Speech Communication 124 (2020) 46–67

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars

Ramiro H. Gálvez [a,b,*], Agustín Gravano [a,b], Štefan Beňuš [c,d], Rivka Levitan [e], Marian Trnka [d], Julia Hirschberg [f]

[a] Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina
[b] Instituto de Ciencias de la Computación, CONICET-UBA, Buenos Aires, Argentina
[c] Constantine the Philosopher University in Nitra, Slovakia
[d] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[e] Department of Computer and Information Science, Brooklyn College CUNY, USA
[f] Department of Computer Science, Columbia University, New York, USA

ARTICLE INFO

Keywords:
Entrainment
Prosody
Spoken dialogue systems
Sociolinguistics

ABSTRACT

Entrainment is the tendency of interlocutors to become more similar to each other in their way of speaking. This phenomenon has been repeatedly documented and is associated with multiple social aspects of human-human conversations. However, there is a dearth of research on the effects of spoken dialogue systems (SDSs) with implemented acoustic-prosodic (dis)entrainment policies. The goal of the present work is to provide further empirical evidence on how acoustic-prosodic (dis)entraining policies affect users. In particular, this article focuses on its effects on users' trust toward the SDSs. In the experiments reported here we analyze if and how different acoustic-prosodic (dis)entrainment policies affect users' perception of a system's ability. We collected data from 98 unique participants, all native speakers of Argentine Spanish. Our results suggest that acoustic-prosodic (dis)entrainment in spoken dialogue systems is effectively associated with the way users perceive the capabilities of such systems. Characterizing these effects remains a challenging task. Overall, we observe a positive effect on trust of entrainment on intensity and a negative effect of entrainment on pitch. Estimated effect sizes are far from negligible.

1. Introduction

Voice assistants such as Google Assistant, Amazon Alexa, Microsoft Cortana and Apple Siri have revolutionized the way in which humans and computers interact. Such has been the advance in these natural language user interfaces that they have been embedded not only into several popular operating systems (e.g. Google Assistant in Android, Cortana in Windows, Siri in IOS), but in what, at the time, mass-media called the "*next big arms race in tech*,"[1] were also incorporated into a wide range of dissimilar products, such as smart speakers (e.g. Google Voice, Amazon Echo), cars,[2] and smart appliances.[3]

This rise in the usability and popularity of voice assistants was fueled largely by dramatic improvements in critical subsystems involved in their operation, such as automatic speech recognition (ASR) systems, natural language understanding (NLU) systems, and text-to-speech (TTS) synthesis. These improvements were driven mainly by advances in deep neural networks trained on large corpora. The development of these critical subsystems up to reliable production levels, leads to the importance of studying other complementary components of speech communication.

In *spoken dialogue systems* (SDSs), such as voice assistants, a feature believed to be associated with improvement in user experience is their *naturalness* (Crumpton and Bethel, 2016). Measuring naturalness in dialogue involves a high degree of subjectivity (Hung et al., 2009), but, in the context of SDSs, it is commonly associated with the degree in which SDSs replicate behaviors and patterns observed in human-human

* Corresponding author.
E-mail address: rgalvez@dc.uba.ar (R.H. Gálvez).

[1] See https://www.fastcompany.com/3066831 (Fast Company), http://time.com/4624067 (TIME).
[2] See https://www.bbc.com/news/technology-38526807 (BBC).
[3] See https://technology.inquirer.net/57441 (Inquirer.net).

https://doi.org/10.1016/j.specom.2020.07.007
Received 28 November 2019; Received in revised form 13 May 2020; Accepted 26 July 2020
Available online 30 July 2020

conversations (Marge et al., 2010). That is, it is believed that SDSs which replicate such human-human behaviors will lead to better interactions with users, and thus to better conversation outcomes as well.

A phenomenon that has been repeatedly documented in human-human conversations is the tendency of interlocutors to become more similar to each other in the way they speak. This behavior, known in the literature as *entrainment, accommodation* or *adaptation*, has been shown to occur along several dimensions during human-human interaction, including: pronunciation (Pardo, 2006); choice of referring expressions (Brennan and Clark, 1996); syntactic structure (Reitter et al., 2011); turn-taking cues (Levitan et al., 2015b); choice of intonational contour (Gravano et al., 2015); and *acoustic-prosodic* behavior (Ward and Litman, 2007; Levitan and Hirschberg, 2011). Although prevalent in human-human conversations, the question of why entrainment occurs is still an active research topic, and several theories have been developed to explain it — many of which differ in the degree of control speakers have over the behavior (see, for example, Natale, 1975; Giles et al., 1991; Chartrand and Bargh, 1999; Pickering and Garrod, 2004; 2013).

Entrainment has been associated with multiple social aspects in human-human conversations (Beňuš, 2014), such as degree of success in completing tasks (Nenkova et al., 2008; Reitter and Moore, 2014), perception of competence and social attractiveness (Street Jr, 1984; Levitan et al., 2011; Beňuš et al., 2014; Michalsky and Schoormann, 2017; Schweitzer and Lewandowski, 2014), and degree of speaker engagement (De Looze et al., 2014; Gravano et al., 2015). *Disentrainment* — speakers actively adapting to become more dissimilar to each other (Healey et al., 2014; De Looze et al., 2014; Reichel et al., 2018a) — has also been correlated with social aspects of conversations. Early research documents evidence suggesting that speakers disentrain to show dislike and to distance themselves from their interlocutor. For example, Welsh subjects broadened their Welsh accent significantly when interviewed by an arrogant interviewer with a strong English accent who called Welsh "a dying language with a dismal future" (Bourhis and Giles, 1977). However, more recent research shows that disentrainment may also be related to positive social outcomes. For example, Pérez et al. (2016) show that metrics which consider entrainment and disentrainment behavior capture perceived positive and negative social outcomes of conversations (e.g. engagement, boredom) in a better way than metrics which only consider entrainment behavior.

Even when acoustic-prosodic (dis)entrainment has consistently been reported to occur and correlate with social outcomes across different types of dialogues (e.g. competitive, cooperative), languages (see Levitan et al., 2015a), and tasks, previous research suggests that the phenomenon has many subtleties. For example, evidence suggests the following: (1) People generally entrain more to those with high levels of power than with low ones (see Danescu-Niculescu-Mizil et al., 2012), which might lead to asymmetrical behaviors in entrainment. (2) Entrainment on some features of language does not necessarily translate into speakers converging in all features (Giles et al., 1991; Reichel et al., 2018a). In fact it may be the case that entrainment on an acoustic-prosodic feature might be associated with disentrainment on another. (3) Entrainment in excess may even be perceived negatively. For example, in an empirical study aimed at finding optimal levels of entrainment, Giles (1979) found that simultaneously entraining on three levels of language — pronunciation, speech rate, and message content — was found to be perceived as patronizing. (4) Entrainment may be stronger at the dialog-act level (see Reichel et al., 2018b; Gauder et al., 2018), which can be taken as an indication that entrainment may not be an automatic process but that it may be actively controlled, at least partially. Subtleties like these make the characterization of acoustic-prosodic (dis)entrainment and its effects quite challenging.

The effects of SDSs entraining to a user's way of speaking is a topic which has been little discussed in the literature. Previous research on entraining SDSs focused mainly on the effects of systems which entrain on lexical or syntactic features (see, for example, Brockmann et al., 2005; Buschmeier et al., 2009; Hu et al., 2016; Lopes et al., 2015) or high-level concepts believed to be conveyed by prosody, such as entraining on emotions and politeness (see, for example, Acosta and Ward, 2011; De Jong et al., 2008). But there is a dearth of research on the effects of systems which follow *acoustic-prosodic* (dis)entrainment policies. Fandrianto and Eskenazi (2012) explore, in the context of an information-driven spoken dialog system, ways to induce users to reduce two particular speaking styles: shouting and hyperarticulation. To do so, they test different strategies. One of these strategies involves disentraining to the way users speak (i.e. reducing the TTS volume if the user shouts, raising the TTS speech rate if the user hyperarticulates). Their results suggest that disentrainment strategies do alleviate these two particular speaking styles, performing better for shouting than for hyperarticulation. Levitan (2014) and Levitan et al. (2016) propose a way of integrating acoustic-prosodic (dis)entrainment into existing SDSs, and present results from a series of pilot studies of the effects of four acoustic-prosodic (dis)entrainment policies. In Lubold et al. (2015) a pitch-adapting dialogue system is proposed, they also study how different ways of matching to users' mean pitch relate with third party perception of naturalness and rapport. In a follow-up study (Lubold et al., 2018), the authors explore how a teachable robot which entrains and introduces social dialogue influences rapport and learning. They find that a robot that entrains and speaks socially results in significantly more learning. Sadoughi et al. (2017) report an approach for online acoustic synchrony on pitch and intensity by using a dynamic Bayesian network learned from prior recordings of child-child play. When testing their system on a robot interacting with children, they report a significant order effect: children that began with a synchronous robot maintained their own synchrony to it and achieved higher engagement than those that did not. Although these efforts already suggest that acoustic-prosodic entrainment may be related to and may even influence users' behavior, results are far from conclusive. Acoustic-prosodic entrainment is a complex phenomenon, and how systems should adapt and which features they should entrain on is far from clear. This is why further empirical evidence on the effects of different acoustic-prosodic entrainment policies is still needed.

The goal of the present work is to provide further empirical evidence of how acoustic-prosodic (dis)entraining SDSs policies affect users. We focus on studying the effects of different acoustic-prosodic entrainment policies on induced *trust*.[4]

To explore this research question, we adapted, implemented, and carried out a large experimental study focused on analyzing if and how different acoustic-prosodic (dis)entrainment policies affect users' perception of SDSs' *ability* and, consequently, their *trustworthiness* (i.e. their quality of being trusted).[5] Studies were carried out in Argentina over the course of two years.

Additionally, as research on the effects of acoustic-prosodic entrainment is based primarily on corpus studies, this article also details on the challenges and nuances of approaching the topic using an experimental setup. We believe these insights may also be of use for future research.

The rest of this article is structured as follows. Section 2 provides details on the experimental task, on the dialogue system used (including how acoustic-prosodic entrainment was implemented), and on the way the data was analyzed. Section 3 presents the main results. Section 4 discusses these results, proposes future work, and concludes.

[4] Trust is defined as the "*willingness of a party to be vulnerable to the actions of another party based on the expectations that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" (Mayer et al., 1995, p. 712).

[5] Ability — "*that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain*" (Mayer et al., 1995, p. 717) — is one of three factors believed to affect trust (the other two being benevolence and integrity).

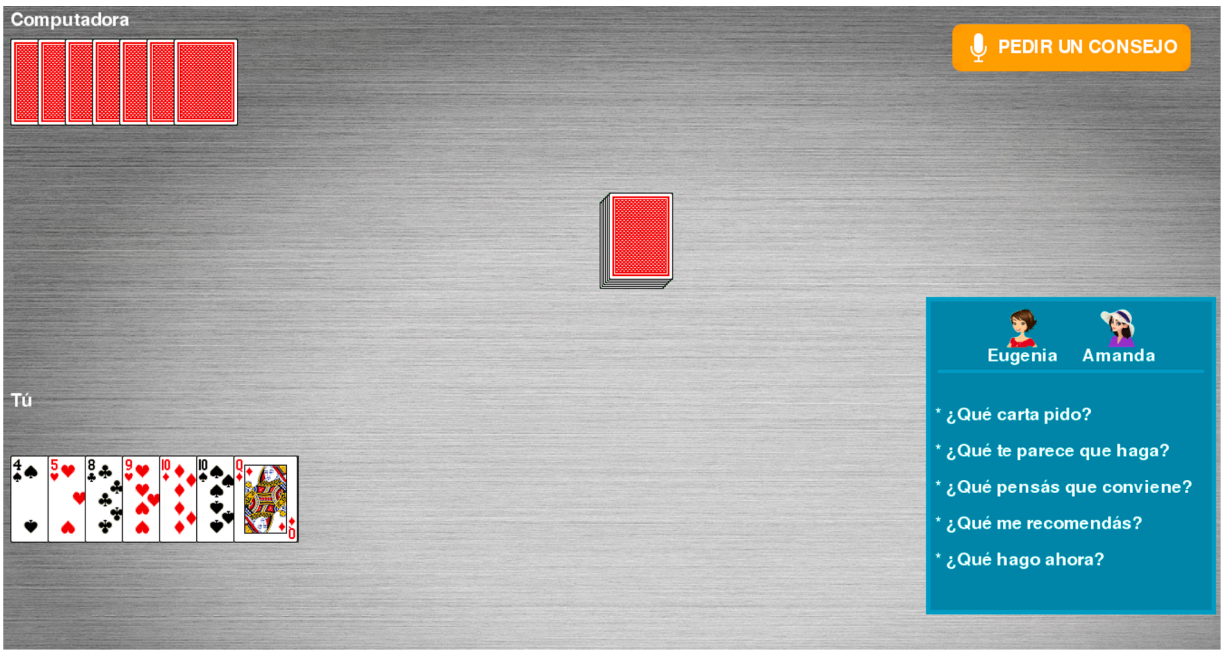

Fig. 1. Screenshot of a non-practice GoFishWithHelpers game.

