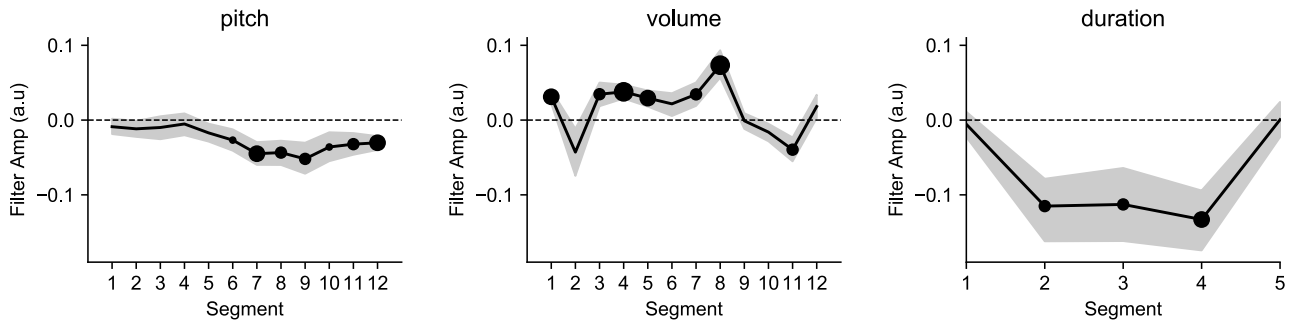
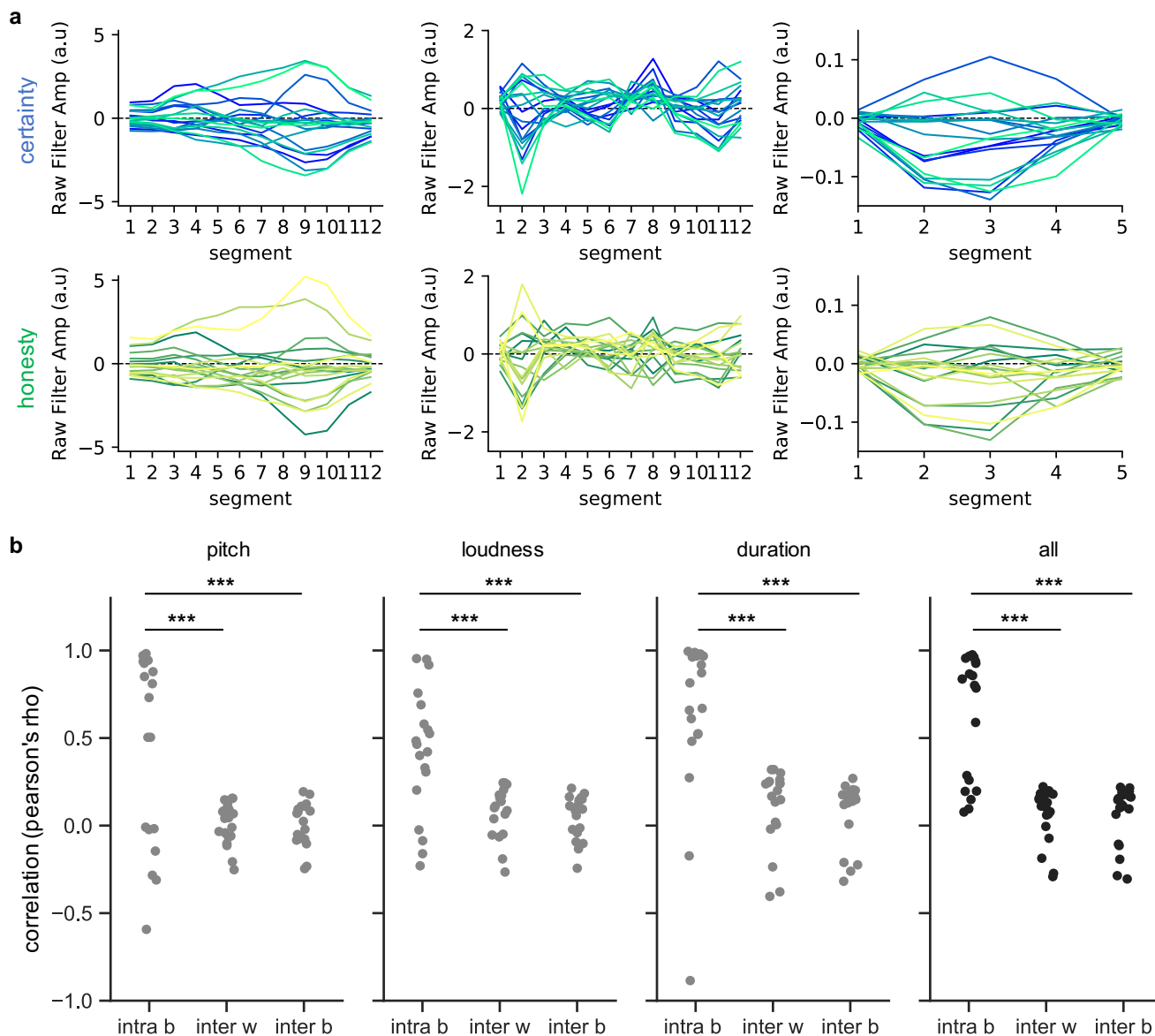


