Distinct signatures of subjective confidence and objective accuracy in speech prosody

Louise Goupil, Jean-Julien Aucouturier

Distinct signatures of subjective confidence and objective accuracy in speech prosody

Louise Goupil[1,2]* & Jean-Julien Aucouturier[1]

1. Laboratoire STMS, UMR 9912 (CNRS/IRCAM/SU), Paris, France.
2. University of East London, London, United Kingdom

* corresponding author: lougoupil@gmail.com

Abstract

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

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

1. Introduction

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

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

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

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

Yet, candidate natural behaviors that can truly convey subjective confidence, over and beyond objective performances, have so far proved surprisingly difficult to identify. Two studies examined observers' ability to rely on response times to infer others' subjective confidence, and revealed that such inferences crucially depend on an observer's own experience with a task (Koriat & Ackerman, 2010; Patel, Fleming, & Kilner, 2012). This may not be surprising given that the relationships between response times, confidence and accuracy is task-dependent, varying in particular with the speed - accuracy trade off (Pleskac & Busemeyer, 2010). More to the point, these results imply that response times are not a good and stable proxy for inferring subjective confidence, and that they can only be exploited to this end when observers have a first-hand experience with observees' task. Similarly, post-decision persistence times have been argued to constitute a directly observable manifestation of confidence in animals (Kepecs, Uchida, Zariwala, & Mainen, 2008) and preverbal infants (Goupil & Kouider, 2016), but other researchers contend that this measure directly reflects the strength or reliability of first-order representations rather than subjective confidence per se (Fleming & Daw, 2017; Insabato, Pannunzi, & Deco, 2016). Thus, so far, a clear behavioral signature of subjective confidence has been lacking, as research focusing on response or persistence times struggled to clearly dissociate genuine behavioral manifestations of subjective confidence from those directly tied to decision-making.

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

First, past research typically relied on methodologies in which actors are asked to deliberately produce utterances with various levels of uncertainty in social contexts. This is known to provide a distorted picture, as requesting participants to produce communicative displays leads them to produce highly stereotypical rather than genuine displays (Juslin, Laukka, & Bänziger, 2018). At a more fundamental level, measuring prosodic displays during social interactions necessarily leads to conflating the contribution of natural, automatic mechanisms, and that of socially induced, deliberate self-presentation mechanisms: speakers do not only show prosodic displays automatically, they can also shape these displays deliberately, for instance in order to persuade (Van Zant & Berger, 2019) or to appear more dominant (Cheng, Tracy, Ho, & Henrich, 2016). Thus, past research leaves open the question of whether epistemic prosody is only displayed when the speaker has a communicative intention, or whether it is constitutively associated with confidence. A first step towards disentangling these influences, and investigating what these prosodic manifestations naturally mean (i.e., a behavior naturally means X when such behavior is typically associated with X; Grice, 1957; Wharton, 2009), can be to measure the relationships between confidence and prosodic features in the absence of an audience, and thus, of self-presentation and socially induced mechanisms. One previous study followed this rationale, and found that confidence impacts speakers' loudness and speech rate even in the absence of an audience (Kimble & Seidel, 1991). This questions the assumption that these prosodic signatures are primarily communicative, and suggests instead that they may reflect confidence constitutively, thereby representing natural signs that the speaker is confident. This study only measured loudness and speech rate however, so it remains unknown whether an important component of epistemic prosody, intonation, is also automatically impacted by confidence in the absence of an audience.

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

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

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

We address these questions by combining a psychophysical paradigm, signal detection theory, automatic classification analysis, and acoustic analysis of verbal reports produced in a non-social context. Isolated participants' verbal responses were recorded during a visual detection task allowing to finely manipulate - and measure - sensory evidence, accuracy and confidence (see Figure 1). By analyzing the pitch, intonation, loudness, and duration of these verbal responses as a function of sensory evidence, accuracy and confidence, we find that these psychological processes have distinct prosodic signatures. We then confirm this result by showing that an automatic classifier is able to decode confidence and accuracy orthogonally from speech prosody alone. Finally, we examine individual factors that modulate the automatic expression of prosodic signatures of confidence and competence.

2. Materials and Methods

2.1. Participants.

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

2.2. Procedure.

Participants ran three experiments during the same session. In the first and third experiment, participants had to memorize spoken pseudo-words, and to judge whether artificially manipulated voices were more or less reliable respectively. The results from these two experiments address a different set of questions related to speakers' reliability and perception, and will thus be reported in a separate article. The second experiment is the focus of the current paper. In this visual detection task, participants first saw a target bi-syllabic pseudo-word (*bazin, bizan, bivan, bavin, bodou, budou, dejon, dojen, dobue, duboe, vagio, vogia, vevon, voven, vizou* or *vuzoi*) that appeared for 16 ms while they were fixating a cross in the middle of the computer screen (see Figure 1). The target could appear at the top or the bottom of the screen, with equiprobable likelihoods. Targets were followed by a surrounding mask after a variable stimulus onset asynchrony (SOA: 16, 50, 83 or 116 ms) in order to induce various level of visibility, and thus, confidence in their verbal response. The mask was presented for 200 ms. Following the mask, the target word (e.g., *bazin*) and an alternative "foil" pseudo-word (e.g., *bazin, bizan*) were presented to the left or right side of the central fixation. Participants were asked to recognize the target word, and to pronounce their verbal response out loud so that it could be recorded. They then reported how well they saw the target on a perceptual awareness (PAS) scale (Ramsøy & Overgaard, 2004), and finally, their confidence in their verbal response on a scale from 1 to 4. The experiment was coded in *python* with the *PsychoPy* toolbox (Peirce, 2007). The target word (16 possibilities), SOA (4 possibilities), position of the response (2 possibilities: left or right) and position of the target word (2 possibilities: top or bottom) were counterbalanced within participants with a Latin square, resulting in 256 trials per participants. At the end of the session participants were asked to provide information regarding their socio-economic status: they were asked about their level of education, income and occupation, and given the fact that a majority of them were students, we also asked them to provide the same information concerning their parents. These data were aggregated to obtain a composite score of socio-economic status (SES). Participants also filled in a questionnaire assessing their level of empathy, which allows computing a general score over three dimensions measuring cognitive empathy, emotional disconnection and emotional contagion (French version of the BESA, Carré, Stefaniak, D'Ambrosio, Bensalah, & Besche-Richard, 2013).

