



**HAL**  
open science

## Distinct signatures of subjective confidence and objective accuracy in speech prosody

Louise Goupil, Jean-Julien Aucouturier

► **To cite this version:**

Louise Goupil, Jean-Julien Aucouturier. Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, 2021, 212, pp.104661. 10.1016/j.cognition.2021.104661 . hal-03181668

**HAL Id: hal-03181668**

**<https://hal.science/hal-03181668v1>**

Submitted on 21 Apr 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33

# **Distinct signatures of subjective confidence and objective accuracy in speech prosody**

Louise Goupil<sup>1,2\*</sup> & Jean-Julien Aucouturier<sup>1</sup>

1. Laboratoire STMS, UMR 9912 (CNRS/IRCAM/SU), Paris, France.

2. University of East London, London, United Kingdom

\* corresponding author: [lougoupil@gmail.com](mailto:lougoupil@gmail.com)

34 **Abstract**

35

36 Whether speech prosody truly and naturally reflects a speaker's subjective confidence is unclear. Here,  
37 using a new approach combining psychophysics with acoustic analysis and automatic classification of verbal  
38 reports, we tease apart the contributions of sensory evidence, accuracy, and subjective confidence to speech  
39 prosody. We find that the loudness, duration and intonation of verbal reports reflect distinct underlying  
40 psychological processes. Strikingly, we show that a speaker's accuracy is encoded in speech prosody beyond  
41 their own metacognitive awareness, and that it can be automatically decoded from this information alone with  
42 performances up to sixty percent. These findings demonstrate that confidence and accuracy have separable  
43 prosodic signatures that are manifested with different timings, and on different acoustic dimensions. Thus,  
44 both subjective mental states of confidence, and objective states related to competence, can be directly inferred  
45 from natural behaviors such as speech prosody.

46

47

48

49 **keywords:** subjective confidence; speech prosody; epistemic vigilance; performance monitoring;  
50 metacognition; social cognition.

51

52

53

54

55

56

57

58

59

## 60 1. Introduction

61

62 Humans' subjective sense of confidence typically reflects an appropriate estimation of the reliability  
63 of their own beliefs and decisions (Bang & Fleming, 2018; Barthelmé & Mamassian, 2010), but whether and  
64 how this information can truly be perceived by social partners remains unclear. This is an important question  
65 because the ability to share subjective states of confidence is crucial for various aspects of human cooperation,  
66 ranging from collective decision-making to cultural transmission (Bahrami et al., 2010; Dunstone & Caldwell,  
67 2018; Heyes, 2016; Sperber et al., 2010). Past research has documented how speakers deliberately and  
68 explicitly communicate their levels of certainty, in particular through language (Aikhenvald, 2018; de Haan,  
69 2001; Fusaroli et al., 2012; Sperber et al., 2010). However, morphosyntactic markers of epistemicity greatly  
70 vary from one language to the next (Aikhenvald, 2018; de Haan, 2001; Roseano, González, Borràs-Comes, &  
71 Prieto, 2016), so such an explicit sharing of subjective confidence requires partners to engage in complex  
72 alignment and calibration processes (Bang et al., 2017; Fusaroli et al., 2012) and extensive cultural learning  
73 (Goupil & Kouider, 2019; Heyes, Bang, Shea, Frith, & Fleming, 2020).

74 It has been argued that receivers' ability to communicate and monitor senders' confidence and  
75 competence is crucial to enable cultures and languages to stabilize in the first place, because mechanisms of  
76 epistemic vigilance ensure that misinformation remains limited, and that stable conventional forms can spread  
77 (Sperber et al., 2010). If this hypothesis is correct, it is likely that basic mechanisms that do not strictly depend  
78 on language and culture should pre-exist to enable humans to detect unreliability from their social partners.  
79 This – along with findings showing that communicating states of uncertainty is highly adaptive (Bahrami et  
80 al., 2010; Dunstone & Caldwell, 2018; Heyes, 2016) and starts relatively early in life (Goupil, Romand-  
81 Monnier, & Kouider, 2016) - suggests that lower-level, more implicit mechanisms allow social partners to  
82 quickly and efficiently share their confidence, without the necessary involvement of voluntary control and  
83 communicative intentions on the side of senders.

84 Yet, whether and how observers may be able to detect subjective states of confidence directly from  
85 their partners' behavior remains unclear. Typically, human adults are able to assess their own performances,  
86 which in turn vary with sensory evidence. This means that the three constructs of sensory evidence, objective  
87 accuracy and subjective confidence tightly correlate (Bang & Fleming, 2018; Barthelmé & Mamassian, 2010).  
88 Thus, whether confidence can truly be perceived from behavior, or only indirectly inferred by observing  
89 behavioral manifestations of underlying constructs such as decision-making or perception, is not immediately  
90 clear.

91 More fundamentally, there is also considerable debate regarding whether or not confidence reduces to  
92 low-level aspects of the decision-making process (Fetsch, Kiani, Newsome, & Shadlen, 2014; Kiani &  
93 Shadlen, 2009), or rather, results from distinct higher-order, inferential processes (Fleming & Daw, 2017;  
94 Hampton, 2004; Koriat, 2012; Moulin & Souchay, 2015; Proust, 2012). In favor of this second hypothesis,  
95 dissociations between objective accuracy and subjective confidence have been observed at the level of the  
96 brain (Bang & Fleming, 2018; Cortese, Amano, Koizumi, Kawato, & Lau, 2016). Furthermore, individuals  
97 differ in their metacognitive ability to assess their own beliefs and performances (Fleming, Weil, Nagy, Dolan,  
98 & Rees, 2010; Navajas et al., 2017), and often show over-confidence biases (Moore & Healy, 2008; Zarnoth  
99 & Sniezek, 1997). Beyond inter-individual variability, specific alterations such as unconscious evidence  
100 accumulation (Vlassova, Donkin, & Pearson, 2014), stress (Reyes, Silva, Jaramillo, Rehbein, & Sackur, 2015),  
101 or targeted pharmacological interventions (Hauser et al., 2017), can lead to dissociations between performances  
102 and confidence. It is therefore important to understand whether behavioral manifestations truly reflect  
103 subjective confidence, over and beyond lower-level processes tightly linked to decision-making.

104 Yet, candidate natural behaviors that can truly convey subjective confidence, over and beyond  
105 objective performances, have so far proved surprisingly difficult to identify. Two studies examined observers'  
106 ability to rely on response times to infer others' subjective confidence, and revealed that such inferences  
107 crucially depend on an observer's own experience with a task (Koriat & Ackerman, 2010; Patel, Fleming, &  
108 Kilner, 2012). This may not be surprising given that the relationships between response times, confidence and

109 accuracy is task-dependent, varying in particular with the speed - accuracy trade off (Pleskac & Busemeyer,  
110 2010). More to the point, these results imply that response times are not a good and stable proxy for inferring  
111 subjective confidence, and that they can only be exploited to this end when observers have a first-hand  
112 experience with observees' task. Similarly, post-decision persistence times have been argued to constitute a  
113 directly observable manifestation of confidence in animals (Kepecs, Uchida, Zariwala, & Mainen, 2008) and  
114 preverbal infants (Goupil & Kouider, 2016), but other researchers contend that this measure directly reflects  
115 the strength or reliability of first-order representations rather than subjective confidence per se (Fleming &  
116 Daw, 2017; Insabato, Pannunzi, & Deco, 2016). Thus, so far, a clear behavioral signature of subjective  
117 confidence has been lacking, as research focusing on response or persistence times struggled to clearly  
118 dissociate genuine behavioral manifestations of subjective confidence from those directly tied to decision-  
119 making.

120 Here, we focus on an alternative candidate: speech prosody. It has long been suggested that prosody  
121 constitutes one of the fundamental ways through which speakers communicate their levels of confidence  
122 (Brennan & Williams, 1995; Scherer, London, & Wolf, 1973; Smith & Clark, 1993). Confident utterances are  
123 generally spoken with a falling intonation and louder volumes as compared to doubtful ones (Brennan &  
124 Williams, 1995; Jiang & Pell, 2017; Kimble & Seidel, 1991), and listeners are able to decode these prosodic  
125 cues to infer a speakers' level of uncertainty (Brennan & Williams, 1995; Goupil, Ponsot, Richardson, Reyes,  
126 & Aucouturier, n.d.; Jiang & Pell, 2017), that are seemingly preserved across languages (Chen &  
127 Gussenhoven, 2003; Goupil et al., 2020). Yet, the determinants of these prosodic manifestations of confidence  
128 in senders (that we hereafter refer to as epistemic prosody) remain unclear, for at least two reasons.

129 First, past research typically relied on methodologies in which actors are asked to deliberately produce  
130 utterances with various levels of uncertainty in social contexts. This is known to provide a distorted picture,  
131 as requesting participants to produce communicative displays leads them to produce highly stereotypical rather  
132 than genuine displays (Juslin, Laukka, & Bänziger, 2018). At a more fundamental level, measuring prosodic  
133 displays during social interactions necessarily leads to conflating the contribution of natural, automatic

134 mechanisms, and that of socially induced, deliberate self-presentation mechanisms: speakers do not only show  
135 prosodic displays automatically, they can also shape these displays deliberately, for instance in order to  
136 persuade (Van Zant & Berger, 2019) or to appear more dominant (Cheng, Tracy, Ho, & Henrich, 2016). Thus,  
137 past research leaves open the question of whether epistemic prosody is only displayed when the speaker has a  
138 communicative intention, or whether it is constitutively associated with confidence. A first step towards  
139 disentangling these influences, and investigating what these prosodic manifestations naturally mean (i.e., a  
140 behavior naturally means X when such behavior is typically associated with X; Grice, 1957; Wharton, 2009),  
141 can be to measure the relationships between confidence and prosodic features in the absence of an audience,  
142 and thus, of self-presentation and socially induced mechanisms. One previous study followed this rationale,  
143 and found that confidence impacts speakers' loudness and speech rate even in the absence of an audience  
144 (Kimble & Seidel, 1991). This questions the assumption that these prosodic signatures are primarily  
145 communicative, and suggests instead that they may reflect confidence constitutively, thereby representing  
146 natural signs that the speaker is confident. This study only measured loudness and speech rate however, so it  
147 remains unknown whether an important component of epistemic prosody, intonation, is also automatically  
148 impacted by confidence in the absence of an audience.

149         Second, typical approaches to this question do not allow discriminating the respective influence of  
150 sensory evidence, accuracy and confidence on prosody, because typically the impact of these distinct variables  
151 are not measured separately (Brennan & Williams, 1995; Dijkstra, Krahmer, & Swerts, 2006; Jiang, Gossack-  
152 Keenan, & Pell, 2020; Jiang & Pell, 2016, 2017; Kimble & Seidel, 1991; Van Zant & Berger, 2019). Thus, it  
153 remains unknown what exact psychological variable these prosodic manifestations reflect: do they reflect  
154 competence (how accurate speakers actually are), or do they genuinely reveal subjective feelings of confidence  
155 (how accurate speakers think they are), thus being akin to non-verbal variants of linguistic expressions such  
156 as "I don't know"?

157         A first possibility is that epistemic prosody truly reflects subjective feelings of confidence or doubt.  
158 Alternatively, it may be that these prosodic signatures actually reflect lower-level underlying psychological

159 processes such as cognitive effort or fluency, noise in the decision-making process, the availability of the  
160 information relevant to the current proposition being uttered (e.g., sensory evidence), or the truth value of the  
161 utterance (i.e. the objective accuracy of the speaker). If such was the case, epistemic prosody would reflect  
162 competence rather than confidence, and constitute a rather loose proxy to subjective metacognitive states.  
163 Finally, a third possibility is that different aspects of prosody (e.g., speech rate, intonation, loudness) reflect  
164 different underlying perceptual, cognitive or metacognitive processes. For instance, it may be that – as is the  
165 case in neural signals (Fleming & Dolan, 2012) – decision making impacts speech prosody earlier in time,  
166 with subjective confidence being reflected only later in the utterance. It may also be that different acoustic  
167 dimensions (e.g., loudness, intonation) reflect distinct underlying mental processes.

168 In the present study, we ask whether epistemic prosody reflects a speaker’s metacognition (i.e.,  
169 subjective confidence), cognition (i.e., accuracy/competence) or perception (e.g., the amount of sensory  
170 evidence that is available to perform a decision), and whether these distinct mental components can be  
171 separated from speech prosody alone. We also examine whether speakers’ competence (i.e., their global level  
172 of accuracy) and metacognitive sensitivity (i.e., their global ability to monitor their accuracy) modulates how  
173 confidence is reflected in their voice, thereby testing the assumptions that explicit metacognition is necessary  
174 for individuals to optimally share their confidence (Shea et al., 2014), and that epistemic prosody constitutes  
175 an efficient way to filter upcoming social information because it depends on an individual’s level of  
176 competence (or meta-competence). Finally, because we are interested in which prosodic signatures naturally  
177 reflect a speaker’s level of confidence or competence, over and beyond social influences and self-presentation  
178 effects, we test participants in isolation.