1 Supplementary Information



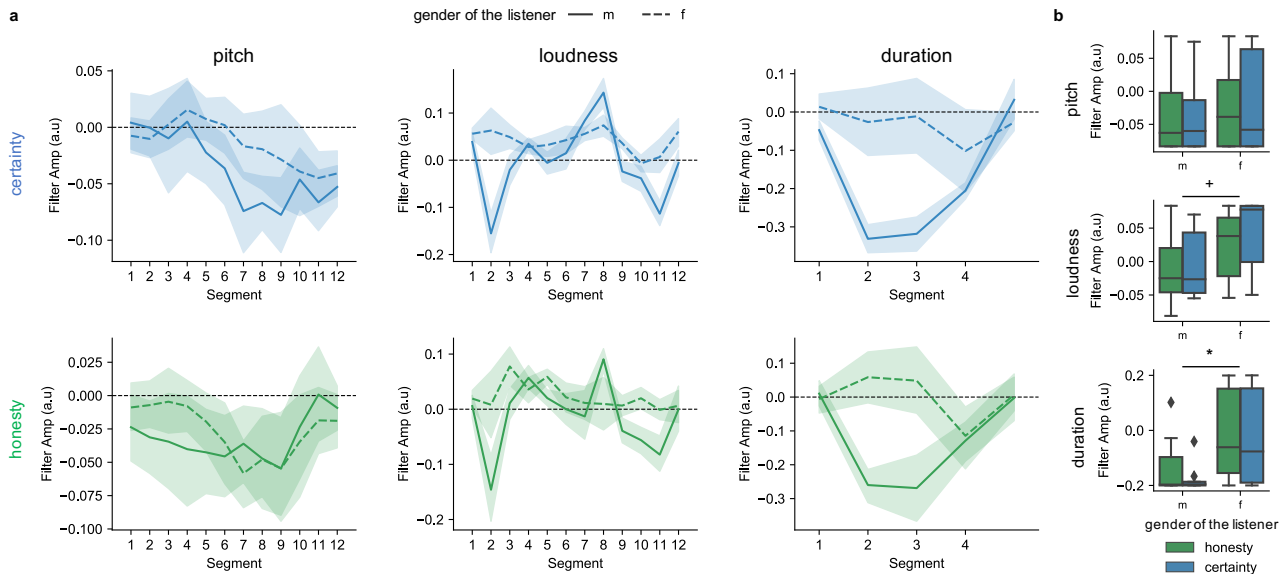
2
3 **Figure SI. Aggregated reverse correlation kernels (study 1).** Normalized kernels derived from the reverse correlation
4 analyses when combining both tasks. Filter amplitudes correspond to the values obtained for each participant, acoustic
5 dimension and segment by subtracting the average (pitch, loudness and duration) values obtained for stimuli judged as
6 certain or honest from the values averaged for the unchosen stimuli, and normalizing these values for each participant by
7 dividing them by the sum of their absolute values. Significant deviations from zero (one-sample t-tests) are indicated at the
8 corresponding segment positions with circles, with increasing sizes corresponding to $p < 0.1$; $p < 0.05$; $p < 0.01$ and $p <$
9 0.001 (p-values per segment for pitch: 0.4, 0.29, 0.53, 0.75, 0.19, 0.08, 0.006, 0.01, 0.02, 0.07, 0.03, 0.004); loudness:
10 0.001, 0.18, 0.048, 0.0002, 0.009, 0.14, 0.03, 0.0004, 0.93, 0.18, 0.02, 0.28; duration: 0.75, 0.01, 0.04, 0.003, 0.98). Shaded
11 areas show SEM. Source data are provided as a Source Data file.
12



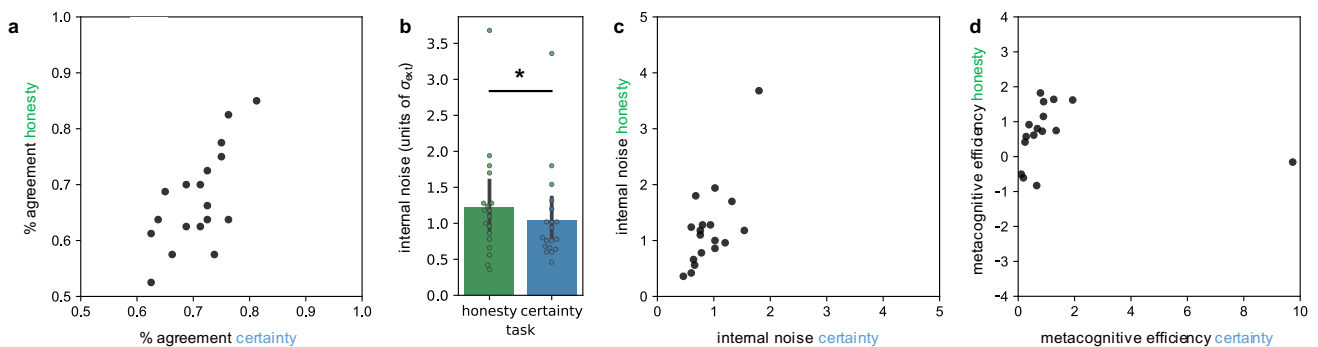
13

14 **Figure SII. Between-task intra-individual consistency despite inter-individual differences (study 1).** **A)** Individual raw
 15 reverse correlation kernels in the two tasks (top, blue: certainty; bottom, green: honesty) across the three acoustic
 16 dimensions (pitch / left, loudness / middle, duration / right). Each individual is depicted by a different shade. This figure
 17 shows non-negligible inter-individual differences. **B)** Comparison of the correlations between the kernels from the same
 18 individuals between both tasks (intra. b) with the correlations between the kernels from different individuals either
 19 computed within the same task (inter. w) or between both tasks (inter. b). Whether these analyses are conducted for the
 20 different dimensions separately (left panels, grey) or pooled together (right panel, black), they clearly show intra-individual
 21 consistency between both tasks, despite the non-negligible inter-individual differences. Dots show individual data; black
 22 asterisks represent the significance of two-sided post-hoc Tukey HSD with Bonferroni correction comparing the three types
 23 of correlations, with *** representing $p < 0.001$ (pitch: intra b vs. inter b $p = 0.00005$, intra b vs. inter w $p = 0.00005$; all
 24 other p -values < 0.00001). Source data are provided as a Source Data file.

