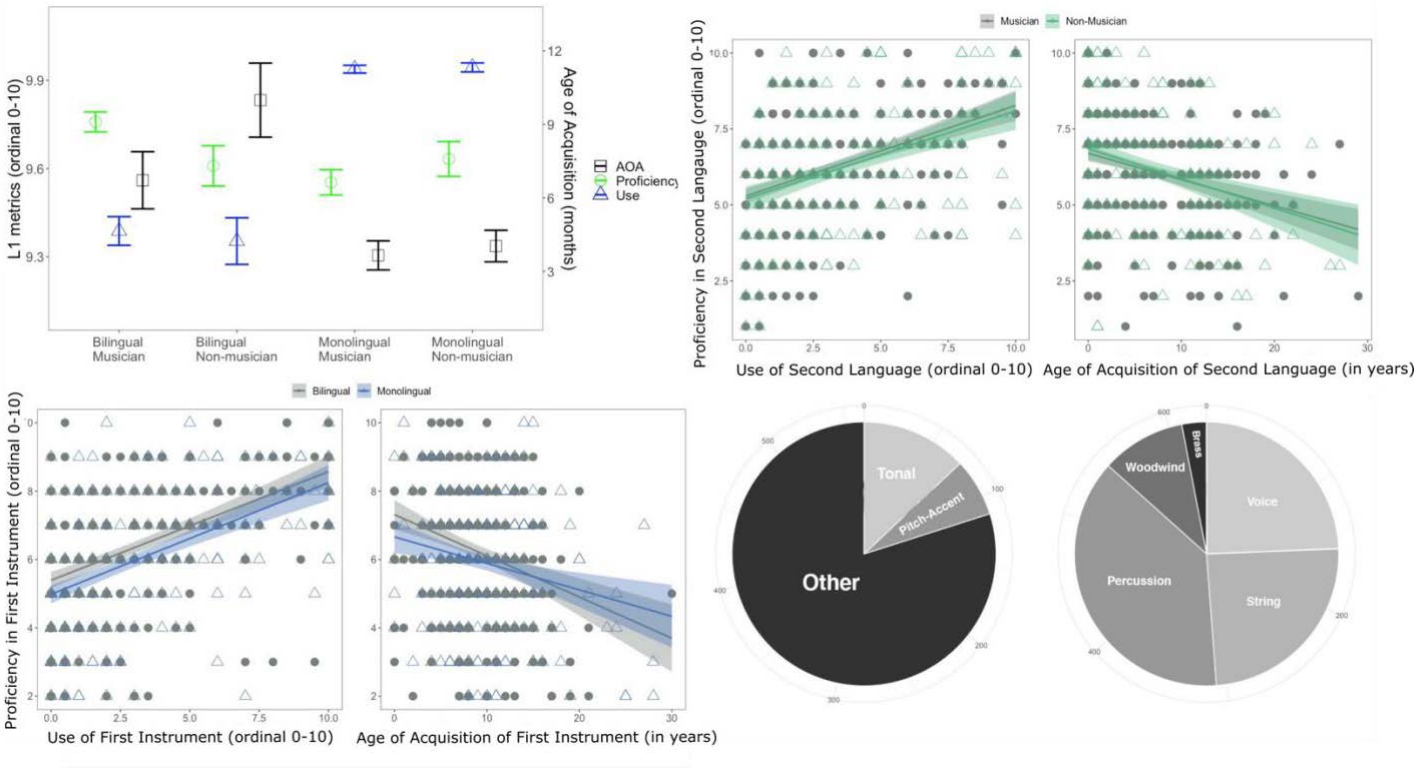
First, we present analysis on the demographic data. They were not the main results of the present study; however, given our large sample size, they are valuable in that they may be generalizable to bilinguals and musicians overall (or least those that can be found online).

Group differences in language and instrument variables. In 2-by-2 ANOVAs (*musicianship* by *bilingualism*) on the combined data from both studies, we analyzed whether the groups differed in three language metrics collected (age of acquisition, proficiency, and use). An interesting observation was that the groups differed in the age of acquisition of English, self-reported as their first language (L1; main effect of bilingualism; F (1, 1037) = 16.65, $p < .001$, $\eta^2 = 0.016$; see Figure 2 top left), where bilinguals learned their first language 0.38 years later than monolinguals ($SE = 0.092$, $p < .001$). This might seem surprising, but perhaps point to a different (more nuanced) understanding of what age of acquisition means for bilinguals than for monolinguals. On the other hand, there was no main effect of *musicianship* (F (1, 1037) = 2.72, $p = .099$, $\eta^2 = 0.003$), and no interaction (F (1, 1037) = 1.71, $p = .192$, $\eta^2 = 0.002$) on age of acquisition of their first language. For proficiency in the first language, there was no main effect of *bilingualism* (F (1, 1037) = 3.35, $p = .067$, $\eta^2 = 0.003$), no main effect of *musicianship* (F (1, 1037) = 0.49, $p = .483$, $\eta^2 < .001$), but surprisingly there was an interaction F (1, 1037) = 5.29, $p = .022$, $\eta^2 = 0.005$). Bilingual musicians rated themselves as more proficient in their L1 than monolingual musicians ($M_{Difference} = 0.21$, $SE = 0.063$, $p = .006$), while no other group comparison reached significance (p ranges from .124-.990). For first language use, there was expectedly a main effect of *bilingualism* (F (1, 1037) = 139.72, $p < .001$, $\eta^2 = 0.12$), where monolinguals used their first language more than bilinguals ($M_{Difference} = 0.57$, $SE = 0.048$, $p <$

.001), but no main effect of *musicianship* (F (1, 1037) = 0.087, p = .77, η^2 = 0.000074), and no interaction (F (1, 1037) = 0.18, p = .676, η^2 < .001). Some of these findings are intuitive (namely monolinguals reporting learning their first language earlier, and using it more, than bilinguals) while others are less so (i.e., bilinguals reporting being more proficient than monolinguals when both groups are also musicians).

Figure 2

Demographic Data



Note. Top left: First language metrics (age of acquisition, proficiency, and use) by group. Top Right: Correlations between proficiency and use, or proficiency and age of acquisition for their second language. Bottom left: Correlations between proficiency and use, or proficiency and age of acquisition of their first instrument. Bottom Right: Pie chart of types of second languages and pie chart of classes of first instruments.

For their second language, bilingual musicians and bilingual non-musicians did not differ in age of acquisition (F (1, 561) = 0.38, p = .54, η^2 = 0.00067) or proficiency (F (1, 561) = 0.46, p = .50, η^2 = 0.00082), but they did differ for use (F (3, 561) = 11.13, p < .001; η^2 = 0.0019) as bilingual non-musicians used their second language more often than bilingual musicians ($M_{Difference}$ = 0.77, SE = 0.23, p < .001). Once again, this finding is far from intuitive, and it may well reflect a property going beyond our online samples. Previous research has shown that musical training positively impacts second language proficiency (see review by Zeromskaite, 2014; Slevc et al., 2006), so it is rather puzzling why it would have opposite effect for use (knowing that proficiency and use are often highly correlated). To our knowledge no observations have been made about the effects of musical training on second language use. However, we could speculate that musicians may be more likely to engage in extracurricular activities, including learning a second language, without necessarily using it consistently.

Next, we analyzed whether the two musician groups (combining both studies) differed in the three metrics collected of their first instrument. Monolingual musicians acquired their first instrument 1.04 years (SE = 0.36) later than bilingual musicians (F (1, 628) = 8.44, p = .004, η^2 = 0.0143). Additionally, bilingual musicians were more proficient in their first instrument than monolingual musicians ($M_{Difference}$ = 0.36, SE = 0.15; F (1, 628) = 6.13, p = .014, η^2 = 0.010). However, they did not differ in the use of their first instrument (F (1, 628) = 0.097, p = .76, η^2 = 0.00016). Thus, bilinguals learned their instrument earlier and were more proficient in it than monolinguals. Once again, we could not identify any similar finding in the literature, but they further support the need to cross-investigate both factors in demographic analyses. We surmise that they might reflect environmental factors (home support, culture, and diligence regarding musicianship) which were not captured here by any other variable.

Group differences on other demographic variables. Other demographic variables, such as age, sex, employment status, and student status, were compared between groups. There were no differences in sex between monolinguals and bilinguals (χ^2 (1, N = 1039) = 2.736, p = .098) nor between musicians and non-musicians (χ^2 (1, N = 1039) = 0.070, p = .791). There was no difference in employment status between musicians and non-musicians (χ^2 (1, N = 1017) = 0.38, p = .54) but more monolinguals (65%) were employed than bilinguals (54%; χ^2 (1, N = 1017) = 12.92, p < .001). A related finding was no difference in student status between musicians and non-musicians (χ^2 (1, N = 1027) = 1.38, p = .24) but more bilinguals (57%) were students than monolinguals (38%; χ^2 (1, N = 1027) = 37.04, p < .001).

Finally, the language groups differed in age (F (1, 1036) = 34.70, p <.001, η^2 = 0.032), such that monolinguals were slightly older than bilinguals ($M_{Difference}$ = 2.02 years, SE = 0.34, p < .001). The music groups did not differ in age (F (1, 1036) = 1.62, p = .203, η^2 = 0.002), but there was an interaction between *musicianship* and *bilingualism* because the difference between monolinguals (older) and bilinguals (younger) was slightly larger in musicians.

Language and instrument variable correlations. All three metrics related to second language (L2; proficiency, use, and age of acquisition) were correlated with each other with an R^2 above .092 (p < .001; See Figure 2 top right). These relationships held within bilingual musicians and within bilingual non-musicians (R^2 above .075, p <.001). In contrast, only some of the first instrument (I1) metrics were correlated with each other. Proficiency was correlated with use and age of acquisition of first instrument (R^2 above .061, p <.001; See Figure 2 bottom left). These relationships held within bilingual musicians and within monolingual musicians (R^2 above .041, p <.001). However, use and age of acquisition of first instrument were not correlated (R^2 =

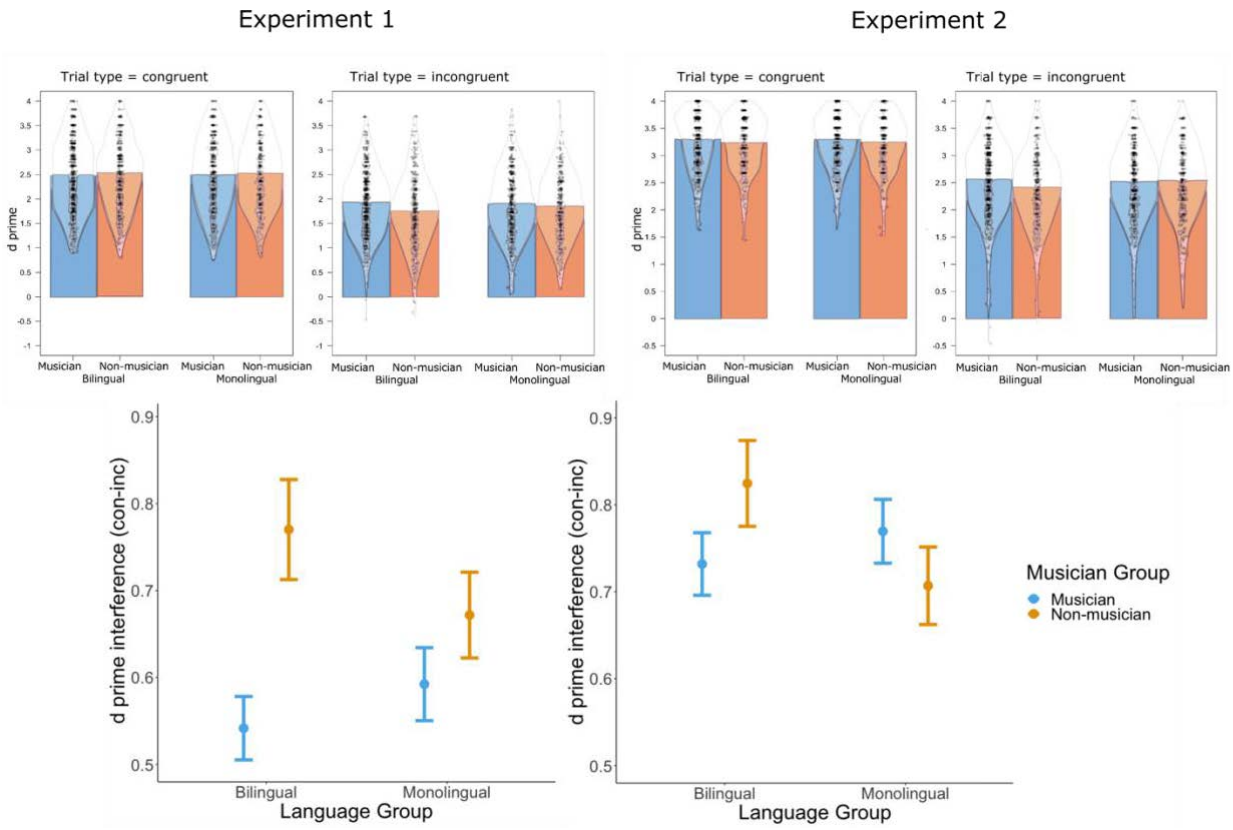
.0044, $p = .100$), and even though this link existed within bilingual musicians, it was weak ($R^2 = .022$, $p = .005$).