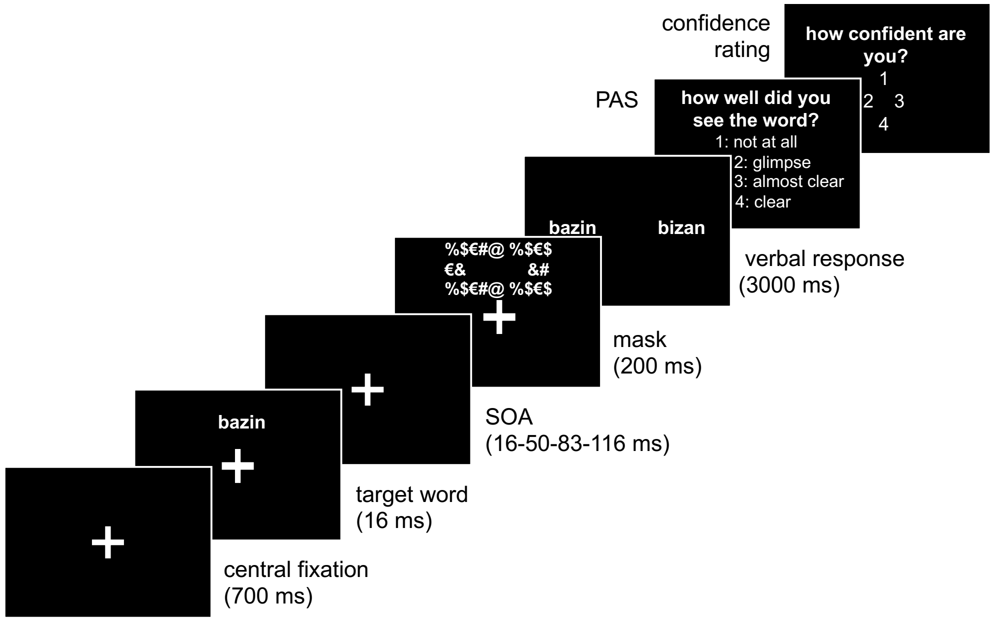

Figure 1. Design of the verbal production task. Participants were asked to fixate the center of the screen while a word was flashed above or below the fixation cross for 16ms. A masked followed the presentation of the word after a variable SOA. Participants were then asked to recognize the flashed word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

2.3. Behavioral analysis.

Unless stated otherwise, analyses were performed, and graphs obtained with *python*. Verbal responses were identified by a coder naive to the experimental conditions. Out of the 10240 trials (256*40 participants), 1207 (~11.8%) were excluded because the verbal response couldn't be reliable identified by the coder (e.g., because of a problem of pronunciation), resulting in a total of 9033 verbal responses. The accuracy of participants' verbal responses were classified as hits, misses, correct rejections or false alarms in order to compute sensitivity, i.e., a d' (Green & Swets, 1966). Metacognitive sensitivity (meta-d') was computed through a hierarchical Bayesian analysis with the *Hmeta* toolbox in *Matlab* (Fleming, 2017). For each participant, a global level of competence was also estimated by averaging their d' over the whole experiment. Confidence bias was estimated for each participant as the average of their confidence rescaled from zero to one, to which we subtracted their average accuracy in order to specifically estimate bias (but similar results were obtained with a simple average of confidence used in previous studies running similar regression analysis, e.g., Rollwage, Dolan, & Fleming, 2018).

2.4. Acoustic analysis.

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

2.5. Statistics.

Hierarchical linear models were run with pitch, RMS or duration as a dependent variable, and with participant and response word as random factors. Fixed factors included SOA, accuracy and confidence for duration, and SOA, accuracy, confidence and segment for pitch and loudness, in order to account for dynamic aspects. Factors were entered into the model in a hierarchical order from the lowest level (i.e., sensory, SOA) to the highest level (i.e., subjective confidence). We report beta estimates, standard errors, t-values, and p-values estimated through hierarchical model comparisons with the *lme4* and *lmerTest* packages in *R* (Kuznetsova, Brockhoff, & Christensen, 2014). To account for the dynamic effect of confidence on intonation, we relied on the *MNE* package in *python* to identify significant clusters with a permutation test providing p-values corrected for multiple comparisons (Gramfort et al., 2014). The permutation test identified 3 clusters: segments 0 to 1 ($p = 0.2$), segments 5 to 11 ($p = 0.012$) and segments 16 to 20 ($p = 0.042$). Pitch was then averaged in the two significant clusters and we examined which variables (SOA, confidence, accuracy) predicted pitch in these two windows separately by running hierarchical linear regressions and mediation analysis with the *mediation* package in *R* (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014).

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

2.6. Machine classification.

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

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

For the method based on k-nearest neighbors, training and testing datasets for each of the two classifications (i.e., decoding accuracy or confidence) were constructed as follows: a balanced subset of 200 speech items was selected pseudo-randomly from the full dataset for each level of the other class: if accuracy was being decoded, a subset was selected for each level of confidence; if confidence was being decoded, a subset was selected for each level of accuracy. The dataset was then randomly divided in 5 folds of 40 items. This set size was chosen so as to allow crossing all combinations of accuracy, SOA and confidence to create balanced datasets (e.g., using training and testing datasets composed of 1/32 of each combinations of accuracy, confidence levels and SOAs). This led to choosing a set size of 100, as the smallest combination of all SOAs/confidence/accuracy was 29. Each fold was thus balanced to contain 50% (i.e., 20 items) of one class level (e.g., correct or high confidence) trials, and 50% of the other class level (e.g., incorrect or low confidence), as well as the same numbers of items for each level of SOA. This equiprobable combinations of conditions ensured that the classifier had to decode the class blindly with respect to the other conditions. Performances were then computed in a 5-fold cross-validation procedure, where one of the folds iteratively served as a “test set”, and the four other folds served as “training test” (Anguita, Ghio, Ridella, & Sterpi, 2009). For each items of the test set, the Euclidean distance between pitch and loudness profiles for this item, and each of the items of the training test, was computed. For duration, a simple difference was computed. For each of the three acoustic dimensions, the 5 smallest distances were then identified, and a prediction of the accuracy or confidence of the test item was made as the most frequent class amongst the nearest neighbors (five for each acoustic dimension). Classifier performance was quantified with the F-value, which is the harmonic mean of the recall and precision of the classifier. In order to allow sufficient resampling of the original dataset, the whole process was repeated and averaged over 20 iterations for each classification. Significance was then assessed with a permutation procedure. For confidence decoding, confidence values were randomly reshuffled for each accuracy level and repetition (i.e., for each fold); for accuracy decoding, accuracy values were randomly reshuffled for each confidence level. Chance-level was then estimated by computing classification performance for these permuted data in the same way as in the real dataset, by computing an F-value. Real and permuted data F-values were then compared by running a rmANOVA with dataset (permuted, randomized) and condition (confidence or accuracy) as independent variables, and repetitions as a repeated measure. Finally, post-hoc differences between permuted and real data were assessed with Tukey post-hoc HSD with false-discovery rate correction for each level of confidence (or accuracy). In order to see if the results would generalize with another classification method, the same analysis was then replicated with SVMs (Figure S4).