25



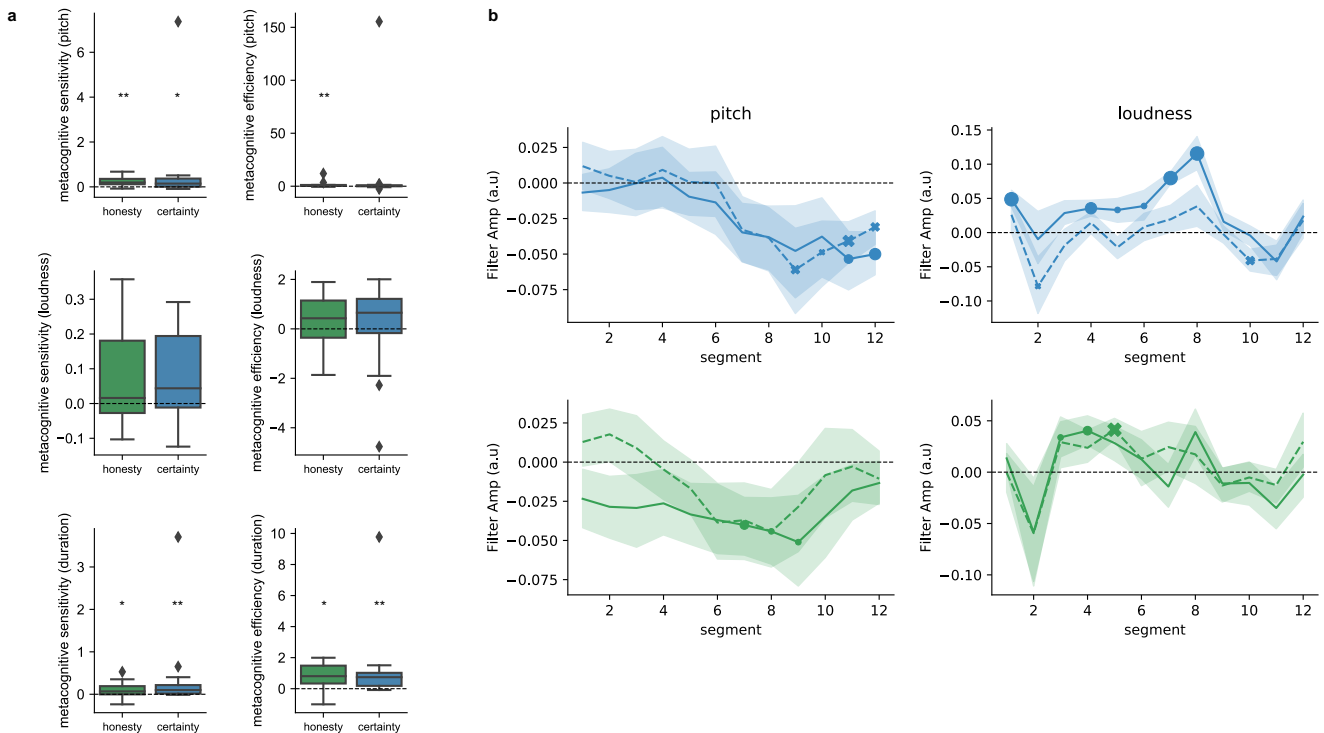
26
 27 **Figure SIII. Reverse correlation results depending on listeners' gender (study 1).** **A) Dynamic effects.** Normalized
 28 reverse-correlation kernels in the two tasks (blue: certainty; green: honesty) depending on listeners' gender (males: solid
 29 lines; females: dotted lines). Male and female listeners used acoustic dimensions in a slightly different fashion. Female
 30 listeners did not rely on duration consistently, as shown by the fact that group-level kernels were flat for this dimension
 31 (interaction between segment and gender: $F(4,68) = 2.98, p = 0.025$ in the certainty task ; $F(4,68) = 3.57, p = 0.01$ in the
 32 honesty task). An inspection of individual data suggests that this was due to idiosyncrasies in using duration rather than
 33 female listeners not relying on duration at all: some ($N = 4$) female listeners tended to judge slower prosodies as more
 34 reliable, while other ($N = 7$) female listeners behaved like male listeners (judging faster prosodies to be more reliable).
 35 Regarding loudness, there was also a significant interaction between segment and gender ($F(11,187) = 4.7, p < 0.001$ in the
 36 certainty task ; $F(4,68) = 1.85, p = 0.05$), reflecting the fact that the filters were less dynamic for female listeners: they
 37 perceived louder voices as more certain and honest (i.e., filters were shifted upwards, but flatter). Regarding pitch, there
 38 was no interaction between gender and segment for pitch kernels in none of the two tasks (certainty: $F(11,187) = 0.5, p =$
 39 0.9 ; honesty: $F(11,187) = 0.36, p = 0.97$), suggesting that both male and female listeners used intonation similarly. **B)**
 40 **Static effects.** The same data averaged across segment. Box plots show the quartiles of the dataset, with the whiskers
 41 extending to show the rest of the distribution, with the exception of outliers (i.e., data points that fall beyond 1.5 times the
 42 inter-quartile range) that are shown by diamonds. There was a main effect of gender on the amplitude of the kernels for
 43 duration ($F(1,17) = 5.6, p = 0.03$, represented by *), a marginal effect for loudness ($F(1,17) = 3.6, p = 0.076$, represented by
 44 +), and no effect for pitch ($F(1,17) = 0.2, p = 0.66$). Strikingly, although there were differences in how female and male
 45 listeners used prosodic dimensions to categorize the stimuli, these variations were reflected similarly in both tasks: there
 46 were no interactions between gender and task for any of the three dimensions (pitch: $F(1,17) = 0.7, p = 0.41$; loudness:
 47 $F(1,17) = 0.18, p = 0.68$; duration: $F(1,17) = 0.05, p = 0.82$). This is in favor of the hypothesis that listeners – despite some
 48 idiosyncratic behaviors in how they rely on specific acoustic dimensions - represent certain and honest prosodies similarly.
 49 Source data are provided as a Source Data file.
 50