179 We address these questions by combining a psychophysical paradigm, signal detection theory, automatic  
180 classification analysis, and acoustic analysis of verbal reports produced in a non-social context. Isolated  
181 participants’ verbal responses were recorded during a visual detection task allowing to finely manipulate - and  
182 measure - sensory evidence, accuracy and confidence (see Figure 1). By analyzing the pitch, intonation,  
183 loudness, and duration of these verbal responses as a function of sensory evidence, accuracy and confidence,



184 we find that these psychological processes have distinct prosodic signatures. We then confirm this result by  
185 showing that an automatic classifier is able to decode confidence and accuracy orthogonally from speech  
186 prosody alone. Finally, we examine individual factors that modulate the automatic expression of prosodic  
187 signatures of confidence and competence.

188

189

## 190 **2. Materials and Methods**

### 191 ***2.1. Participants.***

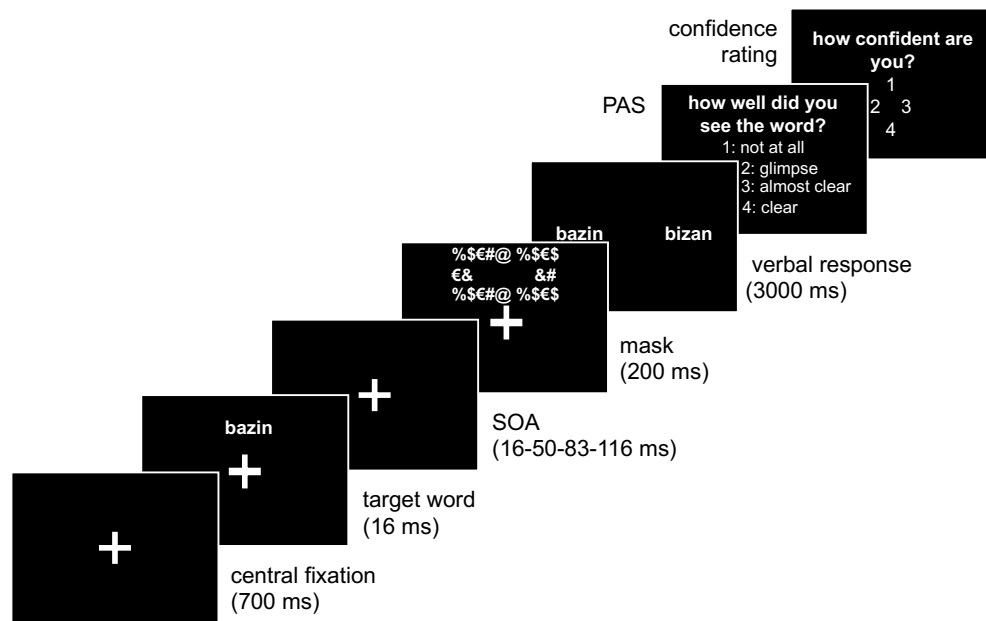
192 We tested 40 participants (21 females, mean age 22.8 +/- 3.42 SD) who had no major hearing or visual  
193 impairments. This sample size was chosen a priori based on previous studies in our group (Goupil et al., 2020;  
194 Ponsot, Burred, Belin, & Aucouturier, 2018), and given constraints associated with other experiments that  
195 were run on the same group of participants (see below). Participants signed informed consents before the study,  
196 and received a financial compensation. Out of the 40 participants, 32 were students, 4 were employees and 4  
197 were unemployed. They were from relatively healthy economic background, with 8 out of 40 participants  
198 reporting a household income below the national median; participant's family income was distributed as  
199 follows: less than 500 euros (N = 1), between 500 and 2000 euros (7), between 2000 and 5000 (N = 23), above  
200 5000 (N=6), not reported (N=3).

201

### 202 ***2.2. Procedure.***

203 Participants ran three experiments during the same session. In the first and third experiment,  
204 participants had to memorize spoken pseudo-words, and to judge whether artificially manipulated voices were  
205 more or less reliable respectively. The results from these two experiments address a different set of questions  
206 related to speakers' reliability and perception, and will thus be reported in a separate article. The second

207 experiment is the focus of the current paper. In this visual detection task, participants first saw a target bi-  
208 syllabic pseudo-word (*bazin, bizan, bivan, bavin, bodou, budou, dejon, dojen, dobue, duboe, vagio, vogia,*  
209 *vevon, voven, vizou* or *vuzoi*) that appeared for 16 ms while they were fixating a cross in the middle of the  
210 computer screen (see Figure 1). The target could appear at the top or the bottom of the screen, with  
211 equiprobable likelihoods. Targets were followed by a surrounding mask after a variable stimulus onset  
212 asynchrony (SOA: 16, 50, 83 or 116 ms) in order to induce various level of visibility, and thus, confidence in  
213 their verbal response. The mask was presented for 200 ms. Following the mask, the target word (e.g., *bazin*)  
214 and an alternative “foil” pseudo-word (e.g., *bazin, bizan*) were presented to the left or right side of the central  
215 fixation. Participants were asked to recognize the target word, and to pronounce their verbal response out loud  
216 so that it could be recorded. They then reported how well they saw the target on a perceptual awareness (PAS)  
217 scale (Ramsøy & Overgaard, 2004), and finally, their confidence in their verbal response on a scale from 1 to  
218 4. The experiment was coded in *python* with the *PsychoPy* toolbox (Peirce, 2007). The target word (16  
219 possibilities), SOA (4 possibilities), position of the response (2 possibilities: left or right) and position of the  
220 target word (2 possibilities: top or bottom) were counterbalanced within participants with a Latin square,  
221 resulting in 256 trials per participants. At the end of the session participants were asked to provide information  
222 regarding their socio-economic status: they were asked about their level of education, income and occupation,  
223 and given the fact that a majority of them were students, we also asked them to provide the same information  
224 concerning their parents. These data were aggregated to obtain a composite score of socio-economic status  
225 (SES). Participants also filled in a questionnaire assessing their level of empathy, which allows computing a  
226 general score over three dimensions measuring cognitive empathy, emotional disconnection and emotional  
227 contagion (French version of the BESA, Carré, Stefaniak, D’Ambrosio, Bensalah, & Besche-Richard, 2013).



228  
229

230 **Figure 1. Design of the verbal production task.** Participants were asked to fixate the center of the screen while a word was flashed  
 231 above or below the fixation cross for 16ms. A masked followed the presentation of the word after a variable SOA. Participants were  
 232 then asked to recognize the flashed word in between two options, before reporting upon the visibility of the flashed word on the  
 233 PAS scale, and reporting how sure they were that they pronounced the correct word on a scale from 1 to 4.

234

235 **2.3. Behavioral analysis.**

236 Unless stated otherwise, analyses were performed, and graphs obtained with *python*. Verbal responses were  
 237 identified by a coder naive to the experimental conditions. Out of the 10240 trials (256\*40 participants), 1207  
 238 (~11.8%) were excluded because the verbal response couldn't be reliable identified by the coder (e.g., because  
 239 of a problem of pronunciation), resulting in a total of 9033 verbal responses. The accuracy of participants'  
 240 verbal responses were classified as hits, misses, correct rejections or false alarms in order to compute  
 241 sensitivity, i.e., a  $d'$  (Green & Swets, 1966). Metacognitive sensitivity (meta- $d'$ ) was computed through a  
 242 hierarchical Bayesian analysis with the *Hmeta* toolbox in *Matlab* (Fleming, 2017). For each participant, a  
 243 global level of competence was also estimated by averaging their  $d'$  over the whole experiment. Confidence  
 244 bias was estimated for each participant as the average of their confidence rescaled from zero to one, to which  
 245 we subtracted their average accuracy in order to specifically estimate bias (but similar results were obtained

246 with a simple average of confidence used in previous studies running similar regression analysis, e.g.,  
247 Rollwage, Dolan, & Fleming, 2018).

248

#### 249 **2.4. Acoustic analysis.**

250 Recordings were segmented to extract isolated spoken pseudo-words. The fundamental frequency (pitch for  
251 short hereafter, in Hz) of each verbal response was then extracted in 20 successive temporal windows using  
252 *Praat*, equally dividing the duration of the recording to allow comparisons across trials and participants. Root-  
253 Mean-Square (RMS) amplitude was also computed in 20 windows, as well as word durations. Pitch and RMS  
254 profiles were then normalized for each participant, word and segment, and duration was normalized for each  
255 participant and word (z-scored). To construct the profiles shown in Figure 2, these measures were then  
256 averaged separately for each participant, each target word and each level of confidence (high: 3 and 4  
257 confidence judgments / low: 1 and 2), and the measures for confident responses were subtracted from the  
258 measures for doubtful responses. A similar analysis contrasted correct versus incorrect responses, and short  
259 (16 and 50 ms) versus long (83 and 116 ms) SOAs.

260

#### 261 **2.5. Statistics.**

262 Hierarchical linear models were run with pitch, RMS or duration as a dependent variable, and with participant  
263 and response word as random factors. Fixed factors included SOA, accuracy and confidence for duration, and  
264 SOA, accuracy, confidence and segment for pitch and loudness, in order to account for dynamic aspects.  
265 Factors were entered into the model in a hierarchical order from the lowest level (i.e., sensory, SOA) to the  
266 highest level (i.e., subjective confidence). We report beta estimates, standard errors, t-values, and p-values  
267 estimated through hierarchical model comparisons with the *lme4* and *lmerTest* packages in *R* (Kuznetsova,

268 Brockhoff, & Christensen, 2014). To account for the dynamic effect of confidence on intonation, we relied on  
269 the *MNE* package in *python* to identify significant clusters with a permutation test providing p-values corrected  
270 for multiple comparisons (Gramfort et al., 2014). The permutation test identified 3 clusters: segments 0 to 1  
271 ( $p = 0.2$ ), segments 5 to 11 ( $p = 0.012$ ) and segments 16 to 20 ( $p = 0.042$ ). Pitch was then averaged in the two  
272 significant clusters and we examined which variables (SOA, confidence, accuracy) predicted pitch in these  
273 two windows separately by running hierarchical linear regressions and mediation analysis with the *mediation*  
274 package in *R* (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014).

275 For the regression analysis presented in Figure 5, we ran three (one for each acoustic dimension) linear  
276 regressions according to the following formula: Dependent Variable (Euclidean Distance, Loudness or  
277 Duration difference score)  $\sim$  (Gender + Age + BESA + SES (composite) + Competence + Confidence Bias +  
278 Metacognitive Sensitivity) \* Measure (Accuracy or Confidence). We report Bonferroni corrected p-values to  
279 account for the fact that there were three comparisons (i.e., three acoustic dimensions). Note that similar  
280 conclusions were reached with a regression analysis involving as Dependent Variables z-scored Pitch,  
281 Duration and RMS values and testing the interaction between all factors and Confidence/Accuracy signaling,  
282 although this analysis is less sensible than the one we present here (which relies on Euclidean distance to also  
283 consider temporal aspects of intonation).

284

## 285 **2.6. Machine classification.**

286 We used two types of classification algorithms: k-nearest neighbors (kNN, Figure 4), which were run  
287 using a custom-made script, and as a confirmatory method, support-vector machines (SVM, Figure S4) with  
288 a radial basis function (RBF) kernel, which were run with the *scikit-learn* toolbox for *python* (Pedregosa et al.,  
289 2011). Both types of classifiers have been used extensively in previous research to classify vocalizations in  
290 both humans and animals (e.g., see Dezechache, Zuberbühler, Davila-Ross, & Dahl, 2019; Laukka, Neiberg, &

291 Elfenbein, 2014; Piazza, Jordan, & Lew-Williams, 2017...). The classifiers aimed to decode the confidence or  
292 the accuracy of the participants from the acoustics properties of their verbal reports, based on distances  
293 computed between their pitch, loudness and duration. For each classification method, we conducted two  
294 separate classifications for the task of estimating accuracy, and estimating confidence.

295 For the method based on k-nearest neighbors, training and testing datasets for each of the two  
296 classifications (i.e., decoding accuracy or confidence) were constructed as follows: a balanced subset of 200  
297 speech items was selected pseudo-randomly from the full dataset for each level of the other class: if accuracy  
298 was being decoded, a subset was selected for each level of confidence; if confidence was being decoded, a  
299 subset was selected for each level of accuracy. The dataset was then randomly divided in 5 folds of 40 items.  
300 This set size was chosen so as to allow crossing all combinations of accuracy, SOA and confidence to create  
301 balanced datasets (e.g., using training and testing datasets composed of 1/32 of each combinations of accuracy,  
302 confidence levels and SOAs). This led to choosing a set size of 100, as the smallest combination of all  
303 SOAs/confidence/accuracy was 29. Each fold was thus balanced to contain 50% (i.e., 20 items) of one class  
304 level (e.g., correct or high confidence) trials, and 50% of the other class level (e.g., incorrect or low  
305 confidence), as well as the same numbers of items for each level of SOA. This equiprobable combinations of  
306 conditions ensured that the classifier had to decode the class blindly with respect to the other conditions.  
307 Performances were then computed in a 5-fold cross-validation procedure, where one of the folds iteratively  
308 served as a “test set”, and the four other folds served as “training test” (Anguita, Ghio, Ridella, & Sterpi, 2009).  
309 For each items of the test set, the Euclidean distance between pitch and loudness profiles for this item, and  
310 each of the items of the training test, was computed. For duration, a simple difference was computed. For each  
311 of the three acoustic dimensions, the 5 smallest distances were then identified, and a prediction of the accuracy  
312 or confidence of the test item was made as the most frequent class amongst the nearest neighbors (five for  
313 each acoustic dimension). Classifier performance was quantified with the F-value, which is the harmonic mean  
314 of the recall and precision of the classifier. In order to allow sufficient resampling of the original dataset, the  
315 whole process was repeated and averaged over 20 iterations for each classification. Significance was then

316 assessed with a permutation procedure. For confidence decoding, confidence values were randomly reshuffled  
317 for each accuracy level and repetition (i.e., for each fold); for accuracy decoding, accuracy values were  
318 randomly reshuffled for each confidence level. Chance-level was then estimated by computing classification  
319 performance for these permuted data in the same way as in the real dataset, by computing an F-value. Real  
320 and permuted data F-values were then compared by running a rmANOVA with dataset (permuted,  
321 randomized) and condition (confidence or accuracy) as independent variables, and repetitions as a repeated  
322 measure. Finally, post-hoc differences between permuted and real data were assessed with Tukey post-hoc  
323 HSD with false-discovery rate correction for each level of confidence (or accuracy). In order to see if the  
324 results would generalize with another classification method, the same analysis was then replicated with SVMs  
325 (Figure S4).