Experiment 1 – Performance in emotional prosody

Figure 3 depicts the *d'* results of Experiments 1 and 2. As a reminder, for experiment 1, participants were instructed to respond to the prosody that they heard, while ignoring semantics. There was a main effect of *trial type*, confirming that *d'* decreased for incongruent stimuli compared to congruent stimuli, thus demonstrating that the task worked (See Table 1 for all model results). There was no main effect of *bilingualism*, no main effect of *musicianship*, and no interaction between the two. There was a two-way interaction between *trial type* and *musicianship*, but no interaction between *trial type* and *bilingualism*. These two-way interactions are subsumed in the three-way interaction and will, therefore, be interpreted within the three-way interaction.

Figure 3

d' results



Note. d' data by group and trial type for Experiment 1 (top left panel) and Experiment 2 (top right panel). Interaction between musicianship and bilingualism on the interference effect (congruent minus incongruent trials) expressed in *d'* units in Experiment 1 (bottom left panel) and Experiment 2 (bottom right panel), where lower *d'* units indicate better performance.

Table 1

Model Results of the logistic mixed effects models with d' as the dependent variable

Fixed Effects:	χ^2	DF	*p*
Experiment 1			
Intercept			
Trial Type	1387.5	1	<.001***
Bilingualism	0.49	1	.484
Musicianship	3.14	1	.0766
Trial Type x Bilingualism	0.039	1	.844
Trial Type x Musicianship	26.7	1	<.001***
Bilingualism x Musicianship	1.40	1	.236
Trial Type x Bilingualism x Musicianship	5.68	1	.0172*
Experiment 2			
Intercept			
Trial Type	2216.6	1	<.001***
Bilingualism	0.22	1	.639
Musicianship	4.34	1	.0373*
Trial Type x Bilingualism	0.67	1	.414
Trial Type x Musicianship	0.42	1	.519
Bilingualism x Musicianship	2.28	1	.131
Trial Type x Bilingualism x Musicianship	7.67	1	.00562**

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

There was a statistically significant 3-way interaction between *trial type*, *bilingualism*, and *musicianship*. Dissecting the three-way interaction, there were no differences in performance between any of the groups on the congruent trials (p always above .963); differences were only seen on the incongruent trials. This confirms the idea that the factors of interest (*bilingualism* and *musicianship*) acted upon the resistance to semantic interference (i.e., correctly attending to prosody), but not on basic emotion recognition. The three-way interaction was caused by a differential effect of *musicianship* among monolinguals compared to bilinguals: bilingual musicians were better able to resist the semantic interference than bilingual non-musicians ($p < .001$) whereas *musicianship* had no effect among monolinguals ($p = .948$). On the other hand, there was no effect of *bilingualism* among non-musicians ($p = .338$) or among musicians ($p = .993$), suggesting that, controlling for musicianship, *bilingualism* had no role. To summarize, musicians were (as we hypothesized) good at attending to prosody and could thus resist semantic interference compared to non-musicians, but this effect appeared to be driven by bilinguals.

Experiment 2 – Performance in emotional semantics

For experiment 2, participants responded to the semantics that they heard, while ignoring prosody. There was a main effect of *trial type*, confirming that *d'* decreased for incongruent stimuli compared to congruent stimuli, thus demonstrating that the task worked (See Table 1 for all model results). There was no main effect of *bilingualism*, but the main effect of *musicianship* was statistically significant. Additionally, there was no interaction between *bilingualism* and *musicianship*. There was no two-way interaction between *trial type* and *musicianship*, nor between *trial type* and *bilingualism*.

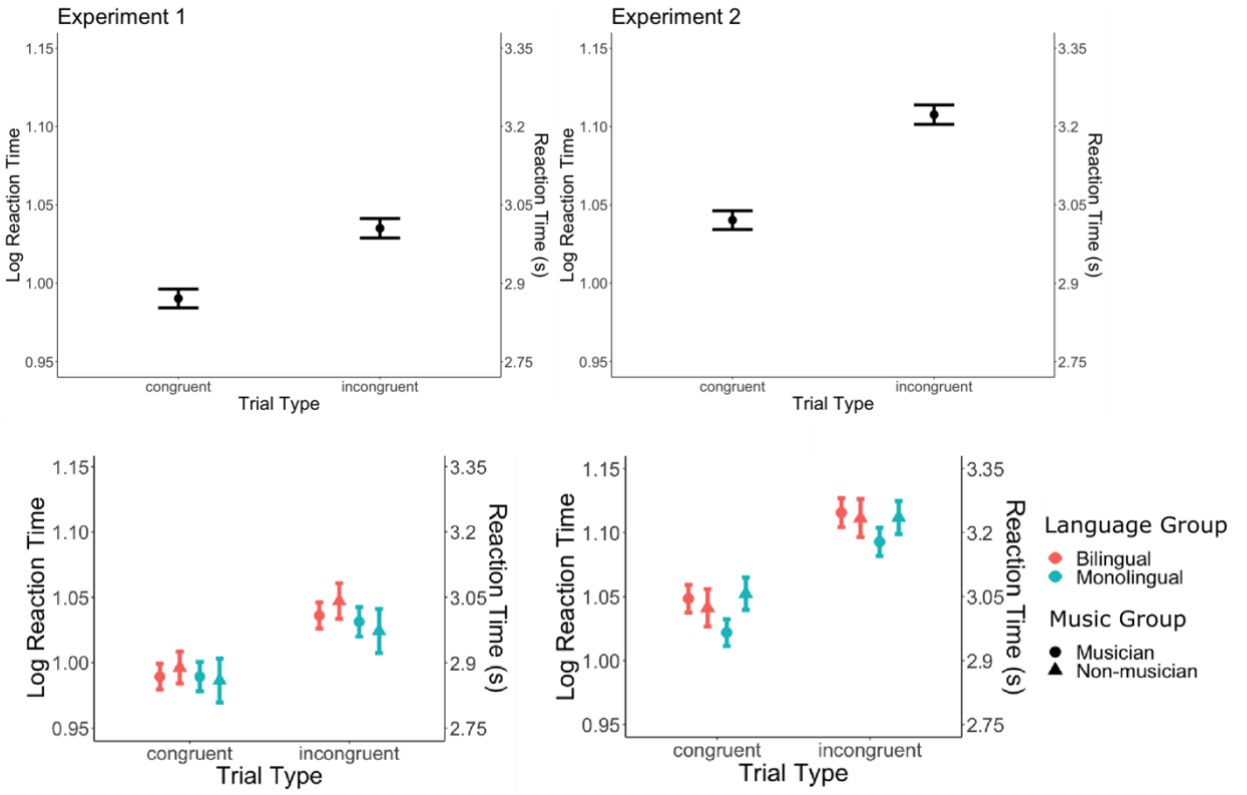
There was a statistically significant 3-way interaction between *trial type*, *bilingualism*, and *musicianship*. Similar to Experiment 1, there were no differences between groups on the congruent trials (p is always above .915), but group differences on the incongruent trials, confirming the idea that the factors of interest (*bilingualism* and *musicianship*) acted upon the resistance to prosodic interference (i.e., correctly attending to semantics). More specifically, there was, a differential effect of *musicianship* among bilinguals and not among monolinguals. That is, bilingual musicians were better able to resist the prosodic interference than bilingual non-musicians (p = .0194), whereas *musicianship* had no role among monolinguals (p > 0.999). On the other hand, there was no effect of *bilingualism* among non-musicians (p = .179) or among musicians (p = .983). To summarize, musicians were also good at attending to semantics and could thus resist prosodic interference compared to non-musicians, but this effect appeared rather exclusive to bilinguals.

Reaction Time

In both experiments, reaction time was delayed in incongruent related to congruent trials (see Table A4.2 in Appendix). Specifically, 3.00 versus 2.87 seconds in Experiment 1 and 3.24 versus 3.00 seconds in Experiment 2 (see Figure 4), but this 130-240-ms delay was not sensitive to group allocation.

Figure 4

Reaction time results by trial type



Note. Reaction time by trial type shown both with log reaction time and reaction time in seconds in Experiment 1 (top left) and Experiment 2 (top right) and by group in Experiment 1 (bottom left) and Experiment 2 (bottom right).

Discussion

The goal of the present study was to examine how bilinguals and musicians would recognize vocal emotions based on prosodic or semantic cues compared to monolinguals and non-musicians. As intended, all groups showed a performance reduction accompanied by a delayed reaction time in incongruent compared to congruent trials. However, consistent with the literature, we found a musician advantage in both experiments, whereby musicians were less prone to interference of the distracting cue (be it prosodic or semantic), although this advantage was only found when also bilingual. As for bilingualism on its own, we failed to observe a prosodic bias like it had been seen in children (i.e., advantage in using prosodic cues and disadvantage in ignoring them). Taken together, these results do not point to differences in cue weighting across these four groups, rather differences in executive functioning among musicians and non-musicians, that are exacerbated when also bilingual.

Regarding the protocol as an emotional Stroop task, it worked as expected and successfully created interference in processing in the incongruent trials. This was demonstrated by a reduction in accuracy of 10-20% in incongruent compared to congruent trials and a delayed (by about 200 ms) reaction time in incongruent trials. These trials are of interest as they require listeners to pit two cues against each other, similar to situations of sarcastic speech encountered in everyday life. Previous studies on vocal emotion recognition have shown similar interference effects, where performance suffered and reaction times were delayed with incongruent versus congruent stimuli (Dupuis & Pichora-Fuller, 2010; Nygaard & Queen, 2008; Wurm et al., 2001). Interestingly, the participants' ability to detect the target emotion differed depending on experience with language and music.

Previous research in children

Based on findings in bilingual children, we had hypothesized that even in adults, experience with multiple languages may influence which domain (semantics or prosody) is recruited in conflicting situations. Previous research had shown that bilingual children begin using prosodic cues earlier than monolingual children (Yow & Markman, 2011) and show a prosodic bias in situations where prosodic and semantic cues to emotions conflict (Champoux-Larsson & Dylman, 2019). The present findings did not replicate the same pattern, suggesting that in young adulthood, bilingualism alone does not lead to greater reliance on prosodic cues in vocal emotion recognition. In other words, we speculate that with greater cognitive maturation and language development, bilinguals can offset their early bias towards prosody and change their listening strategies to make an appropriate (or say a more traditional) use of semantic cues in speech. However, the current results clearly show the importance of controlling for both language and musical experience in these types of designs.

Previous research on the effects of bilingualism and musicianship individually

The present study accounts for both bilingualism and musicianship individually, as well as their combined effects. We added a group of bilingual musicians for a fully orthogonal sampling structure, which had not been done in previous studies on vocal emotion recognition. This was important as our findings generally support a musician advantage effect in vocal emotion recognition that is largely exaggerated among bilinguals. Previous studies looking at the individual effects of bilingualism and musical experience on vocal emotion recognition, had revealed a musician advantage effect in a prosody task (Bialystok & DePape, 2009; Graham & Lakshmanan, 2018) but not in a semantics task (Bialystok & DePape, 2009) relative to monolingual non-musicians. These differences in results may be due to the rudimentary nature of

the semantic material used in these previous studies (i.e., using the words "high" vs "low" and not the use of emotionally loaded sentences), explaining why group differences were not observed in the role of semantics. If this interpretation is correct, it would simply mean that the musician advantage may be found in either domain (prosody or semantics) but would be easier to observe when placing participants in more complex situations which would surely have more relevance to ecological communication settings. In addition, the musician advantage effect that we find among bilinguals is smaller in the semantics task than in the prosody task. This difference is, therefore, going in a direction consistent with the contrast highlighted by Bialystok and DePape (2009). So, it is possible that without accounting for the combined effects of bilingualism and musicianship, group differences may have been missed in their study. As such, the current study aimed to rectify the limitations of previous work in this field by controlling for both musicianship and language experience when evaluating performance in vocal emotion recognition.