51
 52 **Figure SIV. Percentage of agreement and internal noise within each task (study 1).** **A) Percentage of agreement**
 53 between responses given for trials that were presented twice (each dot corresponds to one participant). Agreement was
 54 significantly higher for certainty than honesty ($t(18) = 2.50, p = 0.022$) and was significantly correlated between the two

55 tasks (Spearman $\rho(18) = 0.69$, $p = 0.001$). **B) Internal noise** values inferred in both tasks ($N = 18$; only one out of
 56 nineteen listeners did not fall in the interpretable range, and was thus removed from the analysis; see methods). Bar plots
 57 show the mean with the 95% confidence intervals, and dots show individual data. Internal noise values ($M \pm SD$: honesty:
 58 1.2 ± 0.76 ; certainty: 0.92 ± 0.35) were within the typical range observed in most low-level psychophysical tasks, and
 59 consistent with what was recently observed in a similar high-level cognitive auditory tasks. Internal noise was lower for
 60 certainty as compared to honesty (* represents the result of a two-sided paired t-test comparing the two tasks, $t(17) = 2.23$,
 61 $p = 0.039$, $d = 0.25$), and **C)** these values were correlated between both tasks (Spearman $\rho(17) = 0.55$, $p = 0.019$). **D)**
 62 Correlation between metacognitive efficiency computed separately in the certainty and honesty task for each participant.
 63 Despite the presence of an outlier (one participant had a really high score in the certainty task, but was at zero in the
 64 honesty task), there was a significant correlation between metacognitive efficiency in the two tasks (Spearman's $\rho(15) =$
 65 0.53 , $p = 0.034$). Source data are provided as a Source Data file.