2. Material and methods

This section first describes the task chosen for making participants interact with virtual assistants which adapt to the users' way of speaking. We then describe the procedure for experimental data collection, as well as the (dis)entrainment policies followed by the virtual assistants. Finally, we describe the statistical analyses strategy.

2.1. Experimental task

We tackle our research question through an experimental approach. In particular, we make use of an experimental setup in which participants must interact with virtual helpers that entrain to their speech following predefined entrainment policies.

We chose *GoFishWithHelpers* (Levitan, 2014; Levitan et al., 2016) as the experimental task for this study. GoFishWithHelpers is an adaptation of the canonical game of *Go Fish*, a multiplayer card game. In GoFishWithHelpers, instead of playing against human opponents, each participant is instructed to play against a computer system. At the beginning of each game the player and the system are each dealt a hand of seven cards. The player's goal is to acquire cards from the system's hand to earn points. In the canonical Go Fish game, the player can ask her opponent for cards of any rank that she already has in her own deck, and the opponent must then give her all the cards of that same rank in his hand. If the opponent has no cards of the requested rank, the player has to "Go Fish," selecting a card from the top of the deck.

In GoFishWithHelpers the participant, instead of freely choosing any rank in her deck, must ask for advice from one of possibly multiple virtual helpers. The helper then suggests a rank and the user is forced to follow the helper's advice. The player's goal is to gain as many points as possible. The player receives 10 points for each card she gets from the system and 100 points for completing a "set" (a rank in all four suits). She loses 50 points for "Go Fish".[6] Importantly, as we will explain below, the more points a player earns, the higher the monetary prize she will receive.

Each participant plays four games, first a practice game and then three non-practice games — which are relevant ones for our analysis. In the practice game, which consists of five requests for advice, or *turns*, the participant can only request advice from a single helper (Verónica). In addition, any points earned during this game do not count for the monetary compensation. The goal of this practice game is to introduce the subject to the system and the game rules, as well as measuring the acoustic-prosodic features' base levels of the participant's speech. During this practice game the helper does not adapt its speech in any way.

The following three games consist of fifteen turns each, and in each turn the participant has to choose to ask advice from one of two helpers, named Amanda and Eugenia. More precisely, at the beginning of each turn, the player's hand is disabled and she cannot ask the system for a rank directly. Instead, she presses a button and verbally requests advice from the avatar she specifies by name. Importantly, during these games, helpers may adapt their speech to the way participants ask for advice. Once a game ends, the sum of points collected in its fifteen turns is recorded. In between these non-practice games, participants are shown on-screen the amount of points they have earned in each game. Finally, based on all points collected in the three non-practice games, monetary prizes are awarded.

We followed the strategy presented in Levitan (2014) and Levitan et al. (2016) to choose what advice the helper would provide. To encourage participants to rely subconsciously on paralinguistic cues to choose their helper, it was important to prevent participants from

[6] Note that a single request for advice may lead to both "Go Fish" and a "set", as the card selected from the top of the deck may eventually lead to completing a set. In this case the end result is that she earns 50 points.

deciding whom to trust based on performance. Choosing among possible helpers' advice is not trivial, given that for each turn there are several different outcomes, depending on which rank is requested. To ensure that helpers behave as similarly as possible to each other, each helper is programmed to give advice according to an algorithm that keeps a "persona's global advice score" — the overall perceived quality of the advice it has given so far, corresponding to the number of points earned by following that advice — as close to zero as possible. This is done by assigning each rank a score based on what its outcome would be. If the rank would complete a set, its score is 5; if it would result in "Go Fish," its score is − 15; otherwise, its score is the number of cards the system has of that rank (1–3). Using this scheme, at each turn, the helper giving advice selects the rank whose score would bring the helper's global advice score closer to zero.[7] The subject's score is reset to 0 at the beginning of each game. Additionally, to further obscure the quality of the advice received, the system is dealt a new hand after every turn, so that the player cannot infer the contents of a hand based on responses to her previous requests.

Fig. 1 shows a screenshot of a non-practice GoFishWithHelpers game. The top left and bottom left corners of the screen contain the system's and participant's cards respectively. The top right corner contains a push-to-talk button. The bottom right corner contains the suggested phrases the participant may use to request advice as well as the helpers' avatars and names.

2.2. Data collection procedure

Experiments were carried out in the city of Buenos Aires, Argentina. Participants were publicly recruited, and, during recruitment, were notified that they would be paid for participating. The payment was of up to roughly 9 US dollars per hour in local currency; $4.5 per hour plus up to $4.5 based on their performance in the proposed task. Participants were required to be native Argentine Spanish speakers and to be between 18 and 65 years old.

Upon arrival at the lab, participants were instructed to read and sign an informed consent form.[8] Although all helpers a given participant interacts with generate their advice following exactly the same strategy (as reported above), they were told that they would be playing a computer game in which competing Artificial Intelligence (AI) algorithms were being tested. They then sat in front of a desktop computer wearing a headset with microphone (Genius HS-400A headset) and were handed written instructions describing GoFishWithHelpers. Importantly, the instructions stated that each helper was going to give advice using one of two particular AI algorithms, one being "more advanced" than the other, and that, even though both helpers tend to give good advice, they also make mistakes once in a while, which translates into occasional bad advice. Additionally, these instructions explicitly stated that, to gain more points during the game, their goal as participants was to discover which helper was driven by the "smarter" AI algorithm. Participants were also notified that they would receive additional money based on the number of points they gained. In this way participants were given strong incentives to search for the competent helper — even when in fact both behaved in the same way — and would consider it risky to delegate their choice to the helper perceived as "less advanced."

Once participants declared that they understood the rules, they played the practice game. In between the practice and the three non-practice games, the computer screen indicated that, if they had any doubts regarding the task rules, the lab assistant could be asked for help and they were also reminded that their ultimate goal was to discover which helper was driven by the "smarter" AI algorithm. They then proceeded to play the three non-practice games. Next, they were handed a questionnaire with questions related to the helper voices, and a second one with sociodemographic questions. Finally, they received payment for the points earned, were handed a debriefing form, were given the chance to ask questions regarding the experiment, and left the lab. The whole procedure lasted nearly an hour on average. Sessions were ran in groups of four participants in parallel in a quiet, large computer laboratory, so in an hour, data from four subjects could be acquired.

Importantly, each participant was randomly assigned to one of two different types of helpers:

1. **Mirrored format**: Each of the two helpers follows an opposing (dis) entrainment strategy — for example, if one helper entrains on pitch, the other one disentrains on pitch.
2. **Against-static format**: One helper (dis)entrains on a given acoustic-prosodic feature, and the other helper does not adapt its speech at all.

In this article we present results coming from four consecutive sessions of games, each consisting of around 72 games, for a total of 98 unique participants. We adopted a strategy of running a session of games testing the effects of a particular (dis)entrainment policy, and, based on the analysis of the results obtained in it, deciding which policies to test in the following session. For example, as we will see below, during the first session of games we tested only the effects of (dis)entrainment on speech rate. We then analyzed the data we obtained from these experiments and, based on that analysis, decided to run a session of games allowing helpers to also (dis)entrain on pitch and intensity.

2.3. Dialogue system

This section describes the ways users and system interact by means of an acoustic-prosodic entraining dialogue system. We focus on how the system measures participants' acoustic-prosodic features, and how the helpers entrain to participants' speech.

Given the task design, the way participants interact with helpers is limited, since they only request and receive advice from helpers. Both interactions are done through spoken dialogue. To ask for advice, participants use the computer mouse to select a microphone icon placed in the top right corner of the game screen (see Fig. 1). The button works using the well-known *push-to-talk* paradigm. Pressing the button triggers the recording and releasing it stops it. Once the recording is complete, it is sent to an ASR module and, in parallel, to an acoustic-prosodic feature extraction module.

Based on the time-aligned transcription produced by the ASR module, the identity of the requested helper is identified, and, based on its global advice score, the requested helper provides its advice. Importantly, the acoustic-prosodic feature values of this response will depend on the entrainment policy assigned to the helper (as will be described in Section 2.3.2).

Participants are instructed and required by the system to ask for help using one of a fixed number of request options (which they can toggle freely between turns). As seen in Fig. 1, these options are shown in the bottom right corner of the screen. Additionally, each request for advice must either start or end with a helper's name. For example, a participant can say "*¿Amanda, qué carta pido?*" ("*Amanda, which card should I ask for?*") or "*¿Qué carta pido Amanda?*" ("*Which card should I ask for, Amanda?*"). Subsequently, the selected helper is highlighted and her advice synthesized — e.g. "*Te recomiendo pedir un nueve*" ("*I recommend asking for a nine*"). The player completes the turn by clicking the suggested card, which serves as a request for that rank of card to be provided by the opponent.

[7] Scores were empirically determined in Levitan (2014) based on observing and discussing the game with subjects of a pilot study. There they expressed frustration at losing points to "Go Fish" that far outweighed their satisfaction at receiving cards from a successful request.

[8] Our protocol and all forms were evaluated and approved by the Research Ethics Committee at the Centro de Educación Médica e Investigaciones Clínicas (CEMIC) "Norberto Quirno", Buenos Aires, Argentina, on July 18, 2014, valid through August 31, 2017.

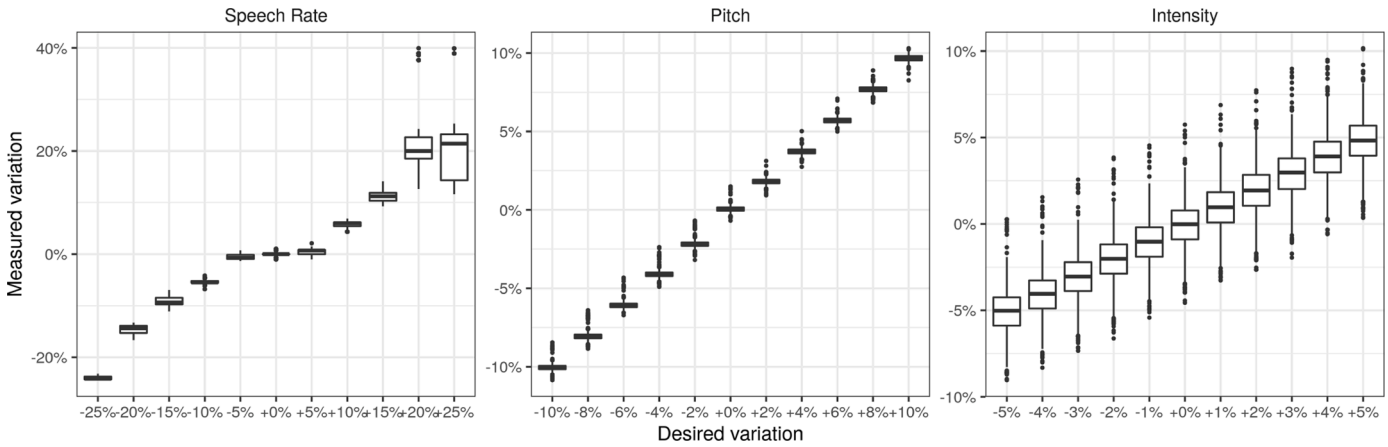

Fig. 2. Desired vs. measured variations across acoustic-prosodic features.