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

3. Results

3.1. Relationship between sensory evidence, accuracy and confidence.

First, we checked that our experimental paradigm was efficient in inducing various levels of confidence in our participants. A hierarchical linear regression revealed that confidence (four levels) was predicted both by SOA (beta = 0.007 +/- 0.0006 se, $t = 10$, $p < 0.001$) and accuracy (beta = 0.85 +/- 0.06 se, $t = 13$, $p < 0.001$), and that there was no interaction between these two factors ($p > 0.2$; see Figure S1.B. and supplementary materials for further details). The fact that confidence increased with SOA over and beyond accuracy is consistent with previous reports suggesting that confidence is also directly impacted by the visibility of the stimulus (Rausch, Hellmann, & Zehetleitner, 2018). We also computed an index of metacognitive sensitivity reflecting the extent to which participants' confidence ratings tracked the reliability of their decisions (Fleming, 2017). Meta-d' was better than chance for every SOA (all p-values < 0.001, see Figure S1.D), and increased with SOA ($F(1,39) = 74$, $p < 0.001$, $\eta p2 = 0.65$), a finding that is consistent with previous research relying on similar visual paradigms (Charles, Van Opstal, Marti, & Dehaene, 2013; Kunimoto, Miller, & Pashler, 2001). Meta-d' was significantly above chance for seen stimuli (glimpse: $M = 1.36 +/- 0.88$, $t(39) = 6.2$, $p < 0.001$, Cohen's $d = 1.4$; almost clear: $M = 1.19 +/- 0.72$, $t(39) = 5.97$, $p < 0.001$, Cohen's $d = 1.35$, clear: $M = 2.55 +/- 1.27$, $t(39) = 10.12$, $p < 0.001$, Cohen's $d = 2.29$), but it was not significantly better than chance for unseen stimuli ($M = 0.59 +/- 1.24$, $t(39) = 0.46$, $p = 0.64$, Cohen's $d = 0.1$). This result is in line with research suggesting that metacognitive sensitivity depends on conscious access (Persaud, McLeod, & Cowey, 2007), but contrasts with other studies reporting that metacognitive sensitivity can be better than chance even for unseen stimuli (Charles et al., 2013). This may be due to the fact that we rely on verbal reports here, and this hypothesis could be specifically explored in further studies.

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

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

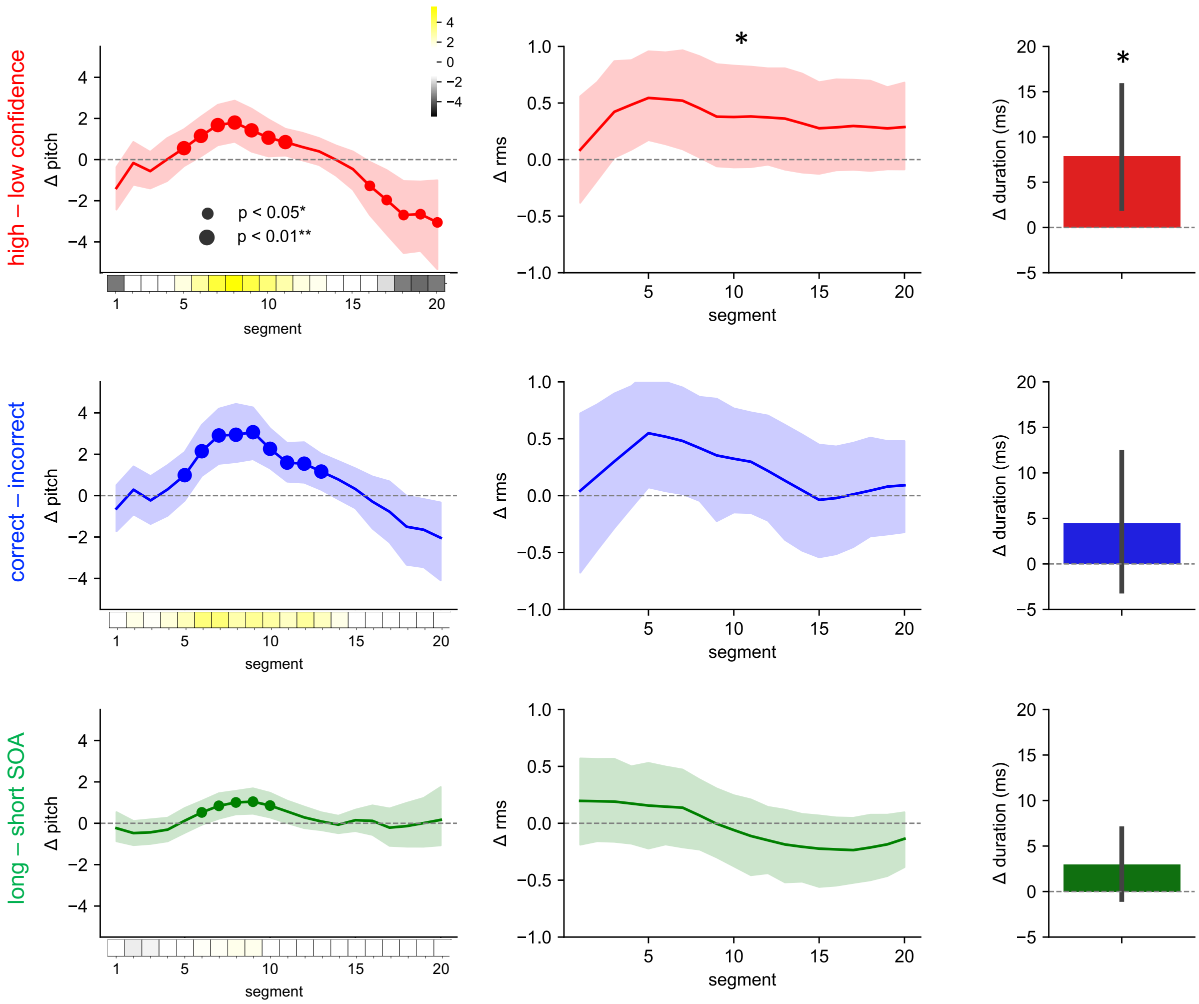

Figure 2. Acoustic analysis of verbal responses. Pitch, loudness (RMS) and duration values for high minus low confidence trials (1-2 versus 3-4; top – red), correct minus incorrect trials (middle – blue) and long (85-116) minus short (16-50ms) SOAs (bottom– green). Pitch: for the contrast between high and low confidence, the permutation test revealed two significant clusters: the first one ranging from the 5th to the 11th segment ($p = 0.008$), and the second ranging from the 16th to the 20th segment $p = 0.036$). For the contrast between correct and incorrect responses, the permutation test revealed one significant cluster ($p = 0.002$) from the 5th to the 13th segment. For the contrast between high and low SOAs, the permutation test revealed one significant cluster ($p = 0.017$) from the 6th to the 10th segment. RMS: the permutation test revealed no significant clusters with the threshold of $p < 0.05$. Circles represents 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.

Regarding mean pitch, there was no significant differences between confident and doubtful responses (mean difference in pitch = -0.23 +/- 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 attitudes such as confidence, that are more related to dynamic aspects of pitch (i.e., intonation, Goupil et al., n.d.; McAleer, Todorov, Belin, Taylor, & Iredell, 2014; Ponsot et al., 2018). Mean pitch may also be easier to manipulate than intonation for speakers asked to persuade or simulate confidence, which would provide a distorted picture of what "confident" prosodies naturally sound like due to social influences and self-presentation effects.

