Effects of the islamic headscarf on vocal arousal and intimacy: a field experiment in the Paris metro

Alban Lemasson, Manon Toutain, Francesco Madrisotti, Martin Aranguren

Effects of the islamic headscarf on vocal arousal and intimacy: a field experiment in the Paris metro

Alban Lemasson[1,2], Manon Toutain[1], Francesco Madrisotti[3], and Martin Aranguren[3*]

[1] Univ Rennes, Normandie Univ, CNRS, EthoS (Éthologie animale et humaine) - UMR 6552, F-35000 Rennes, France

[2] Institut Universitaire de France

[3] Université de Paris, Centre National de la Recherche Scientifique, URMIS (Unité de Recherches Migrations et Sociétés)

* Corresponding author: martin.aranguren@cnrs.fr

Abstract. In recent years physiological indexes of emotion have made their comeback as indicators of prejudice, but vocal measures have lagged behind. The aim of this study is to examine the vocal changes in arousal and intimacy that occur in real-life interactions with a woman who wears the Islamic headscarf or hijab. The study is based on a field experiment performed in the Paris metro. Assuming that vocal behavior serves an expressive function, the hijab is expected to provoke higher arousal and lower intimacy. Assuming that vocal behavior can alternatively function to manage impressions, and in view of observed gender differences, the hijab is expected to elicit lower arousal and higher intimacy among women but to have the opposite effect among men. The data indicates that in response to the hijab arousal decreases among women but increases among men, while intimacy decreases both among women and men. The historically most reliable acoustic measures of arousal, in particular the fundamental frequency, conform to the expectations, arguing for the internal validity of this naturalistic experiment.

Keywords: vocal acoustics, prejudice, social interaction, arousal, intimacy, gender differences

Acknowledgements: The data have been collected as part of the MIDI project, funded by MITI-Centre National de la Recherche Scientifique with a Momentum grant to Martin Aranguren. The funder had no role in the conduct of this research.

We thank Marie Bénédicte Cazeneuve for serving as the confederate, Veronique Biquand for preliminary statistical work, Antoine L'Azou for helping with the literature survey, Maxime Choblet for methodological input, Corentin Monmasson for technical assistance with Praat, and Gaëtan Roisné-Hamelin for acting as the second coder.

Authors contributions: Alban Lemasson conceived the study, contributed methods, performed the measurements and drafted the manuscript. Manon Toutain performed the measurements and explored the data. Francesco Madrisotti collected the data. Martin Aranguren designed the experiment, collected the data, performed the statistical analyses and drafted the manuscript.

After a relative withdrawal in the 1980s and 1990s (Guglielmi, 1999), in the first two decades of the 21st century physiological indexes of emotion have made their comeback as indicators of prejudice. In part, this renewed interest reflects the development of new psychophysiological variables (e.g. neuroimaging, Amodio, 2014; Chekroud et al., 2014), or the novel application of existing ones to the phenomenon of prejudice (e.g. the startle eyeblink response, Amodio et al., 2003; Brown et al., 2006; Mahaffey et al., 2005; March & Graham, 2015; Paulus et al., 2019; Phelps et al., 2000; Vanman et al., 2013). But to a significant extent, it also represents the revival of old measures, rejuvenated by a theoretical return of the pendulum to emotion (vs. cognition) as the key component of prejudice, supplemented by gains both in methodological lucidity and technical sophistication. This resurgence, however, has not concerned all pre-existing psychophysiological measures of prejudice to the same extent. Of these, the center stage has been accorded to responses thought to be controlled by the autonomous nervous system such as facial and electrodermal activity, heart rate, or blood pressure (Amodio, 2009; Dambrun et al., 2003; Graves et al., 2005; Greenland et al., 2012; Kiebel et al., 2017; Littleford et al., 2005).

One remarkable absent in the list of resurrected psychophysiological measures of prejudice is the voice. This is particularly surprising when one considers that an impressive amount of evidence has cumulated in the field of emotion indicating that well-defined vocalizations such as laughs or cries and the acoustic characteristics of human speech vehicle information about underlying emotional (reviewed in Koolagudi & Rao, 2012; Russell et al., 2003). On the plane of theory, this empirical development finds a parallel in the formulation of the "Motivation Structural Rule Theory" (Morton, 1977) and the "Vocal Affect Expression Model" (Scherer, 1986) as plausible conceptualizations of vocal affect signaling. Additionally, compared to placing sensors on participants' head or trunk, or asking them to provide samples of saliva, the apparatus needed to record vocal signals stands out as remarkably unobtrusive, minimizing the potential awkwardness of the data collection procedure.

One possible reason why vocalizations and speech acoustics, as relatively unobtrusive indicators of emotion, have not spilled over from the field of emotion to that of prejudice is that beyond general arousal these cues do not systematically differentiate between positively and negatively valenced affective states (Russell et al., 2003). One way to infer the valence associated with vocally expressed arousal is to supplement measures of emotional vocalizations and voice acoustics with verbal indexes of positive or negative evaluation. This is the strategy that we adopt in the present paper, by treating verbal and conversational behaviors indicative of intimacy as indexes of positive valence. Intimacy makes reference to the degree of interest in, or openness toward, or liking of, the interaction partner that an actor's behavior communicates (Patterson, 1982). For simplicity, we reduce interest, openness and liking to dimensions of the underlying variable of

positive evaluation, so that intimacy boils down to the degree of positivity towards the interaction partner that an actor's behavior functions to signal (a definition that comes close to the one given to the cognate idea of "immediacy", Mehrabian, 1972). Simply put, behaviors characterized as intimate signal a positive evaluation of and to the interaction partner.

A more compelling reason why prejudice research has not resuscitated the voice as an indicator of stress, arousal, or emotion, lies in the uncertainties that surround the degree to which the acoustic features of vocalizations are subject to voluntary control. Measuring prejudice on the basis of psychophysiological indexes is notoriously costly. The main motivation to accept that cost is the hope that by circumventing or supplementing other indicators of prejudice that are under voluntary control (chiefly: self-reports), the researcher can gain access to attitudes that respondents would be otherwise unwilling or unable to express (Guglielmi, 1999). The extent to which the voice is able to fulfill this goal is unknown.

But one may wonder why we ask this from the voice in the first place. The overwhelming majority of the above cited studies focus on a particular group as the target of prejudice: Blacks in the United States. Following the Civil Rights movement in 1960s, the law protects African Americans from various forms of discrimination. Similarly, ordinary morality condemns the overt expression of anti-Black prejudice in the United States. In this particular context, social desirability biases pose a serious problem to the validity of self-reported measures of prejudice against African Americans. It is this particular context that has created the motivation to use indirect measures, including psychophysiological ones, as a "bona fide pipeline" (Fazio et al., 1995) to prejudice.

Rather than using the voice to reveal a prejudiced attitude that people would otherwise hide, in this study we start from the fact of prejudice and investigate the changes that it induces in vocal behavior, broadly understood to cover the acoustic, lexical, pragmatic, and conversational dimensions of speech, as well as emotional vocalizations. This shift is justified by the specific group under investigation, namely women who wear the Islamic headscarf or hijab in France. The practice of hijab-wearing is widely disapproved in France (CNCDH, 2019; Ersanilli & Koopmans, 2013; Pew Research Center, 2015). Since people do not distinguish well between disapproving a Muslim practice and disapproving the Muslim person who enacts that practice (van der Noll et al., 2018), it follows that French residents generally disapprove, or hold negative views of, hijab-wearing women. The aim of this study is to examine the vocal changes in arousal and intimacy that occur in real-life interactions with a woman who wears the Islamic headscarf or hijab, compared to a control condition in which the same woman appears with uncovered hair.

One lingering concern in the psychophysiological study of prejudice is the artificiality of the stimuli and therefore the ecological validity of the results (Guglielmi, 1999; Mendes et al., 2002). In applications of the startle modification paradigm to prejudice research, for example, the typical

study uses photographs of Blacks and Whites as the stimulus. A more realistic setting is to provoke intergroup dyadic interactions in the laboratory (e.g. cite intergroup anxiety studies). While moving from photographs (or vignettes, Vanman et al., 2013) to live interactions is undoubtedly a progress in ecological validity, it remains that the encounter takes place in a laboratory and, perhaps more importantly, that the demographic profile of participants tends to limit itself to university students. The same is true of bioacoustics studies which are typically based on intense stress-provoking experiments (Laukka et al., 2008), on known strong correlations between voice quality and self-scored anxiety (Almeida et al., 2014), or on the identification of emotions purposefully enacted by actors (Banse & Scherer, 1996; Juslin & Laukka, 2001). Using live interactions with a hijab-wearing confederate as the stimulus, here we make a further step towards ecological validity both by provoking the interactions in the "natural" context of a public place, and by sampling participants randomly from the wider population.