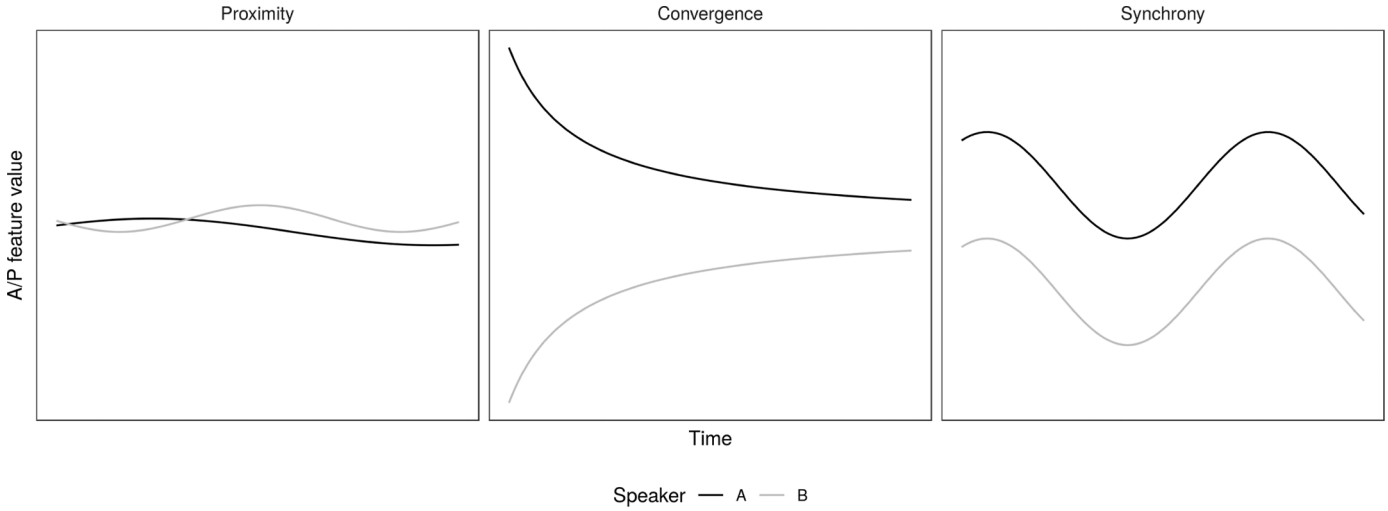

Fig. 3. Different types of acoustic-prosodic entrainment.

2.3.1. Measuring acoustic-prosodic features and synthesizing speech with varying prosody

The computer helpers implemented in our experiments either entrain or disentrain on three acoustic-prosodic features: *speech rate* (measured in syllables per second), *pitch* (measured as F0 mean in Hz) and *intensity* (measured as mean energy in dB). The proposed (dis)entrainment policies rely on three processes: (1) Being able to measure and keep a record of the acoustic-prosodic features of the interlocutor's speech, (2) determining how to adapt the acoustic-prosodic features of the synthesized speech to those of the participants, (3) synthesizing speech with the desired acoustic-prosodic features. Here we cover (1) and (3); we will discuss (2) in Section 2.3.2.

Measuring acoustic-prosodic features values We use the PocketSphinx toolkit (Huggins-Daines et al., 2006) with a restricted grammar to obtain a time-aligned transcription of each utterance and then estimate its syllable count using a pre-defined dictionary. PocketSphinx returns an error when a piece of audio cannot be matched with high confidence to the proposed grammar — in which case the participant is asked to repeat their request for advice. Tests carried out before beginning data collection indicated that the system was quite strict when accepting an utterance as valid. When looking at the collected data, in 4.8% of all non-practice turns participants were asked to repeat their request for advice. (This number equals 15.7% for practice turns, suggesting that participants learned how to ask for advice during the training game.) Participants' hesitations and (slightly) different formulations than the prompted text may introduce acoustic-prosodic features measurement errors (particularly when measuring speech rate). To get an estimate of the prevalence of these effects, we randomly sampled 200 audio clips from the collected data and checked for the presence of disfluencies and formulations which differ from the prompted text. We almost did not register disfluencies (only in 1 of all sampled clips). The prevalence of slight variations in the formulation was infrequent. We found slight variations in 6.5% of all sampled clips. We found no evidence suggesting that this prevalence varied across helpers with differing acoustic-prosodic entrainment policies.

PocketSphinx provides timestamps at the word level in seconds. We define *utterance duration* as the time elapsed between the beginning of the utterance's first word and the end of its last. We thus estimate *speech rate* as the ratio between syllable count and utterance duration,

capturing the *average* number of syllables per second in an utterance. Additionally, we use Praat (Boersma and Weenink, 2018) to estimate mean pitch[9] (in Hz) and mean intensity[10] (in dB), extracted from the beginning of the first word to the end of the last one.

Speech synthesis To synthesize the helpers' responses we use an HMM-based voice built with a corpus of read speech recorded by a female professional speaker of Argentine Spanish (further details in Violante et al., 2013). This TTS system uses MaryTTS as its front-end,[11] which allows us to modify speech rate and pitch level on a percent basis using Speech Synthesis Markup Language (SSML) tags. SSML is a markup language which provides a standard way to annotate text for the generation of synthetic speech. For example, when +10% is introduced in the pitch tag, the system is instructed to synthesize speech with a 10% higher pitch (in Hz) relative to the voice's default value. We modify intensity using the open-source sound processing toolbox SoX.[12] In this case, as we use decibels to measure intensity, a +4% tag indicates that the final audio will have a +4% higher intensity measured in decibels.[13]

Before running our first session of games, we ensured that the desired acoustic-prosodic variations given as input to the TTS system were achieved accurately. We synthesized 11,978 audio files, each containing one of nine possible helper responses, with one of the multiple combinations of values for the target acoustic-prosodic features. We then followed the same procedure described above for measuring the values of such features. Fig. 2 shows the distributions of measured vs. desired variations for each acoustic-prosodic feature. Ideally, one would want the boxes to be as short as possible and to progress along a 45° line.

Fig. 2 shows that, for pitch and intensity, the targets were met quite accurately — and in the case of pitch almost exactly. For speech rate, the degree of accuracy is lower, but there is a strong positive relation between desired and measured variation.[14]

2.3.2. Entrainment algorithm

According to Levitan et al. (2011), three forms of acoustic-prosodic entrainment can be distinguished: *proximity* (acoustic-prosodic features having similar values across interlocutors over the entire conversation), *convergence* (acoustic-prosodic features increasing in similarity across interlocutors over time), and *synchrony* (speakers adjusting the values of their acoustic-prosodic features in accordance to that of their interlocutor). Note that synchrony may occur without proximity. Fig. 3 illustrates these phenomena.

In our experimental task we make use of entrainment policies designed to (dis)entrain according to (anti-)synchrony.[15] In particular, the proposed (dis)entrainment policies build upon the one presented in Levitan et al. (2016). These policies measure how much the acoustic-prosodic feature values of a user utterance deviate from their respective base values (which depend on the user), and give a response in which the TTSs' acoustic-prosodic features deviate accordingly. For example, in the case of an entrainment policy on acoustic-prosodic feature k, if the user produces an utterance with feature k 10% higher than his/her own base value for k, the policy instructs the TTS system to synthesize speech with a value of k 10% higher relative to the TTS default value (10% lower for the case of a policy disentraining on feature k). The rest of this section details how the system achieves this.

Before calculating how to adapt to the users' way of speaking, the algorithm must keep track of the changes in the user's acoustic-prosodic features. To do so, and assuming g is the game being played at the moment, it keeps track of $\overline{\phi^k_{g-1}}$: the average value of acoustic-prosodic feature k in game $g-1$. It does so according to the following formula:

$$\overline{\phi^k_{g-1}} = \sum_{t \in g-1} \frac{\phi^k_t}{|g-1|} \quad (1)$$

where t stands for a given turn, $|g-1|$ for the number of turns in game $g-1$, and ϕ^k_t for the value of acoustic-prosodic feature k in turn t.

Knowing the value of $\overline{\phi^k_{g-1}}$ and having processed ϕ^k_t for an ongoing turn (of game g), the *desired variation* in turn t of acoustic-prosodic feature k in the helper's response (ψ^k_t) is calculated as follows:

$$\psi^k_t = policy_k \cdot \left(\frac{\phi^k_t - \overline{\phi^k_{g-1}}}{\overline{\phi^k_{g-1}}} \right) \% \quad (2)$$

where $policy_k$ equals 1, 0 or −1 if the helper follows an entrainment, static or disentrainment policy, respectively. In other words, a helper entraining on acoustic-prosodic feature k adapts the value of k in synchrony with the subject; a static helper does not adapt in any way; and a disentraining helper adapts in the opposite direction. For example, assume that k corresponds to speech rate, $g = 3$ (the second non-practice game) and $\overline{\phi^k_{3-1}} = 4$ syl/sec. If $\phi^k_t = 4.8$ syl/sec (i.e. the speech rate measured on the current request for advice), then the desired variation in speech rate will be +20% for an entraining helper, 0% for a static helper, and −20% for a disentraining one. Note that a similar calculation is repeated for the remaining acoustic-prosodic features.

During the practice game the helper's voice is always synthesized using the TTS system's default pitch, intensity, and speech rate levels (i.e. $policy = 0$). Additionally, in the three non-practice games, we differentiate the voices of the two helpers by means of distinct base pitch levels. One of the helpers uses the TTS system's default pitch level; the other, a 10% lower pitch level. Speech rate and intensity had the same base levels across helpers. Importantly, to separate the effect of (dis) entrainment policies from the different base pitch levels, helper names and faces, these characteristics were counterbalanced for the three policies across participants.

Lastly, to preserve the naturalness of the synthesized voices and avoid the occurrence of glitches and artifacts, we clipped maximum/minimum values of ψ^k_t (+25%/−25% for speech rate, +5%/−5% for intensity, and +10%/−10% for pitch). These upper and lower bounds were chosen perceptually by the authors. Post-hoc analyses indicate that these ranges were wide enough as to include subjects' variation in nearly 91.4%, 92.3%, 83.6% of all turns for speech rate, intensity and pitch, respectively.

2.4. Data analysis

This section details the statistical techniques used to study the data collected. Note that, instead of following a single approach for analyzing the data, we follow complementary strategies in order to verify results robustness. In addition, note that we focus on the way participants effectively chose between helpers (i.e. their actual behavior) and not on variables derived from subjective perceptions (such as answers to follow-up questionnaires).

2.4.1. Binomial tests

We first analyze associations between entraining policies and trust by running two-sided exact binomial tests. These tests take as input the number of successes c (the times the helper following the entrainment policy being studied is chosen for advise), the total number of turns ($15*n$, 15 being the number of turns in a game and n the number of

[9] http://fon.hum.uva.nl/praat/manual/Sound_To_Pitch__.html.
[10] http://fon.hum.uva.nl/praat/manual/Sound_To_Intensity__.html.
[11] http://mary.dfki.de.
[12] http://sox.sourceforge.net.
[13] Note that decibels is a logarithmic scale.
[14] During the development of the experimental task we tested alternative proprietary TTS systems with Spanish trained voices. Neither achieved better accuracy than the one used in this work.
[15] Anti-synchrony stands for the tendency of speakers to distance their speech from the other's, resulting in mirrored or anti-correlated patterns (Looze and Rauzy, 2011)

games under analysis), and the null-hypothesis probability of success. We set this probability to 0.5 — the probability of choosing the helper of interest in a given turn if participants choose completely at random.