326 All data and codes are available on the Open Science Framework (Goupil & Aucouturier, 2020).

327

328

### 329 **3. Results**

330

#### 331 *3.1. Relationship between sensory evidence, accuracy and confidence.*

332 First, we checked that our experimental paradigm was efficient in inducing various levels of confidence in  
333 our participants. A hierarchical linear regression revealed that confidence (four levels) was predicted both by  
334 SOA (beta = 0.007 +/- 0.0006 se, t = 10, p < 0.001) and accuracy (beta = 0.85 +/- 0.06 se, t = 13, p < 0.001),  
335 and that there was no interaction between these two factors (p > 0.2; see Figure S1.B. and supplementary  
336 materials for further details). The fact that confidence increased with SOA over and beyond accuracy is  
337 consistent with previous reports suggesting that confidence is also directly impacted by the visibility of the  
338 stimulus (Rausch, Hellmann, & Zehetleitner, 2018). We also computed an index of metacognitive sensitivity  
339 reflecting the extent to which participants' confidence ratings tracked the reliability of their decisions  
340 (Fleming, 2017). Meta-d' was better than chance for every SOA (all p-values < 0.001, see Figure S1.D), and

341 increased with SOA ( $F(1,39) = 74, p < 0.001, \eta p^2 = 0.65$ ), a finding that is consistent with previous research  
342 relying on similar visual paradigms (Charles, Van Opstal, Marti, & Dehaene, 2013; Kunitomo, Miller, &  
343 Pashler, 2001). Meta- $d'$  was significantly above chance for seen stimuli (glimpse:  $M = 1.36 \pm 0.88, t(39) =$   
344  $6.2, p < 0.001, \text{Cohen's } d = 1.4$ ; almost clear:  $M = 1.19 \pm 0.72, t(39) = 5.97, p < 0.001, \text{Cohen's } d = 1.35,$   
345 clear:  $M = 2.55 \pm 1.27, t(39) = 10.12, p < 0.001, \text{Cohen's } d = 2.29$ ), but it was not significantly better than  
346 chance for unseen stimuli ( $M = 0.59 \pm 1.24, t(39) = 0.46, p = 0.64, \text{Cohen's } d = 0.1$ ). This result is in line  
347 with research suggesting that metacognitive sensitivity depends on conscious access (Persaud, McLeod, &  
348 Cowey, 2007), but contrasts with other studies reporting that metacognitive sensitivity can be better than  
349 chance even for unseen stimuli (Charles et al., 2013). This may be due to the fact that we rely on verbal reports  
350 here, and this hypothesis could be specifically explored in further studies.

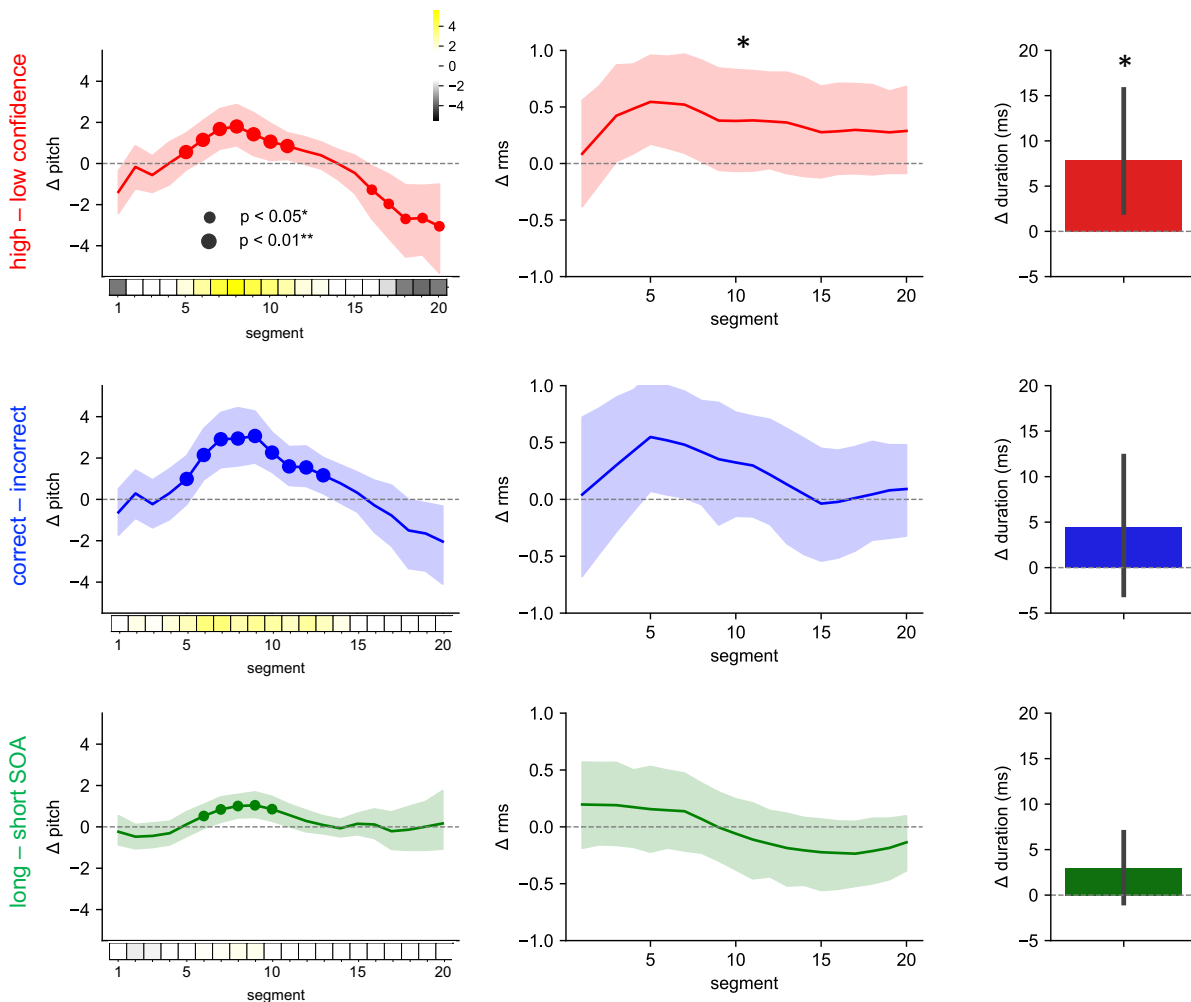
351

### 352 ***3.2. Speech prosody reflects subjective confidence, even in the absence of an audience.***

353 We then turned to the analysis of vocal productions. First, we wanted to compare the prosody of doubtful  
354 and confident responses, to confirm that prosodic markers of confidence are present in speech even in a non-  
355 social context, as expected from a previous study that only examined global loudness and speech rate (Kimble  
356 & Seidel, 1991). To this end, we extracted the duration, pitch profiles and loudness profiles of each verbal  
357 response. As can be seen in Figure 2 and Figure S2, compared to doubtful responses, confident responses were  
358 characterized by rising - falling intonation (LHL%), longer duration, and increased volume - mostly  
359 concentrated at the beginning of the word.

360





361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374

**Figure 2. Acoustic analysis of verbal responses.** Pitch, loudness (RMS) and duration values for high minus low confidence trials (1-2 versus 3-4; top – red), correct minus incorrect trials (middle – blue) and long (85-116) minus short (16-50ms) SOAs (bottom–green). Pitch: for the contrast between high and low confidence, the permutation test revealed two significant clusters: the first one ranging from the 5<sup>th</sup> to the 11<sup>th</sup> segment ( $p = 0.008$ ), and the second ranging from the 16<sup>th</sup> to the 20<sup>th</sup> segment  $p = 0.036$ ). For the contrast between correct and incorrect responses, the permutation test revealed one significant cluster ( $p = 0.002$ ) from the 5<sup>th</sup> to the 13<sup>th</sup> segment. For the contrast between high and low SOAs, the permutation test revealed one significant cluster ( $p = 0.017$ ) from the 6<sup>th</sup> to the 10<sup>th</sup> segment. RMS: the permutation test revealed no significant clusters with the threshold of  $p < 0.05$ . Circles represent the significant clusters obtained with the permutation test (small circles significance threshold of  $p < 0.05$ , bigger circles:  $p < 0.01$ ). Shaded areas and error bars show 95% confidence intervals. \* represents the significant difference between the average acoustic features of high versus low confidence responses (paired t-test, threshold of  $p < 0.05$ ). Heatmaps show the t-values of the hierarchical regression computed separately in each of the twenty temporal windows and including all three (SOA, accuracy and confidence) factors.

375  
376  
377  
378  
379  
380

Regarding mean pitch, there was no significant differences between confident and doubtful responses (mean difference in pitch =  $-0.23 \pm 2.16$ ,  $t(39) = -0.7$ ,  $p = 0.5$ , Cohen's  $d = 0.1$ ). This contrasts with previous research involving actor-produced speech (Jiang & Pell, 2017), or speakers whose intention is to persuade their interlocutors (Van Zant & Berger, 2019), that have produced discrepant findings concerning the relation between mean pitch and confidence. Our result suggests that such discrepancy may be due to focusing on mean pitch, that is likely to be associated to social traits (e.g., dominance, trustworthiness), rather than to

381 attitudes such as confidence, that are more related to dynamic aspects of pitch (i.e., intonation, Goupil et al.,  
382 n.d.; McAleer, Todorov, Belin, Taylor, & Iredell, 2014; Ponsot et al., 2018). Mean pitch may also be easier to  
383 manipulate than intonation for speakers asked to persuade or simulate confidence, which would provide a  
384 distorted picture of what “confident” prosodies naturally sound like due to social influences and self-  
385 presentation effects.

386 By contrast, as expected intonation (i.e., evolutions of the pitch over time) was impacted by confidence: a  
387 rmANOVA revealed an interaction between the level of confidence (including the full range of responses from  
388 1 to 4) and segment ( $F(1,39) = 7.3$ ,  $p = 0.013$ ,  $\eta^2 = 0.01$ ), as well as main effects of both segment ( $F(1,39) =$   
389  $4.1$ ,  $p < 0.05$ ,  $\eta^2 = 0.08$ ) and confidence level ( $F(1,39) = 5.5$ ,  $p < 0.03$ ,  $\eta^2 = 0.01$ ). As can be seen in Figure  
390 2 and S2 this interaction reflects the fact that confident responses present a rise and fall pattern, while doubtful  
391 responses present the opposite fall and rise pattern.

392 Regarding loudness, there was a static effect such that confident responses were louder than doubtful ones  
393 (mean difference =  $0.36 \pm 1$ ,  $t(39) = 2.15$ ,  $p = 0.038$ ,  $d = 0.34$ ). A rmANOVA also revealed a main effect of  
394 segment ( $F(1,39) = 183$ ,  $p < 0.001$ ,  $\eta^2 = 0.78$ ) and confidence level ( $F(1,39) = 5.25$ ,  $p < 0.03$ ,  $\eta^2 = 0.02$ )  
395 but no interaction ( $F < 1$ ), suggesting that contrary to pitch, the effect was global rather than dynamic.

396 Overall, the pattern of intonation and loudness observed in participants’ verbal productions was consistent  
397 with previous results obtained in social contexts (Brennan & Williams, 1995; Dijkstra et al., 2006; Jiang &  
398 Pell, 2017). These results confirm that these two acoustic parameters are consistent indices that can be used  
399 by listeners to infer the confidence of a speaker, and show that these prosodic manifestations of confidence  
400 are constitutively present even in the absence of an audience. The fact that loudness and duration still reflect  
401 confidence in the absence of an audience was known (Kimble & Seidel, 1991), but our results extend this  
402 finding to intonation.