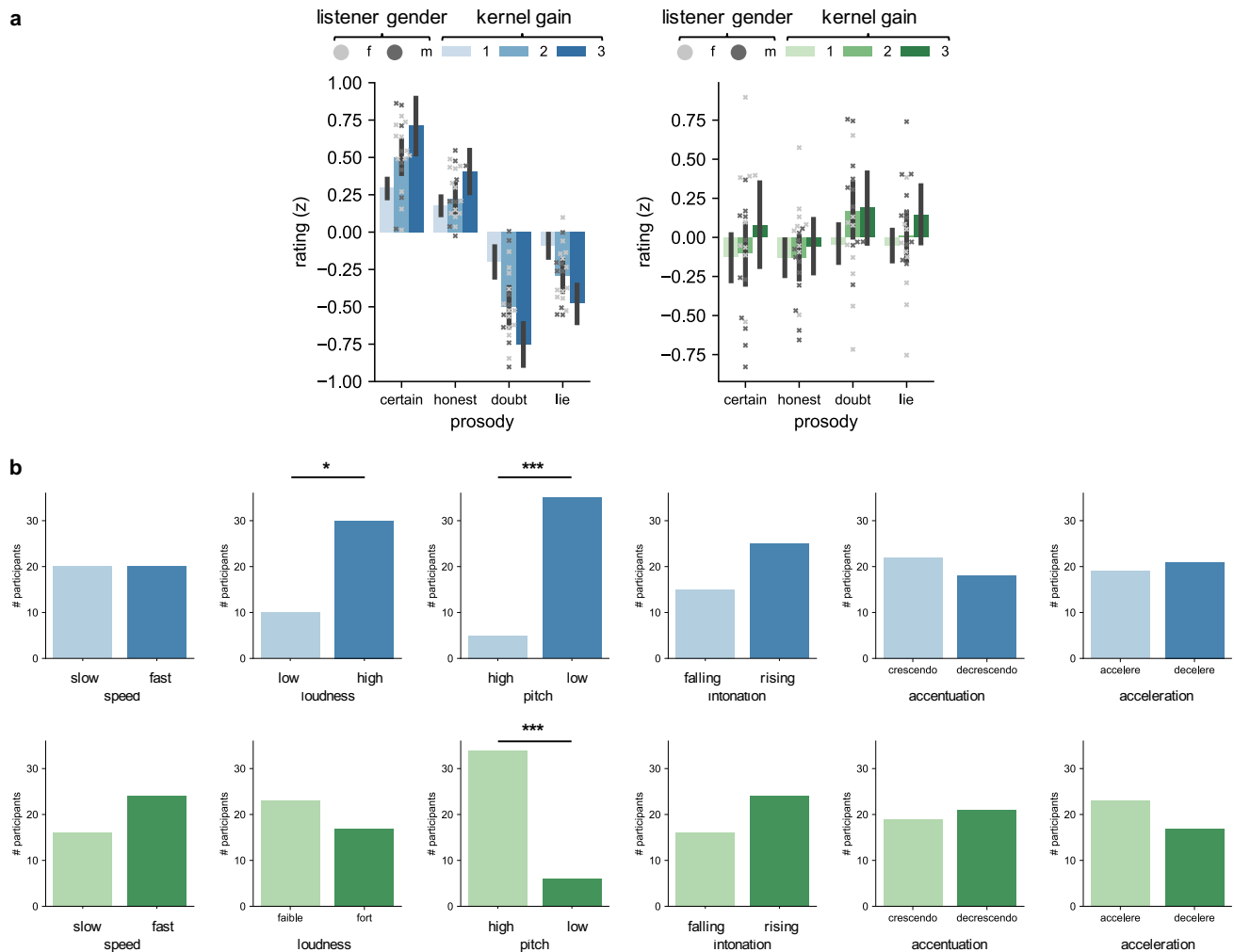
66
 67



68
 69 **Figure SV. A) Metacognitive sensitivity and efficiency computed separately for pitch, loudness and duration (study**
 70 **1).** Metacognitive sensitivity (computed by subtracting slopes of the psychometric functions constructed from high
 71 confidence responses from the slopes constructed from low confidence responses) was significantly above chance in both
 72 tasks (blue certainty; green honesty) for pitch (certainty: $Z(15) = 21$, $p = 0.015$, $d = 0.34$; honesty: $Z(15) = 7$, $p = 0.0016$, d
 73 $= 1.13$) and duration (certainty: $Z(15) = 6$, $p = 0.0013$, $d = 0.41$; honesty: $Z(15) = 26$, $p = 0.03$, $d = 0.6$), but not
 74 significantly so for loudness (certainty: $Z(15) = 32$, $p = 0.06$, $d = 0.57$; honesty: $Z(15) = 42$, $p = 0.18$, $d = 0.54$). Differences
 75 between tasks were not significant (pitch: $p = 0.5$; loudness: $p = 1$; duration: $p = 0.53$). Metacognitive efficiency
 76 (metacognitive sensitivity divided by sensitivity) was above chance for duration in both tasks (certainty: $Z(15) = 6$, $p =$
 77 0.0013 , $d = 0.53$; honesty: $Z(15) = 25$, $p = 0.026$, $d = 0.72$), and for pitch in the honesty task ($Z(15) = 10$, $p = 0.003$, $d =$
 78 0.5), but not in the certainty task ($Z(15) = 39$, $p = 0.13$, $d = 0.26$). It did not significantly differ from chance for loudness
 79 (certainty: $p > 0.3$; honesty: $p > 0.4$). Differences between tasks were not significant (pitch: $p = 0.72$; loudness: $p = 0.84$;
 80 duration: $p = 0.3$). Finally, sensitivity ($X^2 = 4.505$, $p > 0.10$), metacognitive sensitivity ($X^2 = 2.9734$, $p > 0.2$) and
 81 efficiency ($X^2 = 2.2607$, $p > 0.3$) did not significantly vary depending on acoustic dimensions in a linear mixed regression.
 82 Still, the results suggest that pitch and duration impacted confidence, while loudness did not, which may be due to the fact
 83 that this analysis does not account for dynamic effects. Box plots show the quartiles of the distribution, with the whiskers
 84 extending to show the rest of the distribution with the exception of outliers (i.e., data points that fall beyond 1.5 times the
 85 inter-quartile range) that are shown by diamonds. Black asterisks show the significance of one-sample Wilcoxon signed-
 86 ranked tests against zero, with *: $p < 0.05$; **: $p < 0.01$. **B) Reverse correlation kernels depending on confidence.** To
 87 account for dynamic effects, we also computed reverse correlation kernels separately for high (solid lines) versus low
 88 confidence (dotted lines) judgments for pitch and loudness (the two acoustic dimensions that critically varied across time)
 89 in each task (blue certainty; green honesty). A linear mixed regression was then conducted to assess the impact of
 90 confidence level, segment and task on the kernels. As reported in the main text, there was a main effect of segment for
 91 pitch (linear) and loudness (quadratic), no effect of task and no interaction between segment and task. Entering confidence

92 level in the model revealed a main effect of confidence level for pitch ($X^2 = 4.33$, $p = 0.037$, $t = -2.22$, $\beta = -0.011 \pm 0.005$ sem) reflecting the fact that pitch was lower for confident responses overall (there was no interaction between confidence level and segment nor task). For loudness, there was no significant main effect of confidence level ($X^2 = 2.6$, $p = 0.1$), but a significant interaction between confidence level and task ($X^2 = 5.7$, $p = 0.017$, $t = 2.55$, $\beta = 0.02 \pm 0.008$ sem), reflecting the fact that loudness increased with confidence in the certainty but not the honesty task. Significant deviations from chance are indicated at the corresponding segment positions by circles (high confidence) or crosses (low confidence), with increasing sizes corresponding to $p < 0.1$; $p < 0.05$; $p < 0.01$ and $p < 0.001$. Shaded areas show SEM. Source data and exact p-values for Figure SVb are provided as a Source Data file.