Previous research on the combined effects of bilingualism and musicianship

In the few studies that did investigate bilingualism and musicianship simultaneously, findings are rather consistent with the present study. Namely, it is musical training and not bilingualism that is more likely associated with benefits, specifically in task switching and dual-task performance (Moradzadeh et al., 2015). Furthermore, Schroeder and colleagues (2016) disambiguated a "true" interference (by looking at a neutral condition minus incongruent trials) from a facilitation effect (congruent minus neutral trials), and Simon effects (congruent minus incongruent trials, as in the present study) but on a non-linguistic visual-spatial Simon Task in bilingual musicians, bilingual non-musicians, monolingual musicians, and monolingual non-musician young adults. They found that bilingual musicians had a smaller Simon effect

compared to all other groups, consistent with the present findings. However, bilingual musicians, bilingual non-musicians, and monolingual musicians had all smaller interference effects compared to monolingual non-musicians. There were no differences in facilitation effects once confounding variables such as IQ and age were accounted for. Their results suggest an enhanced ability to suppress interfering cues shared among bilinguals, musicians, and bilingual musicians, but they propose that the Simon effect (congruent minus incongruent) is a more convoluted metric encompassing both facilitation and interference effects making it harder to interpret. In the present study, we did not include semantically neutral sentences or sentences spoken with a neutral prosody, so we are unable to disentangle these different effects. It would be interesting to see whether the unique advantage of the combined musician and bilingual advantage taps more into the facilitation than the interference effect. It is important to note though that these studies did not focus on vocal emotion recognition, but rather executive functioning among these groups. However, based on the results of these studies, we could speculate that the present results may be due to better executive functioning among bilingual musicians.

The potential role of executive functions

While we see differences in performance between groups, the current results do not reflect group difference in cue weighting, rather differences in executive functioning. A difference in cue weighting would have resulted in bilingual musicians outperforming the other groups on one task and performing worse on the other task. For example, if they weighted prosody more heavily in such situations, then their performance would have been best when asked to use prosody to detect vocal emotions because they could easily ignore anything unrelated to prosody such as the semantic meaning of the sentence. Additionally, their performance would have been worse when prosody is used as a distracting cue because they

would still rely on these salient prosodic cues that do not necessarily help in deciphering the semantic content of the sentences. Rather, the effects seen here are more in line with an advantage in executive functioning when making judgements about vocal emotions as bilingual musicians were able to use the correct cue regardless of the task and did not favour one cue over another. This may reflect better response inhibition, cognitive control, or cognitive flexibility that have been previously shown to be advantages associated with being bilingual (Bialystok & Craik, 2010; Costa et al., 2008; Krizman et al., 2012; Wiseheart et al., 2016) or a musician (Bialystok & DePape, 2009; Strong & Mast, 2019; Zuk et al., 2014). However, previous research has been somewhat inconclusive on whether bilingualism and musicianship have benefits that extend beyond the realm of language and music, respectively, into other executive functions. Neither bilingualism nor musical experience (D'Souza et al., 2018; Lehtonen et al., 2018) has been consistently shown to be beneficial for executive functioning in adults. Based on the current results, we speculate that this might be partly because the other factor (bilingualism or musicianship) was not controlled for. Given that musicians and bilinguals separately have been shown to have better executive functioning skills than monolinguals and non-musicians, it makes sense that the interaction between these two skills may provide additional benefits to their executive functioning in certain situations. Thus, the effects may be additive. However, executive functioning was not specifically measured in the present study, so this interpretation is only speculative. An alternative interpretation is that the musician advantage in executive functioning transfer to the language domain more easily in bilinguals (although our data suggests exclusively).

Transfer effects

Overlap between music and language has been noted in acoustic properties (Besson et al., 2011; Hausen et al., 2013; Peretz et al., 2015) and the communication of emotions (Paquette et al., 2018), and there is also overlap in brain regions that process language and music (Fedorenko et al., 2009; Levitin, 2003; Maess et al., 2001; Patel & Iversen, 2007). So, one could *speculate* that the benefits of experience in one would transfer to the other. Cross-domain transfer effects have been reported from music to language (Besson et al., 2011; Bidelman et al., 2011; Moreno, 2009; Patel, 2011) and language to music (Deroche, et al., 2019b; Krishnan & Gandour, 2009), but the causality of music training – as opposed to inherent perceptual or cognitive aptitudes – is highly debated (Mankel & Bidelman, 2018; Penhune, 2019, and also McKay, 2021 for a review of this question within the hearing-impaired world). Patel (2011) argues that musical training leads to neuroplasticity in brain networks responsible for speech processing resulting in the better encoding of speech. Patel (2011) hypothesises that this occurs only under several specific conditions. Specifically, there are overlapping brain networks essential for both music and speech perception, where music training must allow for precise processing and discrimination of auditory information, and music training must provide emotional rewards, be associated with focused attention, and repeated frequently, which are all also common to language learning. This could explain why musical training may be beneficial to the acquisition of a second language (Chobert & Besson, 2013) and why individuals who receive musical training and learn multiple languages perform better than those without these experiences in auditory tasks that include linguistic components, such as in the present study. In sum, our findings may demonstrate that a dual training experience of musicianship and bilingualism can enhance the processing of speech and result in the enhanced ability to detect vocal emotions. However, once again, this is only

speculative and further research needs to be done to better understand why such a transfer from music training to the language domain would not occur (or not as easily) in monolinguals.

Emotional intelligence

There are other potential variables that may account for, or mediate, some of the current results: emotional intelligence. Not surprisingly, higher emotional intelligence has been linked to better recognition of emotions. Alqarni & Dewaele (2020) found that participants who have higher trait emotional intelligence (i.e., the construct that relies more on perception of one's own emotions) were better at perceiving and interpreting emotions from audio-visual recordings. Crucially, they found that bilinguals had higher trait emotional intelligence than monolinguals. However, the effect sizes for each of these results was small (Cohens' d of about 0.30). Furthermore, Trimmer & Cuddy (2008) found that emotional prosody discrimination was related to emotional intelligence scores but not musical training (contradicting other reports – see introduction). Also, musical training has not been linked to higher emotional intelligence (Glenn Schellenberg, 2011; Trimmer & Cuddy, 2008) Additionally, to our knowledge, there are no studies on emotional intelligence in individuals who are both a musician and bilingual. Thus, if differences in emotional intelligence were a concern for this study, one might have expected it to enhance performance among bilinguals but not musicians, which was not what we observed. Also, we would expect this variable to affect performance on congruent trials as well, whereas group differences were exclusive to incongruent trials here. For these reasons, we suspect that it is unlikely to be a confound here.

Socioeconomic status

We might equally wonder whether socioeconomic status (SES) could partially explain the results as SES is known to affect research on bilingualism particularly. Some studies have found

SES to be a potential confound when assessing a bilingual advantage in the Simon task (Morton & Harper, 2007), while others have controlled for SES and continued to find a bilingual advantage in inhibitory control (Emmorey et al., 2008; Filippi et al., 2022; Nair et al., 2017). For example, Naeem and colleagues (2018) found that being a bilingual compared to a monolingual had no effect on performance (in the Simon task) among individuals with high SES, but bilinguals outperformed monolinguals (on both congruent and incongruent trials) among individuals with low SES. Additionally, musicians are likely to have higher SES than non-musicians (Swaminathan & Schellenberg, 2018). The present results suggest that bilingualism may be beneficial in tasks of vocal emotion recognition among musicians, who may be of higher SES, but not among non-musicians, who may be of lower SES. These results are not in line with those from Naeem and colleagues (2018) and if differences in emotional intelligence were a concern for this study, one might have expected it to enhance performance among those of low SES (i.e., potentially monolingual and bilingual non-musicians) presumably not just on incongruent trials but also on congruent trials, yet this was not observed. Thus, the present findings do not align easily with an interpretation based on SES differences, though more research should be done to account for this variable.

Limitations

Some limitations to the current study should be acknowledged. Given the nature of online studies: 1) there was generally a lack of control over the audio/stimulus delivery, 2) the quality of bilinguals and musicians and the reliability of their self-reports could be questioned, and 3) the generalizability of online findings should be verified. Here we respond to each of these concerns. In response to the first concern, we asked participants to rate the quality of their audio and did not find any group difference in this regard. Also, performance on congruent trials (including

reaction times) was overall decent. Thus, this concern seems relatively negligible. Second, we relied on participants self-reports to allocate them as either a bilingual or monolingual and musician or non-musician. Tomoschuk and colleagues (2019) found that objective measures of language proficiency (e.g., picture naming or proficiency interviews) are better than self-rating of language proficiency. However, self-report measures have also been shown to be as reliable as objective measures (Lim et al., 2008; Shameem, 1998). Finally, our analytical approach did not rely on precise estimates of age of acquisition, proficiency, and use, since we followed a categorical approach for the groups' definition. In other words, inaccuracies in self-reports would have had little consequence (except for Appendix E where continuous variables were used).

Lastly, the validity of online studies has been investigated in recent years. As outlined in the review by Chandler and Shapiro (2016), there are notable differences between the general population and online convenience samples. Several issues are relevant here such as the observation that online samples tend to be younger than the general population and some samples may be either over- or under-represented (i.e., more participants tend to be Caucasian and Asian and are more educated). In this study specifically, we found that bilinguals were younger and more of them were students and unemployed compared to monolinguals. Of note, Eyal and colleagues (2021) found that the online platform Prolific (the one used in the current study) provided higher quality data in terms of comprehension, attention, and dishonesty, than MTurk (the online platform used in Chandler & Shapiro's (2016) study). We also see certain advantages to conducting the present study online, namely, having a very large sample size, accurately reflecting the heterogeneity of musicians and bilinguals, and being able to easily recruit English-speaking monolinguals (a fairly difficult thing to do in person in Québec). Thus,

we believe that the benefits outweigh the potential disadvantages of using online platforms for the present study.

Conclusions and future directions

In conclusion, musical training appears to benefit the recognition of vocal emotions, either when semantic cues or when prosodic cues are providing conflicting information, but only among bilinguals. Thus, we do not see a difference in cue weighting when identifying vocal emotions among the groups as previously seen in bilingual and monolingual children. Nor do we see a prosodic bias among bilinguals. Instead, differences may be due to enhanced executive functioning in bilingual musicians that results in better performance both when asked to attend to prosodic or semantic cues. We speculate that this is because the enhanced executive functions of musicians are somehow strengthened, or transfer more easily to the language domain, in bilinguals than they do in monolinguals. This may be due to bilinguals being more flexible in their listening strategies or still figuring out the different ways to resolve conflictual situations of communicative intent.

This research has implications for educational and linguistic fields, but also for clinical areas such as in individuals growing up with degraded hearing. For example, school-aged children with cochlear implants or with hearing aids perform worse than their normal hearing counterparts on tasks of emotional prosody (Barrett et al., 2020; Chatterjee et al., 2015; Lin et al., 2022; Most & Peled, 2007). Deficits in these tasks are often linked to poor pitch perception, but it may well be that these children also develop alternative strategies to recognize emotions in sentences. Some of these strategies could involve a stronger reliance on semantics and weaker reliance on prosody, or a different weighting among prosodic cues (e.g., using temporal and intensity cues more than pitch cues). Thus, understanding the particular circumstances or

participant profiles that result in enhanced vocal emotion recognition may be beneficial to understanding how to improve these abilities in hearing-impaired and cochlear implanted children and adults. Experiments are under way to run this exact paradigm in cochlear implant users.

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A Systematic Review and Meta-Analysis of the Cognitive Correlates of Bilingualism. *Review of Educational Research*, *80*(2), 207–245. https://doi.org/10.3102/0034654310368803

Aguert, M., Laval, V., Lacroix, A., Gil, S., & Bigot, L. L. (2013). Inferring emotions from speech prosody: Not so easy at age five. *PLoS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0083657

Aguert, M., Laval, V., Le Bigot, L., & Bernicot, J. (2010). *Understanding Expressive Speech Acts: The Role of Prosody and Situational Context in French-Speaking 5-to 9-Year-Olds* (pp. 1629–1641).

