Appendix S1: 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 S2: Confirming adequate semantics

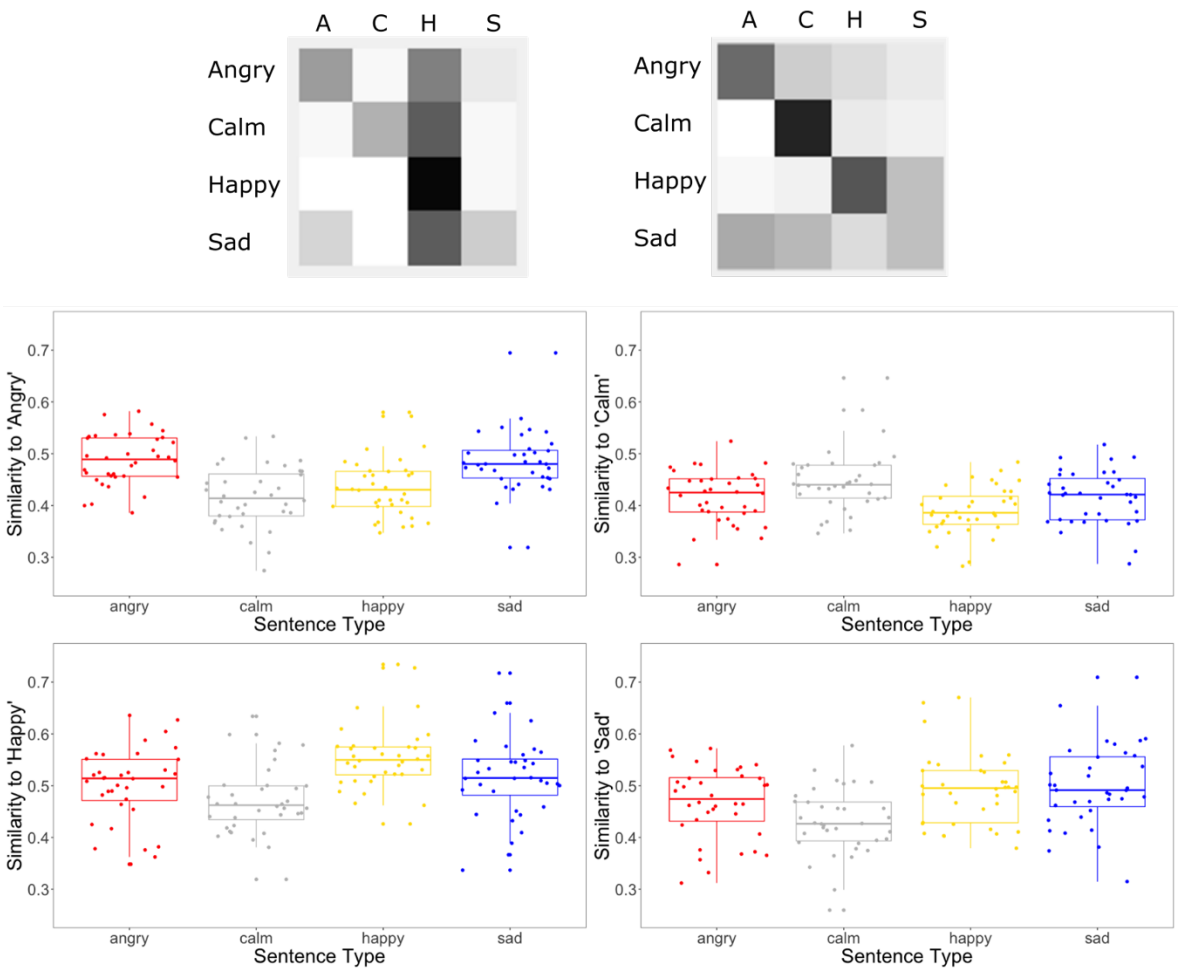
To confirm that each sentence was semantically close to the intended emotion, we ran a semantic similarity analysis using the *word2vec* package in R (Mikolov et al., 2013). The word2vec algorithm is a predictive model that derives semantic relationships between words based on the co-occurrence of words 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 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 S1). Perhaps more recent packages could be better at demonstrating the adequacy of the emotional content in full sentences.