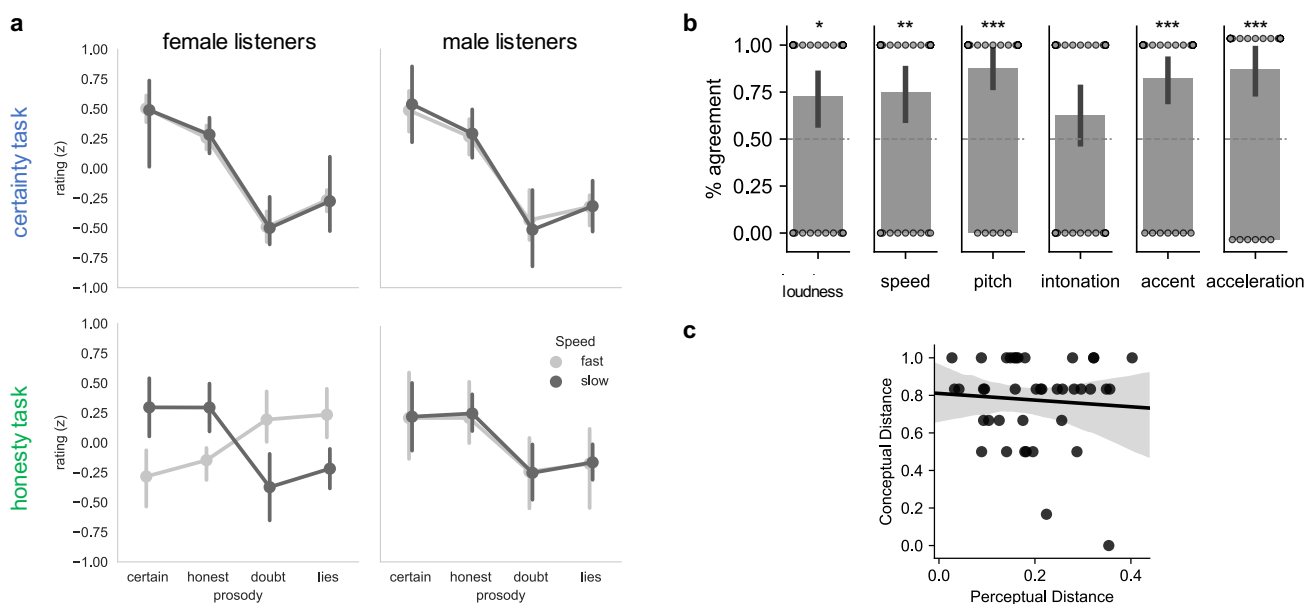
100



101

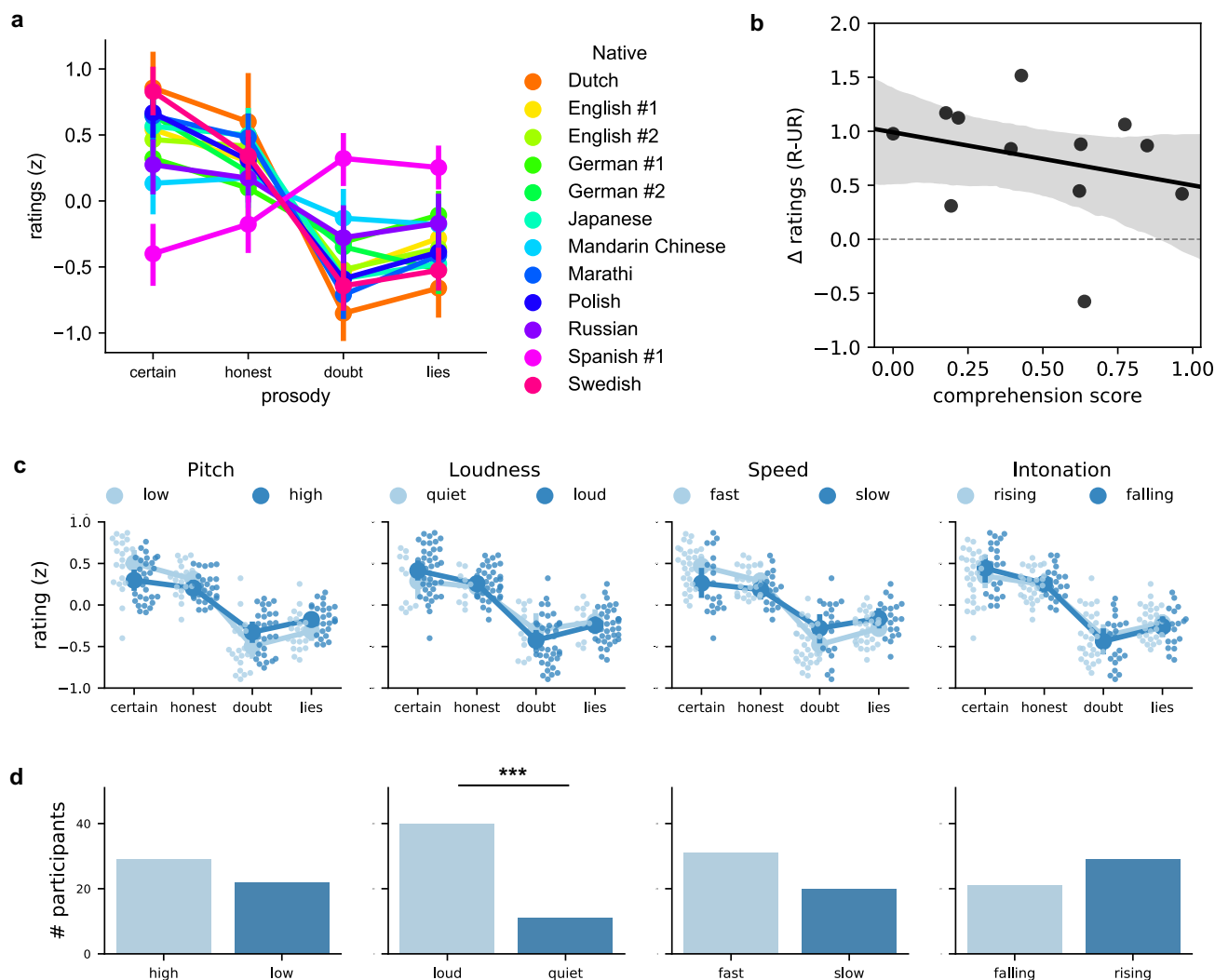
102 **Figure SVI (study 2). A) Group level results depending on gain in the two tasks.** Normalized (z-scored) ratings in the
 103 certainty (left, blue; $N = 20$) and honesty (right, green; $N = 20$) tasks depending on the gain of the acoustic transformations
 104 (represented by the three levels of shades). Bars show the mean with 95% confidence intervals, and crosses represent
 105 individual data for female (grey) and male participants (black), averaged across all levels of gain. In the confidence task,
 106 both honest and confident prosodies were judged as more confident than doubtful and lying prosodies. In the honesty task,
 107 greater inter-individual differences were observed. Certain/honest prosodies were perceived as more certain than lying
 108 prosodies (post-hoc Tukey HSD with Bonferroni correction, gain 1: $p < 0.00001/0.00001$, $d = 2.52/1.81$; gain 2: $p < 0.00001/0.00001$,
 109 $d = 3.23/2.45$; gain 3: $p < 0.00001/0.00001$, $d = 3.3/2.96$) and doubtful prosodies (gain 1: $p < 0.00001/0.00001$,
 110 $d = 2.59/2$; gain 2: $p < 0.00001/0.00001$, $d = 3.61/3.92$; gain 3: $p < 0.00001/0.00001$, $d = 3.93/3.71$) for
 111 each level of gain. In the honesty task, doubtful prosodies were perceived as more dishonest than honest prosodies for
 112 intermediate ($p = 0.0001$, $d = 0.83$) and high levels of gain ($p = 0.003$, $d = 0.54$), and more dishonest than certain prosodies
 113 for intermediate levels of gain ($p = 0.003$, $d = 0.65$; results were not significant for the lowest gain: doubt vs. certain $p = 1$;
 114 doubt vs. honest $p = 1$). By contrast, judgments were not significantly different between lying prosodies and honest or
 115 certain prosodies at the group level for any level of gains (all p-values > 0.2 , and d s < 0.52 , all other comparisons were non-
 116 significant after Bonferroni correction). Beyond confirming the hypothesis of a common signature, in the certainty task the
 117 results also show subtle differences between confidence and honesty for intermediate and high levels of gain: although
 118 honest prosodies were judged as more certain than doubtful and lying prosodies for every level of gain (see above), there