By contrast, as expected intonation (i.e., evolutions of the pitch over time) was impacted by confidence: a rmANOVA revealed an interaction between the level of confidence (including the full range of responses from 1 to 4) and segment ($F(1,39) = 7.3$, $p = 0.013$, $\eta p2 = 0.01$), as well as main effects of both segment ($F(1,39) = 4.1$, $p < 0.05$, $\eta p2 = 0.08$) and confidence level ($F(1,39) = 5.5$, $p < 0.03$, $\eta p2 = 0.01$). As can be seen in Figure 2 and S2 this interaction reflects the fact that confident responses present a rise and fall pattern, while doubtful responses present the opposite fall and rise pattern.

Regarding loudness, there was a static effect such that confident responses were louder than doubtful ones (mean difference = 0.36 +/- 1, $t(39) = 2.15$, $p = 0.038$, $d = 0.34$). A rmANOVA also revealed a main effect of segment ($F(1,39) = 183$, $p < 0.001$, $\eta p2 = 0.78$) and confidence level ($F(1,39) = 5.25$, $p < 0.03$, $\eta p2 = 0.02$) but no interaction ($F < 1$), suggesting that contrary to pitch, the effect was global rather than dynamic.

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

Regarding duration, we found that confident responses were longer than doubtful responses (mean difference = 7.85 +/- 21.4, $t(39) = 2.3$, $p = 0.027$, $d = 0.37$). This is inconsistent with previous reports that confident responses are produced with a faster speech rate (Jiang & Pell, 2017; Scherer et al., 1973), and also with some results obtained in perception (Goupil et al., n.d.). Thus, like response times, speech rate may not be a stable index enabling listeners to infer the reliability of a speaker. This is potentially due to the fact that the relationship between response speed, accuracy and confidence greatly varies depending on task characteristics such as the speed accuracy trade off (our task here was speeded, which would typically lead to slower response speed for correct and confidence responses) (Pleskac & Busemeyer, 2010). Interestingly, previous research has also shown that experience with the contingencies of a task is required to make accurate inferences about how response times relate to confidence in others (Koriat & Ackerman, 2010; Patel et al., 2012). In order to further elucidate the precise relationship between speech rate and confidence, further research relying on the method that we develop here could systematically vary the speed accuracy trade-off.

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

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

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

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
		accuracy:confidence	0.0007	0.002	0.03	0.98
		SOA:accuracy	-0.0006	-0.0008	-0.81	0.42
	pitch	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
		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
first cluster (segments 5 to 11)	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)	pitch	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 segment (all p-values > 0.8) however, revealing that the effects were mostly non-dynamic for this acoustic dimension; we therefore reduced the model to the static model used for duration above. This static model revealed a main effect of accuracy (beta = 0.07 +/- 0.03 se, $t = 2.7$, $p = 0.007$), while the main effect of confidence ($p = 0.21$) and SOA ($p = 0.36$) were not significant when entering the three co-variates into the model. Furthermore, there were no interactions between the three variables (all p-values > 0.2). Hence, it appears that loudness primarily reflects accuracy rather than confidence per se, or sensory evidence.

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

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

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

+/- 0.02 se, $t = -2.4$, $p = 0.015$), while the effect of SOA was not significant when entering all three variables in the model (beta = 0.0003 +/- 0.0002 se, $t = 1.27$, $p = 0.2$). In addition, a mediation analysis revealed that the effect of confidence on pitch was mediated at 12% (95% ci [-0.07, 0.30]) by accuracy in this temporal window, which was not significantly different from chance level ($p = 0.18$). Confidence still had a significant direct effect after taking this mediation into account ($p < 0.001$). Conversely, the effect of accuracy on pitch was partially mediated by confidence (38%, 95% ci [0.23, 0.61], $p < 0.001$), but was still significant after taking this mediation into account ($p < 0.001$). In the second time window, there was a main effect of confidence (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 and accuracy did not mediate the effect of confidence on pitch ($p > 0.7$). Thus, in the beginning of the word, pitch was determined by a mixture of sensory evidence, accuracy and confidence; however, it depended exclusively on confidence towards the end of the word.

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

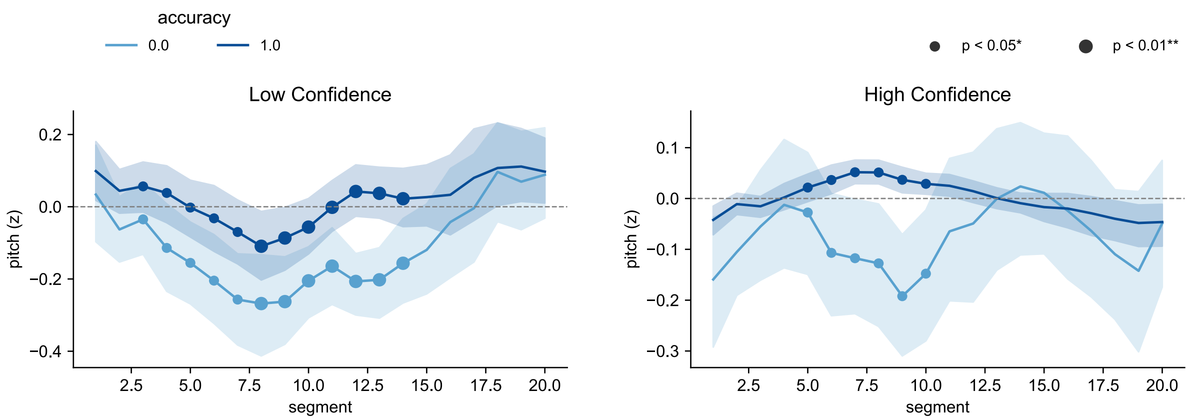

Figure 3. Intonational profiles depending on accuracy and confidence. Normalized pitch is shown separately for low (left) versus high (right) confidence, and accurate (dark blue) and inaccurate trials (light blue). Markers' sizes show significant clusters identified by running a permutation test on the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

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

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

Machine classifiers were able to detect speakers' accuracy with a performance of 60.2% (SD = 3.7) when they reported being 'fully confident' (rating of 4), and with a performance of 54.6 % (SD = 4.4) when they reported being 'confident' (rating of 3). By contrast, the accuracy of the speaker could not be reliably decoded for low levels of confidence: classification performance only reached 53.2% (SD = 4) for the lowest level of confidence, and 50% (SD = 3.8; $p = 0.5$) for the second level of confidence. To assess the significance of this result, these classification performances in decoding accuracy were compared with classification performances obtained with randomly permuted data (Ojala & Garriga, 2010). A rmANOVA with the accuracy of the classifications as a dependent variable, and confidence (four levels) and dataset (real vs. permuted) as independent variables, revealed a main effect of confidence ($F(1,19) = 22.5$, $p < 0.001$, $\eta p2 = 0.33$), a main effect of dataset ($F(1,19) = 58.51$, $p < 0.001$, $\eta p2 = 0.52$) and a significant interaction (F(1,19)

$= 40.81$, $p < 0.001$, $\eta p2 = 0.33$). This interaction reflected the fact that classification performances in decoding a speaker's accuracy were significantly higher than the chance-level estimated in the permuted dataset when participants were confident (post-hoc Tukey HSD with FDR correction, confidence = 4: $p < 0.001$; confidence = 3, $p = 0.004$), but only marginally so for the lowest level of confidence (confidence = 1: $p = 0.07$) and not significantly so for the second level (confidence = 2, $p = 0.78$).