For example, if 36 games are being analyzed ($n = 36$) and the helpers following the entrainment policy being studied were chosen 305 ($c = 305$) out of 540 times (15*36) — that is, in 56% of all turns, a two-sided exact binomial test would reject the null-hypothesis that the probability of choosing the helper is equal to 0.5 (random) with a significance inferior to 1%. This would suggest that a user is more likely to choose a helper carrying out a certain entrainment policy. On the other hand, if entraining helpers were chosen 280 times, the null-hypothesis would not be rejected at standard levels of significance and no claim regarding any effect could be made.[16]

2.4.2. Regression analysis

It is important to note that simply hearing advice coming from an adapting helper does not necessarily translate into listening to synthesized speech differing from its base acoustic-prosodic feature values. Take for example an extreme case, a participant who speaks in all games using a monotone speech with no alteration of her acoustic-prosodic feature values. In this case, whether the helper follows an entraining, static or disentraining policy, the participant will hear the same acoustic-prosodic dynamics during all turns.

To better capture these subtleties, during the analysis we not only check for associations between trusting a helper and that helper's entrainment policy, but also relate trust to a measure of exposure to entrainment and disentrainment in a given turn. Concretely, given a participant request for advice in turn t, we define *exposure to entrainment* on acoustic-prosodic feature k in that turn as:

$$exp\_ent_t^k = \begin{cases} |\psi_t^k| & \text{if advice is given by a helper } \textbf{entraining} \text{ on k} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In parallel, we also define *exposure to disentrainment* on feature k as:

$$exp\_disent_t^k = \begin{cases} |\psi_t^k| & \text{if advice is given by a helper } \textbf{disentraining} \text{ on k} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Note that if the advice is given by a helper which entrains on feature k, the value of $exp\_ent_t^k$ will differ from 0 (as long as $\psi_t^k \neq 0$) but $exp\_disent_t^k$ will be equal to 0, whereas if the advice is given by a helper which disentrains on feature k the opposite will occur. Also note that, if the advice is given by a helper which follows a static policy on feature k, both $exp\_ent_t^k$ and $exp\_disent_t^k$ will be equal to 0 in every turn.

Three potential drawbacks of the binomial test analysis are that (1) it treats games coming from a given player as independent when they may not be so, (2) it completely ignores exposure to entrainment, and (3) it ignores other variables which may affect participant choices. An alternative and more versatile strategy for analyzing the data consists of using regression analysis to model whether participants continue asking the *same* helper for advice after following its advice in a given turn. In doing this, we associate keeping a helper in the next turn to considering its advice trustworthy.

We report results on regression models considering (dis)entrainment policies on the one hand, and exposure to (dis)entrainment on the other. When focusing simply on entrainment policies we estimate models according to the following specification:

$$keeps\_helper_t = \alpha + \beta \cdot policy\_of\_interest_t + \gamma \cdot X_t + \epsilon_t, \tag{5}$$

where:

- $keeps\_helper_t$ is an indicator variable which takes value 1 if a participant chooses the same helper after receiving and seeing the effects of its advice in turn t, and 0 otherwise.
- $policy\_of\_interest_t$ takes a value of 1 if a particular entrainment policy being studied is selected in turn t, and 0 if another policy is selected.
- X_t is a vector which contains a series of control variables (which may or not depend on the turn t). More on this below.
- ϵ_t is an error term.

The parameter of interest is β. Note that $keeps\_helper_t$ is undefined for the last turn in each game, and consequently these turns are not included.

When exposure to (dis)entrainment is analyzed, the model specification varies slightly. Instead of introducing an indicator variable related to an entrainment policy, a series of variables reflecting exposure to (dis)entrainment on each acoustic-prosodic feature are introduced. Concretely, the estimated models have the following specification:

$$keeps\_helper_t = \alpha + \sum_{k \in AP} \beta_k^{ent} \cdot exp\_ent_t^k + \sum_{k \in AP} \beta_k^{disent} \cdot exp\_disent_t^k + \gamma \cdot X_t + \epsilon_t, \tag{6}$$

where $exp\_ent_t^k$ and $exp\_disent_t^k$ indicate exposure to entrainment and disentrainment on feature k in turn t respectively, and the set AP contains all features being analyzed. The parameters of interest are β_k^{ent} and β_k^{disent} for each feature k. Note that, for the case in which neither helper adapts to a given feature k (i.e. $policy_k = 0$) both $exp\_ent_t^k$ and $exp\_disent_t^k$ will be equal to 0 for all turns. This translates into a problem of perfect multicollinearity, which makes it impossible to estimate β_k^{ent} and β_k^{disent} for advice given by a helper static on feature k. For this reason, coefficients associated to static acoustic-prosodic features are not reported.

Before running all of the main body regressions analyzing exposure to (dis)entrainment, we standardize exposure to (dis)entrainment using z-scores. We do this to facilitate the interpretation of effect sizes. In this way, coefficients should be interpreted as the estimated variation observed in the outcome when exposure to acoustic-prosodic (dis)entrainment in a given feature k increases in one standard deviation. Appendix tables report the estimated coefficients without standardizing exposure metrics.

To increase the precision of our estimates of interest, we include as control variables (X_t) a series of variables which may affect the probability of keeping a helper in the next round but are not related to acoustic-prosodic (dis)entrainment policies. These are:

- An indicator variable equal to 1 if the advice given in turn t resulted in Go Fish, and 0 otherwise.
- An indicator variable equal to 1 if the advice given in turn t resulted in a completed deck, and 0 otherwise.
- An indicator variable equal to 1 if the advice was requested by a female participant, and 0 otherwise.
- A helper fixed effect variable equal to 1 if Eugenia (one of the two non-practice helpers) is selected (no matter its entrainment policy), and 0 otherwise. Given that the helpers' look and base pitch level differ, and that entrainment policies are counterbalanced across helpers, this aims at controlling for helper fixed effects.
- A continuous variable indicating the turn number in a game (i.e. t) which ranges from 1 to 14. This variable aims at capturing the expected positive association between exploitation strategies and being close to a game ending. More on this in Section 3.1.
- Two game number indicator variables, one for the second non-practice game ($g = 3$) and one for the last non-practice game ($g = 4$). Concretely, these variables equals 1 if the turn t belongs to game $g = 3$ or $g = 4$ respectively, and 0 otherwise. Note that practice game turns ($g = 1$) are not considered in the analysis and the first non-practice game ($g = 2$) is left as the base category.

[16] Note that small values of c would also result in the rejection of the null-hypothesis, but would indicate a negative association.

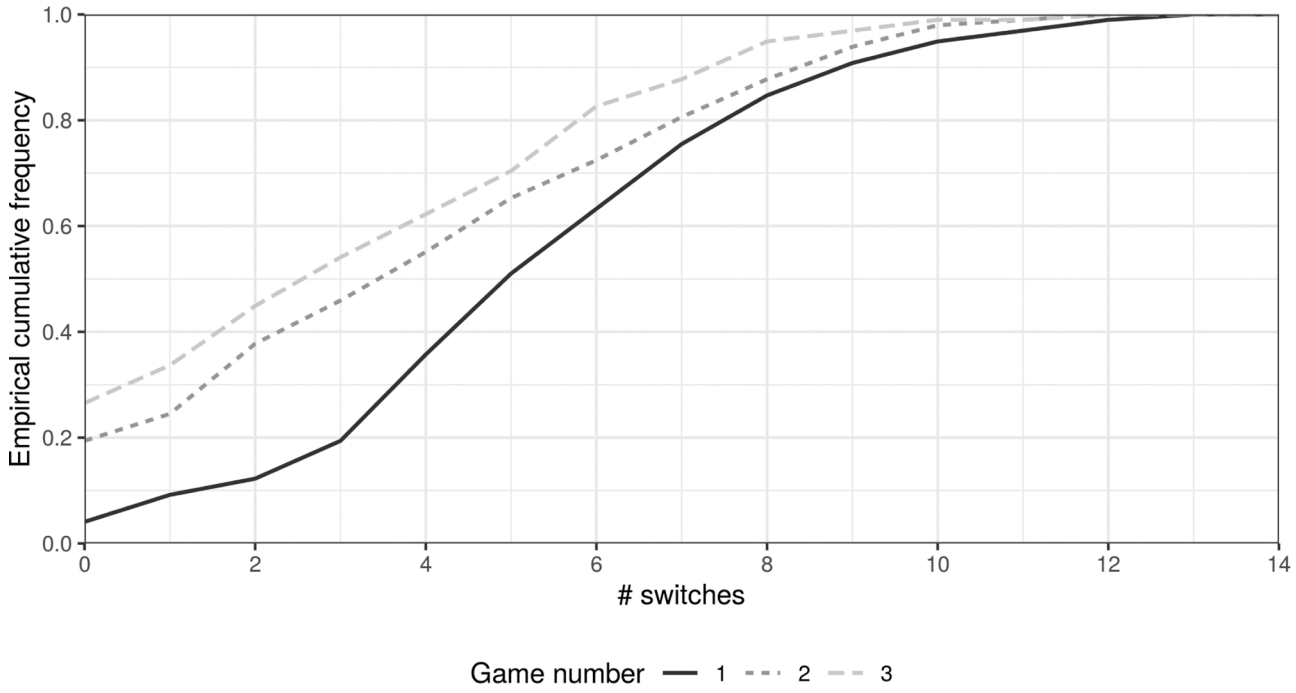

Fig. 4. Empirical cumulative frequency of helper switches for the 294 non-practice games, by game number. *Note*: Detailed information available in Table A1.

We estimate Eqs. 5 and 6 through both ordinary least squares (OLS) — known as the *linear probability model* — and generalized linear mixed-effects (GLME) models. Linear mixed-effects models are commonly used when independence in the data cannot be guaranteed, as they allow one to introduce random-effects which provide a robust analytic approach for addressing problems associated with hierarchical data (West et al., 2014). GLME models are an extension of the linear mixed-effects ones which allow to consider response variables coming from different distributions. When estimating GLME models, we model the response variable as a dichotomous one using the "*logit*" link function and incorporate a participant random intercept.

One disadvantage of GLME models not using the identity link function is that effect sizes become hard to interpret. For this reason, in the main draft, instead of reporting estimated GLME coefficients, we report estimated average marginal effects (AMEs) (see Leeper, 2017). AMEs should be interpreted as the average change in probability when the independent variable being considered increases one unit. Estimated GLME coefficients are reported in the Appendix tables.[17]

3. Results

This section is structured as follows. We first provide an aggregated analysis of the participants behavior during the task. Then we present the patterns observed in each session of games. Finally, we present a meta-analysis considering all the against-static sessions as a whole.

3.1. Participants high-level behavior during the task

A total of four sessions of games testing several acoustic-prosodic entrainment policies were carried out from December 2015 to November 2017. In each session of games, emphasis was placed on testing different hypotheses regarding the relationship between acoustic-prosodic entrainment and trust. Before focusing on acoustic-prosodic entrainment, in this section we analyze how participants behaved in aggregate terms.

Fig. 4 plots the empirical cumulative distributions of the times participants switched the helper they requested advice from.[18] Cumulative frequencies are disaggregated by game number. As an example, this figure indicates that around 19.4% of all participants switched helpers three or less times in the first game, 45.9% in the second one, and 54.1% in the third one.

Fig. 4 shows some interesting patterns. First, participants tended to switch helpers more in the first game relative to the second and third ones. In particular, nearly 49% of all participants switched helpers more than five times in the first game, while only 34.7% and 29.6% did so in the second and third games respectively. Second, many more participants kept choosing the same helper for all turns in games 2 and 3 (this is reflected in the curves intercepts). Both of these patterns suggest that some participants may have perceived each successive game as a continuation of the preceding one.[19] Finally, the most frequent behavior consisted in switching helpers an intermediate number of times (Table A1 shows this more clearly).