Predictions

We divide our analysis into indexes of arousal and indexes of positive valence or intimacy. We derive our predictions from two opposite hypotheses about the function of these signals. If behavior is assumed to be a spontaneous expression of evaluation, from the fact that French residents hold negative views of hijab-wearing women we predict that participants' vocal behavior will show more arousal and less intimacy when the partner wears the Muslim scarf. Now, it may be argued that behavior can serve a different function in interpersonal relations. In particular for the so-called intimacy behaviors, it has been proposed that they may alternatively function to manage impressions in a deliberate manner (Patterson, 1982). That is, an actor, regardless of how positively or negatively the interaction partner is actually evaluated, might display intimacy behaviors intentionally in a conscious effort to produce a desired result. In this sense, there is some evidence that when the interaction begins with a negative evaluation of the interaction partner, compared to interactions in which that starting impression is more positive, actors show not less but *more* signs of intimacy, in an effort to make the interaction less unpleasant (Bond, 1972; Coutts et al., 1980; Ickes et al., 1982). We extend the observation to indexes of arousal, leading to the alternative prediction that the hijab will give rise to less arousal and more intimacy in vocal behavior.

The expected effect of the hijab, then, will depend on the function that signals of arousal and intimacy are postulated to serve. But which function will prevail? The available evidence from the field of intergroup relations suggests that the function might depend on the participant's gender. A study of intergroup interactions examined the relationship between the level of social distance that participants reported with respect to an outgroup and the friendliness that they showed in interaction with a male or female member of the same outgroup (Littleford et al., 2005). It was found that

whereas greater felt distance decreased friendliness among male participants, it *increased* the display of friendly behaviors among their female counterparts. Similarly, a field experiment on intergroup interactions in the metros of Brussels, Paris, and Vienna analyzed the effect of the Islamic headscarf on interpersonal distance as an indicator of intimacy (or "involvement"). The results show that, across the three cities, female passengers *increased* intimacy by interacting closer in response to the hijab, whereas male passengers did not (Aranguren et al., 2021). The same study also considered eye contact as an additional index of intimacy. Men in Paris, but not women, turned out to decrease intimacy in interaction with the covered woman, a result replicated by a follow-up study using a different procedure (Aranguren, 2021).

In other words, there is some evidence to the effect that in an nonoppositional interaction with a more socially distant or more negatively viewed unacquainted person men tend to show colder behavior, in accordance with the spontaneous expression function, but women *warmer* behavior, in accordance with the impression management function. Summing up the discussion, these are the predictions that we set out to test in relation to vocal behavior. In interaction with a hijab-wearing woman,

Main effect: all participants will show P1a) more arousal and P1b) less intimacy;
Simple effect among women: regardless of male participants, female participants will show P2a) less arousal and P2b) more intimacy;
Simple effect among men: regardless of female participants, male participants will show P3a) more arousal and P3b) less intimacy;
Cross-over interaction: P4a) arousal will decrease among females but increase among males, and P4b) intimacy will increase among females but decrease among males.

Method

Design

The experiment follows a between-subjects randomized design with roughly balanced proportions of male and female participants in each of the two experimental conditions and across the six metro stations in which interactions were observed. The goal was to collect at least ten samples representing each combination of sex, experimental condition, and station, that is a total of 240 experimental assays.

Stations selection

Stations were selected at random using a set of filters. The first filter consisted in eliminating all the stations in the upper and lower quartiles by number of passengers, which was a convenient way of taking into account the fact that packed and deserted stations would not offer a suitable environment for the experiment. With the stations in the mid quartiles a random list was then created. The second filter involved visiting the stations in the order stipulated by the random list and ascertaining that the platform was assigned to a single direction (not two) and physically arranged in such a way that there was a single entrance (not many) placed on one of the two longitudinal extremes (not in the middle) of the platform.

Sampling

After a pilot study in March 2018, Martin Aranguren and Francesco Madrisotti performed the experiment between May and June of the same year. The CNRS correspondent of the French commission for the protection of privacy and confidentiality CNIL approved the study and the transportation authority RATP gave us formal clearance to conduct the experiment in the metro. The experimenters made five data collection visits within each of the six selected stations. All visits, scheduled at different weekdays within the same station, had a duration of two hours. Of these, the first hour was assigned to one experimental condition and the second hour to the other condition, balancing for the entire experiment the number of times that each condition was placed first or second in chronological order. During the hour devoted to each condition, in order to recruit an equal number of randomly selected male and female passengers, a method for approximating random selection and another one for stratifying the sampling of men and women was employed. Random selection was approximated with a method of systematic selection: during the time period comprised between the departure of the last train and the arrival of the following one, the confederate approached the first passenger who arrived at the platform. The stratification technique consisted in starting with the method of systematic selection regardless of the sex of the passenger, recruiting one passenger (for example, a man), and then reapplying the method of systematic selection but only to passengers of the opposite sex (women). The third passenger was again selected regardless of sex, the fourth by stratifying by sex, and so on. This means that, in stratifying our sample, the experimenters relied on their own commonsensical understandings of sexual dimorphism to identify passengers as men or women, and not on passengers' self-reported sexual identity. Data collection visits took place on regular weekdays between 12pm and 2pm. In Paris, this is the only period of the working day in which waiting times are in the range of 3-5 minutes (instead of 1-2), maximizing the chances that the confederate will get to complete the script before the incoming train arrives.

Procedure

On a platform of the local metro, a non-immigrant confederate actress approaches the selected passenger asking for help, on the basis of a standardized script. In one experimental condition, she appears with a hijab; in the other, with uncovered hair. The rest of the clothing is identical, as is the script she follows while interacting with the passengers. To register the conversation with the passenger, the confederate carries a discreetly mounted portable microphone (VT506 Voice Technologies) and an audio recorder (DR-22WL linear PCM recorder of Tascam). Before approaching, she waits until the selected passenger stops walking and stays standing somewhere on the platform. The passenger stands typically in a position that is perpendicular, on the frontal or coronal plane, to the rails. The confederate, carrying a portable metro map, approaches walking parallel to the rails and stops when the tip of her shoe is at a rough 10 cm distance from the passenger's. The result is a side-by-side arrangement in which confederate and participant form an approximate right angle on the frontal plane. The script divides the interaction in two stages involving different verbal contents and body postures. The first stage consists in locating items on a portable map with confederate and passenger side-by-side. In the second stage, the confederate shifts to a close face-to-face position, asking the passenger to estimate the duration of the trip ahead of her. After the passenger's reply, the confederate laments being late for an important appointment, emphasizes that she needs to contact the person she has to meet, but regrets that her cell phone has run out of battery. After the passenger's reply to this indirect request, a researcher intervenes to unmask the plot and inform the passenger that the interaction has been recorded, requesting consent to process the collected image and audio files. The passenger is then invited to answer to a short questionnaire on sociodemographics.

Measurements and outcome variables

The demographic variables that were measured with the questionnaire are age, educational achievement, income, and religion.

The outcomes reported in the present article rely exclusively on the audio recordings collected in the experiment. The outcomes describing helping and involvement behaviors, measured from the video files, have been reported elsewhere (Aranguren et al., 2021). The sound files were produced in WAV 16-bit format at a sampling rate of 44.1 kHz per second. Since the hijab represents mainly a visual stimulus, the fact of using only audio files to take the measurements guarantees complete blindness to the experimental condition, as neither the content of the conversations nor the name of the sound files provide any clues to it.

Manon Toutain and Alban Lemasson performed the measurements. The observation period of acoustic measurements are single words that were found to be recurrent in the audio files

pertaining to passengers from different groups by condition and sex. Using the program PRAAT (Boersma, 2001), acoustic measurements were performed on a corpus of the following frequently occurring (and emotionally neutral) words: “là”, “ligne”, “minute”, “oui”, and “voilà”. The common observation period for all the non-acoustic outcomes is the entire dialog, from the confederate’s opening to the last sentence interpretable as the closure of the exchange. Acoustic measurements were only performed on audio signals in which low background noise permitted measurements of satisfactory quality. Except for speaking time, which was directly measured, all the other non-acoustic outcomes result from ratings. One fourth of the sample was recorded by a second, independent coder, resulting in satisfactory reliability coefficients for all ratings (ordinal and nominal Cohen’s kappas above 0.7).

Indexes of arousal

1) *Speech acoustics*. We operationalize vocally signaled arousal as “tense voice” (Frick, 1985; Juslin & Laukka, 2001; Scherer, 2003; Sulter & Wit, 1996). The source–filter theory states that vocal signals result from a two-stage production, with the glottal wave generated in the larynx (the source), being subsequently filtered in the supralaryngeal vocal tract (the filter, Briefer, 2012; Taylor & Reby, 2010) Tense voice is characterized, among others, by shorter vocalizations, by increased fundamental frequency and amplitude and their respective perturbations (“source”), as well as by a rise in the frequency of all formants (“filter”).