The confidence of the speaker could also be decoded above chance, with a performance of 55.4% in incorrect trials (SD = 4.4), and 57.1% (SD = 3.8) in correct trials. A rmANOVA with classification performances as a dependent variable, and accuracy (two levels) and dataset (real vs. permuted) as independent variables, revealed a main effect of dataset ($F(1,19) = 60.95$, $p < 0.001$, $\eta p2 = 0.48$), no effect of accuracy ($F(1,19) = 2.43$, $p = 0.14$, $\eta p2 = 0.03$) and no interaction ($F(1,19) = 0.4$, $p = 0.54$, $\eta p2 = 0.01$). Classification performances in decoding speakers' confidence were significantly higher than the chance-level estimated in the permuted dataset both when participants were accurate (post-hoc Tukey HSD with FDR correction, $p < 0.001$), and when they were inaccurate ($p < 0.001$).

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

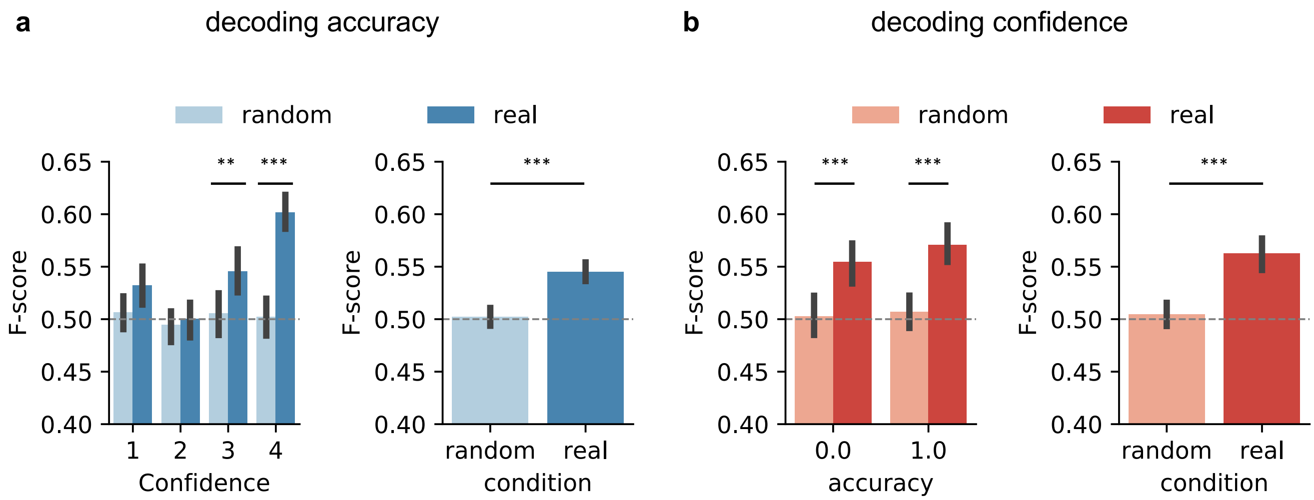

Figure 4. Results of the k-nearest-neighbors classification. A) Classifiers' performances in decoding objective accuracy for each level of confidence (left), and overall (right). To examine whether speech prosody contains enough information to automatically infer a speaker's accuracy, we relied on a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

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

Finally, we wanted to assess whether participants' ability to perform the task (their competence), their general tendency to be confident (their confidence bias), and their global ability to evaluate their performances (their metacognitive sensitivity) related to how accuracy and confidence were automatically reflected in their voice. If epistemic prosody constitutes an adaptive mechanism allowing listeners to filter information coming from unreliable social partners, we may expect that vocal signatures of accuracy and confidence may be more manifest in competent (or meta-competent) speakers.

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

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

As can be seen in Figure 5, after controlling for all other factors, competence significantly predicted higher intonational signaling (beta = 0.39 +/- 0.09 se, $t = 4.27$, Bonferroni corrected $p = 0.002$), with no significant interaction with the type of signaling (i.e., accuracy or confidence, $p > 0.6$). When all other factors including competence were considered, metacognitive sensitivity also significantly predicted increased intonational signaling (beta = 0.28 +/- 0.08 se, $t = 3.32$, $p = 0.049$, here again with no significant interaction with the type of signaling, $p > 0.2$), and it also marginally increased signaling at the level of duration (beta = 0.05 +/- 0.04 se, $t = 1.315$, $p = 0.053$). Thus, speakers' level of competence and metacognitive sensitivity in the task increased their signaling of both confidence and competence. By contrast, there were no significant associations between confidence bias and any of the acoustic dimensions (all p-values > 0.1), which suggests that individuals did not display signs of competence or confidence more or less saliently depending on their metacognitive bias (see Figure 5 and supplementary results for details about additional effects of loudness, age and gender).

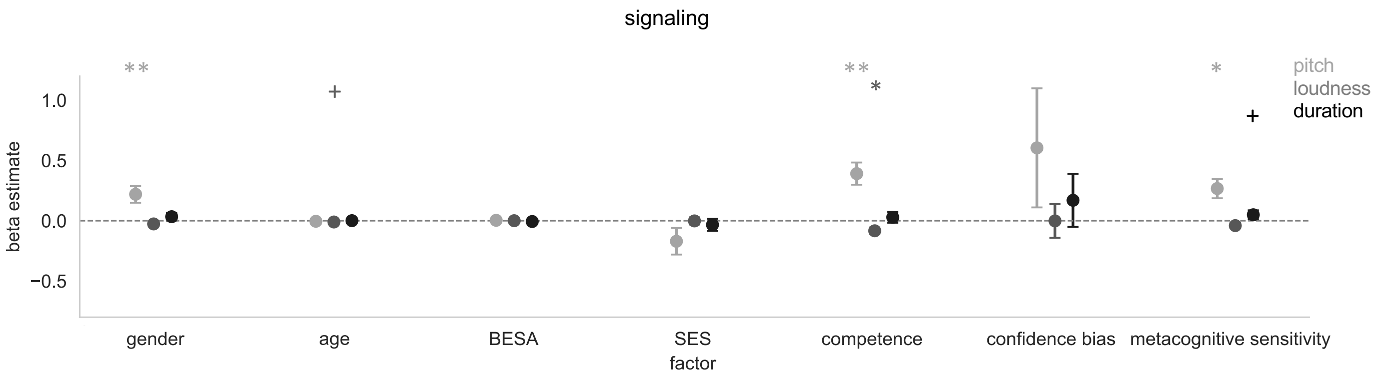

Figure 5. Signaling depending on individual factors. Regression analysis were conducted on each acoustic dimension separately to assess the impact of individual traits on signaling. Signaling for pitch corresponded to the Euclidean distance between intonational profiles computed for high confidence (or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

4. Discussion

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

Our results reveal that intonation, loudness and duration differently reflect the underlying psychological processes leading to the production of a verbal response. While duration and intonation reflect confidence per se, loudness appears to be mostly driven by cognition (i.e., accuracy) rather than metacognition (i.e., confidence). By revealing that various aspects of prosody are associated with different underlying psychological processes, these results go beyond previous research showing simple associations between speech and confidence, without assessing the impact and potentially mediating role of sensory evidence or accuracy.