Alqarni, N., & Dewaele, J.-M. (2020). A bilingual emotional advantage? An investigation into the effects of psychological factors in emotion perception in Arabic and in English of Arabic-English bilinguals and Arabic/English monolinguals. *International Journal of Bilingualism*, *24*(2), 141–158. https://doi.org/10.1177/1367006918813597

Bailey, C., Venta, A., & Langley, H. (2020). The bilingual [dis]advantage. *Language and Cognition*, *12*(2), 225–281. https://doi.org/10.1017/langcog.2019.43

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of Training between Music and Speech: Common Processing, Attention, and Memory. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00094

Bialystok, E. (1999). Cognitive Complexity and Attentional Control in the Bilingual Mind. *Child Development*, *70*(3), 636–644. https://doi.org/10.1111/1467-8624.00046

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin*, *143*(3), 233–262. https://doi.org/10.1037/bul0000099

Bialystok, E., & Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science*, *19*(1), 19–23. https://doi.org/10.1177/0963721409358571

Bialystok, E., & DePape, A. M. (2009). Musical Expertise, Bilingualism, and Executive Functioning. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 565–574. https://doi.org/10.1037/a0012735

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and Cognition*, *77*(1), 1–10. https://doi.org/10.1016/j.bandc.2011.07.006

Botinis, A., Granström, B., & Möbius, B. (2001). Developments and paradigms in intonation research. *Speech Communication*, *33*(4), 263–296. https://doi.org/10.1016/S0167-6393(00)00060-1

Champoux-Larsson, M. F., & Dylman, A. S. (2019). A prosodic bias, not an advantage, in bilinguals' interpretation of emotional prosody. *Bilingualism*, *22*(2), 416–424. https://doi.org/10.1017/S1366728918000640

Champoux-Larsson, M.-F., & Dylman, A. S. (2021). Bilinguals' inference of emotions in ambiguous speech. *International Journal of Bilingualism*, *25*(5), 1297–1310. https://doi.org/10.1177/13670069211018847

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology*, *12*(1), 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chobert, J., & Besson, M. (2013). Musical expertise and second language learning. *Brain Sciences*, *3*(2), 923–940. https://doi.org/10.3390/brainsci3020923

Christoffels, I. K., Kroll, J. F., & Bajo, M. T. (2013). Introduction to Bilingualism and Cognitive Control. *Frontiers in Psychology*, *4*, 1664–1078. https://doi.org/10.3389/fpsyg.2013.00199

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*(1), 59–86. https://doi.org/10.1016/j.cognition.2006.12.013

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141–201. https://doi.org/0.1177/002383099704000203

Daly, H. R., & Hall, M. D. (2018). Not all musicians are created equal: Statistical concerns regarding the categorization of participants. *Psychomusicology: Music, Mind, and Brain*, *28*(2), 117–126. https://doi.org/10.1037/pmu0000213

de Bruin, A. (2019). Not All Bilinguals Are the Same: A Call for More Detailed Assessments and Descriptions of Bilingual Experiences. *Behavioral Sciences*, *9*(3), 33. https://doi.org/10.3390/bs9030033

Deroche, M. L. D., Felezeu, M., Paquette, S., Zeitouni, A., & Lehmann, A. (2019). Neurophysiological Differences in Emotional Processing by Cochlear Implant Users, Extending Beyond the Realm of Speech: *Ear and Hearing*, *40*(5), 1197–1209. https://doi.org/10.1097/AUD.0000000000000701

Deroche, M. L. D., Lu, H.-P., Kulkarni, A. M., Caldwell, M., Barrett, K. C., Peng, S.-C., Limb, C. J., Lin, Y.-S., & Chatterjee, M. (2019). A tonal-language benefit for pitch in normally-

hearing and cochlear-implanted children. *Scientific Reports*, *9*(1), 109. https://doi.org/10.1038/s41598-018-36393-1

Donnelly, S., Brooks, P. J., & Homer, B. D. (2019). Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychonomic Bulletin & Review*, *26*(4), 1122–1147. https://doi.org/10.3758/s13423-019-01567-z

D'Souza, A. A., Moradzadeh, L., & Wiseheart, M. (2018). Musical training, bilingualism, and executive function: Working memory and inhibitory control. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0095-6

Dupuis, K., & Pichora-Fuller, M. K. (2010). Use of affective prosody by young and older adults. *Psychology and Aging*, *25*(1), 16–29. https://doi.org/10.1037/a0018777

Emmorey, K., Luk, G., Pyers, J. E., & Bialystok, E. (2008). The Source of Enhanced Cognitive Control in Bilinguals: Evidence From Bimodal Bilinguals. *Psychological Science*, *19*(12), 1201–1206. https://doi.org/10.1111/j.1467-9280.2008.02224.x

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition*, *37*(1), 1–9. https://doi.org/10.3758/MC.37.1.1

Filippi, R., Ceccolini, A., Booth, E., Shen, C., Thomas, M. S. C., Toledano, M. B., & Dumontheil, I. (2022). Modulatory effects of SES and multilinguistic experience on cognitive development: A longitudinal data analysis of multilingual and monolingual

adolescents from the SCAMP cohort. *International Journal of Bilingual Education and Bilingualism*, 1–18. https://doi.org/10.1080/13670050.2022.2064191

Friend, M. (2000). Developmental changes in sensitivity to vocal paralanguage. In *Developmental Science* (Vol. 3, pp. 148–162). https://doi.org/10.1111/1467-7687.00108

Friend, M. (2001). The transition from affective to linguistic meaning. *First Language*, *21*(63), 219–243. https://doi.org/10.1177/014272370102106302

Friend, M., & Bryant, J. B. (2017). A Developmental Lexical Bias in the Interpretation of Discrepant Messages. *Merrill Palmer Q*, 27.

Graham, R. E., & Lakshmanan, U. (2018). *Tunes and Tones: Music, Language, and Inhibitory Control*. *18*, 104–123. https://doi.org/10.1163/15685373-12340022

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience*, *28*, 219–236. https://doi.org/10.3233/RNN-2010-0499

Hansen, M., Wallentin, M., & Vuust, P. (2013). Working memory and musical competence of musicians and non-musicians. *Psychology of Music*, *41*(6), 779–793. https://doi.org/10.1177/0305735612452186

Hausen, M., Torppa, R., Salmela, V. R., Vainio, M., & Särkämö, T. (2013). Music and speech prosody: A common rhythm. *Frontiers in Psychology*, *4*(SEP). https://doi.org/10.3389/fpsyg.2013.00566

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences*, *106*(16), 6556–6560. https://doi.org/10.1073/pnas.0811323106

Krishnan, A., & Gandour, J. T. (2009). The role of the auditory brainstem in processing linguistically-relevant pitch patterns. *Brain and Language*, *110*(3), 135–148. https://doi.org/10.1016/j.bandl.2009.03.005

Krizman, J., Marian, V., Shook, A., Skoe, E., & Kraus, N. (2012). Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. *Proceedings of the National Academy of Sciences*, *109*(20), 7877–7881. https://doi.org/10.1073/pnas.1201575109

Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, *25*(5), 497–514. https://doi.org/10.1080/20445911.2013.799170

Levitin, D. (2003). Musical structure is processed in "language" areas of the brain: A possible role for Brodmann Area 47 in temporal coherence. *NeuroImage*, *20*(4), 2142–2152. https://doi.org/10.1016/j.neuroimage.2003.08.016

Lim, V. P. C., Liow, S. J. R., Lincoln, M., Chan, Y. H., & Onslow, M. (2008). Determining language dominance in English–Mandarin bilinguals: Development of a self-report classification tool for clinical use. *Applied Psycholinguistics*, *29*(3), 389–412. https://doi.org/10.1017/S0142716408080181

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*(5), 1021–1031. https://doi.org/10.1037/a0024521

Luk, G. (2015). Who are the bilinguals (and monolinguals)? *Bilingualism: Language and Cognition*, *18*(1), 35–36. https://doi.org/10.1017/S1366728914000625

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: An MEG study. *Nature Neuroscience*, *4*(5), 540–545. https://doi.org/10.1038/87502

Mankel, K., & Bidelman, G. M. (2018). Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proceedings of the National Academy of Sciences*, *115*(51), 13129–13134. https://doi.org/10.1073/pnas.1811793115

Mastropieri, D., & Turkewitz, G. (1999). Prenatal experience and neonatal responsiveness to vocal expressions of emotion. *Developmental Psychobiology*, *35*(3), 204–214. https://doi.org/10.1002/(SICI)1098-2302(199911)35:3<204::AID-DEV5>3.0.CO;2-V

Moradzadeh, L., Blumenthal, G., & Wiseheart, M. (2015). Musical Training, Bilingualism, and Executive Function: A Closer Look at Task Switching and Dual-Task Performance. *Cognitive Science*, *39*(5), 992–1020. https://doi.org/10.1111/cogs.12183

Moreno, S. (2009). Can Music Influence Language and Cognition? *Contemporary Music Review*, *28*(3), 329–345. https://doi.org/10.1080/07494460903404410

Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental Science*, *10*(6), 719–726. https://doi.org/10.1111/j.1467-7687.2007.00623.x

Morton, J. B., & Trehub, S. E. (2001). Children's Understanding of Emotion in Speech. In *Child Development* (Vol. 72, pp. 834–843).

Most, T., & Peled, M. (2007). Perception of suprasegmental features of speech by children with cochlear implants and children with hearing aids. *Journal of Deaf Studies and Deaf Education*, *12*, 350–361. https://doi.org/10.1093/deafed/enm012

Naeem, K., Filippi, R., Periche-Tomas, E., Papageorgiou, A., & Bright, P. (2018). The Importance of Socioeconomic Status as a Modulator of the Bilingual Advantage in Cognitive Ability. *Frontiers in Psychology*, *9*, 1818. https://doi.org/10.3389/fpsyg.2018.01818

Nair, V. K., Biedermann, B., & Nickels, L. (2017). Effect of socio-economic status on cognitive control in non-literate bilingual speakers. *Bilingualism: Language and Cognition*, *20*(5), 999–1009. https://doi.org/10.1017/S1366728916000778

Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: Linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(4), 1017–1030. https://doi.org/10.1037/0096-1523.34.4.1017

Paap, K. (2019). The Bilingual Advantage Debate: Quantity and Quality of the Evidence. In J. W. Schwieter & M. Paradis (Eds.), *The Handbook of the Neuroscience of Multilingualism* (1st ed., pp. 701–735). Wiley. https://doi.org/10.1002/9781119387725.ch34

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002

Paquette, S., Takerkart, S., Saget, S., Peretz, I., & Belin, P. (2018). Cross-classification of musical and vocal emotions in the auditory cortex: Cross-classification of musical and

vocal emotions. *Annals of the New York Academy of Sciences*, *1423*(1), 329–337. https://doi.org/10.1111/nyas.13666

Patel, A. D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Frontiers in Psychology*, *2*, 1–14. https://doi.org/10.3389/fpsyg.2011.00142

Patel, A. D., & Iversen, J. R. (2007). The linguistic benefits of musical abilities. *Trends in Cognitive Sciences*, *11*(9), 369–372. https://doi.org/10.1016/j.tics.2007.08.003

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion*, *28*(2), 230–244. https://doi.org/10.1080/02699931.2013.812033

Penhune, V. B. (2019). Musical Expertise and Brain Structure: The Causes and Consequences of Training. In M. H. Thaut & D. A. Hodges (Eds.), *The Oxford Handbook of Music and the Brain* (pp. 417–438). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198804123.013.17

Peretz, I., Vuvan, D., Lagrois, M.-É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140090. https://doi.org/10.1098/rstb.2014.0090

Preti, E., Suttora, C., & Richetin, J. (2016). Can you hear what I feel? A validated prosodic set of angry, happy, and neutral Italian pseudowords. *Behavior Research Methods*, *48*(1), 259–271. https://doi.org/10.3758/s13428-015-0570-7

Raji, S., & de Melo, G. (2020). What Sparks Joy: The AffectVec Emotion Database. *Proceedings of The Web Conference 2020*, 2991–2997. https://doi.org/10.1145/3366423.3380068

Sares, A. G., Foster, N. E. V., Allen, K., & Hyde, K. L. (2018). Pitch and Time Processing in Speech and Tones: The Effects of Musical Training and Attention. *Journal of Speech, Language, and Hearing Research*, *61*(3), 496–509. https://doi.org/10.1044/2017_JSLHR-S-17-0207

Sauter, D. A., Panattoni, C., & Happé, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology*, *31*(1), 97–113. https://doi.org/10.1111/j.2044-835X.2012.02081.x

Schellenberg, E. G. (2011). Examining the association between music lessons and intelligence: Music lessons and intelligence. *British Journal of Psychology*, *102*(3), 283–302. https://doi.org/10.1111/j.2044-8295.2010.02000.x

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92. https://doi.org/10.1177/0022022101032001009

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language: Music and prosody: An ERP study. *Psychophysiology*, *41*(3), 341–349. https://doi.org/10.1111/1469-8986.00172.x

Schroeder, S. R., Marian, V., Shook, A., & Bartolotti, J. (2016). Bilingualism and Musicianship Enhance Cognitive Control. *Neural Plasticity*, *2016*, 1–11. https://doi.org/10.1155/2016/4058620

Shameem, N. (1998). *Validating self-reported language proficiency by testing performance in an immigrant community: The Wellington Indo-Fijians*. *15*(1), 86–108. https://doi.org/10.1177/026553229801500104