The outcomes that we used to investigate these dimensions are as follows: fundamental frequency (Mean fundamental frequency [F_0] in Hz, Maximum F_0 in Hz, Minimum F_0 in Hz, F_0 coefficient of variation), vocal amplitude (Mean intensity amplitude in dB, Maximum intensity amplitude in dB, Minimum intensity amplitude in dB, Intensity coefficient of variation), duration in seconds, vocal tract filtering effects (Highest-pitched frequency in Hz, Temporal position of the maximum amplitude in % of the total duration, First formant in Hz, Second formant in Hz, Third formant in Hz), vocal perturbation or lack of control (Wiener entropy – a measure of tonality/randomness going from 0 to minus infinity with 0 being a White noise -, F_0 disturbance or “jitter” in %, F_0 amplitude disturbance or “shimmer” in %)

2) *Emotional vocalizations*. In this category we consider laughter and conventionalized onomatopoeia indicative of surprise (e.g. “ah!”, “oh!”) or disfluency (“euh” in French; Cook, 1969).

Indexes of intimacy

1) *Lexical intimacy.* The category covers polite words (e.g. “Hello”, “Goodbye”, “You’re welcome”) and apology words (e.g. “I’m sorry”, “Excuse me”, “I beg your pardon”).

2) *Pragmatic intimacy*. Included here are utterances that contextually function to encourage or discourage the continuation of the exchange (ter Maat et al., 2010). We call these "dialog openings" and "dialog closures", respectively. Examples of openings are: "Do you want me to call for you?", "Do you want to text someone?", "The batteries of my cell are dead either but a friend of mine is coming and you can use her phone", "We can go together in the train". Here are some examples of closures: "Ask someone else", "I have to go", "My phone doesn't work", "I can't help you", "It's too complicated".

3) *Conversational intimacy*. As conversational outcomes, we consider speaking time (aggregate length of turns at talk) and the attempt to interrupt the interlocutor, assuming that more time spent speaking and less interruptions reflect higher intimacy (Goldberg, 1990).

Statistical analyses

The models that we estimated are analogous in logic to traditional ANOVAs but computed with Bayesian inference in the context of the Generalized Linear Model (Kruschke, 2015). The common predictors in these models are the experimental condition, the sex of the passenger, and the two-way interaction between these factors. Additional predictors acting as covariates are age and educational achievement, which were in each case discretized into three groups of equal frequency within each sex category. All hyperparameters were given noninformative priors.

In the presence of an important proportion of zeros, all the outcomes measured as counts (that is, all the non-acoustic variables excepting speech duration) were dichotomized. The units of analysis of the corresponding models are unique participant-confederate interactions. The outcome is the probability that the relevant phenomenon occurs at least once in the course of the interaction. The inputs are the dichotomous outcome y, the experimental condition *j*, the sex of the participant *k,* the participant's age group *l*, and the participant's educational level *m*. The model gives the data a Bernoulli distribution and states that

$$(1)\ \text{logit}(\hat{y}_i) = \beta_0 + \beta_1 \text{condition}_{j[i]} + \beta_2 \text{sex}_{k[i]} + \beta_3 \text{ageGroup}_{l[i]} + \beta_4 \text{eduGroup}_{m[i]} + \beta_5 \text{condition.sex}_{j[i],\,k[i],}\ \text{for } i=1\ldots, n.$$

The model on speech duration outcomes is identical to (1) except that the continuous data were given a normal distribution (with the corresponding variance parameter estimated from the data) and the link function is no longer the logit but the identity.

The continuous acoustic outcomes were also assumed to follow a normal distribution with variance estimated from the data. As the acoustic measures pertain to single words with unique characteristics, the challenge posed to the analysis was to find a technique that would at the same time incorporate the singularity of each word (to avoid spurious generalization) while giving an

overall estimation of the effect of the hijab across words (to offer the synoptic view we were after). The solution adopted is a hierarchical version of the Anova-like model that accomplishes "partial pooling" of the estimates (Gelman et al., 2013) and allows for heteroscedasticity across groups defined by word and sex of the passenger. This type of model includes the word as an additional predictor, as well as the two-way interactions between word and sex and between word and condition, and also the three-way interaction between sex, word, and condition. The units of analysis of these models are unique word-participant combinations. The inputs are the continuous outcome y, the experimental condition *j*, the sex of the participant *k,* the participant's age group *l*, the participant's educational level *m*, and the uttered word *o.* The model gives the data a normal distribution, allowing the variance parameter to differ across groups defined by word and sex combinations, and states that

$$(2)\ \hat{y}_i = \beta_0 + \beta_1 \text{condition}_{j[i]} + \beta_2 \text{sex}_{k[i]} + \beta_3 \text{ageGroup}_{l[i]} + \beta_4 \text{eduGroup}_{m[i]} + \beta_5 \text{condition.sex}_{j[i],\,k[i]} + \beta_6 \text{word}_{o[i]} + \beta_7 \text{condition.word}_{j[i],\,o[i]} + \beta_8 \text{sex.word}_{[i],\,o[i]} + \beta_9 \text{condition.sex}_{j[i],\,k[i]} + \beta_{10} \text{condition.sex.word}_{j[i],\,k[i],\,o[i]},\ \text{for i=1…, n.}$$

An additional complication concerns the models on the minimum, maximun and mean values of fundamental frequency and of vocal amplitude, and the individual measures of the three formants. Each of these three groups of variables was subjected to a hierarchical Anova-like model with multiple outcomes. In these more complex models, a first additional predictor estimates varying intercepts for each individual outcome *p* and another additional predictor estimates varying intercepts for each passenger-word combination *q* (that is, the model "knows" that the three measurements come from the same word uttered by the same participant). The resulting model states that

$$(3)\ \hat{y}_i = \beta_0 + \beta_1 \text{condition}_{j[i]} + \beta_2 \text{sex}_{k[i]} + \beta_3 \text{ageGroup}_{l[i]} + \beta_4 \text{eduGroup}_{m[i]} + \beta_5 \text{condition.sex}_{j[i],\,k[i]} + \beta_6 \text{word}_{o[i]} + \beta_7 \text{condition.word}_{j[i],\,o[i]} + \beta_8 \text{sex.word}_{[i],\,o[i]} + \beta_9 \text{condition.sex}_{j[i],\,k[i]} + \beta_{10} \text{condition.sex.word}_{j[i],\,k[i],\,o[i]} + \beta_{11} \text{outcome}_{p[i]} + \beta_{12} \text{passengerWord}_{q[i]},\ \text{for i=1…, n.}$$

To approximate the posterior distribution of the parameters of interest we used Markov chain Monte Carlo (MCMC) sampling as implemented by the software Jags (Plummer, 2003) via the programming language R (R Core Team, 2017). We checked that the samples were representative of the posterior distribution through visual examination of trace plots and density plots, on the one hand, and consideration of the Gelman-Rubin statistic of convergence, on the other. None of these checks gave any signs of unrepresentativeness. On the other hand, we checked that the generated samples were large enough (and therefore accurate and stable) by

considering a measure called the "effective sample size". The estimates of all the parameters reported below rest on effective sample sizes of at over 10,000.

When using Bayesian inference, results come in the form of probability distributions. Every parameter derived from the model (e.g. the simple effect of the experimental condition among women) receives an individual posterior probability distribution, that is a list of all possible parameter values and their corresponding estimated probabilities, which together sum to 1. In a Bayesian framework, testing a null hypothesis amounts to asking if the posterior probability of the relevant parameter is sufficiently different from the parameter value 0. To facilitate this assessment, we offer a graphical display of the posterior 90% (alpha=0.10) and 95% (alpha=0.05) intervals representing the credible values of the parameters of interest. Being central, these intervals provide two-tailed hypothesis tests. The parameters represented in the enclosed plots quantify the main effect of the experimental condition, the simple effects within each sex group, and the difference between these simple effects (i.e. the condition*sex interaction). At a given alpha level, when the posterior interval intersects the value 0 on the x axis of the plots, we accept the null hypothesis of no effect of the hijab. Conversely, when the posterior interval does not intersect 0 on the x axis, we reject the null hypothesis at the corresponding alpha level.

Note on reporting style. It is inherent to Bayesian inference to describe the output from each model, namely parameter values, as intervals (more precisely, as posterior probability distributions) instead of point estimates. The type of interval considered here is known as the "central posterior interval" (ref Gelman), and provides the equivalent of a two-tailed test. Unless otherwise indicated, the default alpha level of all the reported central posterior intervals is the standard 5%. Sample sizes are not provided in separate tables but incorporated to the Figures that present the parameter estimations in graphical form, facilitating the reader's access to the sample size underlying the estimation of every single reported parameter. While the Figures graphically report all the relevant main, simple, and interaction effects, only those that credibly differ from zero are verbally highlighted in the text.