These behaviors can be framed by means of an analogy between GoFishWithHelpers and the well-known formal mathematical optimization problem known as the *multi-armed bandit problem* (Robbins et al.,

[17] We use R for all of the statistical analysis (R Core Team, 2019). For estimating GLME models, we use the *lme4*-package (Bates et al., 2015). GLMEs' *p*-values are calculated using the Satterthwaite approximation as implemented in the *lmerTest*-package (Kuznetsova et al., 2017). GLMEs' AMEs are calculated using the *margins*-package (Leeper, 2017).

[18] Recall that each participant plays three non-practice games and each game has 15 turns, so participants may switch helpers at most 14 times per game.

[19] Participants were neither informed that the assignment of AI algorithms varied across games, nor that it did not.

Table 1
Distribution of the observed variation of speech rate in session of games #1.

Mirrored							
Feature	Min.	*Q*1	Median	*Q*3	Max.	Mean	SD
sp.rate	-0.76	-0.11	-0.02	0.06	0.43	-0.01	0.15
Against-static							
Feature	Min.	*Q*1	Median	*Q*3	Max.	Mean	SD
sp.rate	-0.54	-0.11	-0.02	0.06	0.53	-0.02	0.14

1952; Sutton et al., 1998; Steyvers et al., 2009). In a general *N*-armed bandit problem, there is a set of *N* bandits, each having some fixed but unknown rate of reward. On each trial, a decision-maker selects a bandit and receives as feedback whether or not one unit of probabilistically determined reward was attained. The decision-maker's task is to make a sequence of bandit choices that maximizes their reward using such feedback. The bandit problem faces participants with a trade-off between *exploring* (i.e. acquiring new knowledge regarding bandit payoffs) and *exploiting* (i.e. optimizing the expected total payoff based on their current knowledge). The way people behave in the multi-armed bandit problem scenario has been extensively studied in lab settings (see, for example, Anderson, 2012; Racey et al., 2011; Schulz et al., 2018), and several heuristics — as well as an optimum strategy — have been proposed for solving this problem and have been empirically checked for compliance (see, for example, Lee et al., 2011; Steyvers et al., 2009).

Similar to the multi-armed bandit problem, in GoFishWithHelpers participants receive rewards as a function of choosing among helpers. Moreover, as they are explicitly informed that the helpers' advice is generated by different AI systems, participants are primed to discover which helper provides advice leading to higher rewards and to eventually choose the one believed to be superior. In this way the multi-armed bandit problem analogy provides insights regarding participant atypical behavior, such as having a pre-defined strategy not related to the game development, or not understanding the task instructions. For example, participants who do not explore helpers, especially in the first turns of a game, might be following a strategy not in compliance with the description given of the game. On the other end of the spectrum, a participant who constantly switches between helpers is also contrary to the expected behavior given the game description, as this behavior does not suggest that the participant is exploring to eventually try to earn as many points as possible by exploiting the advice of the helper perceived as less risky.

Taking these facts into account, and with the aim of filtering out anomalous behaviors during our analysis, for the rest of our analysis we discard all games in which participants switched helpers fewer than three times (93 games) or in which they switched 12 or more times (5 games). Thus, we focus on the games showing behavior in accordance to the given the instructions and previous literature.[20]

3.2. Session of games #1: Entrainment on speech rate

Motivation and setup With the aim of validating the proposed paradigm without testing the effects of entrainment on multiple acoustic-prosodic features at a time, in the first session of games we opted to analyze the effects of (dis)entrainment only on speech rate. In the mirrored format participants had to choose between two helpers, one following an entrainment on speech rate policy ($policy\_of\_interest_t = 1$) and the other following a disentrainment on speech rate policy ($policy\_of\_interest_t = 0$) — pitch and intensity did not adapt in any way. In the against-static format, one helper followed an entrainment on speech rate policy ($policy\_of\_interest_t = 1$) while the other did not entrain in any way ($policy\_of\_interest_t = 0$). Participants were assigned randomly to each version. Within versions, the helper which followed the entrainment/disentrainment or entrainment/static policies (Amanda or Eugenia) was counterbalanced across participants. Assignment of policies to helpers remained static across games played by a given participant.

[20] Our results are robust to other less strict limits, such as only excluding games in which the participant did not switch helpers at all (49 games) and games in which they switched 13 or more times (1 game).

Table 2
Estimated marginal effects for session of games #1. *Notes*: Exposure to entrainment metrics were standardized using *z*-scores before being introduced to the regressions. *p < 0.1; **p < 0.05; ***p < 0.01.

A. Dis/Entrainment policy				
	Mirrored		Against-static	
	OLS (1)	GLME (2)	OLS (3)	GLME (4)
$entrainment\ policy_t = 1$	0.0015	0.0014	− 0.0969	− 0.0985*
gofish=1	0.0000	0.0015	− 0.0912	− 0.0900
Num. obs.	322	322	308	308
Num. groups: player		13		10
B. Dis/Entrainment exposure				
	(5)	(6)	(7)	(8)
$exp\_ent_t^{speech\ rate}$	− 0.0341	− 0.0327	− 0.0068	− 0.0096
$exp\_disent_t^{speech\ rate}$	0.0196	0.0203		
gofish=1	0.0062	0.0088	− 0.0919	− 0.0908
Num. obs.	322	322	308	308
Num. groups: player		13		10

Subjects Data from 26 participants was gathered (6 female, 20 male; average age = 28, sd = 8.72), 14 participants were assigned to the mirrored format and 12 to the against-static format. For the reasons described in Section 3.1, 33 games were left out of the analysis as they did not meet our inclusion criteria (19 in the mirrored format, 14 in the against-static format). The sessions were carried out during December 2015.

Distribution of the observed acoustic-prosodic feature variations To check that participants effectively varied the way they asked for advice across turns, Table 1 presents the distribution of measured variation of speech rate in both game formats. It can be seen that, although the median variation of speech rate in both game formats is close to zero (−2% in both cases), there was indeed variation across turns.

Two-sided exact binomial tests When analyzing the data using two-sided exact binomial tests, we observe that in the mirrored format 173 out of 345 times (50.14%) the entraining helper was selected. A two-sided exact binomial test fails to reject the null-hypothesis at standard significance levels ($p = 1$). In the against-static format, the entraining helper was selected 142 out of 330 times (43.03%). In this case, a two-sided exact binomial test rejects at standard significance levels the null hypothesis that the helpers were chosen randomly ($p < .01$).

Regression analysis Table 2 summarizes results from session of games #1 regression analysis. It reports estimates from multiple regressions, in all of which $keeps\_helper_t$ is the dependent variable. Regressions vary in their input data (mirrored or against-static formats), their estimation strategy (OLS or GLME), and the way entrainment is measured (at the policy level or at the exposure level). More precisely, coefficients below the OLS label (regressions 1, 3, 5, 7) come from ordinary least squares estimates, and estimates below the GLME label (regressions 2, 4, 6, 8) report estimated AMEs coming from generalized linear mixed effects model. In the top panels (regressions 1, 2, 3, 4) entrainment is introduced in the analysis as in Eq. 5 (at the policy level) while in the bottom panels (regressions 5, 6, 7, 8) as in Eq. 6 (at the exposure level). It should

Table 3
Distribution of the observed variation of all three acoustic-prosodic features in session of games #2.

Mirrored							
Feature	Min.	Q1	Median	Q3	Max.	Mean	SD
sp.rate	-0.66	-0.12	-0.04	0.06	0.38	-0.04	0.15
pitch	-0.23	-0.02	0.02	0.08	0.38	0.03	0.09
intensity	-0.13	-0.01	0.00	0.02	0.09	0.01	0.03
Against-static							
Feature	Min.	Q1	Median	Q3	Max.	Mean	SD
sp.rate	-0.45	-0.11	-0.01	0.08	0.49	-0.01	0.14
pitch	-0.34	-0.05	0.00	0.05	0.35	0.01	0.09
intensity	-0.09	-0.01	0.01	0.02	0.10	0.01	0.03

Table 4
Estimated marginal effects for session of games #2. *Notes*: Exposure to entrainment metrics were standardized using *z*-scores before being introduced to the regressions. *p < 0.1; **p < 0.05; ***p < 0.01.

A. Dis/Entrainment policy				
	Mirrored		Against-static	
	OLS (1)	GLME (2)	OLS (3)	GLME (4)
entrainment policy$_t$ = 1	0.1358**	0.1384**	− 0.0174	− 0.0154
gofish=1	0.0229	0.0193	− 0.1890***	− 0.1887***
Num. obs.	294	294	378	378
Num. groups: player		12		12
B. Dis/Entrainment exposure				
	(5)	(6)	(7)	(8)
$exp\_ent_t^{speech\ rate}$	− 0.0122	− 0.0082	− 0.0308	− 0.0280
$exp\_disent_t^{speech\ rate}$	0.0362	0.0380		
$exp\_ent_t^{pitch}$	0.0439	0.0409	− 0.0565*	− 0.0556*
$exp\_disent_t^{pitch}$	− 0.0267	− 0.0242		
$exp\_ent_t^{intensity}$	0.0415	0.0498	0.0701**	0.0670*
$exp\_disent_t^{intensity}$	− 0.0042	− 0.0020		
gofish=1	0.0226	0.0169	− 0.1885***	− 0.1882***
Num. obs.	294	294	378	378
Num. groups: player		12		12

be remembered that for all of the main body tables reporting regression estimates analyzing exposure to (dis)entrainment, before running them, exposure to (dis)entrainment metrics were standardized using *z*-scores. Finally, regressions in the left panels (1, 2, 5, 6) take as input data collected from participants playing under the mirrored format, and those in the right panels (3, 4, 7, 8), under the against-static format. To place our focus on the parameters of interest, this table reports estimates associated to the entrainment related variables and the one associated to "*gofish=1*". Estimated coefficients for the remaining X_i covariates are presented in Table A2 and Table A3. Coefficients associated to "*gofish=1*" are reported to interpret the relative magnitude of effect sizes.

In line with the binomial test, at the policy level, we do not find any significant effects of entrainment policies for games under the mirrored format. In against-static games, the GLME coefficient is statistically significant at 10%. Notably, the estimated coefficient almost equals the estimated one for an advice resulting in Go Fish. When focusing on estimates considering the exposure level we find negative coefficients for entrainment on speech rate under both formats, but in neither case can these estimates be considered statistically significant at standard levels.

Discussion This session allowed us to check the effectiveness of the proposed paradigm. For the mirrored format, neither the two-sided exact binomial tests nor the regression analysis results suggest that entrainment on speech rate is preferred over disentrainment on speech rate. For the against-static format, both analysis suggest a negative association between entrainment on speech rate and maintaining a helper. In other words, when participants were forced to choose advice from these helpers, estimates point toward a negative effect of entrainment on speech rate. This last results is reassuring, as it suggest that the experimental setup allows to affect users trust by modifying the way helpers adapt their speech to the users speech.

3.3. Session of games #2: Entrainment on speech rate, pitch, and intensity

Motivation and setup The first session of games allowed us to first test the viability of the proposed paradigm as well as its implementation. Second, it also suggested that associations with trust (as measured by choosing or maintaining a helper) may be induced. Building on these findings, in a second session of games we tested whether, in addition to speech rate, adapting on pitch and intensity levels is associated with trust. The rationale behind this setup lies in the fact that entraining only on a single acoustic-prosodic feature may be perceived as unnatural by participants, while adapting across different acoustic-prosodic features may be more realistic and natural. In this way, in the mirrored format, one helper entrained on speech rate, pitch and intensity ($policy\_of\_interest_t = 1$) while the other disentrained on these features ($policy\_of\_interest_t = 0$). In the against-static format one helper entrained on the three acoustic-prosodic features ($policy\_of\_interest_t = 1$) while the other did not entrain in any way ($policy\_of\_interest_t = 0$). Within formats, the helper that followed each acoustic-prosodic feature behavior was counterbalanced across participants. Assignment of policies to helpers remained static across games played by a given participant.