Figure S1

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 S3: Confirming adequate prosody

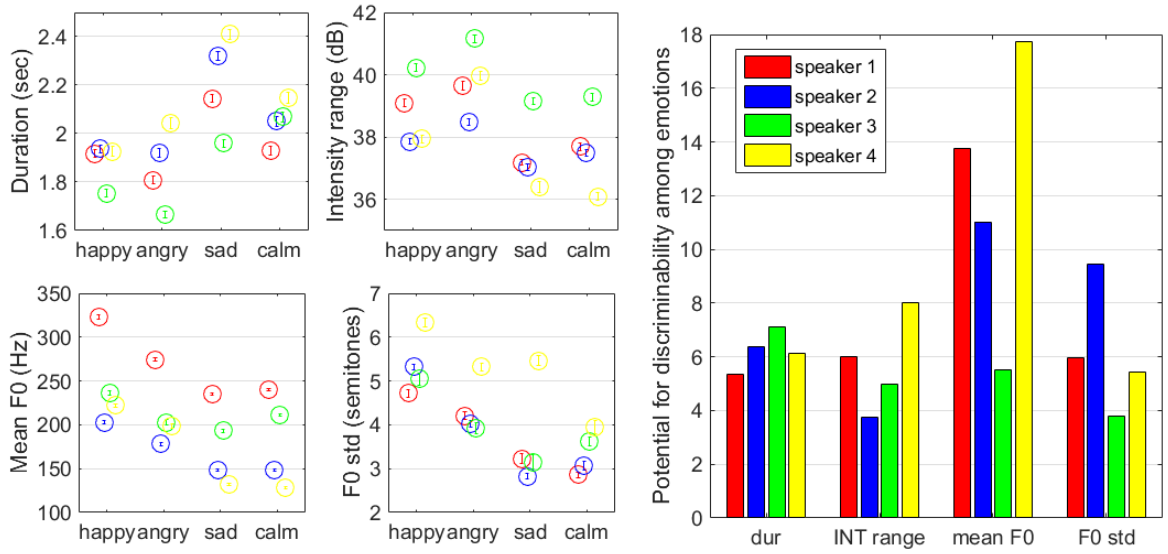
Prosody is often reduced to three acoustic dimensions: duration, intensity, and pitch. Thus, we analyzed these characteristics to ensure that they contained the expected prosodic features of each emotion (see below), using Praat (Boersma & Weenink, 2001).

Appendix S3.1: 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 long 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 S2

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 S3.2: 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 S3.3: 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, F (3,429) = 1386.4, $p < .001$. On average across the four speakers, sentences had mean F0s of 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). Inter-speaker variability is evident on this metric of F0-sd as well, 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 S3.4: Prosodic analyses after randomly shuffling between speakers

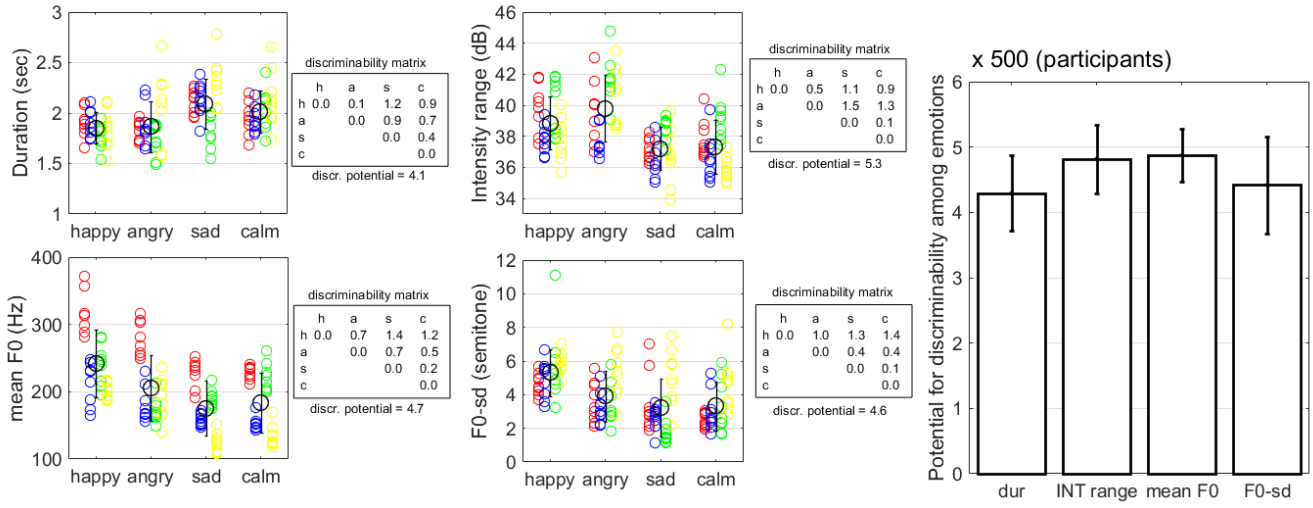
While each emotion was enacted as expected, there was some variability between speakers in each metric. This led not only to a main effect of speaker ($p < .001$) but also an interaction between speaker and emotion ($p < .001$) in every single metric. Rather than delving into the specific patterns exhibited 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 A2). 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). This means 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 roughly the number of participants in each experiment), 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 S3