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

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

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

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

Finally, the fact that such epistemic prosodic markers were observed in the absence of an audience is consistent with past research (Kimble & Seidel, 1991), and shows that they are manifested constitutively and automatically as a function of the speaker's level of confidence and accuracy: i.e., they constitute natural signs of confidence and competence. This is not to say that these displays are never under voluntary control: humans can obviously control the pitch, duration and volume of their voice, making it possible to deliberately use prosodic displays as "social tools" during conversation (Crivelli & Fridlund, 2018; Van Zant & Berger, 2019; Wharton, 2009) and past research has shown that, indeed, similar prosodic signatures as the ones we find here are exploited during communicative interactions: listeners perceive them to infer confidence and honesty in their partners (Goupil et al., 2020; Jiang & Pell, 2017), and speakers manipulate them in order to persuade their interlocutors (Van Zant & Berger, 2019). Thus, it will be important to extend our psychophysical approach to social interactions in future work, for instance by relying on dyadic collective decision-making paradigms (Bahrami et al., 2010; Fusaroli et al., 2012; Pescetelli & Yeung, 2020), in order to examine how specific social settings - such as the fact that the speaker is engaged in a cooperative or competitive interaction - impact how speakers display these prosodic signatures. A particularly interesting question is whether speakers manipulate all prosodic features (intonation, accentuation, global levels of pitch or loudness, duration), or only some of them (e.g., global levels of loudness and pitch, but not intonation). Another open question is how variations in physical attributes (e.g., body size) and social traits (e.g., social dominance) would modulate and interact with the relationships we found here between prosodic signaling and (meta)competence.

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

5. Conclusions

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

References

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency-confidence association in problem solving. *Psychonomic Bulletin and Review*, *19*(6), 1187–1192. https://doi.org/10.3758/s13423-012-0305-z

Aikhenvald, A. (2018). *The Oxford handbook of evidentiality*.

Anguita, D., Ghio, A., Ridella, S., & Sterpi, D. (2009). *K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines*. *Vessels Fuel Consumption Forecast and Trim Optimisation: a Data Analytics Perspective View project K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines*. Retrieved from https://www.researchgate.net/publication/220704948

Bahrami, B., Olsen, K., Latham, P., Roepstorff, A., Rees, G., & Frith, C. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085.

Bang, D., Aitchison, L., Moran, R., Castañón, S. H., Rafiee, B., Mahmoodi, A., … Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(0117), 1–7.

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 201800795.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23.

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences*, *107*(48), 1–6. https://doi.org/10.1073/pnas.1007704107/-/DCSupplemental.www.pnas.org/cgi/

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics*. https://doi.org/10.3758/BF03206939

Brennan, S. E., & Williams, M. (1995). The Feeling of Another′s Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language*, *34*(3), 383–398.

Carré, A., Stefaniak, N., D'Ambrosio, F., Bensalah, L., & Besche-Richard, C. (2013). The basic empathy scale in adults (BES-A): Factor structure of a revised form. *Psychological Assessment*.

Carruthers, P. (2016). Are epistemic emotions metacognitive? *Philosophical Psychology*, 1–15.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94.

Chen, A., & Gussenhoven, C. (2003). Language-dependence in the signalling of attitude in speech. *Proceedings of Workshop on the Subtle Expressivity of Emotion at CHI 2003 Conference on Human and Computer Interaction*.

Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans. *Journal of Experimental Psychology: General*, *145*(5), 1–12. https://doi.org/10.1037/xge0000166

Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, *7*, 13669.

Crivelli, C., & Fridlund, A. J. (2018). Facial Displays Are Tools for Social Influence. *Trends in Cognitive Sciences*, *22*(5), 388–399.

de Haan, F. (2001). The Relation Between Modality and Evidentiality. *Linguistische Berichte*.

Dezecache, G., Zuberbühler, K., Davila-Ross, M., & Dahl, C. (2019). Early vocal production and functional flexibility in wild infant chimpanzees. *BioRxiv*, 848770. https://doi.org/10.1101/848770

Dijkstra, C., Krahmer, E., & Swerts, M. (2006). Manipulating Uncertainty: the contribution of different audiovisual prosodic cues to the perception of confidence. In *Speech Prosody*.

Dunstone, J., & Caldwell, C. A. (2018). Cumulative culture and explicit metacognition: a review of theories, evidence and key predictions. *Palgrave Communications*, *4*(1), 145.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000302

Fetsch, C., Kiani, R., Newsome, W., & Shadlen, M. (2014). Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron*. https://doi.org/10.1016/j.neuron.2014.07.011

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1338–1349.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science*, *329*(5998), 1541–1543.

Fleming, Stephen M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1).

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms. *Psychological Science*, *23*(8), 931–939.

Galvez-Pol, A., Salome, A., Li, C., & Kilner, J. (n.d.). Direct perception of other people's heart rate. https://doi.org/10.31234/OSF.IO/7F9PQ

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jiang, X., Gossack-Keenan, K., & Pell, M. D. (2020). To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology (2006)*, *73*(1), 55–79.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

McAleer, P., Todorov, A., Belin, P., Taylor, L., & Iredell, N. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154.

Reyes, G., Silva, J. R., Jaramillo, K., Rehbein, L., & Sackur, J. (2015). Self-Knowledge Dim-Out: Stress Impairs Metacognitive Accuracy. *PLOS ONE*, *10*(8), e0132320.

Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2016). Communicating Epistemic Stance: How Speech and Gesture Patterns Reflect Epistemicity and Evidentiality. *Discourse Processes*.

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2017.12.017

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/0096-3445.134.1.124

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, *7*(1), 31–44. https://doi.org/10.1016/0092-6566(73)90030-5

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.

Smith, V. L., & Clark, H. H. (1993). On the Course of Answering Questions. *Journal of Memory and Language*, *32*(1), 25–38.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393.

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration Trumps Confidence as a Basis for Witness Credibility. *Psychological Science*, *18*(1), 46–50.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v059.i05

Van Zant, A. B., & Berger, J. (2019). How the Voice Persuades. *Journal of Personality and Social Psychology*, *118*(4), 661–682.