Results

Indexes of arousal: speech acoustics (Figures 1 and 2)

Fundamental frequency (Figures 1a and 1b). Supporting Prediction 4a, the model estimates a credible interaction effect between the experimental condition and the sex of the passenger, so that

the hijab raises fundamental frequency among males but lowers it among females. Further, confirming Prediction 3a, the simple effect of the hijab among men is to increase the fundamental frequency by between slightly more than 0 and 25 Hz (alpha=0.10).

Vocal amplitude (Figures 1c and 1d). Supporting Prediction 1a, the main effect of the hijab is to increase vocal amplitude by between 0.5 dB and 1.9 dB. An accompanying interaction effect indicates that the rise is credibly larger among men then among women. Confirming Prediction 3a, the estimated simple effect among men is an increase in amplitude ranging from 0.7 dB to 2.8 dB, whereas the corresponding simple effect among women is not credible. Contradicting Prediction 1a, the model on the intensity coefficient of variation estimates a credible negative main effect of the hijab (alpha=0.10). In accordance with Prediction 2a, however, the model estimates a credible simple effect among women, consisting in a decrease in this coefficient ranging from 0.01 to 0.07.

Vocal tract filtering effects (Figure 2a, 2b and 2c). Confirming Prediction 1a, the model on the three first formants yields a positive main effect of the hijab ranging from 40 Hz to 640 Hz. Additionally, a credible simple effect among women roughly in the same interval contradicts Prediction 2a. However, supporting Prediction 4a, the model on the highest-pitched frequency estimates a cross-over interaction effect in the expected direction, indicating that the difference in effect between men and women ranges between 10 Hz and 110 Hz. Last, this model also yields a simple effect among women consisting in a decrease going from over 2 Hz to nearly 60 Hz (alpha=0.10), confirming Prediction 2a.

Vocal perturbation (Figures 2d, 2e and 2f). Disconfirming Prediction 1a, the model on the Wiener entropy estimates a negative main effect of the hijab ranging from 0.2 to 0.7, indicating that in the hijab condition signal randomness is lower. Further, a credible simple effect among women in the expected direction confirms Prediction 2a, but an equally credible simple effect among men of same sign contradicts Prediction 3a.

Indexes of arousal: emotional vocalizations (Figure 3)

Supporting Prediction 3a, the simple effect of the hijab among men is to increase the probability of hearing the arousal-related onomatopoeia “ah” or “oh” ranging from slightly more than 0% to nearly 40% (alpha=0.10). The model also estimates a credible cross-over interaction of the expected sign, confirming Prediction 4a.

Pragmatic intimacy (Figures 4c and 4d)

Confirming Prediction 1b, the hijab elicits an overall decrease in the probability of an opening and a similar increase in the probability of a closure roughly ranging from 10% to 40%. On both outcomes, a simple effect is credible among women but not men, further disconfirming Prediction 2b. Similarly, a credible interaction effect between the experimental condition and the sex of the passenger indicates that women's decrease in pragmatic intimacy in response to the hijab is larger than men's.

Conversational intimacy (Figures 4e and 4f)

Supporting Prediction 3b, the simple effect of the hijab among men on the probability of interrupting the confederate is a credible increase ranging from 5% to 40% (Figure 4e). Confirming Prediction 4a, the model also estimates a credible cross-over interaction effect between condition and sex in the expected direction.

Post hoc model on pragmatic intimacy outcomes

Rationale. The results from the analysis of dialog openings and closures, while they confirm Prediction 1b, clearly contradict the expectation that women should show greater intimacy in the hijab condition (Prediction 2b). To assist the interpretation of this partly unexpected result, a post hoc model was estimated by adding one control predictor to formula (1), namely whether the passenger offers assistance or not. The latter variable was measured independently of the present research on vocal arousal and intimacy and has been previously analyzed as an outcome, using the hijab, sex, and other factors as predictors (Aranguren et al., 2021). The rationale for reusing this variable as a control predictor in a model estimating the probability of an opening or a closure lies in the fact that the typical instances of these behaviors (see "Measurements and outcome variables" above) seem to illustrate, respectively, the manner in which the decision to help or not to help the confederate is verbally communicated. The aim of the post hoc model is to quantify this association. If openings are strongly associated with helping, and closures with its absence, interpretation can be facilitated by considering these three outcomes together.

Results. The first post hoc model estimates that the probability of a dialogue opening raises by between 57% and 77% when the passenger offers assistance. Similarly, the second post hoc model calculates that the probability of a dialog closure increases by between 75% and 90% when the passenger *does not* offer assistance. Importantly, the main effects and the simple effects among women found in the planned models reported above remain credible in spite of the addition of this powerful control predictor. There is evidence, then, of a strong association

between openings and the decision to help, and between closures and the decision not to help, but no evidence of redundancy between these outcomes.

Discussion

Given that we have considered multiple outcomes that do not necessarily converge in the same qualitative result, we need some rationale for assessing the degree to which the predictions find support in the data. We proceed in two steps. Assuming that all measures are equally informative, we first provide a purely quantitative assessment of the number of outcomes that confirm or disconfirm each prediction. When the evidence is equivocal, we then qualify this assessment, in Bayesian style, by according greater credibility to the results that agree with prior knowledge.

We assess the degree to which the starting predictions are supported by the data using the following scale. If the number of outcomes contradicting the prediction is lower than the number of outcomes confirming it, we deem the support to the prediction "favorable". When the numbers of contradictions and confirmations are equal, we say that the evidence is "mixed". If contradictions outnumber confirmations, we consider the data to be "unfavorable".

Prediction 1 stated that participants should show 1a) more arousal and 1b) less intimacy in the hijab condition. The support for 1a) is mixed, whereas that for 1b) is favorable. Prediction 2 expected women to 2a) decrease arousal and 2b) increase intimacy in response to the hijab, irrespective of what men do. The evidence for 2a) is favorable, while that for 2b) is unfavorable. Prediction 3 anticipated the opposite simple effect of the hijab for men, expecting them to 3a) increase arousal and 3b) decrease intimacy in the hijab condition, irrespective of what women do. The evidence is mostly favorable to 3a) and favorable to 3b). Last, Prediction 4 foresaw a cross-over interaction such that 4a) arousal should decrease among women but increase among men and 4b) intimacy should increase among women but decrease among men. The support for 4a) is favorable, while that for 4b) is mostly unfavorable.

From this purely quantitative assessment, a difference appears to emerge between arousal and intimacy indexes. Arousal outcomes tend to conform to the hypothesis of a sex contrast in the sign of the response to the hijab, whereas intimacy indexes exhibit consistence in sign between sex groups. Men show higher arousal and, less clearly, women show lower arousal. At the same time, both men and women indicate by their vocal behavior lower intimacy.

It is no novelty that multiple acoustic measures designed to capture stress or arousal do not converge (Giddens et al., 2013). However, from a historical perspective not all acoustic measures appear to be equally informative, as we have so far methodologically assumed. As a matter of fact, if there is a single vocal change that has consistently been found to reflect

underlying stress or arousal, that is an increase in the fundamental frequency of vocalizations (Giddens et al., 2013; Russell et al., 2003). From this point of view, in the face of conflicting results across outcomes, it is reasonable to lend greater credence to the analysis of the fundamental frequency. As we saw, the data related to this outcome support the prediction of a cross-over interaction indicating that in response to the hijab men show more but women less arousal. Over and above the consistent effects of stress or arousal on the fundamental frequency, on the perceptual side increases on this acoustic variable have been repeatedly found to give rise to negative attributions. For example, Apple and colleagues (1979) played normal and artificially modified voices (raising or lowering the pitch by 20%) to subjects who were then invited to make personal attributions to speakers. Subjects judged higher-pitched voices less truthful, less empathic and more nervous than unmodified voices. Natural voices recorded by a telephone service operator with higher F0, higher formant and higher amplitude were also judged more negatively (Forsell et al., 2007; Laukka et al., 2011).

Interpreting the difference in response to the hijab between the sexes

The vocal response to the hijab, then, is for men to increase arousal and decrease intimacy, and for women to decrease both arousal and intimacy. Men's response shows consistency between arousal and intimacy indexes, whereas women's does not. This observation is in line with the results of a previously completed analysis involving the same sample of participants but dealing with other outcomes, namely *non*verbal intimacy and helping behavior (Aranguren et al., 2021). In that analysis, the observation period of nonverbal intimacy preceded that of helping behavior. In the first period women showed more "proxemic" intimacy by interacting closer to the confederate if she wore the hijab. But in the second period, women helped the confederate more often if she *did not* wear the scarf. Further, among women a higher level of proxemic intimacy during the first period did not predict a higher probability of helping the confederate in the second one. In contrast, men's helping behavior did not vary in response to the hijab. And more importantly, higher proxemic intimacy in one period did predict a higher probability of offering assistance to the confederate in the subsequent one. Bringing together the findings from that study with those of the present one, among men vocal arousal, verbal and nonverbal intimacy, and helping behavior can be regarded as analytically distinct dimensions of an overarching negative attitude towards hijab-wearing women. Among women, verbal intimacy and helping behavior (which our post hoc model shows to be strongly correlated) convey negativity, whereas nonverbal intimacy and vocal arousal tend to signal positivity, making general conclusions more elusive.