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 S4: Trial by Trial analyses

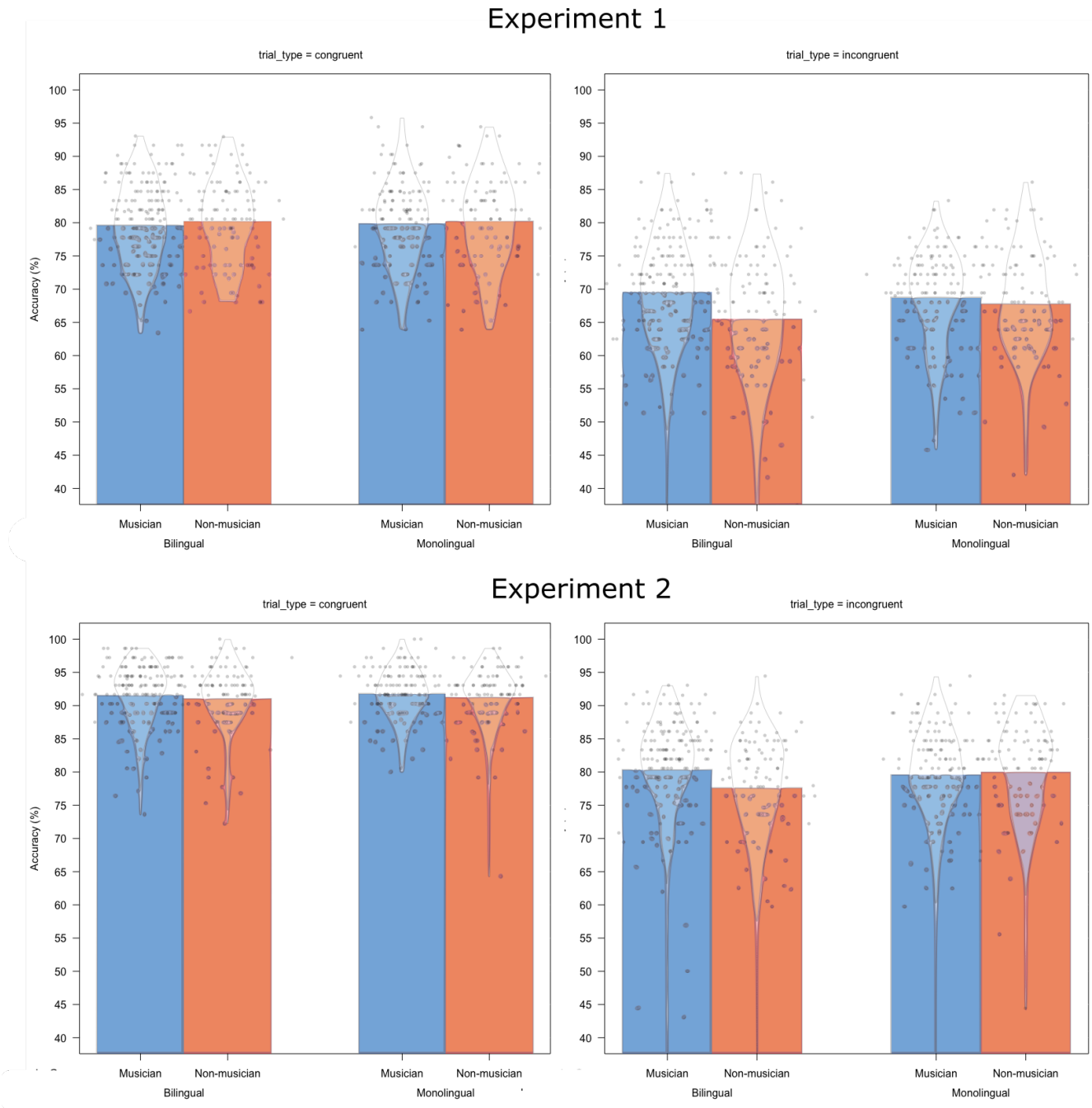
Appendix S4.1: Performance

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 dependent 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 S4

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 S1

Model Results of the logistic mixed effects models analyzing individual trial performance

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 S4.2: 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, p* <.001, such that 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 S2