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences*, *111*(45), 16214–16218. https://doi.org/10.1073/pnas.1403619111

Wharton, T. (2009). *Pragmatics and Non-Verbal Communication*. Cambridge: Cambridge University Press.

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

Acknowledgements.

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

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

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

Supplementary Materials

Behavioral results. As expected from previous research relying on similar visual paradigms (Charles et al., 2013; Kunimoto et al., 2001; Rausch et al., 2018), both visibility ($F(1,39) = 103$, $p < 0.001$, $\eta p2 = 0.72$), sensitivity (i.e., d', $F(1,39) = 169$, $p < 0.001$, $\eta p2 = 0.81$) and confidence ($F(1,39) = 116$, $p < 0.001$, $\eta p2 = 0.74$) increased with SOA. As can be seen in Figure S1.A. below, at the shortest SOA participants rarely reported not seeing anything at all, but often reported seeing only a glimpse of the stimulus. Also of note is the fact that sensitivity (d') remained above chance level even for the shortest SOA (M = 1.48 +/- 0.8, $t(39) = 7.7$, $p < 0.001$) but not for unseen stimuli (M = 0.84 +/- 2.41, $t(39) = 0.87$, $p = 0.38$). This pattern of result contrasts with previous findings showing that objective performances can be better than chance even for unseen stimuli (Charles et al., 2013; Kunimoto et al., 2001). This could be due to the fact that we rely on verbal reports here, rather than less ecological task involving poorly demanding motor responses such as button presses.

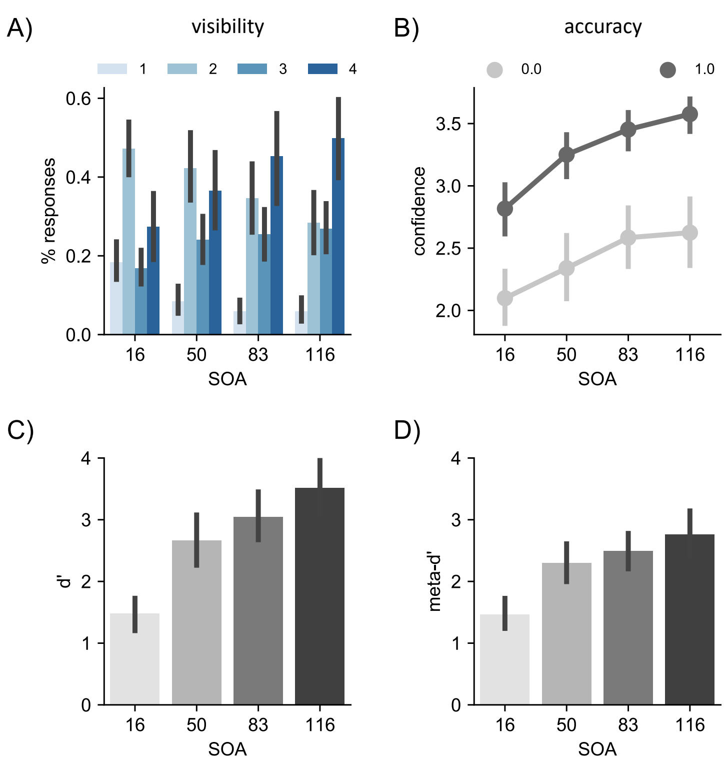

Figure S1. Behavioral results. A) Visibility ratings depending on SOA. The percentage of trials was computed for each level of visibility depending on SOA, and averaged across participants. B) Confidence was averaged for each participant depending on accuracy. B) Sensitivity (d') was computed for each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

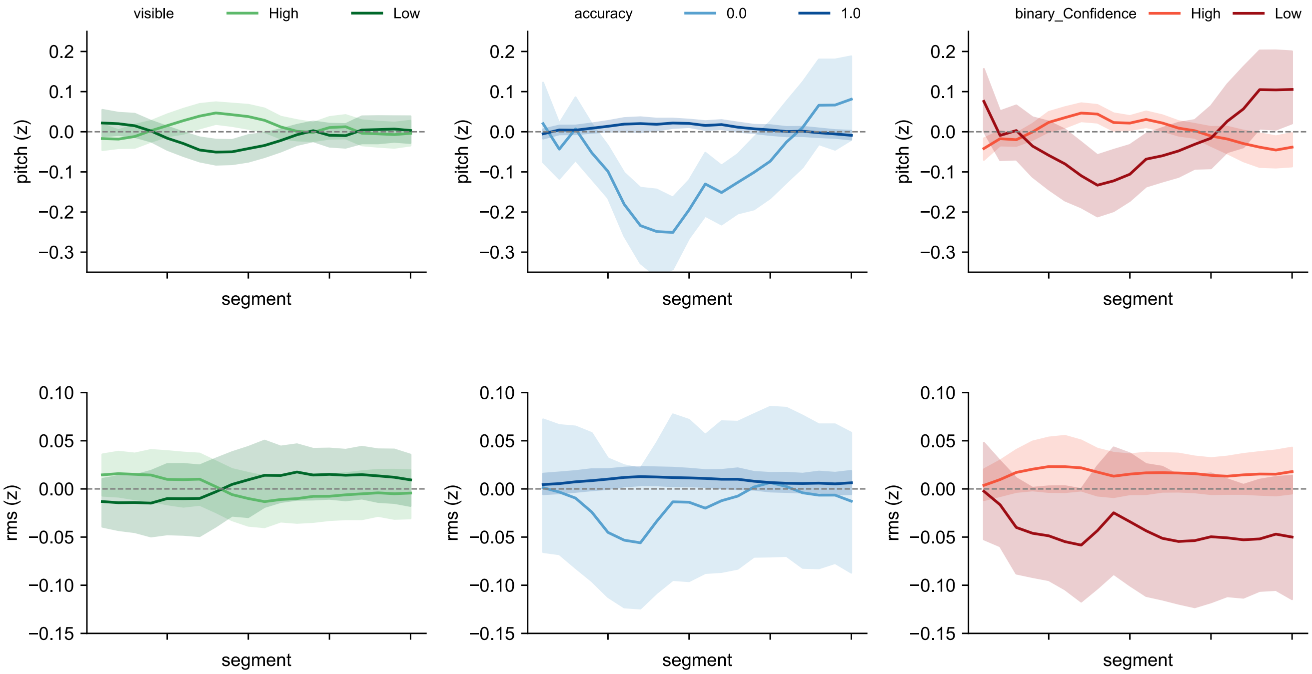

Figure S2. Normalized pitch (top) and RMS (bottom) are shown for each segment, depending on SOA (left / green, low: 16 and 50ms versus high: 83 and 116ms), accuracy (middle / blue) and confidence (right / red). Error bar show the 95% confidence intervals.