One speculation to restore the consistency of women's behavior is to posit that the cost of a

friendly or positive signal moderates the effect of the hijab on impression management. More concretely, in this interpretation we expect women to deliberately show positive behaviors in interaction with the hijab-wearing woman as long as the cost of those behaviors remains below a certain threshold. Women probably perceive that adopting a relaxed tone of voice and/or accepting to interact at a short distance for a few seconds is less costly than lending their mobile and/or encouraging the dialog's continuation. Speaking softly and accepting to stand close do not imply the risk of financial loss; lending the phone does. Those behaviors do not create a commitment to accommodate whatever the interaction partner might request at a later point in time; pursuing the conversation does.

The fact that the confederate is a female might have influenced male passengers differently from female passengers. For example, whether men generally interrupt speakers more often than women is currently disputed in the literature on language and sex (Anderson & Leaper, 1998; James & Clarke, 1993; Yuan et al., 2007), but it has been documented that women are more often overlapped than men (Yuan et al., 2007). Future research could explore if the specific behaviors used to communicate lower intimacy to a negatively evaluated other depend on the sex of the individuals placed in the positions of author and target of those behaviors.

Differential informativeness of acoustic measures

In view of the results, not all acoustic measures have proved equally informative as indexes of arousal. We first discuss the possibility that informativeness might depend on the sex of the vocalizer. We then consider whether the information value of each acoustic measure, for the naturalistic data we have dealt with, might depend on whether the acoustic measure reflects source, filter, perturbation, or variability characteristics of the voice.

Here, as in other studies (Özseven et al., 2018; Tolkmitt & Scherer, 1986), the acoustic measures that best encode arousal changes differ between men and women. Notably the relative importance of the fundamental frequency (source) with respect to the measures explaining the energy distribution along the frequency spectrum (filter) differs to some extent between sexes. The differences in voice acoustics between human males and females are to a large extent a consequence of differences in the morphology of the larynx (Sulter & Wit, 1996; Titze, 1989). The usual approach is to understand sex differences in voice acoustics in terms of a proportional re-scaling of larynx characteristics. For example, the membranous length of the vocal folds is on average 60% larger in males than in females. By virtue of the principle of proportionality, this leads to the prediction that the fundamental frequency of the typical male voice should be, on average, 60% lower than that of the typical female voice, a prediction that is in general agreement

with the data (Titze, 1989). However, the male larynx is not 60% larger in all respects. For example, the gap in overall larynx size between the sexes is only 20%. This raises the question whether all sex differences in acoustic measures can be expected to conform to the 1.6 rescaling factor found to be in correspondence with the average gap in fundamental frequency between male and female voices. For example, it has been shown that sex differences in vocal amplitude are better predicted by a 1.2 rescaling factor (Titze, 1989). If the gap in fundamental frequency between male and female voices is larger than the corresponding gap in amplitude, this is another way of saying that the acoustic effect of larynx dimorphism depends on the acoustic variable under consideration.

This argument has been extended to the relationship between sex differences in general body morphology and various measures of vocal acoustics. It has been suggested that formant frequency, being constrained by the size and shape of the larynx and ultimately by that of the skull, is a better predictor of body morphology than the fundamental frequency (Pisanski et al., 2016). As said, the latter measure depends on the size of the vocal folds, which in turn may develop and grow independently of the rest of the body. A further extension of the argument pleads for a sex difference in the implementation of physiological arousal that might lead to differences in the ability of various acoustic measures to encode bodily activation (Giddens et al., 2013).

The extension we propose imputes differences in the ability of various acoustic measures to index physiological arousal between the sexes to morpho-physiologically constrained sex differences in the degree of plasticity that males and females exhibit with regard to the source and filter characteristics of the voice. Arousal consists in physiological modifications such as cardiovascular alterations, autonomic reactions, neuroendocrine and immunologic as well as psycho-neuro-immunologic changes (Hansen & Patil, 2007). Regarding the “source” characteristics of the voice, by increasing respiration rate, arousal creates greater subglottal pressure while speaking, leading to higher fundamental frequencies. Mouth dryness also plays a role in changing the muscle activity of the larynx. When in comes to the “filter” characteristics, changes in the activity of the muscles controlling the tongue, lips and jaws result in changes that shape the resonant cavities of the vocal system, affecting the distribution of energy along the frequency spectrum. The above-mentioned anatomical dimorphism explains why some physiological changes may impact differently the vocal fold vibration and the tract resonances in men and women (Sulter & Wit, 1996; Titze, 1989). Also, several authors have discussed firstly a sex-determined automatic nervous system (ANS) and sex-differences in the hypothalamic adrenal axis (HPA) response to arousal, and secondly a relationship between ANS and HPA and voice

parameters (Giddens et al., 2013; Kudielka & Kirschbaum, 2005; Pisanski et al., 2016). This could also explain why the vocal implementation of arousal differs between males and females.

Over and above the question of dimorphism, a final assessment of the differential informativeness of acoustic measures is in order. Overall, the measures that yielded predicted differences between the experimental conditions belong either to the source/phonation or the filter/resonance families, in overall agreement with the literature on vocal indexes of stress (Banse & Scherer, 1996; Forsell et al., 2007; Juslin & Laukka, 2001; Laukka et al., 2008, 2011; Menahem, 1983; Özseven et al., 2018; Scherer, 1986; Scherer, 2003). With the notable exception of the Wiener entropy (a measure of tonality), none of the variability and perturbation indexes (e.g. shimmer or jitter) resulted in credible differences. Unexpectedly, the analysis of the Wiener entropy indicated that vocalizations were more tone-like, as opposed to noise-like, when the confederate wore the hijab. Although greater noise or "hoarseness" might suggest itself as a natural vocal concomitant of arousal, reduced noise in response to stressors has been repeatedly documented (reviewed in Giddens et al., 2013; Van Puyvelde et al., 2018). Our study provides an additional confirmation of this counterintuitive tendency.

Although not all vocal indexes appear to be equally informative in a naturalistic setting, this study demonstrates the viability of decoding arousal from spontaneous vocalizations in real-life situations. For male participants, it also demonstrates consistency between vocal and lexical indexes of arousal, and also a negative relationship between these and measures of intimacy.

References

Almeida, L. N. A., Lopes, L. W., Costa, D. B. da, Silva, E. G., Cunha, G. M. S. da, Almeida, A. A. F. de, Almeida, L. N. A., Lopes, L. W., Costa, D. B. da, Silva, E. G., Cunha, G. M. S. da, & Almeida, A. A. F. de. (2014). Características vocais e emocionais de professores e não professores com baixa e alta ansiedade. *Audiology - Communication Research*, *19*(2), 179–185. https://doi.org/10.1590/S2317-64312014000200013

Amodio, D. M. (2009). Intergroup anxiety effects on the control of racial stereotypes: A psychoneuroendocrine analysis. *Journal of Experimental Social Psychology*, *45*(1), 60–67. https://doi.org/10.1016/j.jesp.2008.08.009

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670–682. https://doi.org/10.1038/nrn3800

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and

self-report. - PsycNET. *Journal of Personality and Social Psychology*, *84*(4), 738–753. https://doi.org/10.1037/0022-3514.84.4.738

Anderson, K. J., & Leaper, C. (1998). Meta-Analyses of Gender Effects on Conversational Interruption: Who, What, When, Where, and How. *Sex Roles*, *39*(3), 225–252. https://doi.org/10.1023/A:1018802521676

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. - PsycNET. *Journal of Personality and Social Psychology*, *37*(5), 715–727. https://doi.org/10.1037/0022-3514.37.5.715

Aranguren, M. (2021). *Interactional discrimination against hijab-wearing women in public places: Field experiments* [Working paper]. https://hal.archives-ouvertes.fr/hal-03094580

Aranguren, M., Madrisotti, F., Durmaz-Martins, E., Gerger, G., Wittmann, L., & Méhu, M. (2021). *Responses to the islamic headscarf in everyday interactions depend on sex and locale: A field experiment in the metros of Brussels, Paris, and Vienna on helping and involvement behaviors (submitted)*. https://hal.archives-ouvertes.fr/hal-03107103

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expressoin. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bond, M. H. (1972). Effect of an impression set on subsequent behavior. *Journal of Personality and Social Psychology*, *24*, 301–305. https://doi.org/doi.org/10.1037/h0033716

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1–20. https://doi.org/10.1111/j.1469-7998.2012.00920.x

Brown, L. M., Bradley, M. M., & Lang, P. J. (2006). Affective reactions to pictures of ingroup and outgroup members. *Biological Psychology*, *71*(3), 303–311. https://doi.org/10.1016/j.biopsycho.2005.06.003

Chekroud, A. M., Everett, J. A. C., Bridge, H., & Hewstone, M. (2014). A review of neuroimaging studies of race-related prejudice: Does amygdala response reflect threat? *Frontiers in Human Neuroscience*, *8*, 179. https://doi.org/10.3389/fnhum.2014.00179

CNCDH. (2019). *La lutte contre le racisme, l'antisémitisme et la xénophobie: Rapport 2018* (p. 345). Commission Nationale Consultative des Droits de l'Homme; La documentation française. https://www.cncdh.fr/sites/default/files/23072019_version_corrigee_rapport_racisme.pdf

Cook, M. (1969). Anxiety, Speech Disturbances and Speech Rate. *British Journal of Social and Clinical Psychology*, *8*(1), 13–21. https://doi.org/10.1111/j.2044-8260.1969.tb00580.x

Coutts, S. M., Schneider, F. W., & Montgomery, S. (1980). An investigation of the arousal model of interpersonal intimacy. *Journal of Experimental Social Psychology*, *16*, 545–561.