Shenker, J. J., Steele, C. J., Chakravarty, M. M., Zatorre, R. J., & Penhune, V. B. (2022). Early musical training shapes cortico-cerebellar structural covariation. *Brain Structure and Function*, *227*(1), 407–419. https://doi.org/10.1007/s00429-021-02409-2

Shook, A., Marian, V., Bartolotti, J., & Schroeder, S. R. (2013). Musical Experience Influences Statistical Learning of a Novel Language. *The American Journal of Psychology*, *126*(1), 95–104. https://doi.org/10.5406/amerjpsyc.126.1.0095

Strong, J. V., & Mast, B. T. (2019). The cognitive functioning of older adult instrumental musicians and non-musicians. *Aging, Neuropsychology, and Cognition*, *26*(3), 367–386. https://doi.org/10.1080/13825585.2018.1448356

Swaminathan, S., & Schellenberg, E. G. (2018). Musical Competence is Predicted by Music Training, Cognitive Abilities, and Personality. *Scientific Reports*, *8*(1), 9223. https://doi.org/10.1038/s41598-018-27571-2

Tierney, A., & Kraus, N. (2013). Music Training for the Development of Reading Skills. In *Progress in Brain Research* (Vol. 207, pp. 209–241). Elsevier. https://doi.org/10.1016/B978-0-444-63327-9.00008-4

Tierney, A., Krizman, J., Skoe, E., Johnston, K., & Kraus, N. (2013). High school music classes enhance the neural processing of speech. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00855

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, *22*(3), 516–536. https://doi.org/10.1017/S1366728918000421

Trimmer, C. G., & Cuddy, L. L. (2008). Emotional Intelligence, Not Music Training, Predicts Recognition of Emotional Speech Prosody. *Emotion*, *8*(6), 838–849. https://doi.org/10.1037/a0014080

Vaquero, L., Rousseau, P.-N., Vozian, D., Klein, D., & Penhune, V. (2020). What you learn & when you learn it: Impact of early bilingual & music experience on the structural characteristics of auditory-motor pathways. *NeuroImage*, *213*, 116689. https://doi.org/10.1016/j.neuroimage.2020.116689

Wiseheart, M., Viswanathan, M., & Bialystok, E. (2016). Flexibility in task switching by monolinguals and bilinguals. *Bilingualism: Language and Cognition*, *19*(1), 141–146. https://doi.org/10.1017/S1366728914000273

Wurm, L. H., Vakoch, D. A., Strasser, M. R., Calin-Jageman, R., & Ross, S. E. (2001). Speech perception and vocal expression of emotion. *Cognition and Emotion*, *15*(6), 831–852. https://doi.org/10.1080/02699930143000086

Yow, W. Q., & Markman, E. M. (2011). Bilingualism and children's use of paralinguistic cues to interpret emotion in speech. *Bilingualism: Language and Cognition*, *14*(4), 562–569. https://doi.org/10.1017/s1366728910000404

Yow, W. Q., & Markman, E. M. (2015). A bilingual advantage in how children integrate multiple cues to understand a speaker's referential intent. *Bilingualism: Language and Cognition*, *18*(3), 391–399. https://doi.org/10.1017/S1366728914000133

Yow, W. Q., & Markman, E. M. (2016). Children Increase Their Sensitivity to a Speaker's Nonlinguistic Cues Following a Communicative Breakdown. *Child Development*, *87*(2), 385–394. https://doi.org/10.1111/cdev.12479

Zuk, J., Benjamin, C., Kenyon, A., & Gaab, N. (2014). Behavioral and Neural Correlates of Executive Functioning in Musicians and Non-Musicians. *PLoS ONE*, *9*(6), e99868. https://doi.org/10.1371/journal.pone.0099868

Appendix A: Transcripts of all sentences

Angry	He stole my parking spot	She pulled my hair out	My brother left a mess
	My brother will not share	He is constantly late	My boyfriend bothers me
	My parents grounded me	My parents always argue	Stop being so shallow
	I hate people who steal	School is frustrating	You burnt my food
	My mom is shouting at me	They are upset with me	I cannot stand my job
	My brother hit me	Those kids are rude	I spilled water on my desk
	My neighbour smashed my car	She was bothered by that	My dog chewed up my doll
	She tears everyone down	The kids fight all the time	He cracked my tooth in half
	I can never get it right	The thief destroyed our house	My neighbour cursed me
	You ruined my night	I'm often stuck in traffic	My sister gets on my nerves
	Stop being nosy	She punches her brother	My candy is missing
	My coach yells at our team	She never follows the rules	He's a horrible boss
Calm	The flowers are blooming	She found peace in her life	He offered moral support
	The beach is breezy	He is deep in thought	Let go of your fears
	Baths are relaxing	The child spoke softly	He watched the night sky
	The sky is baby blue	They took a walk in nature	His tension melted away
	She sat without making noise	Her environment is pleasant	The stars are nice and bright
	You will get through this	Open your heart	She felt a gentle breeze
	Focus on your breathing	She maintained her focus	The snow lightly fell
	Let go of the bad	The sun rises slowly	All her concerns disappeared
	The baby is sound asleep	Forgetting all your worries	The path became clear
	The waves are soothing	They massaged my shoulders	There was a gentle sound of a stream
	I am centered	Quiet your mind	He rocked softly on the hammock
	Meditation reduces stress	Seek new experiences	The birds sang all around her
Happy	Today is my birthday	You achieved your dream job	Playing sports is fun
	Let's go to Disneyland	She won her soccer game	You have the sweetest heart

	That music is awesome	My dad bought me a new bike	The kids are having fun
	The sun is shining bright	She plays with her best friend	She dances all night long
	He adored that movie	I am glad to see you	Let's go on vacation
	That was a great experience	School is so much fun	Her exams are all done
	I enjoy reading books	You have the cutest dog	He jumps joyfully
	I love my family	I love walking my pet	This is a special day
	My sister got married	Tomorrow is pay day	I love to make pizza
	I scored a brilliant goal	She went to a party	You did a great job
	I'm having a baby	I am so proud of you	The doctors cured her mom
	I just won a contest	I accomplish many things	My husband bought me a rose
Sad	I failed my math test	She was rushed to the hospital	She cried herself to sleep
	He misses his parents	Everyone ignores me	The car killed my cat
	My sister is crying	They cried at the funeral	My wife wants a divorce
	His grandmother died	She lost all her money	I regret my behaviour
	He lost his job last night	They received bad news	Tears roll down her cheeks
	I wrecked my dad’s car	Our trip was cancelled	It hurts to be left behind
	My vacation is over	It never stops raining	Everything is going wrong
	What a gloomy day	He is getting bullied	My life is a mess
	She feels very lonely	She lost her first baby	He left her at the alter
	He hasn't slept well all week	She is disappointed	The country is starving
	We are all stuck inside	We have no more food	They never came back from war
	You are always alone	My girlfriend broke my heart	Let's remember this loss

Appendix B: Confirming adequate semantics

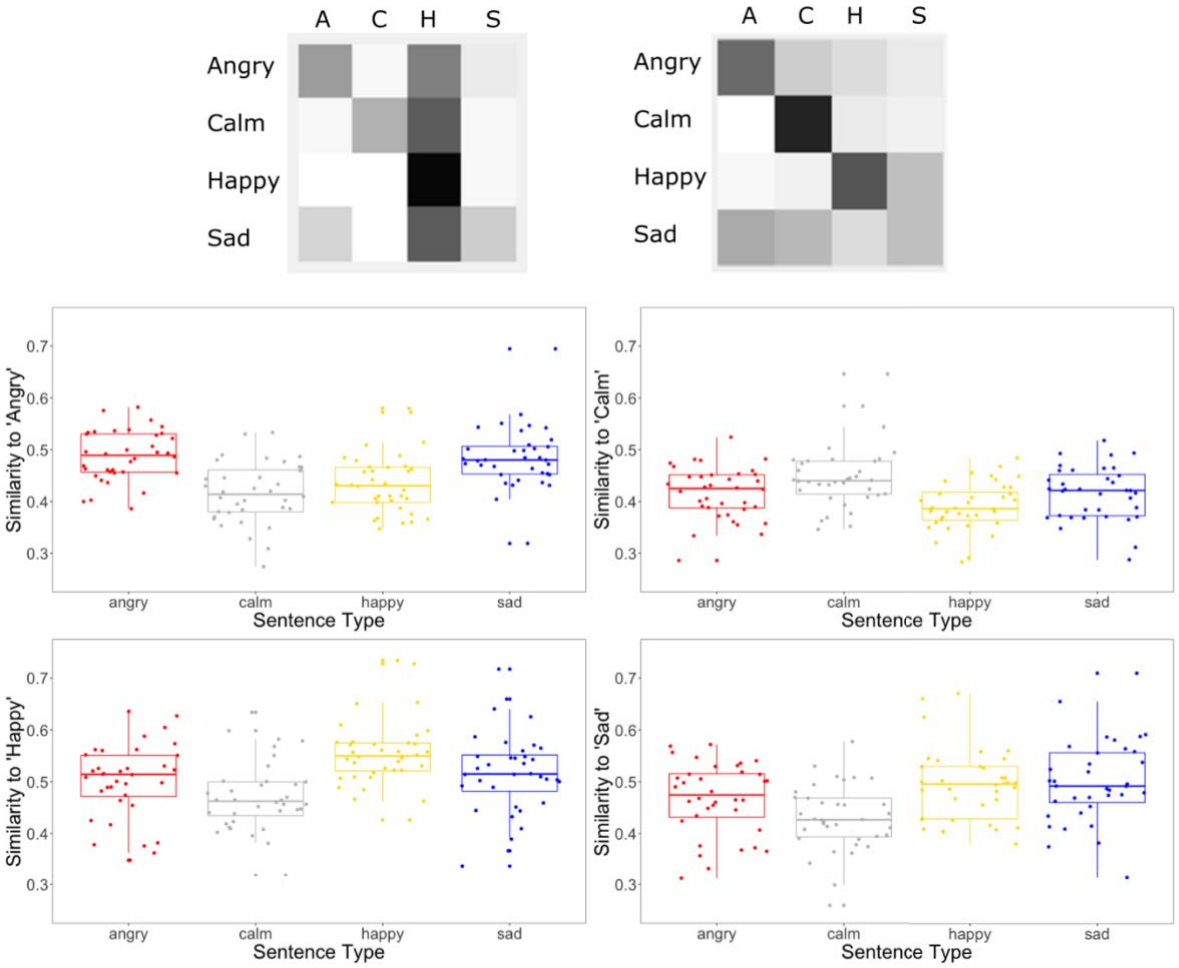
To confirm that each sentence was semantically close to the intended emotion, we ran a semantic similarity analysis using the *word2vec* package in R (Mikolov et al., 2013). The word2vec algorithm is a predictive model that derives semantic relationships between words based on the co-occurrence of words in a set of texts. Rather than training our own word2vec model, we used a pre-trained model by Mikolov and colleagues (retrieved on March 20th, 2022), which applied the skip-gram procedure with negative sampling. This pre-trained model was built from English texts. In our case, we used this model to calculate the similarity of the content of our sentences to their respective emotion (e.g., how co-occurring was the content of the sentence "Let's go to Disneyland" to the single word "happy" in this database).

As illustrated in Figure A1, there was considerable variability within a set of 36 sentences meant to convey the same emotion. The 36 semantically angry sentences were closer to the word "angry" or "happy" than they were to the word "sad" or "calm". The 36 semantically calm sentences were closer to 'calm' and 'happy' than 'angry' or 'sad'. The 36 semantically happy sentences were closer to 'happy' than other words. The 36 semantically sad sentences were closer to 'happy' and roughly equally close 'angry' as they were to 'sad'. To explore this in a more intuitive way, we considered an artificial subject who would respond automatically to the emotion with the highest similarity with a given transcript, and we obtained the confusion matrix shown on the bottom-left panel. Surprisingly, there was a strong bias towards the 'happy' emotion across the four semantic sets. Presumably, this reflects that the word 'happy' occurs disproportionally in the database that fed the pre-trained model that word2vec relied on. To circumvent this problem, we z-scored all the similarity values (top panels) across the 144 sentences and reiterated this classification procedure (top-right). This time, the diagonal

illustrates that semantically angry, calm, and happy sentences would tend to be correctly assigned to their respective emotion, but the semantically sad sentences remained highly confusable with the others. To summarize, we attempted to provide objective support for the semantic choices made in constructing the transcripts, but this proved difficult (even though from a human perspective, there does not seem to be as much ambiguity in the transcripts – see Appendix A). Perhaps more recent packages could be better at demonstrating the adequacy of the emotional content in full sentences

Figure A1

Semantic Similarity of the stimuli to their intended emotion



Note. Top left: Confusion matrix created from the similarity scores. Top right: Confusion matrix created from the z-scored similarity scores. Middle left: Similarity of each sentence to the word 'Angry'. Middle right: Similarity of each sentence to the word 'Calm'. Bottom left: Similarity of each sentence to the word 'Happy'. Bottom right: Similarity of each sentence to the word 'Sad'.