403 Regarding duration, we found that confident responses were longer than doubtful responses (mean  
404 difference =  $7.85 \pm 21.4$ ,  $t(39) = 2.3$ ,  $p = 0.027$ ,  $d = 0.37$ ). This is inconsistent with previous reports that  
405 confident responses are produced with a faster speech rate (Jiang & Pell, 2017; Scherer et al., 1973), and also

406 with some results obtained in perception (Goupil et al., n.d.). Thus, like response times, speech rate may not  
 407 be a stable index enabling listeners to infer the reliability of a speaker. This is potentially due to the fact that  
 408 the relationship between response speed, accuracy and confidence greatly varies depending on task  
 409 characteristics such as the speed accuracy trade off (our task here was speeded, which would typically lead to  
 410 slower response speed for correct and confidence responses) (Pleskac & Busemeyer, 2010). Interestingly,  
 411 previous research has also shown that experience with the contingencies of a task is required to make accurate  
 412 inferences about how response times relate to confidence in others (Koriat & Ackerman, 2010; Patel et al.,  
 413 2012). In order to further elucidate the precise relationship between speech rate and confidence, further  
 414 research relying on the method that we develop here could systematically vary the speed accuracy trade-off.

415 Regardless of these fine-grained aspects, the presence of prosodic markers of confidence in the absence of  
 416 an interlocutor confirms that they constitute natural signs (Kimble & Seidel, 1991), that are present even when  
 417 speakers have no deliberate intention to communicate their uncertainty. Next, we wanted to determine what  
 418 these prosodic markers really reflect: metacognition, cognition, or perception?

419

420 ***3.3. Respective contributions of sensory evidence, accuracy and confidence to speech prosody.***

421 To this aim, we also computed differential prosodic profiles for correct versus incorrect responses, and  
 422 long versus short SOAs. As can be seen in Figure 2, we observed that both accuracy (middle row) and SOA  
 423 (bottom row) were also reflected to some extent in prosody. To elucidate whether prosody is specifically linked  
 424 to confidence or related to other underlying variables, we ran hierarchical linear mixed regressions assessing  
 425 the impact of SOA (four durations), accuracy (two levels) and confidence (four levels) on duration, loudness  
 426 and pitch (see Table 1 for the full outputs of the models).

427

428

429

time window	dependent variable	independent variable	beta	se	t	p
global	duration	SOA	0.0001	0.0003	0.37	0.71
		accuracy	0.007	0.03	-0.22	0.82
		confidence	0.035	0.01	3	0.003

		SOA:confidence	0.0004	0.0003	1.21	0.22	
		accuracy:confidence	0.03	0.027	1.31	0.19	
		SOA:accuracy	0.0008	0.0009	0.9	0.37	
	loudness	SOA	-0.0002	-0.0002	-0.92	0.36	
		accuracy	0.07	0.03	2.7	0.007	
		confidence	0.013	0.01	1.24	0.21	
		SOA:confidence	0.00001	0.0002	0.05	0.96	
	pitch	accuracy:confidence	0.0007	0.002	0.03	0.98	
		SOA:accuracy	-0.0006	-0.0008	-0.81	0.42	
		SOA	-0.0004	0.0002	-1.9	0.052	
		accuracy	0.017	0.016	1.07	0.29	
		confidence	0.08	0.008	10.7	< 0.001	
		SOA:confidence	-0.0002	-0.00006	-3.1	0.002	
		accuracy:confidence	-0.054	0.006	-8.8	< 0.001	
		SOA:accuracy	0.0004	0.0002	1.94	0.053	
	first cluster (segments 5 to 11)	SOA:segment	0.00001	0.000009	1.34	0.18	
		accuracy:segment	-0.001	0.0008	-1.63	0.1	
		confidence:segment	-0.002	0.0004	-5.53	< 0.001	
		pitch	SOA	0.0003	0.0002	1.27	0.2
			accuracy	0.06	0.025	2.4	0.016
	confidence		0.08	0.02	4.2	< 0.001	
	SOA:confidence		-0.0002	0.0002	-0.9	0.37	
	accuracy:confidence		-0.05	0.02	-2.4	0.015	
	SOA:accuracy		-0.0002	0.0006	-0.3	0.77	
	second cluster (segments 16 to 20)	SOA	-0.00006	0.0003	-0.26	0.79	
		accuracy	0.005	0.03	0.18	0.86	
		confidence	-0.03	0.01	-3	0.002	
		SOA:confidence	0.00006	0.0002	0.23	0.81	
		accuracy:confidence	-0.04	0.02	-1.94	0.052	
		SOA:accuracy	0.001	0.0008	2.02	0.044	

**Table 1.** Results of the linear mixed regressions testing the impact of SOA, accuracy and confidence on the duration, loudness and pitch of participants' verbal responses, computed in the whole 20 segments window (top) or in the two significant clusters windows (bottom; this analysis was conducted only for pitch as interactions with segments were not significant for loudness). We also report the interactions between SOA / accuracy / confidence and segments (e.g., SOA:segment), and interactions between variables (e.g., SOA:confidence). Shaded cells show significant results with the lightest shade corresponding to  $p < 0.05$  and the darkest shade to  $p < 0.001$ .

For duration, we included SOA, accuracy and confidence as fixed factors, plus interactions between these factors, and participant and target word as random factors. The regression revealed that duration was significantly predicted by confidence (beta = 0.035 +/- 0.01 se,  $t = 3$ ,  $p = 0.003$ ), but not significantly so by accuracy ( $p > 0.7$ ) and SOA ( $p > 0.8$ ) when the three covariates were present in the model. In addition, there were no significant interactions between the three acoustic dimensions (all  $p$ -values  $> 0.1$ ). Thus, overall, duration was predicted by subjective confidence rather than underlying variables, with confident responses being spoken slower than doubtful responses.

For pitch and loudness, we ran a similar model that also included interactions with segment, since these two acoustic parameters typically vary across time. Regarding loudness, there were no interactions with

446 segment (all p-values > 0.8) however, revealing that the effects were mostly non-dynamic for this acoustic  
447 dimension; we therefore reduced the model to the static model used for duration above. This static model  
448 revealed a main effect of accuracy (beta = 0.07 +/- 0.03 se, t = 2.7, p = 0.007), while the main effect of  
449 confidence (p = 0.21) and SOA (p = 0.36) were not significant when entering the three co-variates into the  
450 model. Furthermore, there were no interactions between the three variables (all p-values > 0.2). Hence, it  
451 appears that loudness primarily reflects accuracy rather than confidence per se, or sensory evidence.

452       Regarding pitch, we found a significant main effect of confidence (beta = 0.08 +/- 0.008 se, t = 10.7, p  
453 < 0.001), but the effects of accuracy (beta = 0.017 +/- 0.016 se, t = 1.07, p = 0.29) and SOA (beta = -0.0004  
454 +/- 0.0002 se, t = -1.9, p = 0.052) were not significant when entering the three co-variates into the model.  
455 Importantly, there was also a significant interaction between segment and confidence (beta = -0.002 +/- 0.0004  
456 se, t = -5.53, p < 0.001), reflecting the fact that this effect was dynamic (the interaction with segment did not  
457 reach significance for accuracy: p = 0.1, nor SOA: p = 0.18). While in low confidence trials participant's  
458 intonation presented a typical fall and rise pattern (HLH%), in high confidence trials it presented the opposite  
459 rise and fall (LHL%) pattern (see Figures 1B and S2). Finally, there was also an interaction between confidence  
460 and accuracy (beta = -0.054 +/- 0.006 se, t = -8.8, p < 0.001) and confidence and SOA (beta = -0.0002 +/-  
461 0.00006 se, t = -3.1, p < 0.01).

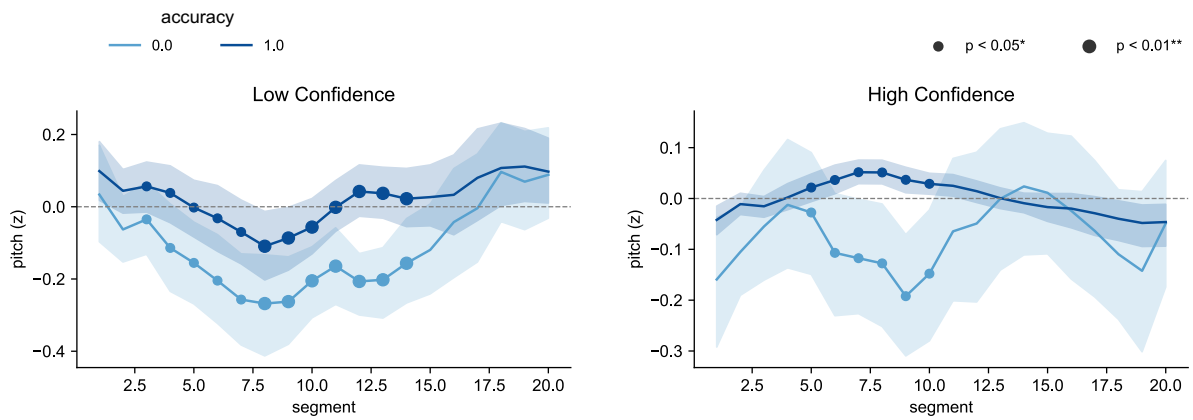
462       In order to further examine these dynamic effects, we identified significant clusters in participant's  
463 intonation by running a permutation test on the differences between confident and doubtful utterances (see  
464 methods). There were two significant clusters: the first one corresponded to segments 5 to 11 (p = 0.008) and  
465 the second one to segments 16 to 20 (p = 0.036, see Figure 2). To examine which underlying variables (SOA,  
466 accuracy or confidence) predicted pitch in these two temporal windows, we ran hierarchical regressions in the  
467 two clusters separately.

468       In the first time window, we found that – as expected – there was a highly significant effect of  
469 confidence (beta = 0.08 +/- 0.02 se, t = 4.2, p < 0.001) on pitch, but there was also a main effect of accuracy  
470 (beta = 0.06 +/- 0.025 se, t = 2.4, p = 0.016) and an interaction between confidence and accuracy (beta = -0.05

471 +/- 0.02 se,  $t = -2.4$ ,  $p = 0.015$ ), while the effect of SOA was not significant when entering all three variables  
472 in the model ( $\beta = 0.0003$  +/- 0.0002 se,  $t = 1.27$ ,  $p = 0.2$ ). In addition, a mediation analysis revealed that the  
473 effect of confidence on pitch was mediated at 12% (95% ci [-0.07, 0.30]) by accuracy in this temporal window,  
474 which was not significantly different from chance level ( $p = 0.18$ ). Confidence still had a significant direct  
475 effect after taking this mediation into account ( $p < 0.001$ ). Conversely, the effect of accuracy on pitch was  
476 partially mediated by confidence (38%, 95% ci [0.23, 0.61],  $p < 0.001$ ), but was still significant after taking  
477 this mediation into account ( $p < 0.001$ ). In the second time window, there was a main effect of confidence  
478 ( $\beta = -0.03$  +/- 0.01 se,  $t = -3$ ,  $p = 0.002$ ), but no effects of SOA ( $p > 0.7$ ) nor accuracy ( $p > 0.8$ ), and SOA  
479 and accuracy did not mediate the effect of confidence on pitch ( $p > 0.7$ ). Thus, in the beginning of the word,  
480 pitch was determined by a mixture of sensory evidence, accuracy and confidence; however, it depended  
481 exclusively on confidence towards the end of the word.

482 Strikingly, the interaction between confidence and accuracy reflected the fact that, when examining  
483 separately high and low confidence trials, intonation still reflected accuracy (Figure 3; see also Figure S3 for  
484 a detail of the four levels of confidence). In particular, when participants reported being confident in their  
485 responses, their pitch was still higher in correct trials than in incorrect trials in a temporal window ranging  
486 from the 5<sup>th</sup> to the 10<sup>th</sup> segment (see Figure 3). Similarly, when participants reported low confidence, their pitch  
487 was still higher in correct trials as compared to incorrect trials in a temporal window ranging from the 3<sup>rd</sup> to  
488 the 14<sup>th</sup> segment (corresponding to two successive significant clusters ranging from the 3<sup>rd</sup> to the 7<sup>th</sup> and 8<sup>th</sup> to  
489 the 14<sup>th</sup> segment). This analysis shows that speakers' accuracy is still manifested in their intonation, over and  
490 beyond their own metacognitive awareness.

491



492

493 **Figure 3. Intonational profiles depending on accuracy and confidence.** Normalized pitch is shown separately for low (left) versus  
 494 high (right) confidence, and accurate (dark blue) and inaccurate trials (light blue). Markers' sizes show significant clusters identified  
 495 by running a permutation test on the differences between accurate and inaccurate responses in low and high confidence trials  
 496 separately ( $p < 0.05$ : small circles;  $p < 0.01$ : big circles). For low confidence responses, the permutation test revealed two significant  
 497 clusters: the first one ranging from the 3<sup>rd</sup> to the 7<sup>th</sup> segment ( $p = 0.04$ ), and the second ranging from the 8<sup>th</sup> to the 14<sup>th</sup> segment  $p =$   
 498 0.005). For high confidence responses, the permutation test revealed one significant cluster ( $p = 0.013$ ) from the 5<sup>th</sup> to the 10<sup>th</sup> segment.  
 499 Shaded areas show the 95% confidence intervals.