Dambrun, M., Desprès, G., & Guimond, S. (2003). On the multifaceted nature of prejudice: Psychophysiological responses to ingroup and outgroup ethnic stimuli. *Current Research in Social Psychology (University of Iowa)*, *8*(14), 12 p.

Ersanilli, E., & Koopmans, R. (2013). *The Six Country Immigrant Integration Comparative Survey (SCIICS)—Technical report*. WZB Discussion Paper SP IV 2013-102.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. http://dx.doi.org/10.1037/0022-3514.69.6.1013

Forsell, M., Elenius, K., & Laukka, P. (2007). Acoustic correlates of frustration in spontaneous speech. *TMH-QPSR*, *50*(1), 37–40.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. - PsycNET. *Psychological Bulletin*, *97*(3), 412–429. https://doi.org/10.1037/0033-2909.97.3.412

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal Indices of Stress: A Review. *Journal of Voice*, *27*(3), 390.e21-390.e29. https://doi.org/10.1016/j.jvoice.2012.12.010

Goldberg, J. A. (1990). Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, *14*(6), 883–903. https://doi.org/10.1016/0378-2166(90)90045-F

Graves, R. E., Cassisi, J. E., & Penn, D. L. (2005). Psychophysiological evaluation of stigma towards schizophrenia. *Schizophrenia Research*, *76*(2), 317–327. https://doi.org/10.1016/j.schres.2005.02.003

Greenland, K., Xenias, D., & Maio, G. (2012). Intergroup anxiety from the self and other: Evidence from self-report, physiological effects, and real interactions. *European Journal of Social Psychology*, *42*(2), 150–163. https://doi.org/10.1002/ejsp.867

Guglielmi, R. S. (1999). Psychophysiological Assessment of Prejudice: Past Research, Current Status, and Future Directions. *Personality and Social Psychology Review*, *3*(2), 123–157. https://doi.org/10.1207/s15327957pspr0302_3

Hansen, J. H. L., & Patil, S. (2007). Speech Under Stress: Analysis, Modeling and Recognition. In C. Müller (Ed.), *Speaker Classification I: Fundamentals, Features, and Methods* (pp. 108–137). Springer. https://doi.org/10.1007/978-3-540-74200-5_6

Ickes, W., Patterson, M. L., Rajecki, D. W., & Tanford, S. (1982). Behavioral and cognitive consequences of reciprocal versus compensatory responses to preinteraction expectancies. *Social Cognition*, *1*(2), 160–190. https://doi.org/10.1521/soco.1982.1.2.160

James, D., & Clarke, S. (1993). Women, men, and interruptions: A critical review. In *Gender and conversational interaction* (pp. 231–280). Oxford University Press.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. - Abstract—Europe PMC. *Emotion*, *1*(4), 381–412. https://doi.org/10.1037/1528-3542.1.4.381

Kiebel, E. M., McFadden, S. L., & Herbstrith, J. C. (2017). Disgusted but not afraid: Feelings toward same-sex kissing reveal subtle homonegativity. *The Journal of Social Psychology*, *157*(3), 263–278. https://doi.org/10.1080/00224545.2016.1184127

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*, 99–117. https://doi.org/10.1007/s10772-011-9125-1

Kruschke, J. K. (2015). *Doing Bayesian data analysis*. Academic Press.

Kudielka, B. M., & Kirschbaum, C. (2005). Sex differences in HPA axis responses to stress: A review. *Biological Psychology*, *69*(1), 113–132. https://doi.org/10.1016/j.biopsycho.2004.11.009

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Frans, Ö., Faria, V., Michelgård, Å., Appel, L., Fredrikson, M., & Furmark, T. (2008). In a Nervous Voice: Acoustic Analysis and Perception of Anxiety in Social Phobics' Speech. *Journal of Nonverbal Behavior*, *32*(4), 195. https://doi.org/10.1007/s10919-008-0055-9

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, *25*(1), 84–104. https://doi.org/10.1016/j.csl.2010.03.004

Littleford, L. N., Wright, M. O., & Sayoc-Parial, M. (2005). White Students' Intergroup Anxiety During Same-Race and Interracial Interactions: A Multimethod Approach. *Basic and Applied Social Psychology*, *27*(1), 85–94. https://doi.org/10.1207/s15324834basp2701_9

Mahaffey, A. L., Bryan, A., & Hutchison, K. E. (2005). Using Startle Eye Blink to Measure the Affective Component of Antigay Bias. *Basic and Applied Social Psychology*, *27*(1), 37–45. https://doi.org/10.1207/s15324834basp2701_4

March, D. S., & Graham, R. (2015). Exploring implicit ingroup and outgroup bias toward Hispanics. *Group Processes & Intergroup Relations*, *18*(1), 89–103. https://doi.org/10.1177/1368430214542256

Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton.

Menahem, R. (1983). La voix et la communication des affects. *L'année Psychologique*, *83*(2), 537–560.

Mendes, W. B., Blascovich, J., Lickel, B., & Hunter, S. (2002). Challenge and Threat During Social Interactions With White and Black Men. *Personality and Social Psychology Bulletin*, *28*(7), 939–952. https://doi.org/10.1177/014616720202800707

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869. https://doi.org/10.1086/283219

Özseven, T., Düğenci, M., Doruk, A., & Kahraman, H. İ. (2018). Voice Traces of Anxiety: Acoustic Parameters Affected by Anxiety Disorder. *Archives of Acoustics*, *43*(4), 625–636. https://doi.org/10.24425/aoa.2018.125156

Patterson, M. L. (1982). A sequential functional model of nonverbal exchange. *Psychological Review*, *89*(3), 231–249. https://doi.org/10.1037/0033-295X.89.3.231

Paulus, A., Renn, K., & Wentura, D. (2019). One plus one is more than two: The interactive influence of group membership and emotional facial expressions on the modulation of the affective startle reflex. *Biological Psychology*, *142*, 140–146. https://doi.org/10.1016/j.biopsycho.2018.12.009

Pew Research Center. (2015). *Five facts about the Muslim population in Europe*. http://www.pewresearch.org/fact-tank/2015/11/17/5-facts-about-the-muslim-population-in-europe/

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. https://doi.org/10.1162/089892900562552

Pisanski, K., Nowak, J., & Sorokowski, P. (2016). Individual differences in cortisol stress response predict increases in voice pitch during exam stress. *Physiology & Behavior*, *163*, 234–238. https://doi.org/10.1016/j.physbeh.2016.05.018

Plummer, M. (2003). *JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling*. *124:10*, 1–10.

R Core Team. (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Russell, J. A., Bachorowski, J. A., & Fernández-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, *54*(1), 329–349.

Scherer, K. (1986). Voice, Stress, and Emotion. In M. H. Appley & R. Trumbull (Eds.), *Dynamics of stress: Physiological, psychological and social perspectives*. Springer.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256.

Sulter, A. M., & Wit, H. P. (1996). Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age. *The Journal of the Acoustical Society of America*, *100*(5), 3360–3373. https://doi.org/10.1121/1.416977

Taylor, A. M., & Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, *280*(3), 221–236. https://doi.org/10.1111/j.1469-7998.2009.00661.x

ter Maat, M., Truong, K. P., & Heylen, D. (2010). How Turn-Taking Strategies Influence Users' Impressions of an Agent. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents* (pp. 441–453). Springer. https://doi.org/10.1007/978-3-642-15892-6_48

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Tolkmitt, F. J., & Scherer, K. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(3), 302–313. https://doi.org/10.1037/0096-1523.12.3.302

van der Noll, J., Saroglou, V., Latour, D., & Dolezal, N. (2018). Western anti-muslim prejudice: Value conflict or discrimination of persons too? *Political Psychology*, *39*(2), 281–301. https://doi.org/10.1111/pops.12416

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01994

Vanman, E. J., Ryan, J. P., Pedersen, W. C., & Ito, T. A. (2013). Probing prejudice with startle eyeblink modification: A marker of attention, emotion, or both? *International Journal of Psychological Research*, *6*, 30–41. https://doi.org/10.21500/20112084.717

Yuan, J., Liberman, M., & Cieri, C. (2007, January 1). Towards an integrated understanding of speech overlaps in conversation. *Paper Presented at the 16th International Conference of Phonetic Sciences*.