Appendix C: Confirming adequate prosody

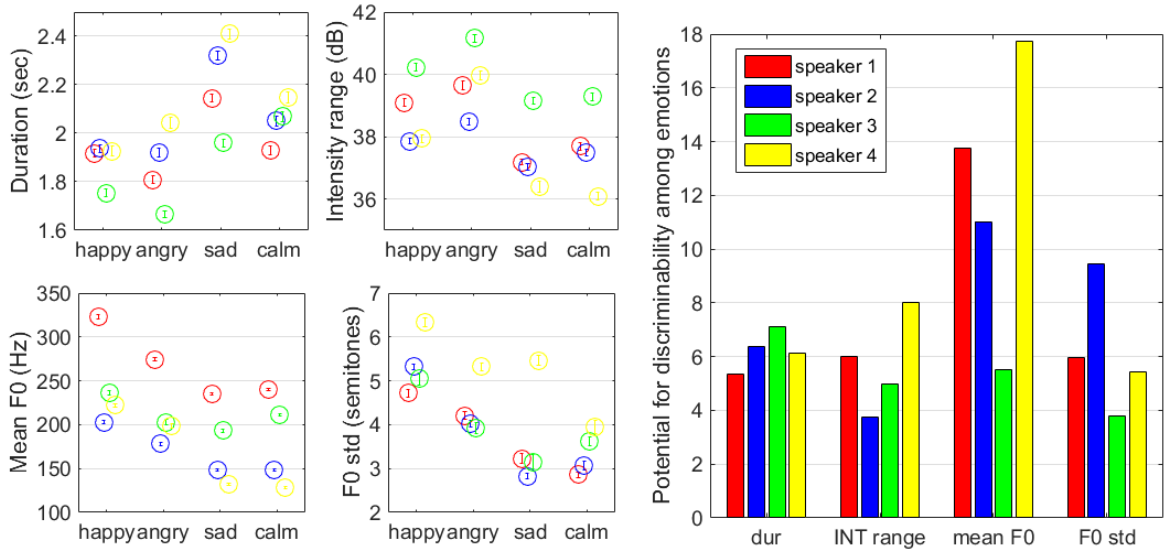
Prosody is often reduced to three acoustic cues: duration, intensity, and pitch. Thus, we analyzed these characteristics in the present stimuli to ensure that they contained the expected prosodic features of each emotion (see below), using Praat (Boersma & Weenink, 2001).

Appendix C.1: Duration cues

Results revealed a main effect of emotion on the duration cue (F (3,429) = 633.40, p < .001). On average across the four speakers, sentences were 1859, 1883, 2050, and 2209 ms for angry, happy, calm, and sad sentences respectively (all pairwise comparisons p < .045 with Bonferroni correction; see Figure A2). Note, however, that these differences in duration (e.g., 350 ms shorter in angry than in sad stimuli) varied by speaker, suggesting that listeners would need to perform these comparisons within a given speaker for this cue to be reliable. The random shuffling of each trial across the four speakers would, therefore, make it harder for listeners to follow such a strategy.

Figure A2

Prosodic features of each stimulus by emotion and speaker



Note. Illustration of the prosodic features expected in sentences (left) with means (circles) and standard errors (error bars) across 144 productions of each speaker. These features have different potential for discriminability among the four emotions (right), with pitch often being the dominant one.

Appendix C.2: Intensity cues

There was roughly a 10-12 dB difference in mean intensity between happy/angry and sad/calm in the initial recordings, confirming that emotions were adequately enacted. However, we presumed that this loudness cue would be too salient and could inflate performance in certain incongruent trials. For this reason, we dampened this cue by equalizing all stimuli at 65 dB SPL. This did not affect the change in dynamic range that occurred throughout the sentences, so listeners could still use intensity cues but in a more subtle manner. To analyze this cue, we extracted intensity contours and subtracted each minimum from its maximum. Results revealed a main effect of emotion on the intensity cue of the sentences (F (3,429) = 393.1, $p < .001$). On average across the four speakers, sentences had a dynamic range of 37.4, 37.6, 38.8, and 39.8 dB for sad, calm, happy and angry sentences respectively (all pairwise comparisons $p < .001$ with Bonferroni correction except sad vs. calm $p = .079$; see Figure A2). These differences in intensity range were relatively small (< 2.4 dB) but also varied by speaker to a small degree.

Appendix C.3: Pitch cues

There is no single metric within the fundamental frequency (F0) contours that can perfectly summarize a given emotion, so we chose the mean F0 and F0 standard deviation (F0-sd) to tap into voice pitch height and contour. Results revealed a main effect of emotion on the mean F0 as a pitch cue (F (3,429) = 1386.4, $p < .001$). On average across the four speakers, sentences were 177.0, 182.1, 213.5, and 246.3 Hz for sad, calm, angry, and happy sentences

respectively (all pairwise comparisons p < .001 with Bonferroni correction; see Figure A2). Here, there was expectedly a great amount of variability between speakers, especially between males (speakers 2 and 4) and females (speakers 1 and 3). Additionally, results revealed a main effect of emotion on the F0-sd (F (3,429) = 301.80, p < .001). On average across the four speakers, sentences had F0-sd of 3.4, 3.7, 4.4, and 5.4 semitones for calm, sad, angry, and happy sentences respectively (all pairwise comparisons p < .001 with Bonferroni correction; see Figure A2). There is also inter-speaker variability on this metric of F0-sd, where speaker 4 exhibited a range of 1-2 semitones larger than the other speakers. Overall, these pitch cues allowed listeners to discriminate at least certain pairs of emotions (e.g., happy vs. sad/calm), but the speaker variability made it harder for listeners to rely on this cue exclusively.

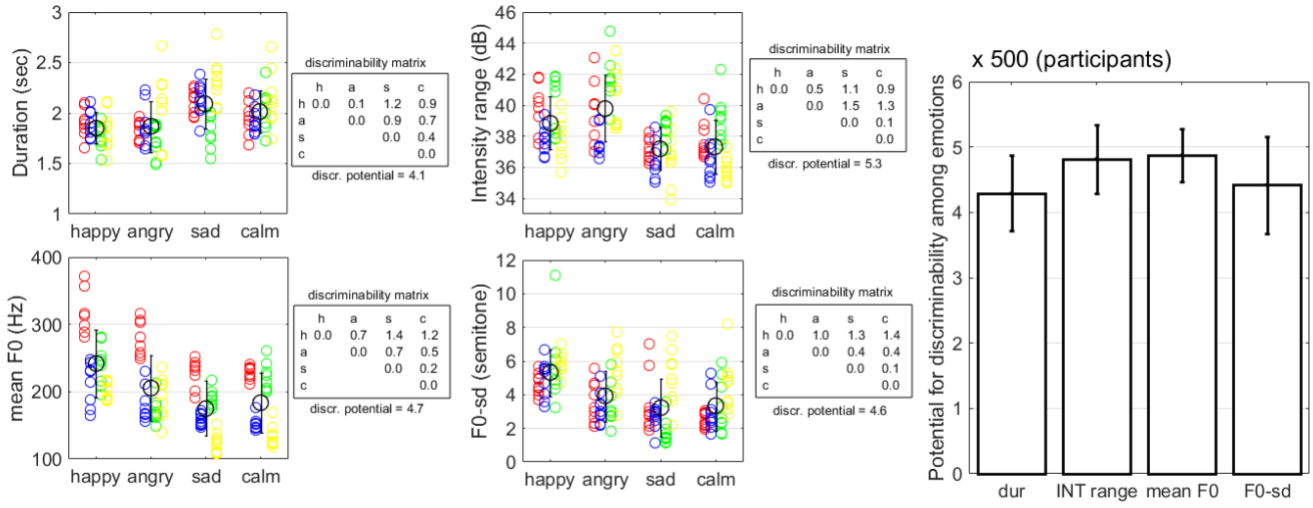
Appendix C.4: Prosodic analyses after randomly shuffling between speakers

While each emotion was enacted as expected, there was some variability between speakers in each metric. This led not only to a main effect of speaker (p < .001) but also an interaction between speaker and emotion (p < .001) in every single metric. Rather than delving into the specific patterns exhibited by each speaker, we calculated a discriminability measure for each pair of emotions (as the absolute value of the difference between mean values divided by their averaged standard deviation). We could then sum all the values in a discriminability matrix (i.e., considering all possible pairs) to provide a single metric reflecting the potential for discriminability offered by a given prosodic feature (right panel of Figure A3). For example, speaker 4's mean voice pitch was by far the most beneficial for discriminability. To a smaller degree, this was also the case for speaker 1 and 2, while speaker 3 exhibited a relatively balanced potential for discriminability across the prosodic features.

If listeners were exposed to each speaker one after another and were given time to learn their peculiar speaking styles, the contrasts between emotions would become quite salient for certain cues (e.g., high pitch for happy/angry versus low pitch for calm/sad, within a given speaker). This was not the case in the present study since all trials were randomly shuffled across speakers. In other words, these cues were not easy to spot as they were swamped by inter-subject variability, as well as intra-subject variability to some degree. To reflect the discriminability potential in a manner that was identical to how it occurred during the study, we did the following procedure 500 times. In each iteration, 36 items were chosen for each emotion, equally drawn from each speaker. The means and standard deviations (across the 36 items) of duration, intensity range, mean F0, and F0-sd, were computed for the same random items in each emotion and a discriminability matrix was calculated from all pairs of emotions, eventually summed to provide an estimate of discriminability potential (Figure A3, left/middle panels). On average across all iterations, the mean F0 was no longer as discriminable as it was without this speaker shuffling. This is precisely because the inter-subject variability was considerable in comparison with the emotion-induced differences, hindering the reliance on a given cue when swapping from one speaker to another randomly throughout the study. As a result, all prosodic features were now more comparable in their discriminability potential (Figure A3, right panel), meaning that depending on the random allocation of speakers into each emotion, different participants might have preferentially used one cue over the others. Out of 500 iterations (simulating participants), intensity range was the dominant cue in 38.4% of cases, followed by mean F0 in 34.0% of cases, F0-sd in about 17.2% of cases, and duration in only 10.4% of cases.

Figure A3

Prosodic features and discriminability pattern for 500 iterations



Note. Prosodic features shown for each emotion enacted by four speakers (colors), and their respective discriminability matrix from which a discriminability potential was derived (left/middle panels). Replicating this procedure for 500 iterations (each iteration representing a different participant receiving a random set of stimuli drawn equally from each speaker) results in a relatively homogeneous discriminability potential across duration, intensity, and pitch cues (right panel).