500

### 501 3.4. Subjective confidence and objective accuracy can be extracted from speech prosody algorithmically

502 To further examine this dissociation, we used automatic classification algorithms to test whether  
 503 speakers' accuracy and confidence can be decoded separately from the pitch, loudness and duration of their  
 504 voice (see methods). We found that both accuracy and confidence could be separately decoded from this  
 505 information only (see Figure 4 and S4).

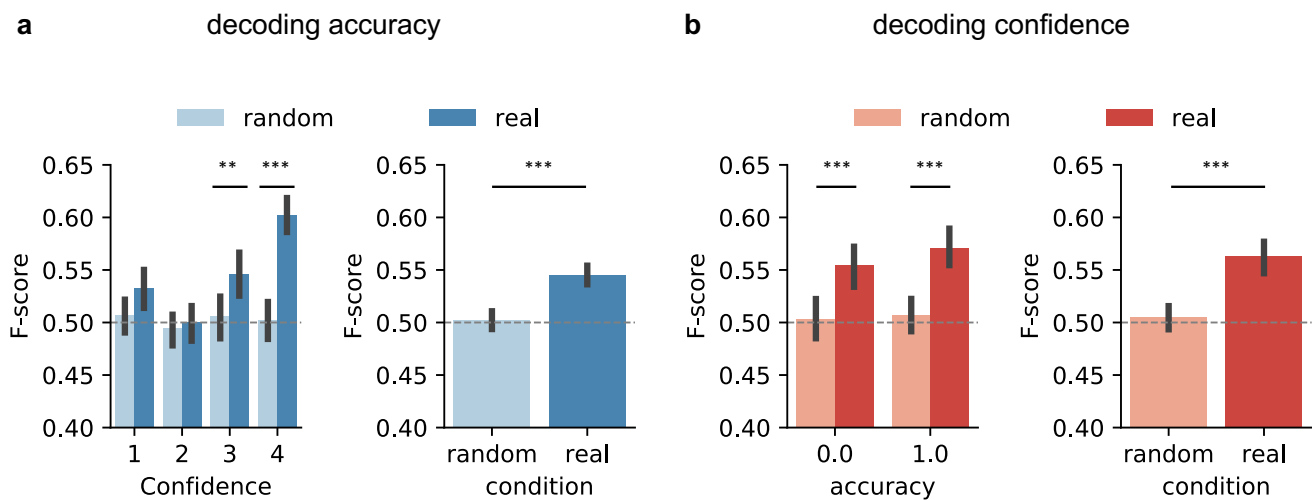
506 Machine classifiers were able to detect speakers' accuracy with a performance of 60.2% (SD = 3.7)  
 507 when they reported being 'fully confident' (rating of 4), and with a performance of 54.6 % (SD = 4.4) when  
 508 they reported being 'confident' (rating of 3). By contrast, the accuracy of the speaker could not be reliably  
 509 decoded for low levels of confidence: classification performance only reached 53.2% (SD = 4) for the lowest  
 510 level of confidence, and 50% (SD = 3.8;  $p = 0.5$ ) for the second level of confidence. To assess the significance  
 511 of this result, these classification performances in decoding accuracy were compared with classification  
 512 performances obtained with randomly permuted data (Ojala & Garriga, 2010). A rmANOVA with the  
 513 accuracy of the classifications as a dependent variable, and confidence (four levels) and dataset (real vs.  
 514 permuted) as independent variables, revealed a main effect of confidence ( $F(1,19) = 22.5$ ,  $p < 0.001$ ,  $\eta p^2 =$   
 515 0.33), a main effect of dataset ( $F(1,19) = 58.51$ ,  $p < 0.001$ ,  $\eta p^2 = 0.52$ ) and a significant interaction ( $F(1,19)$

516 = 40.81,  $p < 0.001$ ,  $\eta^2 = 0.33$ ). This interaction reflected the fact that classification performances in decoding  
517 a speaker's accuracy were significantly higher than the chance-level estimated in the permuted dataset when  
518 participants were confident (post-hoc Tukey HSD with FDR correction, confidence = 4:  $p < 0.001$ ; confidence  
519 = 3,  $p = 0.004$ ), but only marginally so for the lowest level of confidence (confidence = 1:  $p = 0.07$ ) and not  
520 significantly so for the second level (confidence = 2,  $p = 0.78$ ).

521 The confidence of the speaker could also be decoded above chance, with a performance of 55.4% in  
522 incorrect trials (SD = 4.4), and 57.1% (SD = 3.8) in correct trials. A rmANOVA with classification  
523 performances as a dependent variable, and accuracy (two levels) and dataset (real vs. permuted) as independent  
524 variables, revealed a main effect of dataset ( $F(1,19) = 60.95$ ,  $p < 0.001$ ,  $\eta^2 = 0.48$ ), no effect of accuracy  
525 ( $F(1,19) = 2.43$ ,  $p = 0.14$ ,  $\eta^2 = 0.03$ ) and no interaction ( $F(1,19) = 0.4$ ,  $p = 0.54$ ,  $\eta^2 = 0.01$ ). Classification  
526 performances in decoding speakers' confidence were significantly higher than the chance-level estimated in  
527 the permuted dataset both when participants were accurate (post-hoc Tukey HSD with FDR correction,  $p <$   
528  $0.001$ ), and when they were inaccurate ( $p < 0.001$ ).

529 Overall, this analysis confirms that the intonation, loudness and duration of a spoken utterance  
530 separately reflect accuracy and confidence, since both constructs could be decoded automatically, across all  
531 conditions in the case of confidence, and in a subset of the data (i.e., high confidence responses) for accuracy.  
532 Note that an alternative classification method (support vector machines) lead to essentially the same  
533 conclusions (see Figure S4).





534

535 **Figure 4. Results of the k-nearest-neighbors classification. A) Classifiers' performances in decoding objective accuracy for**  
 536 **each level of confidence (left), and overall (right).** To examine whether speech prosody contains enough information to  
 537 automatically infer a speaker's accuracy, we relied on a 5-fold cross-validation k-nearest neighbors (kNN) classification procedure.  
 538 Over 20 independent iterations, a balanced subset of the data was selected pseudo-randomly from the full dataset for each levels of  
 539 confidence, and divided into five folds containing 50% of correct trials, and 50% of incorrect trials (see methods for full details).  
 540 One of the folds served as a "test set", and the four other fold served as a "training set". For each items of the test set, the Euclidean  
 541 distance between the pitch and loudness profiles of this item, and the pitch and loudness profiles of each of the items of the training  
 542 test, was computed. For duration, a simple difference was computed. For each acoustic dimension, the 5 training test items with the  
 543 smallest distance to the test item were identified. The supposed accuracy of the test item was then classified as the most frequent  
 544 class amongst these fifteen nearest neighbors (five for each acoustic dimension). Finally, the classifier's performance was estimated  
 545 by computing an F-value, which is the harmonic mean of the recall and precision of the classifier (see methods). We present the F-  
 546 values averaged across the 20 repetitions. Bar plots show the average performances of the classifier for real (darker shades) and  
 547 permuted (lighter shades) data, with error bars showing the 95% confidence intervals estimated over the 20 repetitions. Dashed lines  
 548 show the theoretical chance-level (50%, black). Asterisks show the results of the post-hoc Tukey HSD with FDR correction  
 549 comparing real and permuted data allowing to estimate chance-level (see methods), with \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$   
 550 (exact p-values are reported in the main text). The chance-level estimated with permuted data was 50.2% overall (SD = 2; confidence  
 551 = 1: 50.7% (3.5); confidence = 2: 49.5% (3.3); confidence = 3: 50.6% (4.6); confidence = 4: 50.2% (4.2)). The performance of the  
 552 classifier over all confidence levels was 54.5% (SD = 2), which was highly significantly above chance level ( $t(19) = 7.65$ ,  $p < 0.001$ ).  
 553 **B) Classifiers' performances in decoding subjective confidence for each level of accuracy (left) and overall (right).** To assess  
 554 whether speech prosody contains enough information to infer a speaker's level of confidence, we applied the same method, now  
 555 decoding binary confidence (High vs. Low) for each level of accuracy and SOA (see methods). The chance-level estimated with  
 556 permuted data was 50.3% (SD = 4.2) for incorrect trials, 50.7 (3.5) for correct trials, and 50.5 (2.6) overall. The performance of the  
 557 classifier over all accuracy levels was 56.3% (SD = 3.5), which was highly significantly above chance level ( $t(19) = 7.81$ ,  $p < 0.001$ ).

558

559 **3.5. Impact of competence, confidence bias and metacognitive sensitivity on prosodic signatures of**  
 560 **confidence.**

561 Finally, we wanted to assess whether participants' ability to perform the task (their competence), their  
 562 general tendency to be confident (their confidence bias), and their global ability to evaluate their performances  
 563 (their metacognitive sensitivity) related to how accuracy and confidence were automatically reflected in their  
 564 voice. If epistemic prosody constitutes an adaptive mechanism allowing listeners to filter information coming

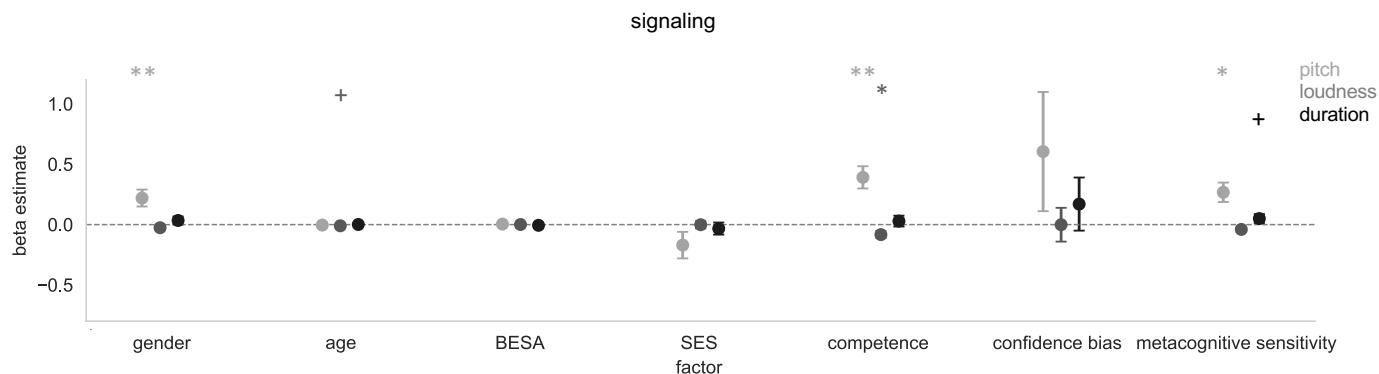
565 from unreliable social partners, we may expect that vocal signatures of accuracy and confidence may be more  
566 manifest in competent (or meta-competent) speakers.

567 To test this idea, we computed for each participant their global performances (mean  $d'$  over all trials,  
568 reflecting how competent they were in the perceptual task), their confidence bias (mean confidence over all  
569 trials corrected for performances, see methods), and their metacognitive sensitivity (approximated through  
570 meta- $d'$ , a measure that reflects how well participants confidence judgements' track their performance,  
571 independently of their general biases to be more or less confident, see methods and Fleming, 2017). We then  
572 examined how these measures related to signaling (after controlling for several other individual factors, see  
573 below), by computing three metrics that reflected the extent to which confidence and accuracy affected pitch,  
574 loudness and duration.

575 For pitch, we quantified this difference by taking the Euclidean distance between pitch profiles  
576 extracted from high versus low confidence (or correct versus incorrect) responses for each participant. For  
577 loudness and duration, we computed the mean difference between high (or correct) and low confidence (or  
578 incorrect) trials. Three linear regressions including global performance, confidence bias, metacognitive  
579 sensitivity, as well as several individual factors (gender, age, socioeconomic status, and empathic traits, see  
580 methods), and interactions between these factors and signaling type (accuracy or confidence) were then  
581 conducted separately for each acoustic dimension (see methods for the exact formula).

582 As can be seen in Figure 5, after controlling for all other factors, competence significantly predicted  
583 higher intonational signaling (beta = 0.39 +/- 0.09 se,  $t = 4.27$ , Bonferroni corrected  $p = 0.002$ ), with no  
584 significant interaction with the type of signaling (i.e., accuracy or confidence,  $p > 0.6$ ). When all other factors  
585 including competence were considered, metacognitive sensitivity also significantly predicted increased  
586 intonational signaling (beta = 0.28 +/- 0.08 se,  $t = 3.32$ ,  $p = 0.049$ , here again with no significant interaction  
587 with the type of signaling,  $p > 0.2$ ), and it also marginally increased signaling at the level of duration (beta =  
588 0.05 +/- 0.04 se,  $t = 1.315$ ,  $p = 0.053$ ). Thus, speakers' level of competence and metacognitive sensitivity in  
589 the task increased their signaling of both confidence and competence. By contrast, there were no significant

590 associations between confidence bias and any of the acoustic dimensions (all p-values > 0.1), which suggests  
 591 that individuals did not display signs of competence or confidence more or less saliently depending on their  
 592 metacognitive bias (see Figure 5 and supplementary results for details about additional effects of loudness,  
 593 age and gender).