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 S5: 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 (as opposed to categorical for the results reported in the article). 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 cube and tend to have lower interference effect (lower on the vertical axis).

These results (see Table A3) mirror those presented in the main article. 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 S3

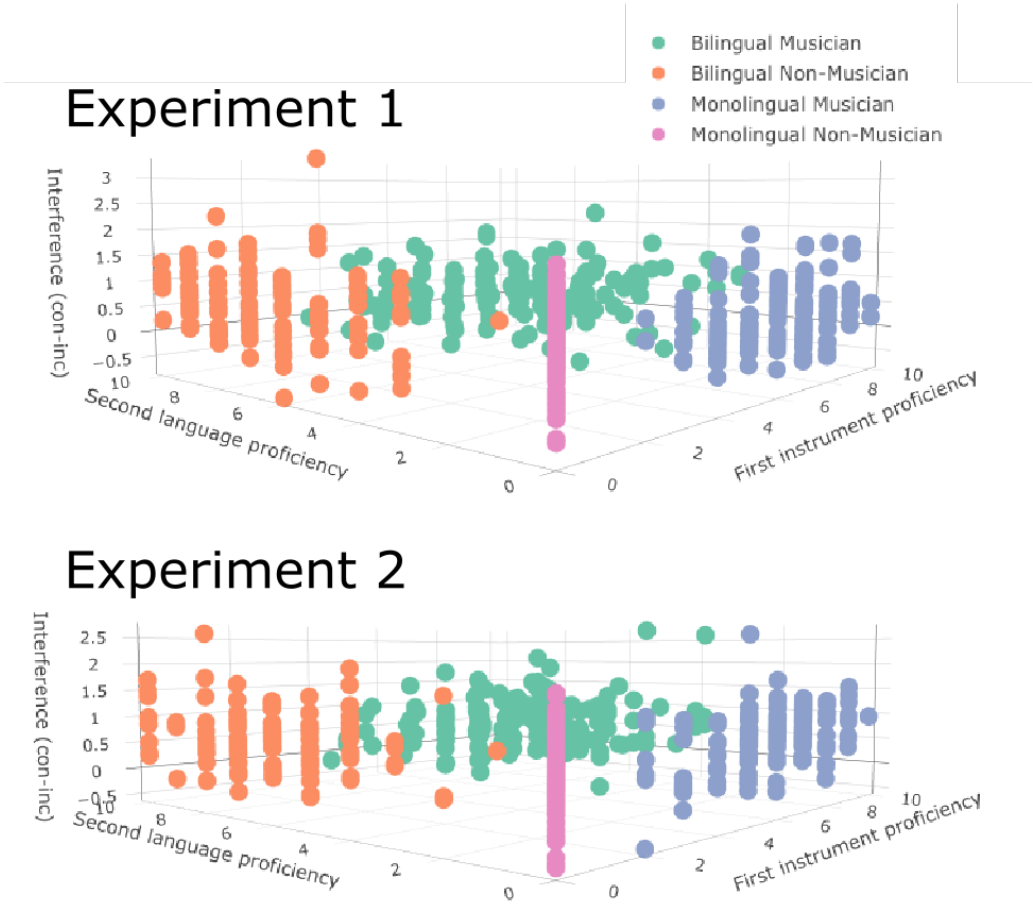
Model Results of the linear mixed effects models using continuous bilingualism (second language proficiency) and musicianship (first instrument proficiency) variables and raw performance per trial (1 or 0) as the dependent variable