Subjects Data from 24 participants was gathered (12 female, 12 male; average age = 22.71, sd = 3.25), 12 participants played in the mirrored format, while 12 played in the against-static format. Again, for the reasons described in Section 3.1, 24 games were left out of the analysis as they did not meet our inclusion criteria (15 in the mirrored format, 9 in the against-static format). The sessions were carried out during September 2016.

Distribution of the observed acoustic-prosodic feature variations Table 3 presents the distribution of the variation of all three acoustic-prosodic features across both game formats. Again, although the median of most acoustic-prosodic features is close to zero, they still varied across turns.

Two-sided exact binomial tests In the mirrored games the entraining helper was selected 180 out of 315 times (57.14%). A two-sided exact binomial test rejects the null hypothesis that helpers were chosen randomly ($p \approx .01$). In the against-static games the entraining helper was chosen 193 out of 405 times (47.65%); in this case a two-sided exact binomial test fails to reject the null hypothesis ($p = .37$).

Regression analysis Table 4 summarizes results from session of games #2 regression analysis. Tables A4 and A5 present detailed information on each regression. When entrainment on all three acoustic-prosodic features is introduced in the regressions at the policy level (top

Table 5
Distribution of the observed variation of pitch and intensity in session of games #3.

Mirrored							
Feature	Min.	Q1	Median	Q3	Max.	Mean	SD
pitch	-0.17	-0.03	0.01	0.05	0.35	0.01	0.07
intensity	-0.08	-0.02	0.00	0.01	0.08	0.00	0.02
Against-static							
Feature	Min.	Q1	Median	Q3	Max.	Mean	SD
pitch	-0.21	-0.04	0.00	0.05	0.65	0.01	0.09
intensity	-0.13	-0.02	0.00	0.01	0.09	0.00	0.03

Table 6
Estimated marginal effects for session of games #3. *Notes*: Exposure to entrainment metrics were standardized using *z*-scores before being introduced to the regressions. *p < 0.1; **p < 0.05; ***p < 0.01.

A. Dis/Entrainment policy				
	Mirrored		Against-static	
	OLS (1)	GLME (2)	OLS (3)	GLME (4)
$entrainment\ policy_t = 1$	0.0616	0.0628	0.0153	0.0160
gofish=1	-0.2692^{***}	-0.2753^{***}	-0.1308^{*}	-0.1309^{*}
Num. obs.	364	364	392	392
Num. groups: player		11		12
B. Dis/Entrainment exposure				
	(5)	(6)	(7)	(8)
$exp\_ent_t^{pitch}$	0.0204	0.0210	−0.0348	−0.0346
$exp\_disent_t^{pitch}$	0.0144	0.0110		
$exp\_ent_t^{intensity}$	0.0198	0.0167	0.0531*	0.0536*
$exp\_disent_t^{intensity}$	0.0108	0.0099		
gofish=1	-0.2600^{***}	-0.2666^{***}	-0.1271^{*}	-0.1269^{*}
Num. obs.	364	364	392	392
Num. groups: player		11		12

panels) we find a positive effect of entrainment in the mirrored format — i.e. players tended to maintain the helper when it followed an entrainment policy relative to a disentrainment one. But we do not find statistically significant effects in the against-static format.

When the focus is placed on analyzing exposure to entrainment (bottom panels) we do not see any significant positive effect of exposure to entrainment or disentrainment in the mirrored format. However, it should be noted that exposure to entrainment on both pitch and intensity have positive coefficients, while exposure to disentrainment on pitch and intensity have negative ones. We do find statistically significant effects in the against-static format when analyzing exposure to entrainment. Interestingly, these tend to go in opposite directions: the coefficient for pitch entrainment exposure is negative, while the one for intensity is positive.[21] Notably, comparing these estimates to the Go Fish one, suggests that a one standard deviation rise in exposure to entrainment on pitch lowers the probability of keeping the avatar 29.5% as much as an advice leading to Go Fish reduces it, while a one standard deviation rise in exposure to entrainment on intensity rises this probability 35.6% as much as an advice leading to a Go Fish reduces it.

Discussion Taken as a whole, and, in particular, when compared to the results from the first session, this second session of games illustrates the challenges behind setting up entrainment policies which may influence the player's predisposition to choose a helper. Results from our first session of games suggested a negative effect of entrainment on speech rate in the against-static format, but results from this session of games suggest that, when entrainment on pitch and intensity is added to the entrainment policy, no effect is observed at the policy level. Notably, when analyzing exposure to entrainment, we find that this might be driven by the fact that effects of entrainment on different acoustic-prosodic features do not necessarily occur in the same direction. Moreover, estimated effect sizes are non-negligible when compared to the one of an advice leading to Go Fish.

Results of the mirrored format also illustrate the complexity of identifying acoustic-prosodic entrainment effects. Even when a positive effect of entrainment is observed at the policy level, the fact that no single coefficient is statistically significant when analyzing exposure to entrainment makes it hard to pinpoint this effect to a particular behavior of an acoustic-prosodic feature.

[21] It should be noted that, for this session of games, in nearly half of the non-practice turns analyzed (48.3%) participants' intensity and pitch variation relative to their base values went in the same direction (i.e. both varied simultaneously above or below their respective means), and in the other half it did not. This suggests that this result is not driven entirely by the fact that the these effects are constantly cancelling each other.

3.4. Session of games #3: Entrainment on pitch and intensity

Motivation and setup With the aim of better isolating the effects found in the previous sessions, in a third session of games we experimented with a setup in which only pitch and intensity were adapted to the user's way of speaking. Concretely, in the mirrored format one helper entrained on pitch and intensity ($policy\_of\_interest_t = 1$) while the other one disentrained on pitch and intensity ($policy\_of\_interest_t = 0$). In the against-static format, a helper entrained on these features ($policy\_of\_interest_t = 1$) and the other one followed a static policy ($policy\_of\_interest_t = 0$). Helpers and entrainment behaviors were counterbalanced as usual. Assignment of policies to helpers remained static across games played by a given participant.

Subjects Data from 24 participants was gathered (6 female, 18 male; average age = 22.38, sd = 2.39), 12 participants played in the mirrored format, while 12 player in the against-static one. Again, for the reasons described in Section 3.1, 18 games were left out of the analysis as they did not meet our inclusion criteria (10 in the mirrored format, 8 in the against-static format). The sessions were carried out during December 2016.

Distribution of the observed acoustic-prosodic feature variations Table 5 presents the distribution of the variation of pitch and intensity across both game formats. The patterns are almost identical to those reported in the session of games #2.

Two-sided exact binomial tests When analyzing the data, we observed that, in the mirrored format, 204 out of 390 times the entraining helper was selected (52.31%). A two-sided exact binomial fails to reject the null-hypothesis at standard significance levels ($p = .39$). In the against-static format, the entraining helper was selected 214 out of 420 times (50.95%). An equivalent two-sided exact binomial test also fails to reject at standard significance levels the null hypothesis that the helpers were chosen randomly ($p = .73$).

Regression analysis Table 6 summarizes results from session of games #3 regression analysis. Tables A6 and A7 present detailed information on each regression. In contrast with session of games #2, when only (dis) entrainment policies on pitch and intensity are effective, regressions at the policy level (top panels) fail to find statistically significant coefficients under any format. Results go in the same direction for the case

Table 7
Distribution of the observed variation of all acoustic-prosodic features in session of games #4.

Mirrored							
Feature	Min.	*Q*1	Median	*Q*3	Max.	Mean	SD
sp.rate	-0.50	-0.13	0.00	0.10	0.63	0.00	0.17
pitch	-0.19	-0.04	0.00	0.05	0.23	0.01	0.07
intensity	-0.10	-0.02	0.00	0.02	0.09	0.00	0.03
Against-static							
Feature	Min.	*Q*1	Median	*Q*3	Max.	Mean	SD
sp.rate	-0.75	-0.10	0.00	0.09	0.77	0.00	0.16
pitch	-0.19	-0.03	0.01	0.05	0.20	0.01	0.06
intensity	-0.15	-0.02	0.00	0.02	0.10	0.00	0.03

Table 8
Estimated marginal effects for session of games #4. *Notes*: Exposure to entrainment metrics were standardized using *z*-scores before being introduced to the regressions. *p < 0.1; **p < 0.05; ***p < 0.01.

A. Dis/Entrainment policy				
	Mirrored		Against-static	
	OLS	GLME	OLS	GLME
	(1)	(2)	(3)	(4)
$tailored\ policy_t = 1$	− 0.0522	− 0.0302	− 0.0595	− 0.0608
gofish=1	− 0.3064***	− 0.3212***	− 0.1950***	− 0.1981***
Num. obs.	308	308	378	378
Num. groups: player		11		11
B. Dis/Entrainment exposure				
	(5)	(6)	(7)	(8)
$exp\_ent_t^{speech\ rate}$	− 0.0236	− 0.0107		
$exp\_disent_t^{speech\ rate}$	0.1064***	0.1256***	0.0006	0.0017
$exp\_ent_t^{pitch}$	− 0.0639*	− 0.0713*	− 0.0065	− 0.0120
$exp\_disent_t^{pitch}$	0.0410	0.0145		
$exp\_ent_t^{intensity}$	− 0.0464	− 0.0364	− 0.0089	− 0.0039
$exp\_disent_t^{intensity}$	0.0004	0.0140		
gofish=1	− 0.2968***	− 0.3189***	− 0.1965***	− 0.1988***
Num. obs.	308	308	378	378
Num. groups: player		11		11

Table 9
Estimated average marginal effects for all against-static sessions taken as a whole. *Notes*: Exposure to entrainment metrics were standardized using *z*-scores before being introduced to the regressions. *p < 0.1; **p < 0.05; ***p < 0.01.

	OLS (1)	GLME (2)
$exp\_ent_t^{speech\ rate}$	− 0.0152	− 0.0157
$exp\_disent_t^{speech\ rate}$	− 0.0084	− 0.0079
$exp\_ent_t^{pitch}$	− 0.0282*	− 0.0281*
$exp\_ent_t^{intensity}$	0.0355**	0.0348**
gofish=1	− 0.1566***	− 0.1580***
Num. obs.	1456	1456
Num. groups: player		45

of exposure to entrainment under the mirrored format. However, for the case of the against-static format, we still find statistically significant coefficients for exposure to intensity entrainment. The ones associated to entrainment on pitch under this format were not found to be statistically significant but, as in session of games #2, remain negative. Once again, the estimated effect size is considerable, a one standard deviation rise in exposure to entrainment on intensity rises the probability of keeping the avatar 42.3% as much as an advice leading to Go Fish reduces it.

Discussion Overall, this session showed less association between the experimental setup and helpers' trustworthiness. Still, these are quite consistent with results from the second session for the against-static format: Once again the effects of entrainment on pitch and intensity went in opposite directions. In the mirrored format, where we previously found evidence of positive effects at the policy level of entrainment on the three acoustic-prosodic features taken altogether, we do not observe evidence suggesting significant effects of entraining only on pitch and intensity.

3.5. Session of games #4: Disentrainment on speech rate and entrainment on pitch and intensity

Motivation and setup For our last session of games we opted to experiment with a procedure incorporating as many insights gathered in previous games as possible. In particular, results from the first session of games suggested that entrainment on speech rate might have a negative effect on the players' choice to maintain a helper, while the second suggested that entrainment on pitch and intensity, in addition to entrainment on speech rate, might have a positive effect. Results from the third session pointed toward a similar direction, although effects were much smaller. Taking all this into account, we opted to test the effects of a helper which disentrains on speech rate and entrains on pitch and intensity.

It should be mentioned that, even when our results pointed to opposing effects of entrainment on intensity and pitch (the effect of entrainment on intensity being positive in general and the effect of entrainment on pitch negative or null), we decided to keep varying both acoustic-prosodic features in synchrony as empirical evidence points toward a tendency for these two features to be positively correlated (see Gramming et al., 1988).