Appendix D: Trial by Trial analyses

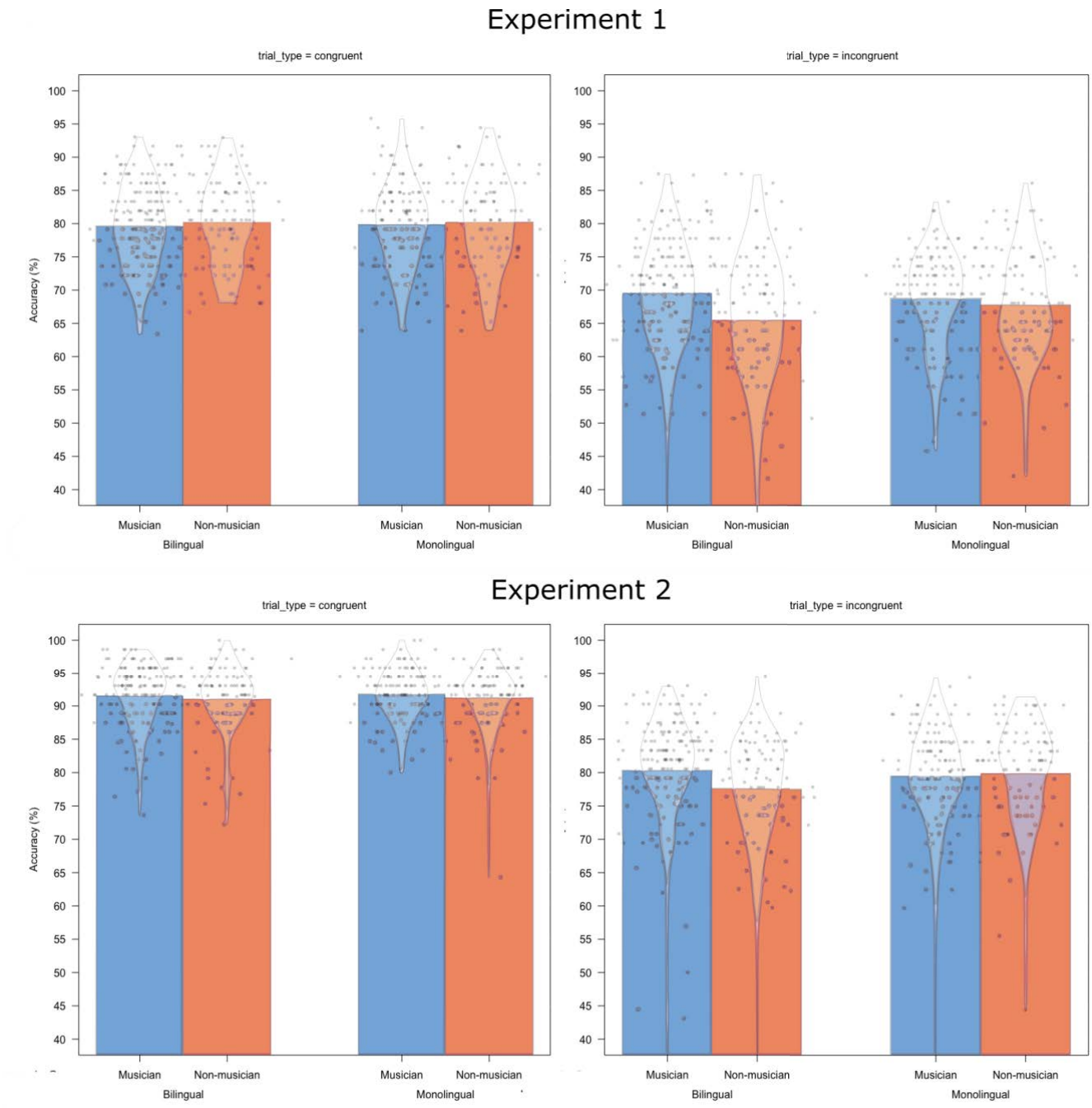
Appendix D.1: Performance

In these analyses, we used performance on each trial (1 = correct and 0 = incorrect) as the dependent variable in logistic mixed effects models to examine the recognition of emotional prosody in Experiment 1 and the recognition of emotional semantics in Experiment 2 (See Table A1 for all model results). These results mirror those presented in the main article using *d'* as the outcome variable. In Experiment 1, there was a main effect of *trial type,* such that performance suffered in the incongruent trials compared to the congruent trials (p <.001; see Figure A4). There was a main effect of *musicianship*, whereby musicians outperformed non-musicians, but this effect was very modest (p = .0453), not so different from the main analysis (p = .0766). The two-way interaction between *musicianship* and *trial type* revealed no difference between musicians and non-musicians on congruent trials (p = .793), whereas musicians outperform non-musicians on incongruent trials (p = .002). Finally, there was a three-way interaction between *bilingualism*, *musicianship*, and *trial type*, such that musicians only outperform non-musicians on the incongruent trials when also a bilingual (p < .001) and not a monolingual (p = .990). There were no other significant main effects or interactions. In Experiment 2, there was also a main effect of *trial* type, such that performance suffered in the incongruent trials compared to the congruent trials (p <.001). The main effect of musicianship (p = .0549) was technically lost but not so different from the main analysis (p = 0.0373). Additionally, there was a three-way interaction between *bilingualism*, *musicianship*, and *trial type*. Once again, musicians only outperformed non-musicians on the incongruent trials provided that they were also a bilingual (p = .0509) and not a monolingual (p = .999). There were no other significant main effects or interactions. To

summarize, this logistic analysis was conducted on a trial basis and the findings were largely in line with those presented in the article.

Figure A4

Trial by trial performance data



Note. Interaction between musicianship and bilingualism on performance (% score) by trial type (congruent on the left panels, and incongruent on the right panels) in Experiment 1 (Top) and Experiment 2 (Bottom).

Table A1

Model Results of the individual trial performance logistic mixed effects models

Fixed Effects:	χ^2	DF	*p*
Experiment 1			
Intercept			
Trial Type	1391.2	1	<.001***
Bilingualism	0.15	1	.703
Musicianship	4.01	1	.0453*
Trial Type x Bilingualism	0.035	1	.853
Trial Type x Musicianship	18.68	1	<.001***
Bilingualism x Musicianship	1.96	1	.162
Trial Type x Bilingualism x Musicianship	5.22	1	.0223*
Experiment 2			
Intercept			
Trial Type	2313.0	1	<.001***
Bilingualism	0.27	1	.606
Musicianship	3.68	1	.0549
Trial Type x Bilingualism	0.16	1	.689
Trial Type x Musicianship	0.059	1	.809
Bilingualism x Musicianship	3.27	1	.0706
Trial Type x Bilingualism x Musicianship	3.86	1	.0495*

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Appendix D.2: LogRT

In these analyses, we used log reaction time on each trial as the dependent variable in linear mixed effects models to examine the processing speed of emotional prosody in Experiment 1 and the processing speed of emotional semantics in Experiment 2 (See Table A2 for all model results). In both experiments, there was a main effect of *trial type* ($\chi^2(1) = 382.91$, $p < .001$; and $\chi^2(1) = 925.09$, $p < .001$, respectively), such that log reaction times were longer for incongruent trials compared to congruent trials. But there were no other significant main effects or interactions. In other words, all participants took more time to process the incongruent stimuli (generally a good

sign that they paid attention to the task). However, this interference-delay did not differ among the groups.

Table A2

Model Results of the individual trial log reaction time linear mixed effects models

Fixed Effects:	χ^2	DF	*p*
Experiment 1			
Intercept			
Trial Type	382.9	1	<.001***
Bilingualism	0.44	1	.507
Musicianship	0.051	1	.821
Trial Type x Bilingualism	3.11	1	.078
Trial Type x Musicianship	0.0054	1	.941
Bilingualism x Musicianship	0.33	1	.567
Trial Type x Bilingualism x Musicianship	0.77	1	.380
Experiment 2			
Intercept			
Trial Type	925.1	1	<.001***
Bilingualism	1.09	1	.296
Musicianship	0.52	1	.473
Trial Type x Bilingualism	0.19	1	.667
Trial Type x Musicianship	0.77	1	.380
Bilingualism x Musicianship	1.54	1	.214
Trial Type x Bilingualism x Musicianship	2.53	1	.112

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Appendix E: Bilingualism and Musicianship as continuous variables

In these analyses, we used the interference effect (congruent minus incongruent trials) in *d'* units as the dependent variable in linear mixed effects models with *bilingualism* and *musicianship* as continuous variables. These models were run separately for Experiments 1 and 2, and always contained random intercepts by subject, and random intercepts by emotion, as in the main article. Second language (always non-English) proficiency was used as the continuous metric of *bilingualism* (on a scale from 0-10, where 0 is not proficient at all and 10 is the most proficient). First instrument proficiency was used as the continuous metric of *musicianship* (on a scale from 0-10, where 0 is not proficient at all and 10 is the most proficient). Note that age of acquisition of the second language/first instrument is difficult to use because monolinguals and non-musicians do not have a value for this metric, and it is questionable what should be used instead (e.g., age at testing or a high arbitrary value). Similarly, use of the second language/first instrument was avoided because it did not sum to 100% and again different standardization procedures could be envisioned (depending on the number of languages/instruments involved). So, we chose proficiency and assigned it to 0 respectively for the monolinguals' L2 or the non-musicians' I1. Figure A5 illustrates a 3D plot of the interference effect varying as a function of the *bilingualism* and *musicianship* metrics chosen. The bilingual musicians (green symbols) spread through this space tends to have lower interference effect (lower on the vertical axis).

These results (see Table A3) mirror those presented in the main article using *bilingualism* and *musicianship* as categorical variables. That is, both experiments successfully generated an interference effect from the incongruency between prosody and semantics, but the size of this interference depended on the group allocation. Musicians had a smaller interference effect compared to non-musicians, but only when also a bilingual and not when a monolingual. This

difference can be seen from a different angle in Figure A6 (top right panel) for Experiment 1, where no difference in interference is seen on the left side of the abscissa between monolingual musicians and monolingual non-musicians while this difference progressively arises between musicians and non-musicians as L2 proficiency increases. A similar departure between the regression lines can be seen for Experiment 2 (Figure A6 bottom right panel).

Table A3

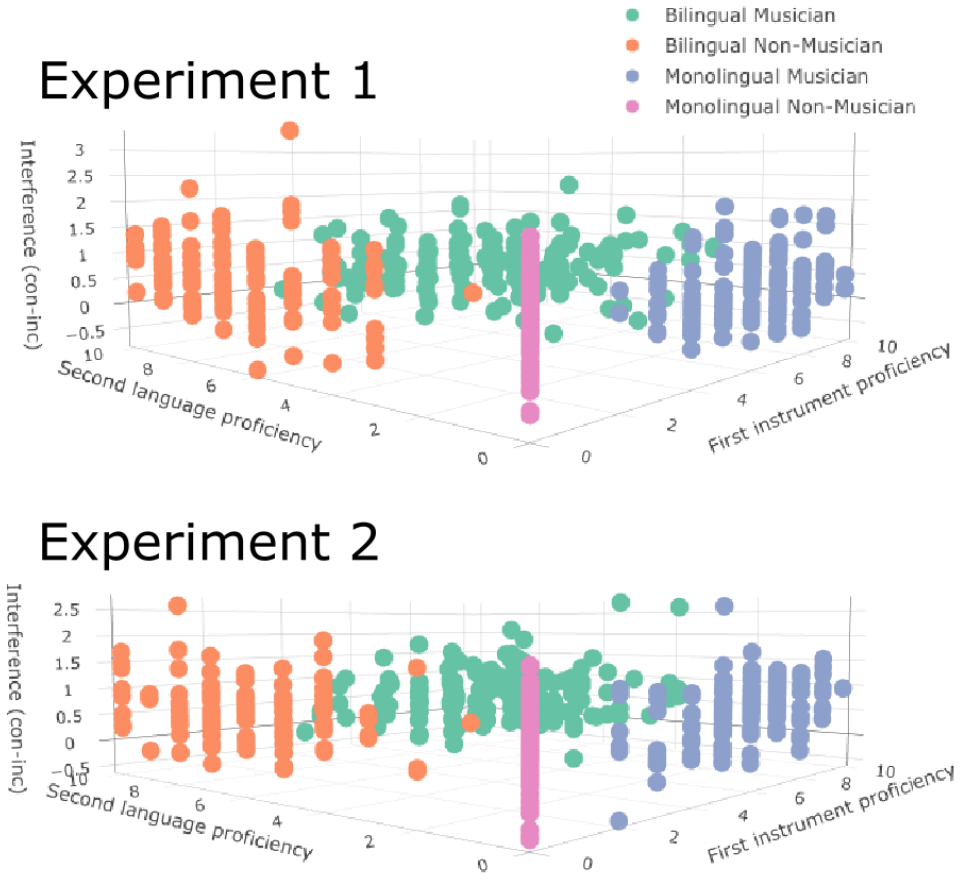
Model Results of the linear mixed effects models using continuous bilingualism (second language proficiency) and musicianship (first instrument proficiency) variables

Fixed Effects:	χ^2	DF	*p*
Experiment 1			
Intercept			
Trial Type	1387.5	1	<.001***
Bilingualism	3.01	1	.0828
Musicianship	1.39	1	.239
Trial Type x Bilingualism	0.00	1	.995
Trial Type x Musicianship	24.09	1	.<.001***
Bilingualism x Musicianship	1.70	1	.192
Trial Type x Bilingualism x Musicianship	8.11	1	.00439**
Experiment 2			
Intercept			
Trial Type	2216.6	1	<.001***
Bilingualism	0.035	1	.852
Musicianship	3.73	1	.0535
Trial Type x Bilingualism	0.0008	1	.978
Trial Type x Musicianship	0.60	1	.440
Bilingualism x Musicianship	1.99	1	.159
Trial Type x Bilingualism x Musicianship	12.04	1	.00052***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Figure A5