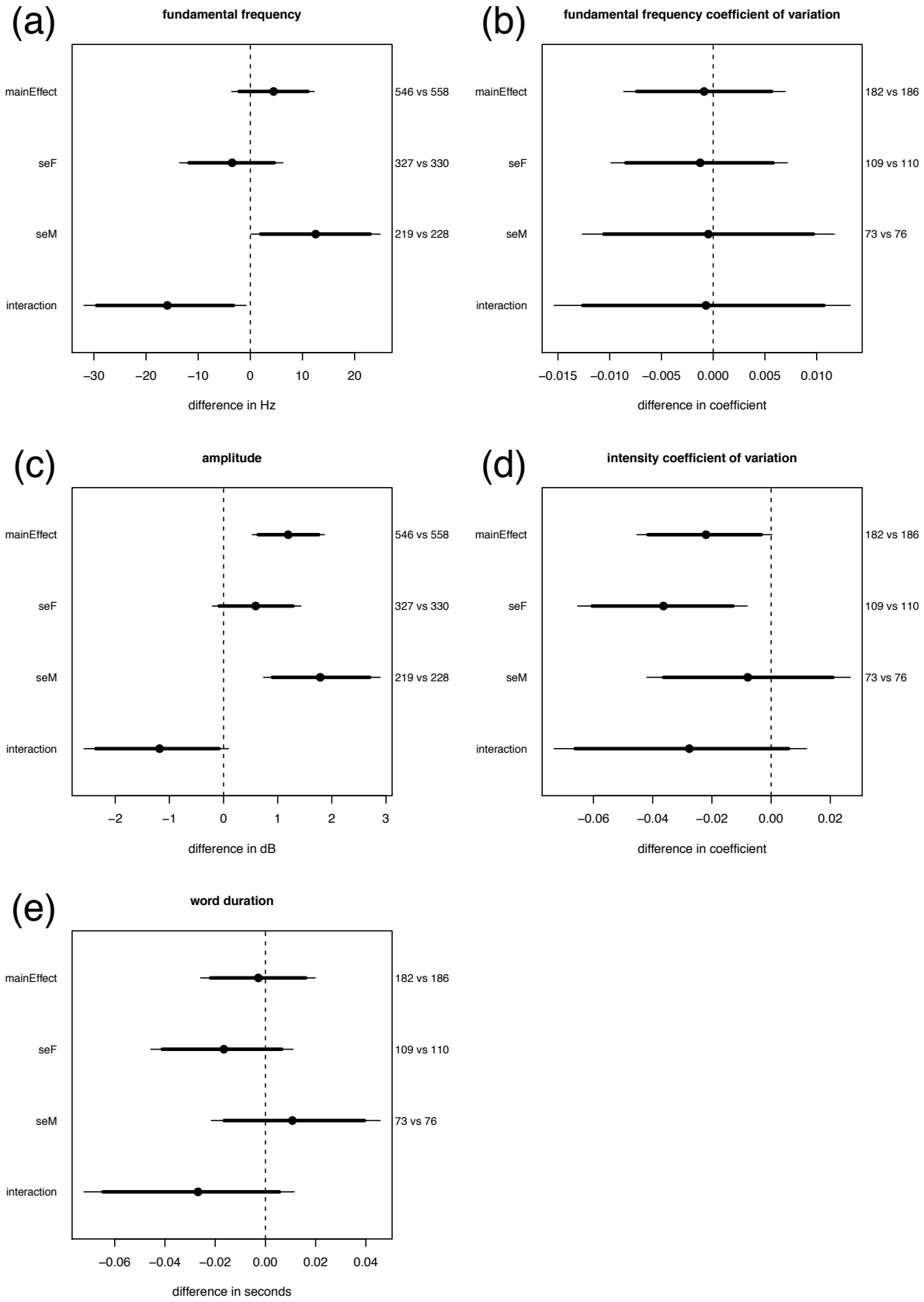

Figure 1. Arousal, speech acoustics 1. Panes (a) and (b): Fundamental frequency. Panes (c) and (d): Amplitude. Pane (e): duration. Within each pane, the graph title specifies the outcome variables under examination. The x axis represents the difference between the hijab and the control condition. The y axis lists the parameters of interest, the main effect, the simple effect among female passengers ("seF"), the simple effect among male passengers ("seM"), and the

difference between these effects, that is the condition*sex interaction. The quantities on the right-hand side of the plot specify the number of observations on which the estimation of the corresponding parameter directly relies; the first number refers to the sample size of the hijab group, the second number to that of the control group. Being the substraction of the simple effects, the interaction effect's sample size equals the sum of the simple effects' sample, which equals the total sample. Within the plot area, the dashed vertical line in the middle indicates the location of the value 0, which signifies no difference in approach between the hijab and the conditions conditions. The horizontal segments represent the central 95% posterior intervals of the parameters. The bolder section of the segment corresponds to the central 90% posterior interval and the solid point indicates the median of the distribution. Our decision rule is to reject the null hypothesis of no effect of the hijab if the 95% or 90% posterior interval of a given parameter excludes the value zero. In graphical terms the null hypothesis is rejected when the thin (95%) or bold (90%) segment estimating the difference between the experimental conditions does not intersect the dashed vertical line.

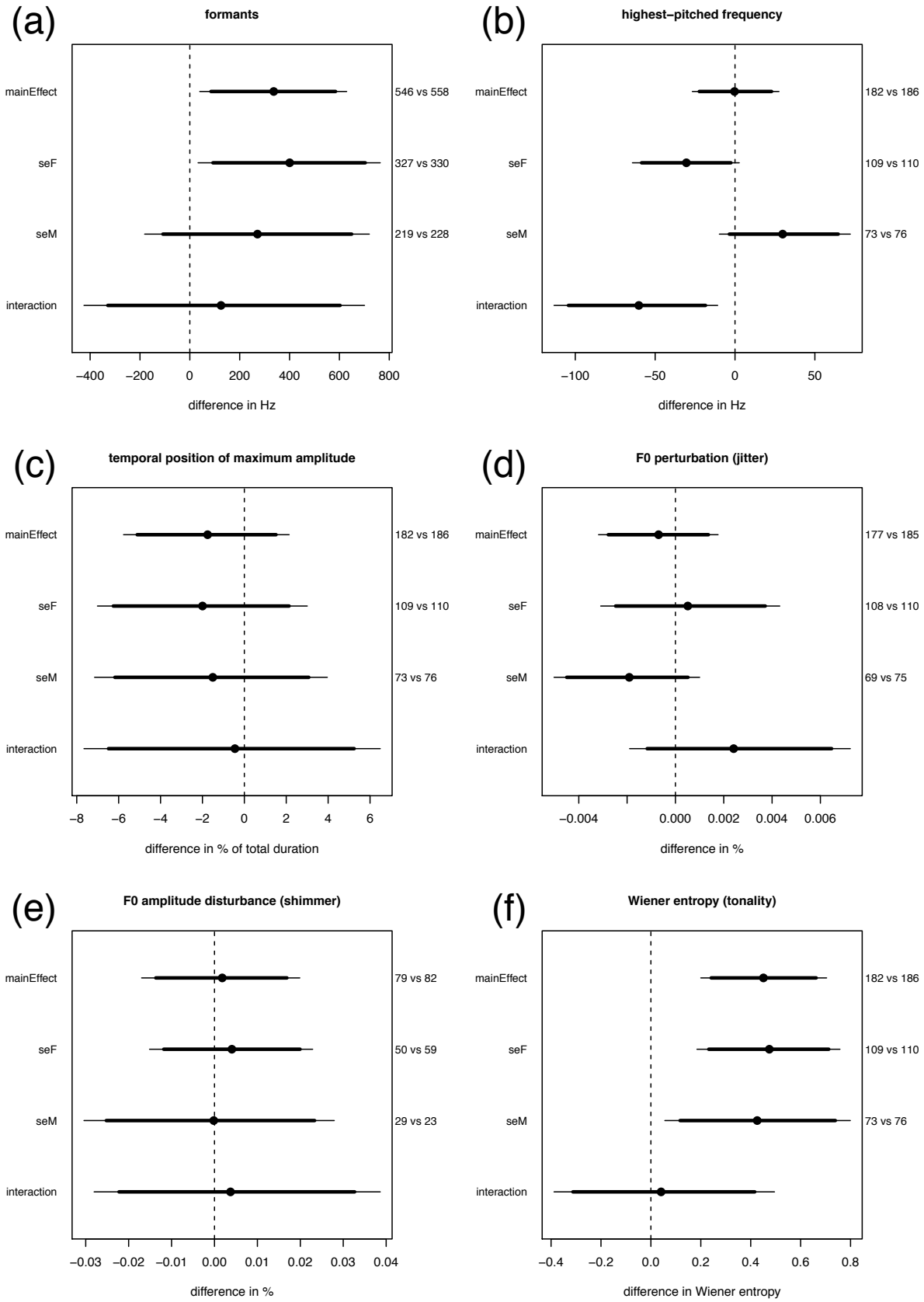

Figure 2: Arousal, speech acoustics 2. Pane (a), (b) and (c): vocal-tract filtering effects. Panes (d), (e) and (f): vocal perturbation. Interpretation: see caption of Figure 1.