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 S5

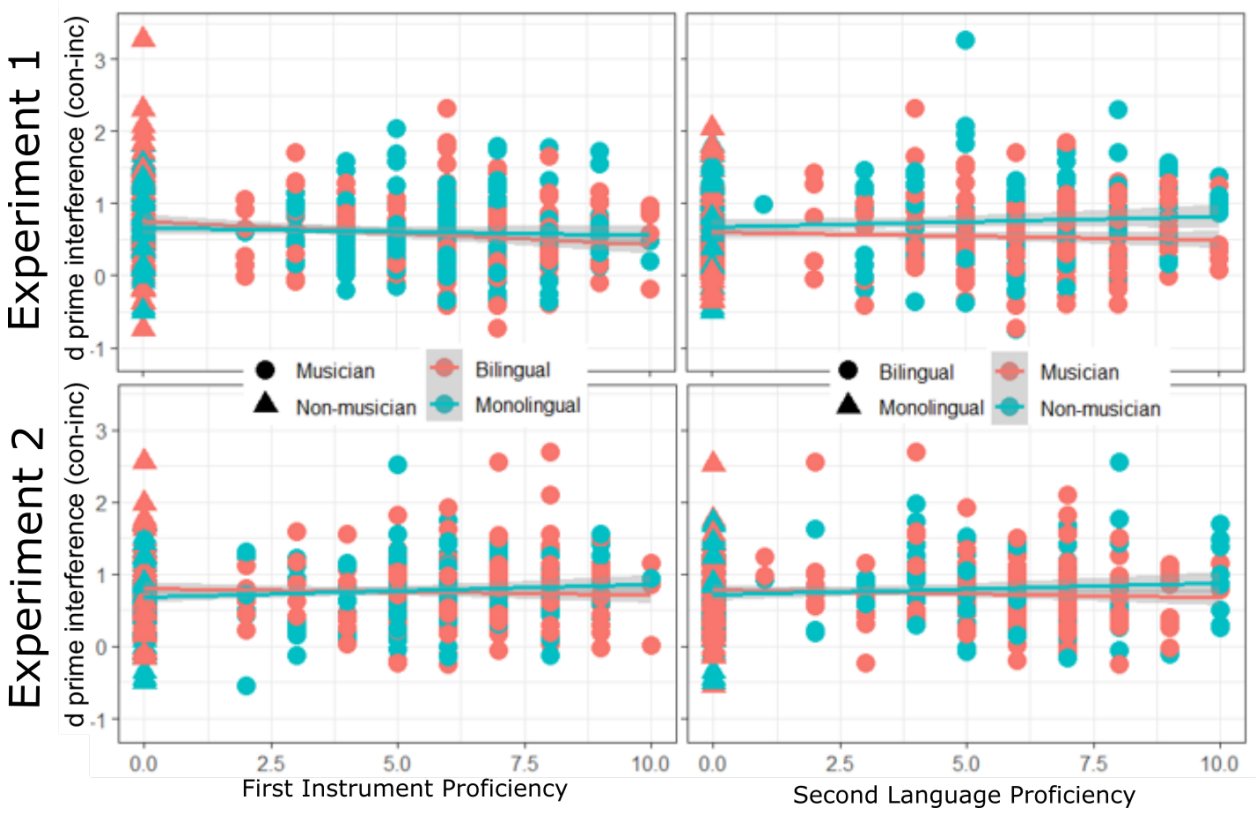
3D plot of the interference effect, second language proficiency and first instrument proficiency by group



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 S6

Correlations between the interference effect (in d' units) 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).