594

595 **Figure 5. Signaling depending on individual factors.** Regression analysis were conducted on each acoustic dimension separately  
 596 to assess the impact of individual traits on signaling. Signaling for pitch corresponded to the Euclidean distance between intonational  
 597 profiles computed for high confidence (or correct responses) minus low confidence (or incorrect) responses. Signaling for loudness  
 598 and duration were computed similarly, but using average values rather than time series. Given that no interactions were observed  
 599 between factors and type of signaling (accuracy and confidence), we show combined effects. We present beta estimates, with error  
 600 bars corresponding to standard errors. + represents Bonferroni corrected  $p < 0.06$ ; \*  $p < 0.05$  and \*\*  $p < 0.01$  for the statistical  
 601 significance of each factor in the three (one for each acoustic dimension) linear regressions.

602

603

#### 604 4. Discussion

605

606 We find that, even in the absence of an audience, speech prosody automatically and separately reflects  
 607 speakers' confidence and accuracy. This finding shows that the subjective confidence and objective  
 608 competence of speakers are naturally manifested in on aspect of their behavior, thus potentially providing a  
 609 low-level, cheap mechanism for detecting whether the information they are communicating should be trusted  
 610 or not.

611 Our results reveal that intonation, loudness and duration differently reflect the underlying  
 612 psychological processes leading to the production of a verbal response. While duration and intonation reflect  
 613 confidence per se, loudness appears to be mostly driven by cognition (i.e., accuracy) rather than metacognition

614 (i.e., confidence). By revealing that various aspects of prosody are associated with different underlying  
615 psychological processes, these results go beyond previous research showing simple associations between  
616 speech and confidence, without assessing the impact and potentially mediating role of sensory evidence or  
617 accuracy.

618         Some aspects of epistemic prosody were not systematically linked to cognitive aspects presumably  
619 associated with fluency, such as sensory evidence and accuracy, but rather, truly reflected subjective aspects  
620 of experience linked to metacognition (i.e., the subjective perception of such fluency, Ackerman & Zalmanov,  
621 2012; Proust, 2012). In particular, intonation was impacted by sensory evidence and accuracy early in the  
622 word, while towards the end of the word it was exclusively determined by subjective confidence. Thus, this  
623 specific intonation pattern, in which pitch falls at the end of the word, naturally means that the speaker is  
624 confident: it is tightly linked to confidence reports per se, and present even when speakers have no deliberate  
625 intention to produce it. Interestingly, this intonation pattern finely overlaps with listeners mental  
626 representations about confident prosodies uncovered with a data driven method (Goupil et al., 2020), which is  
627 in line with our hypothesis that epistemic prosody supports a low-level adaptive mechanism of epistemic  
628 vigilance, with concurrent adaptations on the side of both senders and receivers.

629         Another interesting aspect of this result concerns timing. Intonation was found to reflect the  
630 chronometry of the mental processes used to produce an utterance: cognition is reflected in intonation before  
631 metacognition, just like it is in neural signals where correlates of perceptual and decisional processes are  
632 observable several hundreds of milliseconds before neural correlates of metacognitive processes (Fleming &  
633 Dolan, 2012). This sequence of events is thought to reflect the fact that metacognition, supported by pre-frontal  
634 regions (Bang & Fleming, 2018; Cortese et al., 2016), relies on the integration of several sources of  
635 information coming from downstream associative and perceptual areas. As such, our results are compatible  
636 with the idea that the subjective confidence expressed in explicit reports results from inferential processes that  
637 incorporate various sources of information, over and beyond processes and representations directly responsible  
638 for decisions (Fleming & Daw, 2017; Koriat, 2012; Proust, 2012).

639 We also find that other acoustic features previously associated with confidence in the literature, such  
640 as loudness, are actually not systematically linked to confidence per se, but rather, reflect the speaker's  
641 underlying accuracy. Thus, beyond offering a window into speakers' confidence, speech prosody also directly  
642 provides information about competence. Consistent with this idea, we also found that accuracy can be decoded  
643 from prosody over and beyond confidence (Figure 4). Further research should investigate whether - as is the  
644 case for confidence (Goupil et al., 2020; Jiang & Pell, 2017) - listeners are actually able to exploit these  
645 prosodic signatures to infer the accuracy of a speaker. This could be particularly important given the fact that  
646 explicit confidence reports are highly prone to biases (Moore & Healy, 2008), so being able to infer  
647 interlocutor's competence directly (i.e., without relying on their metacognitive evaluations of confidence)  
648 could be a highly adaptive solution. Notably, individuals' tendency to display their accuracy and confidence  
649 in speech prosody was not related to their confidence bias (Figure 5). Thus, compared to explicit (verbal)  
650 reports, which are highly prone to metacognitive biases, speech prosody may provide a better proxy to  
651 competence, and be less misleading to infer whether a speaker is actually right or wrong, in particular when  
652 interacting with individuals that have an overconfident (Moore & Healy, 2008; Zarnoth & Sniezek, 1997) or  
653 underconfident bias (Björkman, Juslin, & Winman, 1993; Scheck & Nelson, 2005).

654 We also find that epistemic prosody is increased in individuals who are more competent and, to a lesser  
655 extent, in individuals who have higher metacognitive sensitivity (after controlling for the impact of accuracy).  
656 Thus, individuals who are proficient in a task manifest their confidence in speech prosody more than others,  
657 even in the absence of social partners. This is consistent with the idea that epistemic prosody serves an adaptive  
658 function, enabling listeners to infer truth and certainties from proficient partners.

659 Finally, the fact that such epistemic prosodic markers were observed in the absence of an audience is  
660 consistent with past research (Kimble & Seidel, 1991), and shows that they are manifested constitutively and  
661 automatically as a function of the speaker's level of confidence and accuracy: i.e., they constitute natural signs  
662 of confidence and competence. This is not to say that these displays are never under voluntary control: humans  
663 can obviously control the pitch, duration and volume of their voice, making it possible to deliberately use

664 prosodic displays as "social tools" during conversation (Crivelli & Fridlund, 2018; Van Zant & Berger, 2019;  
665 Wharton, 2009) and past research has shown that, indeed, similar prosodic signatures as the ones we find here  
666 are exploited during communicative interactions: listeners perceive them to infer confidence and honesty in  
667 their partners (Goupil et al., 2020; Jiang & Pell, 2017), and speakers manipulate them in order to persuade  
668 their interlocutors (Van Zant & Berger, 2019). Thus, it will be important to extend our psychophysical  
669 approach to social interactions in future work, for instance by relying on dyadic collective decision-making  
670 paradigms (Bahrami et al., 2010; Fusaroli et al., 2012; Pescetelli & Yeung, 2020), in order to examine how  
671 specific social settings - such as the fact that the speaker is engaged in a cooperative or competitive interaction  
672 - impact how speakers display these prosodic signatures. A particularly interesting question is whether  
673 speakers manipulate all prosodic features (intonation, accentuation, global levels of pitch or loudness,  
674 duration), or only some of them (e.g., global levels of loudness and pitch, but not intonation). Another open  
675 question is how variations in physical attributes (e.g., body size) and social traits (e.g., social dominance)  
676 would modulate and interact with the relationships we found here between prosodic signaling and  
677 (meta)competence.

678         Beyond vocal communication, this result is to our knowledge, the first experimental demonstration  
679 that distinct features of a single observable behavior can reflect accuracy and confidence sequentially, and  
680 distinctively. Because accuracy and confidence typically correlate, there is considerable debate concerning  
681 whether or not confidence reduces to objective aspects of the decision-making process (Carruthers, 2016;  
682 Kiani & Shadlen, 2009) or rather, is tied to higher-order, integrative processes (Fleming & Daw, 2017; Koriat,  
683 2012; Moulin & Souchay, 2015). In favor of the second hypothesis, dissociations between objective accuracy  
684 and subjective confidence have been observed at the level of the brain (Bang & Fleming, 2018; Cortese et al.,  
685 2016), but whether this dissociation can also be manifested in overt behaviors, such as response time (Patel et  
686 al., 2012) or post-decision persistence, remained unclear (e.g., see Insabato et al., 2016 vs. Kepecs et al., 2008  
687 for debates concerning animals; Gliga & Southgate, 2016 vs. Goupil & Kouider, 2016 concerning preverbal  
688 children). By showing that decision-making and metacognition have different manifestations at the level of a

689 socially-observable behavior like speech prosody, our results therefore make a key theoretical contribution in  
690 support of distinguishing confidence from decision-making processes.

691

692

## 693 **5. Conclusions**

694

695 In this study, we show that individuals truly and automatically display their subjective confidence in  
696 the absence of an audience, and thus, without the necessary involvement of voluntary control and  
697 communicative intentions. Further research could examine whether this behavioral signature can be used to  
698 assess subjective confidence in pre-verbal populations (Goupil & Kouider, 2016), to discriminate confidence  
699 from accuracy in the context of forensic practices or witness testimonies (Tenney, MacCoun, Spellman, &  
700 Hastie, 2007), improve epistemic vigilance during linguistic interactions to limit the spread of fake news  
701 (Lazer et al., 2018), or as a diagnostic tool, given that explicit metacognition appears to be specifically linked  
702 to psychiatric symptoms, over and beyond the impact of task performances (Rouault, Seow, Gillan, & Fleming,  
703 2018). Beyond confidence, the present methodology of “event-related prosody”, which combines a  
704 psychophysical task with single-trial acoustic analysis, opens new avenues to investigate how subjective  
705 mental states are related to speech prosody. For instance, it is generally assumed that emotional feelings such  
706 as happiness and sadness can be directly perceived from the voice (Juslin & Laukka, 2003), but it remains  
707 unclear whether we can truly and directly perceive feelings from prosody, rather than inferring them indirectly  
708 through the perception of physiological changes typically associated with these feelings (Barrett, 2017;  
709 Galvez-Pol, Salome, Li, & Kilner, 2020).

710

711

712 **References**

- 713  
714  
715 Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency-confidence association in problem solving.  
716 *Psychonomic Bulletin and Review*, 19(6), 1187–1192. <https://doi.org/10.3758/s13423-012-0305-z>
- 717 Aikhenvald, A. (2018). *The Oxford handbook of evidentiality*.
- 718 Anguita, D., Ghio, A., Ridella, S., & Sterpi, D. (2009). *K-Fold Cross Validation for Error Rate Estimate in Support*  
719 *Vector Machines. Vessels Fuel Consumption Forecast and Trim Optimisation: a Data Analytics Perspective*  
720 *View project K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines*. Retrieved from  
721 <https://www.researchgate.net/publication/220704948>
- 722 Bahrami, B., Olsen, K., Latham, P., Roepstorff, A., Rees, G., & Frith, C. (2010). Optimally Interacting Minds.  
723 *Science*, 329(5995), 1081–1085.
- 724 Bang, D., Aitchison, L., Moran, R., Castañón, S. H., Rafiee, B., Mahmoodi, A., ... Summerfield, C. (2017).  
725 Confidence matching in group decision-making. *Nature Human Behaviour*, 1(0117), 1–7.
- 726 Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex.  
727 *Proceedings of the National Academy of Sciences of the United States of America*, 201800795.
- 728 Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and  
729 categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1–23.
- 730 Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence.  
731 *Proceedings of the National Academy of Sciences*, 107(48), 1–6. [https://doi.org/10.1073/pnas.1007704107/-](https://doi.org/10.1073/pnas.1007704107/-/DCSupplemental.www.pnas.org/cgi/)  
732 [/DCSupplemental.www.pnas.org/cgi/](https://doi.org/10.1073/pnas.1007704107/-/DCSupplemental.www.pnas.org/cgi/)
- 733 Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The  
734 underconfidence phenomenon. *Perception & Psychophysics*. <https://doi.org/10.3758/BF03206939>
- 735 Brennan, S. E., & Williams, M. (1995). The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to  
736 Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language*, 34(3), 383–398.
- 737 Carré, A., Stefaniak, N., D'Ambrosio, F., Bensalah, L., & Besche-Richard, C. (2013). The basic empathy scale in  
738 adults (BES-A): Factor structure of a revised form. *Psychological Assessment*.
- 739 Carruthers, P. (2016). Are epistemic emotions metacognitive? *Philosophical Psychology*, 1–15.
- 740 Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus  
741 subliminal error detection. *NeuroImage*, 73, 80–94.
- 742 Chen, A., & Gussenhoven, C. (2003). Language-dependence in the signalling of attitude in speech. *Proceedings of*  
743 *Workshop on the Subtle Expressivity of Emotion at CHI 2003 Conference on Human and Computer Interaction*.