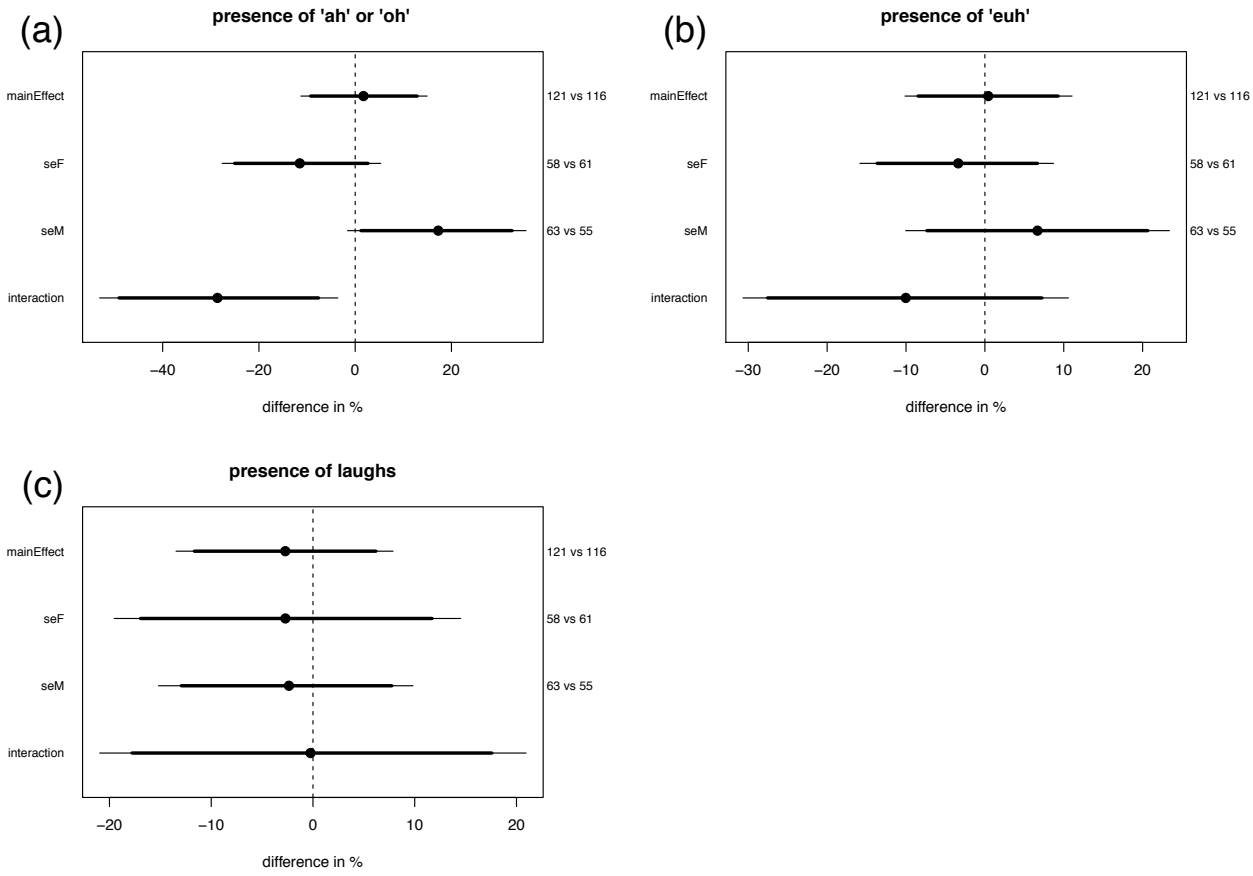

Figure 3. Arousal, emotional vocalizations. Panes (a) and (b): onomatopoeia indicative of arousal. Pane (c): laughs. Interpretation: see caption of Figure 1.

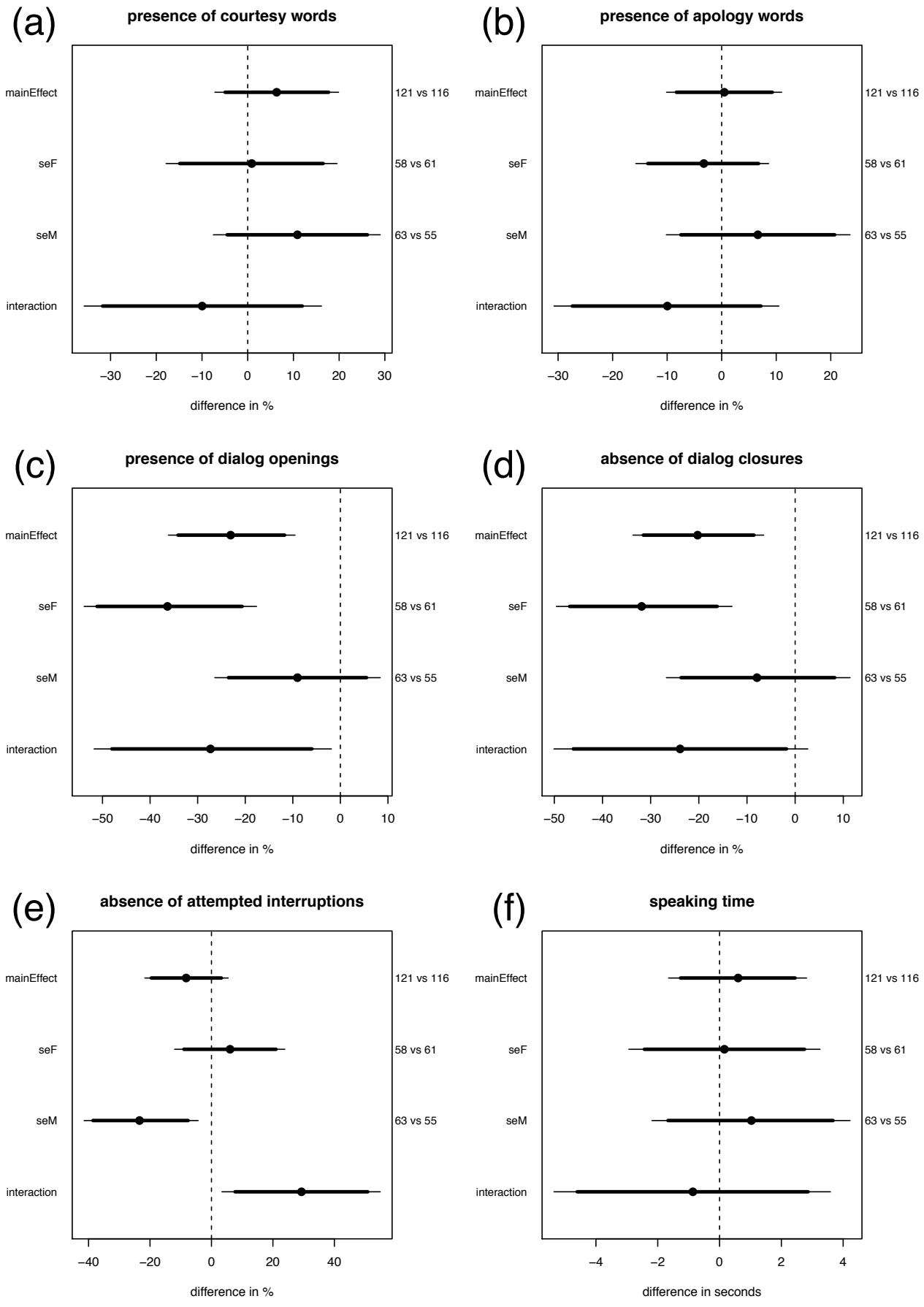

Figure 4: intimacy. Panes (a) and (b): lexical intimacy. Panes (c) and (d): pragmatic intimacy. Panes (e) and (f): conversational intimacy. In all plots, negative values indicate lower intimacy.

Interpretation: see caption of Figure 1.