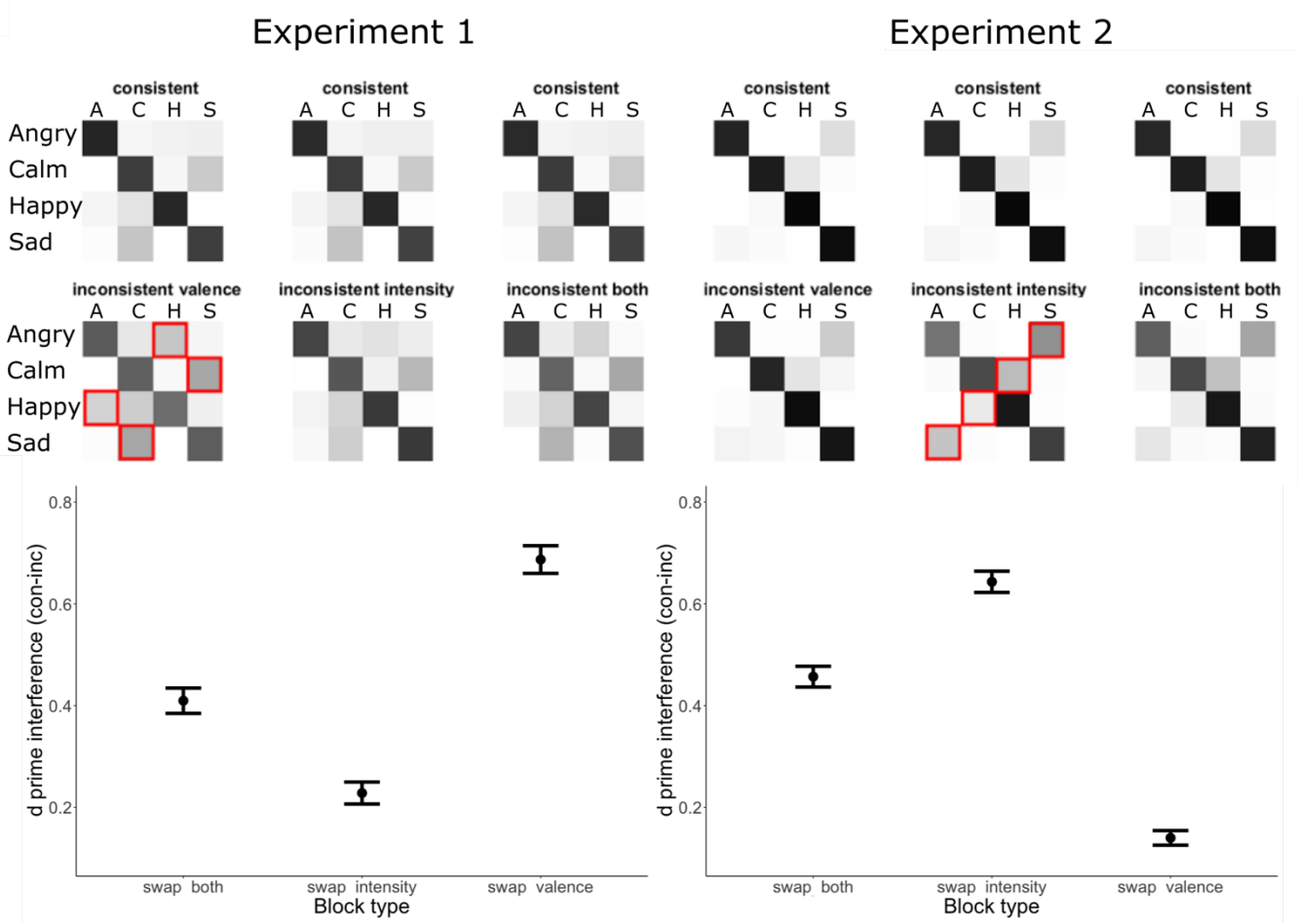
Appendix S6: Block Type

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 an interesting observation which was not discussed in the article, namely that the interference changed quite dramatically based on the type of incongruency and the 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 very good in the congruent trials in similar 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 S7



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 (when asked to attend to the prosodic cues).

In fact, it is interesting to note that these error types existed to a very 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 conflict, 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).

Appendix S6.1: Results by block type

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.