119 were still perceived as less certain than certain prosodies at high ($p < 0.00001$ and $d = 0.84$) and intermediate levels of gain
 120 ($p < 0.00001$ and $d = 1.18$). Similarly, although lying prosodies were judged as less certain than honest and certain
 121 prosodies for every level of gain (all p -values < 0.00001 and d s > 1.8), there were still judged as more certain than doubtful
 122 prosodies for high ($p < 0.00001$ and $d = 0.92$) and intermediate levels of gain ($p = 0.00004$ and $d = 0.79$). For small levels
 123 of gain, certain/honest ($p = 0.08$, $d = 0.86$) and lying/doubtful ($p = 0.93$, $d = 0.52$) prosodies were not significantly
 124 different. These subtle differences were present only in the certainty task but not in the honesty task, and are most likely
 125 due to the fact that there were differences in gain between the archetypes derived from the certainty and honesty tasks (i.e.,
 126 listeners in study 1 were less precise in the honesty task, leading to kernels with a smaller gain). **B) Conceptual knowledge**
 127 **about epistemic prosody in the group of forty French speakers.** We show the number of participants who provided each
 128 of the possible responses for each of the six questions. Chi-squared tests were used to assess whether the frequencies of the
 129 responses given to each of the six questions differed from chance level. We show the significant results with a Bonferroni
 130 corrected threshold for multiple comparisons, with *: $p < 0.008$; ***: $p < 0.0002$. The only aspects where the distribution
 131 differed significantly from chance were the questions about mean pitch for both certainty ($X=22.5$, $p < 0.00001$) and lie
 132 ($X=19.6$, $p < 0.00001$), as well as the question about mean loudness for certainty ($X=10$, $p = 0.0015$) but not lie ($X=0.9$, p
 133 $= 0.34$). The other contrasts did not differ from chance (certainty: speed $p = 1$, intonation $p = 0.11$, accentuation $p = 0.5$,
 134 acceleration $p = 0.75$; honesty: speed $p = 0.2$, intonation $p = 0.2$, accentuation $p = 0.75$, acceleration $p = 0.34$). Source data
 135 are provided as a Source Data file.
 136



137 **Figure SVII. Relationships between conceptual knowledge and gender (study 2).** **A)** Data show the normalized ratings
 138 averaged separately for each prosody, and as a function of listeners' concepts about speed (mean with 95% confidence
 139 intervals). In the honesty task, there was a triple interaction between listener's gender, prosody and concepts about speed
 140 ($F(3, 48) = 3.16$, $p = 0.03$, $\eta^2 = 0.11$), which reflected the fact that females' reports about the speed of lying archetypes
 141 were reflected in their perceptual judgments. Male listeners rated the archetypes in the canonical direction predicted by
 142 study 1 regardless of their concepts: they perceived certain/honest prosodies as more certain and more honest (all p -values
 143 < 0.001 , Bonferroni corrected Tukey HSD). By contrast, female listeners rated the archetypes as predicted in the certainty
 144 task regardless of their concepts (i.e., they perceived certain/honest prosodies as more certain, all p -values < 0.001), but
 145 most of them performed the honesty task in the opposite direction when they thought that lying prosodies are faster (i.e.,
 146 they perceived certain prosodies as more dishonest than lying and doubtful prosodies, all p -values < 0.001 , and honest
 147 prosodies as more dishonest than lying prosodies, p -value < 0.001). Thus, although male listeners perceived epistemic
 148 prosodies in a canonic way in both tasks, and regardless of their conceptual knowledge, females' ascriptions of lie in the
 149 "poker" context of study 2 depended on their conceptual knowledge about speed. This result, is consistent with the idea that
 150 this prosodic signature can be interpreted differently depending on the context, and listeners' identities and concepts. Note
 151 that no variations were found concerning the gender of the speaker (all p -values < 0.1 , words uttered by the male and the
 152 female speaker were judged similarly). Also, note that in the memory experiment (study 4), there were no differences
 153 between male and female listeners (i.e., the impact of reliability on accuracy and confidence did not interact with gender, p -
 154 values > 0.1). **B)** Percentage of agreement in response to conceptual knowledge questions about epistemic prosody across
 155 the two tasks, computed as the percentage of trials in which participants provided the same response to questions about
 156 certainty or honesty; * represents two-tailed t-tests against chance level with $p < 0.008$ (Bonferroni corrected threshold for
 157 multiple comparisons; $N = 40$ listeners); **: $p < 0.001$; ***: $p < 0.0001$; loudness: $p = 0.003$; speed: $p = 0.0009$; pitch: $p <$
 158 0.00001 ; intonation: $p = 0.11$; accent: $p < 0.00001$; acceleration: $p < 0.00001$. Bar plots show mean values with 95%