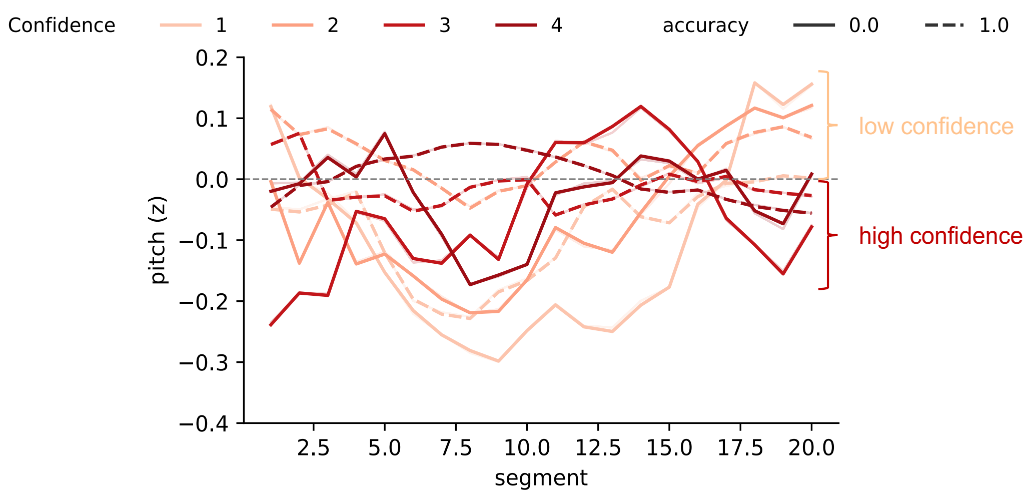

Figure S3. Normalized pitch is shown separately for each level of confidence and accurate (dashed lines) and inaccurate trials (plain lines).

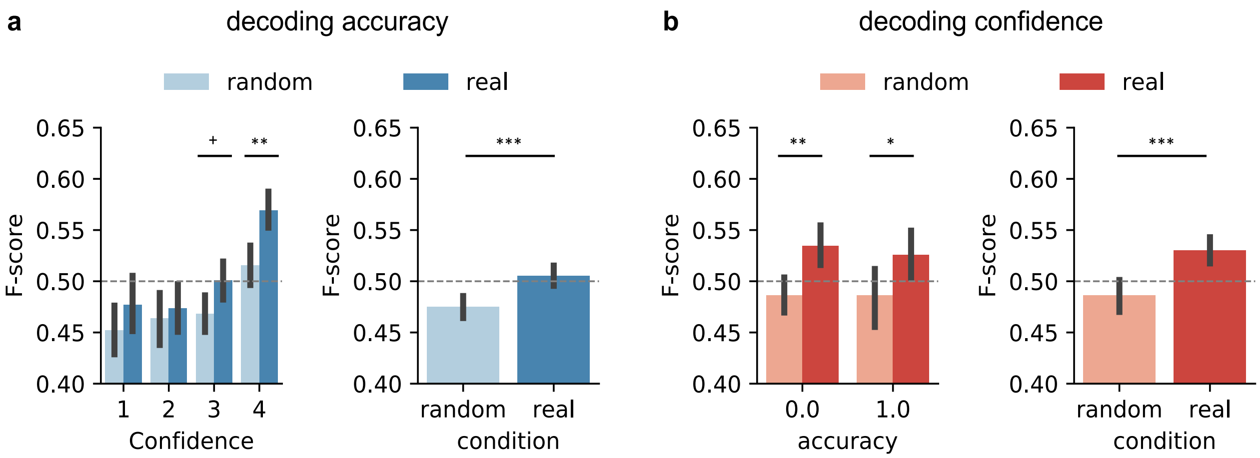

Figure S4. Results of the support-vector machines classification. A) Decoding objective accuracy. The same analysis presented in Figure 4 was repeated with an alternative classification procedure (support vector machines, SVMs). We present the F-values averaged across the 20 repetitions. Bar plots show the average performances of the classifier for real (darker shades) and permuted (lighter shades) data, with error bars showing the 95% confidence intervals estimated over the 20 repetitions. The chance-level estimated with permuted data (see methods) was 47.5% (SD = 2.6) overall (confidence = 1: 45.2 (5.7); confidence = 2: 46.4% (5.8); confidence = 3: 46.8% (4); confidence = 4: 51.5% (4.4)). The performance of the classifier over all confidence levels was 50.5% (SD = 2.3), which was highly significantly above the chance level estimated with permuted data ($t(19) = 5.58$, $p < 0.001$). As for KNNs, a rmANOVA revealed a significant main effect of condition (real vs. permuted, $F(1,19) = 31.2$, $p < 0.001$, $\eta p2 = 0.27$), and a main effect of confidence ($F(1,19) = 47.4$, $p < 0.001$, $\eta p2 = 0.53$) and a marginal interaction between the two factors ($F(1,19) = 3.3$, $p = 0.08$, $\eta p2 = 0.053$). Performances were higher in the dataset as compared to permuted data when participants were highly confident (confidence 4: $p = 0.002$, post-hoc Tukey HSD with FDR correction), but only marginally so for confidence = 3 ($p = 0.068$), and not significantly so for lower levels of confidence (confidence 1: $p = 0.17$; confidence 2: $p = 0.62$). Asterisks show the results of the post-hoc Tukey HSD with FDR correction comparing classification performances with the chance-level estimated with permuted data, with + $p < 0.07$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (see main text for exact p-values). Dashed lines show the theoretical chance-level (50%, black). **B) Decoding subjective confidence.** To assess whether speech prosody contains enough information to infer a speaker's level of confidence, we applied the same method, now decoding binary confidence (High vs. Low) for each level of accuracy and SOA (see methods). The chance-level estimated with permuted data was 48.7.3% (SD = 3.9) for incorrect trials, 48.6 (6.3) for correct trials, and 48.6 (3.5) overall. The performance of the classifier over all accuracy levels was 53% (SD = 2.9), which was highly significantly above chance level ($t(19) = 6.9$, $p < 0.001$). As for KNNs, a rmANOVA revealed a significant main effect of condition (real vs. permuted, $F(1,19) = 47.74$, $p < 0.001$, $\eta p2 = 0.21$), no effect of accuracy ($F(1,19) = 0.08$, $p = 0.76$, $\eta p2 = 0.003$) and no interaction ($F(1,19) = 0.29$, $p = 0.59$, $\eta p2 = 0.002$). Performances were higher in the dataset as compared to permuted data for both levels of accuracy (correct: $p = 0.02$; incorrect $p = 0.006$; post-hoc Tukey HSD with FDR correction).

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 +/- 0.03 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 faster overall, rho =-0.27, $p < 0.001$). Age also marginally decreased signaling at the level of loudness (beta = -0.01 +/- 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 (18- to 30-year-olds). Finally, gender was significantly associated with intonational signaling (beta = 0.22 +/- 0.07 se, t = 3.1, p = 0.003, all other comparisons were not significant), reflecting the fact that intonational variations were stronger in males as compared to females. This could be consistent with previous reports suggesting substantial differences in men and women regarding subjective confidence reports (Lundeberg, Fox, & Puncoha, 1994), but is more likely to be due to general gender differences in the range of pitch variations (Elliott & Theunissen, 2009; Henton, 1989).