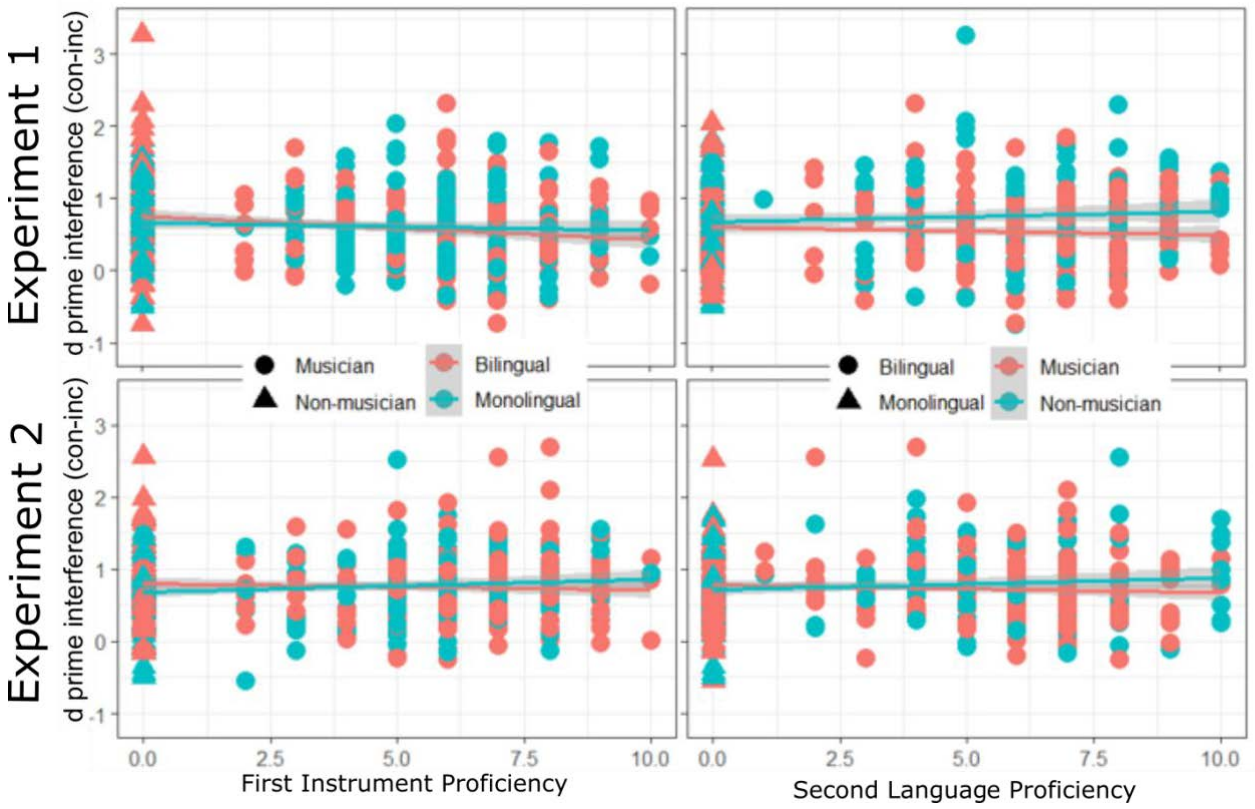
3D plot of the interference effect, second language proficiency and first instrument proficiency by group



Note. 3D plot of the *d'* interference effect (congruent minus incongruent trials; Y-axis), second language proficiency (0-10; X-axis) and first instrument proficiency (0-10; Z-axis) by group in Experiment 1 (top panel) and Experiment 2 (bottom panel).

Figure A6

Correlations between d' interference effect and first instrument proficiency/ second language proficiency by group



Note. Scatterplots of the *d'* interference effect (congruent minus incongruent trials) by first instrument proficiency (0-10; left panels) and by second language proficiency (0-10; right panels) in Experiment 1 (top) and Experiment 2 (bottom).

Appendix F: Block Type

In the current experiments the type of incongruency was not presented randomly, instead it was done systemically in each block type. More specifically, each participant was presented with three blocks of 48 trials, where 24 trials were incongruent. Each of the three blocks differed in the way in which the semantics and prosody were swapped in the incongruent trials. In the *swap valence* block, the valence, or positive-negative dimension, of the emotions was swapped (e.g., a semantically angry sentence enacted with a happy prosody). In the *swap intensity* block, the intensity, or high-low energy dimension, of the emotions was swapped (e.g., a semantically happy sentence enacted with a calm prosody). Finally, in the *swap both* block, both the intensity and valence of the emotions were swapped (e.g., a semantically angry sentence enacted with a calm prosody). In the following section, we discuss that the interference in processing that we found in the main results changed quite dramatically based on the type of incongruency and type of task.

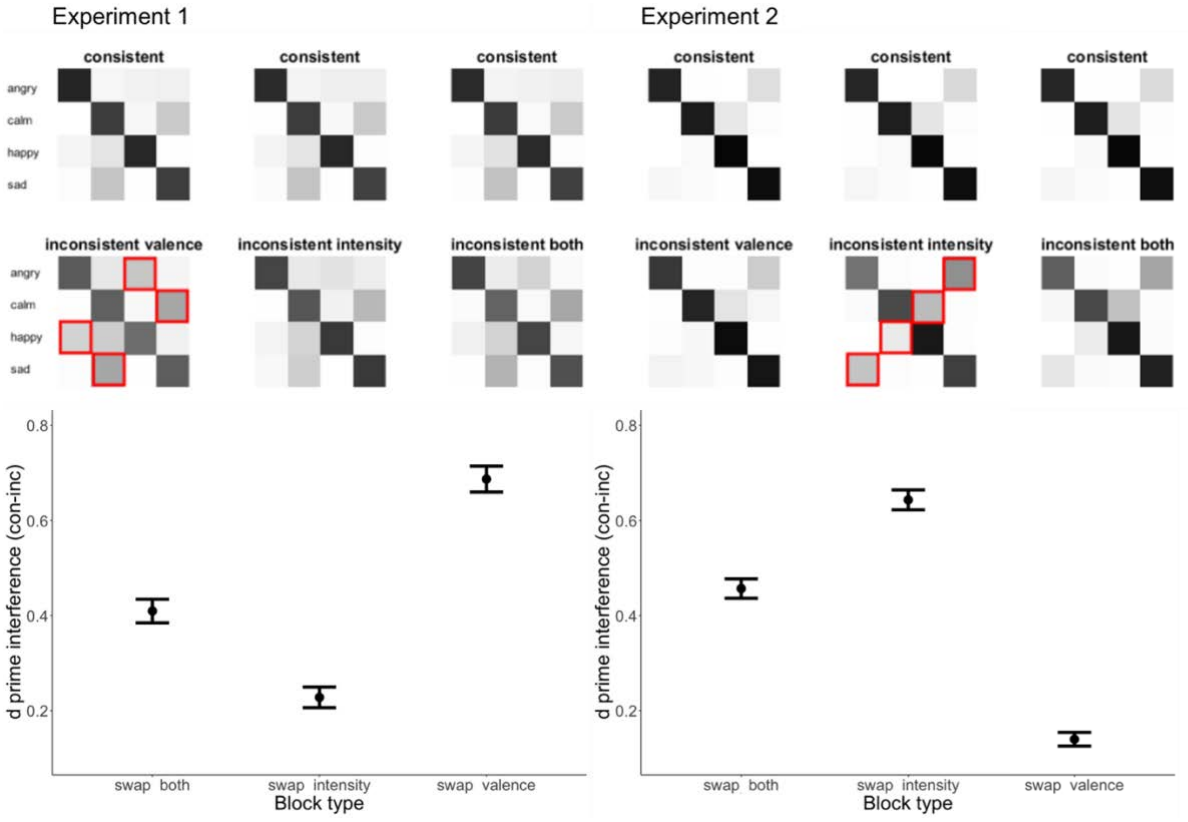
In both experiments, block type had a big role. Figure A7 displays confusion matrices showing correct and incorrect response patterns for each emotion. The congruent trials led to almost identical patterns in each block (Figure A7, middle row). The dark diagonal simply reflects that there were few errors (i.e., participants responded mostly sad for sad stimuli, and similarly for the other emotions). Thus, as expected, performance was good in the congruent trials in all blocks. In incongruent trials, on the other hand, the error patterns were similar across the three blocks, but differed between the two experiments. In Experiment 1, happy and angry were most often confused, and so were sad and calm. These types of errors are considered valence-based, as for example happy and angry are both high intensity emotions, but they are of

opposite valence (i.e., positive vs. negative). In contrast, in Experiment 2, angry and sad were most often confused, and so were happy and calm.

These types of errors are considered intensity-based, as for example angry and sad are both negative emotions, but they are of opposite intensities (i.e., high vs. low). This finding is quite remarkable: we had intended to generate different error patterns in each block type, and instead found the same error patterns (based on valence in Experiment 1 or based on intensity in Experiment 2).

Figure A7

Results by block type



Note. Top panels: Confusion matrices by *trial type* (congruent and incongruent) and *block type* (Experiment 1 top left and Experiment 2 top right). Emotions presented in the sentences are displayed rows and emotions responded by participants displayed as columns. Darker colours

represent larger values. Bottom panels: The interference effect (congruent minus incongruent) by *block type* (Experiment 1 bottom left and Experiment 2 bottom right), where lower *d'* units indicate better resistance to the distracting cue (i.e., better performance).

To illustrate these phenomena in a more compact way, we calculated the interference effect in *d'* units from these confusion matrices by subtracting d' for the incongruent conditions from d' for the congruent conditions (Figure A7, top panels). In Experiment 1, the largest interference occurred for the *swap-valence* block (about 0.7 reduction in *d'*, equivalent to about 18% drop in performance), followed by the *swap-both* block (about 0.4 reduction in *d'*, equivalent to about 11% drop in performance) and the *swap-intensity* block generated the weakest interference (only about 0.2 reduction in *d'*, equivalent to about 7% drop in performance), with each pairwise comparison significant (p <0.001). In Experiment 2, the largest interference occurred for the *swap-intensity* block (about 0.65 reduction in *d'*, equivalent to about 18% drop in performance), followed by the *swap-both* block (> 0.4 reduction in *d'*, equivalent to about 13% drop in performance) and the *swap-valence* block generated the weakest interference (< 0.2 reduction in *d'*, equivalent to about 5% drop in performance), with each pairwise comparison significant (p <0.001).

Let us discuss this observation here briefly because presumably, this tells us about the nature of semantic versus prosodic features in speech. Semantics are powerful at indicating positive versus negative emotions but poor at conveying low versus high emotional intensity. For example, reading the sentence "I am glad to see you." Based on the semantics, it is clearly positive, but it is not clear whether this is low or high in emotional intensity. Therefore, having conflicting semantic and prosodic cues to an emotion that are of the same valence but differ in their intensity is most likely to generate confusion (when asked to respond to the semantic cues).

In contrast, prosodic features strongly discriminate high versus low emotional intensity, but they are weaker in indicating valence. Think of the tone of voice of a sad person; their low and monotonous pitch and volume combined with slow-paced articulation are strongly indicative of a low-energy emotional state but could arguably be a depressed or calm speaker. Therefore, having conflicting semantic and prosodic cues to an emotion that are of the same intensity but differ in their valence is most likely to generate confusion asked to attend to the prosodic cue.

In fact, it is interesting to note that these error types existed to a small degree even within congruent trials, being slightly more valence-based in Experiment 1 and more intensity-based in Experiment 2. This suggests again that it is not about particular emotions conflicting with another, but rather about the general power of prosody versus semantics, and what happens when we rely on only one or the other. Even in the absence of conflicting semantic cues, prosody is more likely to be misrecognized for an emotion with opposite valence (as it does not convey it well), whereas a given semantic content is more likely to be misrecognized for an emotion with opposite intensity (as it does not convey it well).

We further examined whether this effect of block type would depend on group allocation. In both experiments, there was a main effect of *block type* (Experiment 1: $\chi^2(2) = 202.43$, $p < .001$; Experiment 2: $\chi^2(2) = 393.49$, $p < .001$). However, *block type* did not interact with either *musicianship* (Experiment 1: $\chi^2(2) = 2.46$, $p = .292$; Experiment 2: $\chi^2(2) = 1.43$, $p = .490$), or with *bilingualism* (Experiment 1: $\chi^2(2) = 0.410$, $p = .815$; Experiment 2: $\chi^2(2) = 0.167$, $p = .920$). Additionally, there was no three-way interaction between *block type*, *musicianship*, and *bilingualism* in Experiment 1 ($\chi^2(2) = 2.22$, $p = .330$), and a modest one in Experiment 2 ($\chi^2(2) = 7.05$, $p = .030$). The source of this interaction was not particularly interesting as it seemed to come from a floor effect (i.e., the group factors were less likely to matter when there was no

interference to act upon). Since there was little interference in some block types (e.g., swap intensity in Experiment 1) *bilingualism* and *musicianship* did not play much of a role in these instances. To simplify, as a first approximation, this analysis revealed that choosing a particular form of incongruency will have a considerable influence on the size of the interference generated, but not on whether the listener's profile (i.e., whether they have language or musical experience) make them subject to it or not.