During this session of games we continued to use the mirrored and the against-static formats. In the mirrored format one helper, which we refer to as the *tailored helper*, disentrained on speech rate and entrained on pitch and intensity ($policy\_of\_interest_t = 1$), while the other one entrained on speech rate and disentrained on pitch and intensity ($policy\_of\_interest_t = 0$). In the against-static format a helper (the tailored one) disentrained on speech rate and entrained on pitch and intensity ($policy\_of\_interest_t = 1$) while the other followed a static behavior ($policy\_of\_interest_t = 0$). Once again, helper assignment was counterbalanced across participants. Assignment of policies to helpers remained static across games played by a given participant.

Subjects Data from 24 participants was gathered (9 female, 15 male. Average age = 21.13, sd = 2.01), 12 participants played in the mirrored approach, while 12 player in the against-static approach. 23 games were excluded as they did not meet the inclusion criteria described in Section 3.1 (14 in the mirrored approach, 9 in the against-static approach). The

sessions were carried out during November 2017.

Distribution of the observed acoustic-prosodic feature variations Table 7 presents the distribution of the variation of all three acoustic-prosodic features across both game formats. This table report statistics similar to the previous ones. However, when compared to the previous session of games, speech rate shows a much more symmetrical behavior.

Two-sided exact binomial tests In the mirrored games the tailored helper was selected 156 out of 330 times (47.27%). A two-sided exact binomial test fails to reject the null hypothesis that the helpers were chosen randomly ($p = .35$). In the against-static games the tailored helper was chosen 186 out of 405 times (45.93%). A two-sided exact binomial fails again to reject at standard significance levels the null hypothesis that the helpers were chosen randomly ($p = .11$).

Regression analysis Table 8 summarizes results from session of games #4 regression analysis. Tables A8 and A9 present detailed information on each regression. At the policy level, the estimated coefficients do not pass statistical significance tests at standard levels in both the mirrored and against-static formats. At the exposure level, the estimated coefficient of disentrainment on speech rate under the mirrored format shows a significant positive effect (equal to 35.8% the estimated fall associated to an advice leading to Go Fish), while the one for pitch suggests a significant negative effect (equal to 21.5% the estimated fall associated to an advice leading to Go Fish). All estimated coefficients under the against-static format do not pass statistical significance tests at standard levels.

Discussion Notably, results from this session of games are not entirely consistent with the ones found in session of games #2 and session of games #3 (although the effects of speech rate go in hand with the ones found in session of games #1). In this sense, these results reinforce the insight gathered in the second session of games, where we noted that it may be the case that the effects of a (dis)entrainment policy on a given acoustic-prosodic feature may be influenced to some extent by the behavior of the remaining acoustic-prosodic features. In particular, the case of entrainment on intensity in the against-static games is illustrative: In both the second and third sets, we found positive effects of entrainment on intensity under the against-static approach, however when disentrainment on speech rate is added to the mix (as it happened in these last games), we do not find any statistically significant effect associated to entrainment on intensity.

3.6. Meta analysis

An advantage of the against-static format relative to the mirrored one is that no matter which session subjects participated in, they always had to choose between an adapting helper versus a static one, which remained the same across all sessions of games. Note however that this is not true for the mirrored format, where no helper exhibited the same behavior across sessions of games. This property of the against-static format allowed us to run a meta analysis of the data collected across different sessions, as if it had been collected in a single one. Doing this allows us to gain statistical power and to capture more subtle phenomena.

In this section we analyze the data obtained from all sessions played under the against-static format as a whole. Concretely, we estimated Eq. 6 as in previous sections with the sole difference that we added session number indicator variables as covariates to capture any fixed effects which might be attributed to a particular session of games.

Table 9 contains the obtained estimates when running the regression analysis on all against-static games taken as a whole. Table A10 presents detailed results. For the case of pitch and intensity we do find associations between exposure to entrainment and maintaining a helper. Nonetheless, we find effects in opposite directions. For the case of entrainment on pitch, our results suggest that higher exposure to entrainment impacts negatively in keeping a helper, but for the case of entrainment on intensity we find a positive effect. We do not find any association between entrainment or disentrainment on speech rate and participants keeping a helper. Regarding effect sizes, results suggest that a one standard deviation rise in exposure to entrainment on pitch lowers the probability of keeping the avatar 17.8% as much as an advice leading to Go Fish reduces it, while a one standard deviation rise in exposure to entrainment on intensity rises this probability 22% as much as an advice leading to a Go Fish reduces it.

4. Discussion

4.1. Summary of findings

Across different experimental setups we found associations between acoustic-prosodic entrainment policies and trust (as measured by relying on a particular virtual helper for guidance and assistance). In this way, our results suggest and provide further evidence pointing toward an association between acoustic-prosodic (dis)entrainment in spoken dialogue systems and the way users perceive the capabilities of such systems.

Based on a meta-analysis considering data from all sessions of games, we observe, as overall patterns, associations between maintaining a helper and exposure to entrainment on pitch and intensity. But these effects go in opposite directions. For entrainment on pitch a negative association with maintaining a helper is observed, while for entrainment on intensity a positive one is observed. Interestingly, the estimated effect sizes of statistically significant coefficients are non-negligible when compared to the effect size of an advice leading to Go Fish. A one standard deviation rise in exposure to entrainment in pitch/intensity conveys a decrease/increase in the probability of keeping an avatar equivalent to 17.8%/22% the estimated fall of an advice leading to Go Fish.

However, a detailed characterization of these associations stands as a challenging task. Our findings explicitly forefront reasons which make this difficult. In particular:

- Our data show that the way entrainment on a given acoustic-prosodic feature affects users is not completely independent of the way other acoustic-prosodic features behave. For example, this was the case of entrainment on intensity. In sessions of games #2 and #3 the effects found were positive, but when disentrainment on speech rate was added to the mix in session of games #4, we did not find statistically significant effects.
- Finding a particular effect of a given entrainment policy for a particular acoustic-prosodic feature does not necessarily lead to observing the opposite effect if the system adapts in the opposite way (for example, seeing a negative effect of entrainment on speech rate does not imply that disentrainment on speech rate has a positive effect). In general, when significant effects of a given entrainment policy were found (e.g. a negative effect of pitch entrainment) the opposite effect of adapting in the opposite way was not found (i.e. a positive effect of pitch disentrainment).
- Third, and somewhat connected to the first point, at the policy level one may not be seeing an effect of a given entrainment strategy, but this may be due to the effects of exposure to entrainment across acoustic-prosodic features cancelling each other. In particular,

results from the against-static format games played in session of games #2 and #3 suggest this may be the case. For these games, we did not find effects at the policy level, but we did find effects in opposite directions at the exposure level (negative for pitch entrainment and positive for intensity entrainment).

4.2. Limitations and future research

Making a system adapt to the users' way of speaking involves the interaction of different sub-systems, each of them having subtleties. At the moment of implementation most commercial ASR systems did not return word time-stamps, something which is currently done by some at different levels of detail. This led us to use the PocketSphinx toolkit with a restricted grammar. Clearly, a restricted grammar limits the naturalness of speech interaction, and this should be addressed in future work.

TTS system limitations also impacted the task design. TTS systems are quite limited regarding their prosodic modification capabilities, especially when trained in languages other than English. As there is evidence showing that humans entrain on more acoustic-prosodic features than the three we studied (e.g. pause length, voice quality), and given that our results suggest that adapting one acoustic-prosodic feature may impact on the effects of another one, future research could also add into the analysis omitted acoustic-prosodic features and study their relation with conversation social outcomes. Recent developments in expressive speech synthesis (e.g. Wang et al., 2018; Skerry-Ryan et al., 2018) make this line of research particularly promising.

There are also aspects regarding the entrainment algorithm which should be considered. First, as previously reported, humans may entrain in multiple ways (proximity, convergence, synchrony), and it may be the case that not all acoustic-prosodic features entrain in the same manner (for example, pitch entraining according to synchrony and intensity according to proximity). This work placed its focus on the effects of synchrony and anti-synchrony, but future research should also focus on other forms of entrainment (see, for example, Weise and Levitan, 2018).

Second, in our experimental task, a helper entrains as it gives its advice to the way users request it. Given that this is a very common dialogue structure in current virtual assistants, understanding the effects of acoustic-prosodic entrainment under this type of dialogues is clearly important. However, recent research provides preliminary evidence suggesting that entrainment may emerge differently for different dialog acts (Reichel et al., 2018b; Gauder et al., 2018). Future research should also focus on the effects of incorporating dialog acts into the entrainment algorithms.

In this article we followed a purely experimental approach, and, in this way, studied a reduced number of hypotheses which were of interest to us. Nevertheless, we recognize that the collected data is rich and may be hiding patterns that, although not explicitly tested, may guide the design of future acoustic-prosodic entrainment algorithms. In this way, future research should study this data following a corpus study approach.

Regarding the complexity of characterizing this kind of effects, we believe this study, by making explicit how complex the problem is, provides an important indirect methodological result. Well-designed experimental paradigms are regarded as a "gold standard" for characterizing the impact of speech related behavior on social outcomes; however, speech behavior is a very complex phenomenon, and characterizing complex phenomena requires large amounts of data in order to gain statistical power. This poses a dilemma, as collecting in-lab speech data under experimental paradigms is quite costly. Future research should focus on assessing and validating online formats for running these experiments. One first step in this direction could be to replicate well-known and established in-lab speech related experiments and test if their results are robust to an online setup.

Funding sources

This material is based upon work supported by ANPCYT PICT 2014-1561, the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055, the National Science Foundation under Award No. 1845710, and the Scientific Grant Agency of the Slovak Republic under the grant VEGA 2/0161/18.

CRediT authorship contribution statement

Ramiro H. Gálvez: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Agustín Gravano:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Štefan Beňuš:** Conceptualization, Methodology, Investigation, Resources, Writing - review & editing, Supervision, Formal analysis. **Rivka Levitan:** Conceptualization, Methodology, Writing - review & editing. **Marian Trnka:** Conceptualization, Methodology, Software, Resources. **Julia Hirschberg:** Conceptualization, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Table A1
Frequency of helper switches in a game for the 294 non-practice games. Note: Frequencies are disaggregated by game number. For example, the third row indicates that three participants switched helpers twice in the first game, 13 switched twice in the second one, and 11 switched twice in the third one.

	Game number		
# switches	1	2	3
0	4	19	26
1	5	5	7
2	3	13	11
3	7	8	9
4	16	9	8
5	15	10	8
6	12	7	12
7	12	8	5
8	9	7	7
9	6	6	2
10	4	4	2
11	2	1	0
12	2	1	1
13	1	0	0
14	0	0	0