- 744 Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals of dominance  
745 predict emergent social rank in humans. *Journal of Experimental Psychology: General*, *145*(5), 1–12.  
746 <https://doi.org/10.1037/xge0000166>
- 747 Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively  
748 modulates confidence without changing perceptual performance. *Nature Communications*, *7*, 13669.
- 749 Crivelli, C., & Fridlund, A. J. (2018). Facial Displays Are Tools for Social Influence. *Trends in Cognitive Sciences*,  
750 *22*(5), 388–399.
- 751 de Haan, F. (2001). The Relation Between Modality and Evidentiality. *Linguistische Berichte*.
- 752 Dezecache, G., Zuberbühler, K., Davila-Ross, M., & Dahl, C. (2019). Early vocal production and functional  
753 flexibility in wild infant chimpanzees. *BioRxiv*, 848770. <https://doi.org/10.1101/848770>
- 754 Dijkstra, C., Kraemer, E., & Swerts, M. (2006). Manipulating Uncertainty: the contribution of different audiovisual  
755 prosodic cues to the perception of confidence. In *Speech Prosody*.
- 756 Dunstone, J., & Caldwell, C. A. (2018). Cumulative culture and explicit metacognition: a review of theories,  
757 evidence and key predictions. *Palgrave Communications*, *4*(1), 145.
- 758 Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS*  
759 *Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1000302>
- 760 Fetsch, C., Kiani, R., Newsome, W., & Shadlen, M. (2014). Effects of Cortical Microstimulation on Confidence in a  
761 Perceptual Decision. *Neuron*. <https://doi.org/10.1016/j.neuron.2014.07.011>
- 762 Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for  
763 metacognitive computation. *Psychological Review*, *124*(1), 91–114.
- 764 Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the*  
765 *Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1338–1349.
- 766 Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating Introspective Accuracy to Individual  
767 Differences in Brain Structure. *Science*, *329*(5998), 1541–1543.
- 768 Fleming, Stephen M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence  
769 ratings. *Neuroscience of Consciousness*, *2017*(1).
- 770 Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms.  
771 *Psychological Science*, *23*(8), 931–939.
- 772 Galvez-Pol, A., Salome, A., Li, C., & Kilner, J. (n.d.). Direct perception of other people's heart rate.  
773 <https://doi.org/10.31234/OSF.IO/7F9PQ>

774 Gliga, T., & Southgate, V. (2016). Metacognition: Pre-verbal Infants Adapt Their Behaviour to Their Knowledge  
775 States. *Current Biology*. <https://doi.org/10.1016/j.cub.2016.09.065>

776 Goupil, L., & Aucouturier, J. J. (2020). Distinct signatures of subjective confidence and objective accuracy in speech  
777 prosody. Retrieved October 20, 2020, from <https://osf.io/xegfv/>

778 Goupil, Louise, & Kouider, S. (2016). Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal  
779 Infants. *Current Biology*, *26*(22), 3038–3045.

780 Goupil, Louise, & Kouider, S. (2019). Developing a Reflective Mind: From Core Metacognition to Explicit Self-  
781 Reflection. *Current Directions in Psychological Science*, *8*(4). <https://doi.org/10.1177/0963721419848672>

782 Goupil, Louise, Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (n.d.). Hearing reliability: a shared  
783 prosodic code automatically signals confidence and honesty to human listeners. *In Prep*.

784 Goupil, Louise, Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know.  
785 *Proceedings of the National Academy of Sciences of the United States of America*, *113*(13), 3492–3496.

786 Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2014).  
787 MNE software for processing MEG and EEG data. *NeuroImage*.  
788 <https://doi.org/10.1016/j.neuroimage.2013.10.027>

789 Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York Wiley (Vol. 4054).

790 Grice, H. P. (1957). *Meaning*. *Philosophical Review*. Philosophical Review.

791 Hampton, R. R. (2004). Metacognition as Evidence for Explicit Representation in Nonhumans. *Behavioral and Brain*  
792 *Sciences*, *26*(3), 346–347. Retrieved from <http://dx.doi.org/10.1017/S0140525X03300081>

793 Hauser, T. U., Allen, M., Purg, N., Moutoussis, M., Rees, G., & Dolan, R. J. (2017). Noradrenaline blockade  
794 specifically enhances metacognitive performance. *ELife*, *6*, e24901.

795 Henton, C. G. (1989). Fact and fiction in the description of female and male pitch. *Language and Communication*.  
796 [https://doi.org/10.1016/0271-5309\(89\)90026-8](https://doi.org/10.1016/0271-5309(89)90026-8)

797 Heyes, C. (2016). Who Knows? Metacognitive Social Learning Strategies. *Trends in Cognitive Sciences*.

798 Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural  
799 Origins of Metacognition. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2020.02.007>

800 Insabato, A., Pannunzi, M., & Deco, G. (2016). Neural correlates of metacognition: A critical perspective on current  
801 tasks. *Neuroscience and Biobehavioral Reviews*.

802 Jiang, X., Gossack-Keenan, K., & Pell, M. D. (2020). To believe or not to believe? How voice and accent  
803 information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology (2006)*,

804 73(1), 55–79.

805 Jiang, X., & Pell, M. D. (2016). Neural responses towards a speaker’s feeling of (un)knowing. *Neuropsychologia*, 81,  
806 79–93. <https://doi.org/10.1016/j.neuropsychologia.2015.12.008>

807 Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, 88, 106–126.

808 Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance:  
809 Different channels, same code? *Psychological Bulletin*, 129(5), 770–814. [https://doi.org/10.1037/0033-](https://doi.org/10.1037/0033-2909.129.5.770)  
810 2909.129.5.770

811 Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The Mirror to Our Soul? Comparisons of Spontaneous and Posed  
812 Vocal Expression of Emotion. *Journal of Nonverbal Behavior*. <https://doi.org/10.1007/s10919-017-0268-x>

813 Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural  
814 impact of decision confidence. *Nature*, 455(7210), 227–231.

815 Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the  
816 parietal cortex. *Science*, 324(5928), 759–764.

817 Kimble, C. E., & Seidel, S. D. (1991). Vocal signs of confidence. *Journal of Nonverbal Behavior*, 15(2), 99–105.  
818 <https://doi.org/10.1007/BF00998265>

819 Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*.  
820 <https://doi.org/10.1037/a0025648>

821 Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: judgments of learning for Self and Other during  
822 self-paced study. *Consciousness and Cognition*, 19(1), 251–264.

823 Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses.  
824 *Consciousness and Cognition*, 10(3), 294–340.

825 Kuznetsova, A., Brockhoff, P. B., & Christensen, H. B. (2014). lmerTest: Tests for random and fixed effects for  
826 linear mixed effect models (lmer objects of lme4 package). *R*.

827 Laukka, P., Neiberg, D., & Effenbein, H. A. (2014). Evidence for cultural dialects in vocal emotion expression:  
828 Acoustic classification within and across five nations. *Emotion*. <https://doi.org/10.1037/a0036048>

829 Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018).  
830 The science of fake news. *Science*.

831 Lundeberg, M. A., Fox, P. W., & Puncoha, J. (1994). Highly confident but wrong: Gender differences and similarities  
832 in confidence judgments. *Journal of Educational Psychology*, 86(1), 114–121.

833 McAleer, P., Todorov, A., Belin, P., Taylor, L., & Iredell, N. (2014). How do you say “hello”? Personality

834 impressions from brief novel voices. *PLoS ONE*, 9(3), e90779. <https://doi.org/10.1371/journal.pone.0090779>

835 Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502–517.

836 Moulin, C., & Souchay, C. (2015). An active inference and epistemic value view of metacognition. *Cognitive*  
837 *Neuroscience*, 6(4), 221–222. <https://doi.org/10.1080/17588928.2015.1051015>

838 Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of  
839 confidence. *Nature Human Behaviour*, 1. <https://doi.org/10.1038/s41562-017-0215-1>

840 Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine*  
841 *Learning Research*.

842 Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions.  
843 *Proceedings of the Royal Society B: Biological Sciences*.

844 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-  
845 learn: Machine learning in Python. *Journal of Machine Learning Research*.

846 Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–  
847 13.

848 Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature*  
849 *Neuroscience*, 10(2), 257–261.

850 Pescetelli, N., & Yeung, N. (2020). The effects of recursive communication dynamics on belief updating.  
851 *Proceedings of the Royal Society B: Biological Sciences*, 287(1931), 20200025.  
852 <https://doi.org/10.1098/rspb.2020.0025>

853 Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers Consistently Alter Their Unique Vocal  
854 Fingerprints When Communicating with Infants. *Current Biology : CB*, 3162-3167.e3.  
855 <https://doi.org/10.1016/j.cub.2017.08.074>

856 Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and  
857 confidence. *Psychological Review*, 117(3), 864–901.

858 Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using  
859 reverse correlation. *Proceedings of the National Academy of Sciences*, 201716090.

860 Proust, J. (2012). Metacognition and mindreading: one or two functions? In M. J. Beran, J. L. Brandl, J. Perner, & J.  
861 Proust (Eds.), *Foundations of Metacognition* (pp. 234–251). Oxford, UK: Oxford University Press.

862 Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive*  
863 *Sciences*. <https://doi.org/10.1023/b:phen.0000041900.30172.e8>

- 864 Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by  
865 both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154.
- 866 Reyes, G., Silva, J. R., Jaramillo, K., Rehbein, L., & Sackur, J. (2015). Self-Knowledge Dim-Out: Stress Impairs  
867 Metacognitive Accuracy. *PLOS ONE*, *10*(8), e0132320.
- 868 Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2016). Communicating Epistemic Stance: How Speech  
869 and Gesture Patterns Reflect Epistemicity and Evidentiality. *Discourse Processes*.
- 870 Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated  
871 With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*.  
872 <https://doi.org/10.1016/j.biopsych.2017.12.017>
- 873 Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary  
874 conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*.  
875 <https://doi.org/10.1037/0096-3445.134.1.124>
- 876 Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience  
877 evaluation. *Journal of Research in Personality*, *7*(1), 31–44. [https://doi.org/10.1016/0092-6566\(73\)90030-5](https://doi.org/10.1016/0092-6566(73)90030-5)
- 878 Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and  
879 metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.
- 880 Smith, V. L., & Clark, H. H. (1993). On the Course of Answering Questions. *Journal of Memory and Language*,  
881 *32*(1), 25–38.
- 882 Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G., & Wilson, D. (2010). Epistemic vigilance.  
883 *Mind and Language*, *25*(4), 359–393.
- 884 Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration Trumps Confidence as a Basis for  
885 Witness Credibility. *Psychological Science*, *18*(1), 46–50.
- 886 Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation  
887 analysis. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v059.i05>
- 888 Van Zant, A. B., & Berger, J. (2019). How the Voice Persuades. *Journal of Personality and Social Psychology*,  
889 *118*(4), 661–682.
- 890 Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not  
891 confidence. *Proceedings of the National Academy of Sciences*, *111*(45), 16214–16218.  
892 <https://doi.org/10.1073/pnas.1403619111>
- 893 Wharton, T. (2009). *Pragmatics and Non-Verbal Communication*. Cambridge: Cambridge University Press.

894 Zarnoth, P., & Sniezek, J. A. (1997). The Social Influence of Confidence in Group Decision Making. *Journal of*  
895 *Experimental Social Psychology*, 33(4), 345–366.

896

897

898 **Acknowledgements.**

899 The authors thank Gabriel Vogel, Louise Vasa and Lou Seropian for their help with collecting and coding  
900 the data, as well as Emmanuel Ponsot and Pablo Arias for their input regarding data collection and analysis.  
901 Ethical approval was obtained, and experimental data were collected at INSEAD/ Sorbonne University  
902 Center for Behavioural Science. This work was supported by ERC StG CREAM 335536 to J.J.A., and a  
903 Marie Slodowska Curie H2020 grant (845859, JDIL) to L.G.

904

905 **Authors contributions.** L.G., and J.J.A. designed the experiment. L.G. collected, and analyzed the data. L.G.  
906 wrote the paper with comments from J.J.A.

907

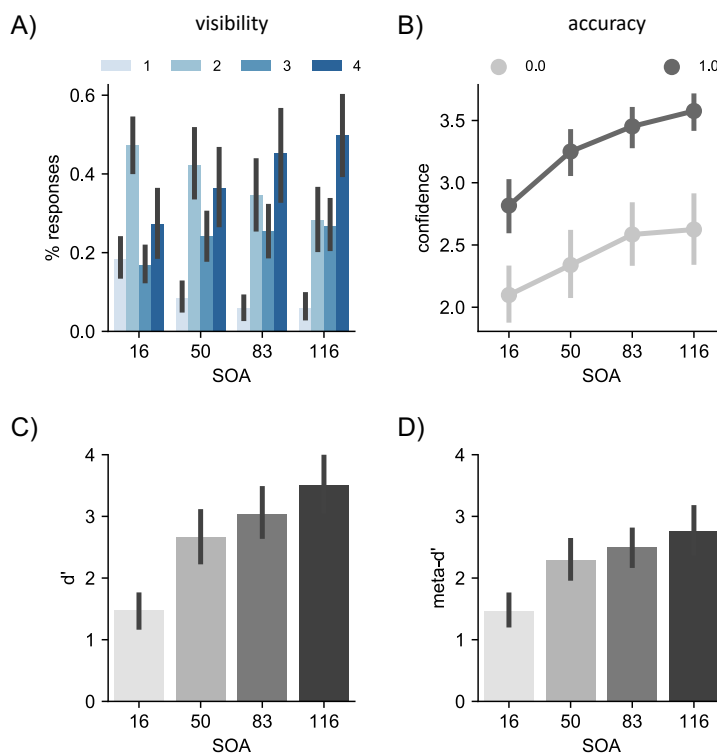
908 **Competing interests.** The authors declare that there is no conflict of interest regarding the publication of this  
909 article.