160 confidence intervals, and dots show individual data points. C) There was no correlation between conceptual distance (i.e.,
 161 the % of agreement for conceptual questions between the two tasks) and perceptual distance (i.e., the average of the
 162 absolute difference between ratings given for certain vs. honest or lying vs. doubtful prosodies (Pearson $\rho(39) = -0.07$, p
 163 $= 0.65$). Source data are provided as a Source Data file.
 164
 165



166
 167 **Figure SVIII. Perception of the common signature across language (study 3).** A) Social perceptions of certainty
 168 depending on listeners' native language for each prosodic archetype in the group of 12 native speakers of German, Dutch,
 169 English, Japanese, Mandarin Chinese, Polish, Marathi, Swedish, Spanish and Russian that were familiar with French (and
 170 English). Data show the mean ratings for each participant with 95% confidence intervals. The pattern of results was similar
 171 to the results obtained with French, English and Spanish speakers (see Figure 4): at the level of the group there was a main
 172 effect of prosody on ratings ($F(3,33) = 20$, $p < 0.001$), and listeners perceived certain and honest archetypes as more certain
 173 than doubtful and lying archetypes (all post-hoc Bonferroni corrected Tukey HSD comparisons < 0.001). They also
 174 perceived certain prosodies to be more certain than honest prosodies ($p = 0.00013$) and lying prosodies to be more certain
 175 than doubtful prosodies ($p = 0.04$). At the individual level, all but one female participant (a Spanish speaker) judged the
 176 stimuli in the congruent direction. B) Ratings as a function of listeners' level of comprehension of spoken French in the
 177 multi-language group. Dots show individual data, and the shaded area shows the best fitting regression line with 95%
 178 confidence interval. There was no significant correlation between difference scores (ratings for reliable minus unreliable
 179 archetypes) and individuals' scores in the objective test of comprehension of spoken French (Pearson $\rho = -0.27$, $p =$
 180 0.39), confirming that the prosodic signature of reliability that we identified is not language specific. There was no
 181 significant relationship between listeners' difference score and their self-reported levels of comprehension of spoken
 182 French ($\rho = -0.49$, $p > 0.1$), and French intonations ($\rho = 0.07$, $p > 0.82$). There was a negative correlation between their
 183 self-reported ability to have a basic conversation and the difference score ($\rho = -0.59$, $p = 0.043$). Overall, there was no
 184 evidence that individual' level of proficiency with the French language positively correlated with their perception of
 185 epistemic prosody from the archetypes. C) Data represent the mean ratings averaged separately depending on concepts and

186 prosodic archetypes, with error bars showing the 95% confidence interval. Dots show individual data. Although there was
187 an interaction between prosody-type and concepts about speed ($F(3,135) = 7.1, p = 0.0002$) and pitch ($F(3,135) = 4.8, p =$
188 0.003), participants perceived certain and honest archetypes to be more certain than lying and doubtful archetypes
189 regardless of their concepts (all post-hoc Bonferroni corrected Tukey HSD p -values < 0.001). These results are highly
190 consistent with the results observed in the group of French speakers, and confirms that conceptual knowledge does not
191 constrain perception in this task. **D)** Responses to conceptual questions about pitch, loudness speed and intonation in the
192 group of English ($N= 19$), Spanish ($N=20$) and multi-language ($N=12$) speakers ($N = 51$; 4 participants did not respond to
193 these questions). The bars show the number of participants who provided each of the possible responses for each of the four
194 questions. The only dimension for which participants provided consistent responses was loudness: more participants
195 reported that a certain voice is louder than a doubtful one (40 versus 11 participants, two-sided $\text{chisq} = 16.5, p < 0.0001$).
196 For the other acoustic dimensions, participants did not significantly favor one or the other option (pitch: 29 vs. 22, two-
197 sided $\text{chisq} = 0.96, p = 0.33$; speed: 31 vs. 20, two-sided $\text{chisq} = 2.37, p = 0.12$; intonation 21 vs. 29, two-sided $\text{chisq} = 1.3,$
198 $p = 0.26$; for intonation one participant was excluded because he responded “both”). Source data are provided as a Source
199 Data file.

200

201

202 **Study 4.**

203

204 To further examine how speaker reliability impacted observer’s confidence we also computed an index
205 of metacognitive sensitivity called meta- d' with the Hmeta- d' toolbox². This analysis revealed that the
206 decrease in confidence observed in the reliable condition was associated with a marginal decrease in
207 metacognitive sensitivity ($t(39) = 1.94, p = 0.059$). However, when correcting for the difference in
208 sensitivity between the two conditions by dividing meta- d' by d' (i.e., computing metacognitive
209 efficiency²), we found no significant differences between the two conditions ($t(39) = 1.4, p = 0.16$).

210

211

212 **References**

213

214 1. Ponsot, E., Burred, J. J., Belin, P. & Aucouturier, J.-J. Cracking the social code of speech
215 prosody using reverse correlation. *Proc. Natl. Acad. Sci.* 201716090 (2018).

216 2. Fleming, S. M. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from
217 confidence ratings. *Neurosci. Conscious.* **2017**, (2017).

218