910

911 **Supplementary Materials**

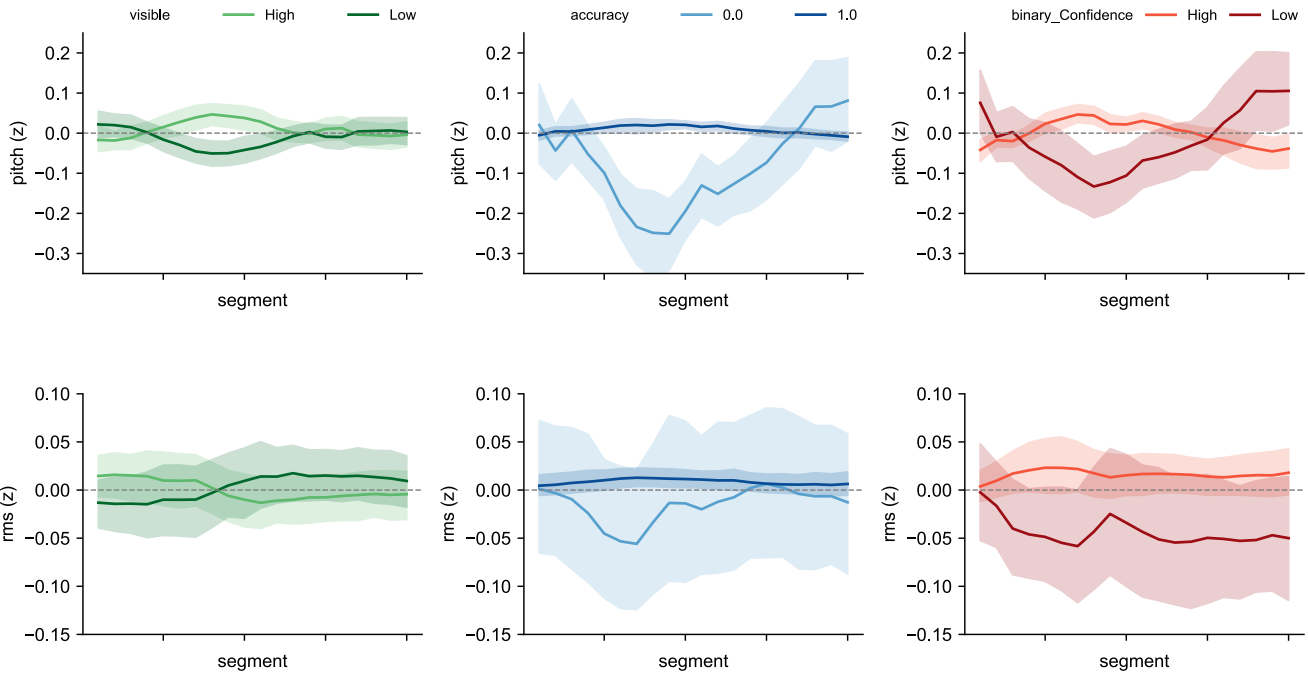
912

913 **Behavioral results.** As expected from previous research relying on similar visual paradigms (Charles et al., 2013;  
914 Kunimoto et al., 2001; Rausch et al., 2018), both visibility ( $F(1,39) = 103, p < 0.001, \eta^2 = 0.72$ ), sensitivity (i.e.,  $d'$ ,  
915  $F(1,39) = 169, p < 0.001, \eta^2 = 0.81$ ) and confidence ( $F(1,39) = 116, p < 0.001, \eta^2 = 0.74$ ) increased with SOA. As  
916 can be seen in Figure S1.A. below, at the shortest SOA participants rarely reported not seeing anything at all, but often  
917 reported seeing only a glimpse of the stimulus. Also of note is the fact that sensitivity ( $d'$ ) remained above chance level  
918 even for the shortest SOA ( $M = 1.48 \pm 0.8, t(39) = 7.7, p < 0.001$ ) but not for unseen stimuli ( $M = 0.84 \pm 2.41, t(39)$   
919  $= 0.87, p = 0.38$ ). This pattern of result contrasts with previous findings showing that objective performances can be  
920 better than chance even for unseen stimuli (Charles et al., 2013; Kunimoto et al., 2001). This could be due to the fact  
921 that we rely on verbal reports here, rather than less ecological task involving poorly demanding motor responses such  
922 as button presses.



923

924 **Figure S1. Behavioral results.** A) **Visibility ratings depending on SOA.** The percentage of trials was computed for  
925 each level of visibility depending on SOA, and averaged across participants. B) Confidence was averaged for each  
926 participant depending on accuracy. B) Sensitivity ( $d'$ ) was computed for each SOA. D) Metacognitive sensitivity  
927 (meta- $d'$ ) was computed for each SOA. Error bars show the 95% confidence interval.

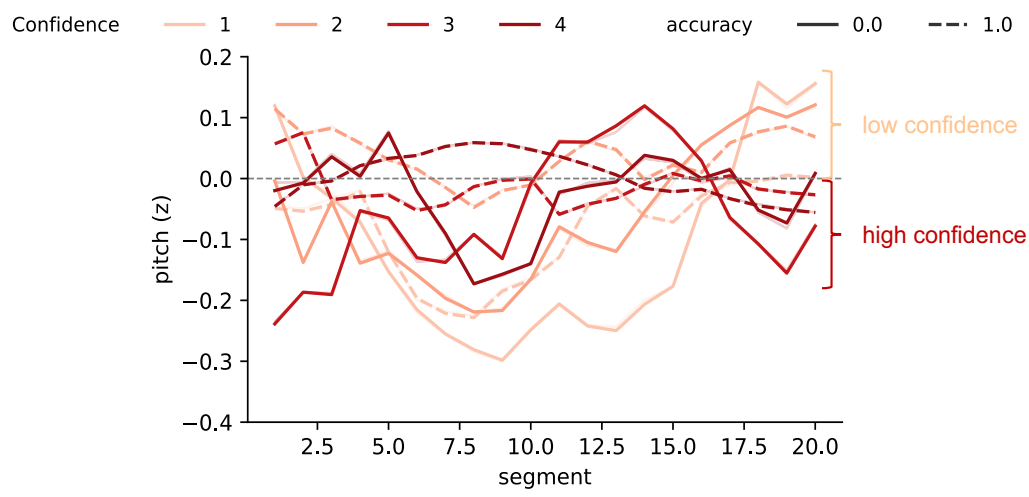


928

929 **Figure S2.** Normalized pitch (top) and RMS (bottom) are shown for each segment, depending on SOA (left / green, low:  
 930 16 and 50ms versus high: 83 and 116ms), accuracy (middle / blue) and confidence (right / red). Error bar show the 95%  
 931 confidence intervals.

932

933



934

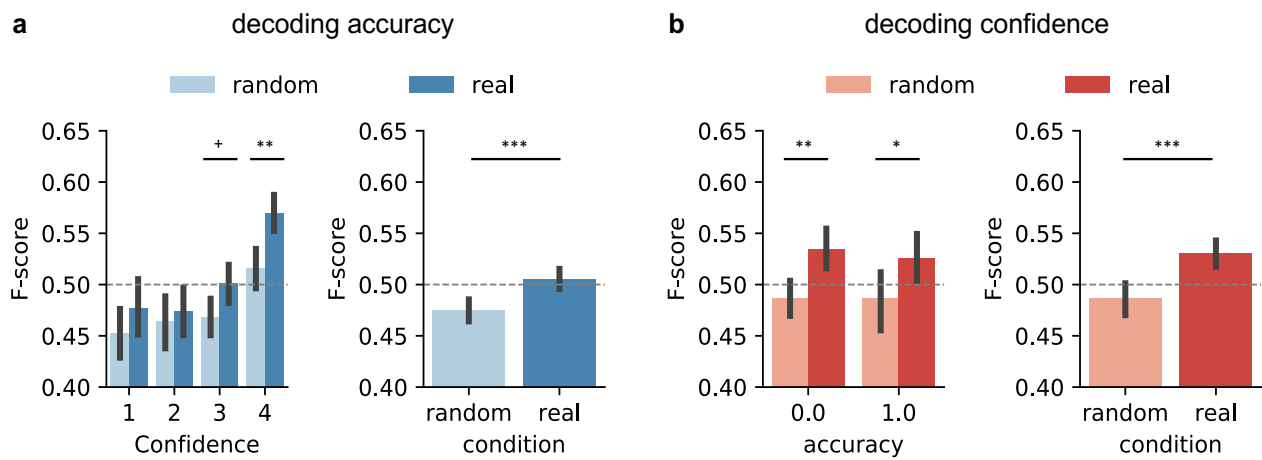
935

936 **Figure S3.** Normalized pitch is shown separately for each level of confidence and accurate (dashed lines) and  
 937 inaccurate trials (plain lines).

938

939





940

941 **Figure S4. Results of the support-vector machines classification. A) Decoding objective accuracy.** The same  
 942 analysis presented in Figure 4 was repeated with an alternative classification procedure (support vector machines,  
 943 SVMs). We present the F-values averaged across the 20 repetitions. Bar plots show the average performances of the  
 944 classifier for real (darker shades) and permuted (lighter shades) data, with error bars showing the 95% confidence  
 945 intervals estimated over the 20 repetitions. The chance-level estimated with permuted data (see methods) was 47.5%  
 946 (SD = 2.6) overall (confidence = 1: 45.2 (5.7); confidence = 2: 46.4% (5.8); confidence = 3: 46.8% (4); confidence = 4:  
 947 51.5% (4.4)). The performance of the classifier over all confidence levels was 50.5% (SD = 2.3), which was highly  
 948 significantly above the chance level estimated with permuted data ( $t(19) = 5.58, p < 0.001$ ). As for KNNs, a rmANOVA  
 949 revealed a significant main effect of condition (real vs. permuted,  $F(1,19) = 31.2, p < 0.001, \eta^2 = 0.27$ ), and a main  
 950 effect of confidence ( $F(1,19) = 47.4, p < 0.001, \eta^2 = 0.53$ ) and a marginal interaction between the two factors ( $F(1,19)$   
 951  $= 3.3, p = 0.08, \eta^2 = 0.053$ ). Performances were higher in the dataset as compared to permuted data when participants  
 952 were highly confident (confidence 4:  $p = 0.002$ , post-hoc Tukey HSD with FDR correction), but only marginally so for  
 953 confidence = 3 ( $p = 0.068$ ), and not significantly so for lower levels of confidence (confidence 1:  $p = 0.17$ ; confidence  
 954 2:  $p = 0.62$ ). Asterisks show the results of the post-hoc Tukey HSD with FDR correction comparing classification  
 955 performances with the chance-level estimated with permuted data, with +  $p < 0.07, * p < 0.05, ** p < 0.01, *** p <$   
 956  $0.001$  (see main text for exact p-values). Dashed lines show the theoretical chance-level (50%, black). **B) Decoding**  
 957 **subjective confidence.** To assess whether speech prosody contains enough information to infer a speaker's level of  
 958 confidence, we applied the same method, now decoding binary confidence (High vs. Low) for each level of accuracy  
 959 and SOA (see methods). The chance-level estimated with permuted data was 48.7.3% (SD = 3.9) for incorrect trials,  
 960 48.6 (6.3) for correct trials, and 48.6 (3.5) overall. The performance of the classifier over all accuracy levels was 53%  
 961 (SD = 2.9), which was highly significantly above chance level ( $t(19) = 6.9, p < 0.001$ ). As for KNNs, a rmANOVA  
 962 revealed a significant main effect of condition (real vs. permuted,  $F(1,19) = 47.74, p < 0.001, \eta^2 = 0.21$ ), no effect of  
 963 accuracy ( $F(1,19) = 0.08, p = 0.76, \eta^2 = 0.003$ ) and no interaction ( $F(1,19) = 0.29, p = 0.59, \eta^2 = 0.002$ ). Performances  
 964 were higher in the dataset as compared to permuted data for both levels of accuracy (correct:  $p = 0.02$ ; incorrect  $p =$   
 965  $0.006$ ; post-hoc Tukey HSD with FDR correction).

966

967

968

969

970

971

972

973

974

**Relationship between individual factors and signaling.**

Beyond the effects related to our main claims reported in the manuscript, we also observed that, at the level of loudness,  
 competence significantly decreased signaling ( $\beta = -0.08 \pm 0.03 \text{ se}, t = -2.94, p = 0.014$ ), mirroring the positive  
 impact observed for duration (there was a negative correlation between loudness signaling and duration signaling:  $\rho$   
 $= -0.35, p = 0.026$ ; and more generally between duration and volume, participants spoke louder when they responded

975 faster overall,  $\rho = -0.27$ ,  $p < 0.001$ ). Age also marginally decreased signaling at the level of loudness (beta =  $-0.01 \pm$   
976  $0.006$  se,  $t = -1.75$ ,  $p = 0.056$ ), but this impact of age is difficult to interpret given the short range included in our study  
977 (18- to 30-year-olds). Finally, gender was significantly associated with intonational signaling (beta =  $0.22 \pm 0.07$  se,  $t$   
978 =  $3.1$ ,  $p = 0.003$ , all other comparisons were not significant), reflecting the fact that intonational variations were stronger  
979 in males as compared to females. This could be consistent with previous reports suggesting substantial differences in  
980 men and women regarding subjective confidence reports (Lundeberg, Fox, & Puncoha, 1994), but is more likely to be  
981 due to general gender differences in the range of pitch variations (Elliott & Theunissen, 2009; Henton, 1989).