Table A2
Detailed regression analysis of dis/entrainment setup for session of games #1.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.6419***	0.5861*		0.5524***	0.2324	
	(0.0742)	(0.3246)		(0.0817)	(0.3484)	
$entrainmentpolicy_t = 1$	0.0015	0.0060	0.0014	− 0.0969	− 0.4084*	− 0.0985
	(0.0563)	(0.2448)		(0.0600)	(0.2475)	
deck=1	− 0.0693	− 0.2915	− 0.0679	− 0.0163	− 0.0697	− 0.0166
	(0.0722)	(0.3086)		(0.0814)	(0.3386)	
gofish=1	0.0000	0.0065	0.0015	− 0.0912	− 0.3736	− 0.0900
	(0.0794)	(0.3450)		(0.0863)	(0.3562)	
Eugenia=1	− 0.0492	− 0.2119	− 0.0484	0.0676	0.2764	0.0664
	(0.0548)	(0.2380)		(0.0595)	(0.2456)	
female=1	0.1197*	0.5410*	0.1192	− 0.0732	− 0.3583	− 0.0863
	(0.0628)	(0.2935)		(0.0657)	(0.3372)	
turn number	− 0.0046	− 0.0203	− 0.0046	0.0077	0.0321	0.0077
	(0.0071)	(0.0309)		(0.0079)	(0.0327)	
game number=3	0.0750	0.3433	0.0772	− 0.0243	− 0.0838	− 0.0200
	(0.0668)	(0.3005)		(0.0641)	(0.2690)	
game number=4	0.0363	0.1648	0.0379	0.0009	0.1048	0.0247
	(0.0711)	(0.3127)		(0.0816)	(0.3633)	
R^2	0.0225			0.0278		
Adj. R^2	-0.0025			0.0018		
Num. obs.	322	322		308	308	
RMSE	0.4840			0.4965		
AIC		436.8280			433.0272	
BIC		474.5735			470.3282	
Log Likelihood		-208.4140			-206.5136	
Num. groups: player		13			10	
Var: player constant		0.0149			0.0557	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A3
Detailed regression analysis of dis/entrainment exposure for session of games #1.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.6429***	0.5899		0.5055***	0.0420	
	(0.0870)	(0.3831)		(0.0784)	(0.3327)	
$exp\_ent_t^{speech\ rate}$	− 0.0045	− 0.0192	− 0.0043	− 0.0010	− 0.0057	− 0.0014
	(0.0041)	(0.0179)		(0.0042)	(0.0173)	
$exp\_disent_t^{speech\ rate}$	0.0026	0.0119	0.0027			
	(0.0043)	(0.0192)				
deck=1	− 0.0652	− 0.2716	− 0.0626	− 0.0183	− 0.0781	− 0.0188
	(0.0720)	(0.3105)		(0.0818)	(0.3375)	
gofish=1	0.0062	0.0391	0.0088	− 0.0919	− 0.3735	− 0.0908
	(0.0794)	(0.3484)		(0.0866)	(0.3548)	
Eugenia=1	− 0.0449	− 0.1905	− 0.0431	0.0896	0.3637	0.0881
	(0.0542)	(0.2372)		(0.0581)	(0.2392)	
female=1	0.1175*	0.5361*	0.1171	− 0.0790	− 0.3773	− 0.0917
	(0.0626)	(0.2975)		(0.0659)	(0.3347)	
turn number	− 0.0049	− 0.0222	− 0.0050	0.0074	0.0310	0.0074
	(0.0071)	(0.0311)		(0.0079)	(0.0327)	
game number=3	0.0893	0.4088	0.0910	− 0.0158	− 0.0519	− 0.0125
	(0.0679)	(0.3074)		(0.0651)	(0.2703)	
game number=4	0.0564	0.2580	0.0586	− 0.0040	0.0793	0.0189
	(0.0710)	(0.3168)		(0.0824)	(0.3622)	
R^2	0.0315			0.0195		
Adj. R^2	0.0035			-0.0067		
Num. obs.	322	322		308	308	
RMSE	0.4826			0.4987		
AIC		435.8939			435.6518	
BIC		477.4140			472.9528	
Log Likelihood		-206.9470			-207.8259	
Num. groups: player		13			10	
Var: player constant		0.0183			0.0542	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A4
Detailed regression analysis of dis/entrainment setup for session of games #2.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.5116***	0.0103		0.5574***	0.2276	
	(0.0891)	(0.3872)		(0.0752)	(0.3337)	
entrainment policy$_t$ = 1	0.1358**	0.5743**	0.1384	− 0.0174	− 0.0666	− 0.0154
	(0.0599)	(0.2506)		(0.0513)	(0.2211)	
deck=1	0.1585*	0.6531*	0.1501	0.0991	0.4739	0.1074
	(0.0836)	(0.3687)		(0.0697)	(0.3131)	
gofish=1	0.0229	0.0817	0.0193	− 0.1890***	− 0.7925***	− 0.1887
	(0.0836)	(0.3463)		(0.0716)	(0.3035)	
Eugenia=1	− 0.0068	− 0.0304	− 0.0072	− 0.0418	− 0.1930	− 0.0447
	(0.0582)	(0.2434)		(0.0514)	(0.2219)	
female=1	− 0.0355	− 0.1202	− 0.0285	− 0.0466	− 0.1776	− 0.0412
	(0.0660)	(0.3048)		(0.0525)	(0.2727)	
turn number	− 0.0045	− 0.0183	− 0.0043	0.0123*	0.0533*	0.0123
	(0.0078)	(0.0325)		(0.0069)	(0.0294)	
game number=3	0.0341	0.1660	0.0389	− 0.0102	− 0.0237	− 0.0055
	(0.0679)	(0.2930)		(0.0589)	(0.2578)	
game number=4	− 0.0683	− 0.2479	− 0.0593	− 0.0165	− 0.0063	− 0.0014
	(0.0894)	(0.3812)		(0.0648)	(0.2911)	
R^2	0.0359			0.0426		
Adj. R^2	0.0088			0.0219		
Num. obs.	294	294		378	378	
RMSE	0.4949			0.4880		
AIC		412.1100			515.7234	
BIC		448.9458			555.0723	
Log Likelihood		-196.0550			-247.8617	
Num. groups: player		12			12	
Var: player constant		0.0402			0.0606	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A5
Detailed regression analysis of dis/entrainment exposure for session of games #2.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.5123***	− 0.0589		0.5403***	0.1722	
	(0.1118)	(0.5001)		(0.0746)	(0.3264)	
$exp\_ent_t^{speech\ rate}$	− 0.0015	− 0.0042	− 0.0010	− 0.0043	− 0.0171	− 0.0039
	(0.0047)	(0.0200)		(0.0047)	(0.0205)	
$exp\_disent_t^{speech\ rate}$	0.0052	0.0230	0.0054			
	(0.0056)	(0.0236)				
$exp\_ent_t^{pitch}$	0.0114	0.0452	0.0106	− 0.0156*	− 0.0672*	− 0.0154
	(0.0108)	(0.0458)		(0.0093)	(0.0408)	
$exp\_disent_t^{pitch}$	− 0.0077	− 0.0298	− 0.0070			
	(0.0116)	(0.0484)				
$exp\_ent_t^{intensity}$	0.0246	0.1253	0.0295	0.0477**	0.1997*	0.0456
	(0.0229)	(0.1013)		(0.0233)	(0.1082)	
$exp\_disent_t^{intensity}$	− 0.0029	− 0.0059	− 0.0014			
	(0.0289)	(0.1192)				
deck=1	0.1590*	0.6437*	0.1473	0.0926	0.4409	0.0990
	(0.0849)	(0.3717)		(0.0695)	(0.3146)	
gofish=1	0.0226	0.0719	0.0169	− 0.1885***	− 0.8000***	− 0.1882
	(0.0848)	(0.3515)		(0.0713)	(0.3057)	
Eugenia=1	− 0.0106	− 0.0368	− 0.0087	− 0.0159	− 0.0825	− 0.0189
	(0.0619)	(0.2579)		(0.0518)	(0.2243)	
female=1	− 0.0247	− 0.0600	− 0.0141	− 0.0309	− 0.1323	− 0.0303
	(0.0674)	(0.3236)		(0.0527)	(0.2504)	
turn number	− 0.0038	− 0.0151	− 0.0036	0.0124*	0.0543*	0.0124
	(0.0079)	(0.0327)		(0.0068)	(0.0296)	
game number=3	0.0459	0.2401	0.0560	− 0.0010	− 0.0094	− 0.0022
	(0.0700)	(0.3077)		(0.0590)	(0.2575)	
game number=4	− 0.0487	− 0.1303	− 0.0310	− 0.0068	− 0.0104	− 0.0024
	(0.0924)	(0.3998)		(0.0651)	(0.2890)	
R^2	0.0394			0.0575		
Adj. R^2	-0.0052			0.0318		
Num. obs.	294	294		378	378	
RMSE	0.4984			0.4855		
AIC		420.7816			514.5521	
BIC		476.0353			561.7708	
Log Likelihood		-195.3908			-245.2760	
Num. groups: player		12			12	
Var: player constant		0.0630			0.0285	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A6
Detailed regression analysis of dis/entrainment setup for session of games #3.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.4454***	− 0.2588		0.4318***	− 0.2923	
	(0.0737)	(0.3277)		(0.0713)	(0.2971)	
entrainment policy$_t$ = 1	0.0616	0.2714	0.0628	0.0153	0.0679	0.0160
	(0.0514)	(0.2213)		(0.0499)	(0.2095)	
deck=1	− 0.0432	− 0.1963	− 0.0455	0.1306*	0.5743**	0.1339
	(0.0689)	(0.2977)		(0.0666)	(0.2900)	
gofish=1	− 0.2692***	− 1.1731***	− 0.2753	− 0.1308*	− 0.5541*	− 0.1309
	(0.0750)	(0.3283)		(0.0723)	(0.3053)	
Eugenia=1	− 0.0158	− 0.0669	− 0.0154	0.0279	0.1202	0.0283
	(0.0514)	(0.2209)		(0.0502)	(0.2110)	
female=1	0.0691	0.3040	0.0689	− 0.0378	− 0.1598	− 0.0376
	(0.0727)	(0.3999)		(0.0539)	(0.2262)	
turn number	0.0149**	0.0659**	0.0152	0.0159**	0.0673**	0.0158
	(0.0070)	(0.0305)		(0.0066)	(0.0279)	
game number=3	0.0313	0.1668	0.0390	− 0.1119*	− 0.4750*	− 0.1129
	(0.0625)	(0.2699)		(0.0612)	(0.2576)	
game number=4	0.0838	0.3188	0.0737	0.0098	0.0399	0.0094
	(0.0631)	(0.2801)		(0.0602)	(0.2536)	
R^2	0.0474			0.0546		
Adj. R^2	0.0259			0.0349		
Num. obs.	364	364		392	392	
RMSE	0.4883			0.4904		
AIC		497.3185			539.0358	
BIC		536.2900			578.7484	
Log Likelihood		-238.6592			-259.5179	
Num. groups: player		11			12	
Var: player constant		0.0809			0.0000	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A7
Detailed regression analysis of dis/entrainment exposure for session of games #3.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.4314***	− 0.2994		0.4266***	− 0.3137	
	(0.0855)	(0.3756)		(0.0694)	(0.2903)	
$exp\_ent_t^{pitch}$	0.0064	0.0285	0.0066	− 0.0106	− 0.0454	− 0.0106
	(0.0104)	(0.0449)		(0.0088)	(0.0372)	
$exp\_disent_t^{pitch}$	0.0044	0.0147	0.0034			
	(0.0111)	(0.0478)				
$exp\_ent_t^{intensity}$	0.0165	0.0601	0.0139	0.0343*	0.1487*	0.0346
	(0.0261)	(0.1165)		(0.0182)	(0.0792)	
$exp\_disent_t^{intensity}$	0.0091	0.0365	0.0084			
	(0.0284)	(0.1215)				
deck=1	− 0.0383	− 0.1820	− 0.0423	0.1273*	0.5657*	0.1308
	(0.0697)	(0.2997)		(0.0664)	(0.2909)	
gofish=1	− 0.2600***	− 1.1306***	− 0.2666	− 0.1271*	− 0.5421*	− 0.1269
	(0.0755)	(0.3279)		(0.0721)	(0.3058)	
Eugenia=1	− 0.0197	− 0.0823	− 0.0190	0.0342	0.1508	0.0351
	(0.0519)	(0.2218)		(0.0503)	(0.2135)	
female=1	0.0689	0.3060	0.0695	− 0.0494	− 0.2148	− 0.0501
	(0.0749)	(0.3972)		(0.0542)	(0.2299)	
turn number	0.0145**	0.0644**	0.0149	0.0161**	0.0686**	0.0160
	(0.0071)	(0.0307)		(0.0066)	(0.0280)	
game number=3	0.0331	0.1721	0.0404	− 0.1111*	− 0.4759*	− 0.1121
	(0.0629)	(0.2702)		(0.0611)	(0.2594)	
game number=4	0.0942	0.3626	0.0839	0.0167	0.0657	0.0152
	(0.0646)	(0.2848)		(0.0600)	(0.2546)	
R^2	0.0465			0.0634		
Adj. R^2	0.0167			0.0413		
Num. obs.	364	364		392	392	
RMSE	0.4906			0.4887		
AIC		503.8718			537.3820	
BIC		554.5348			581.0659	
Log Likelihood		-238.9359			-257.6910	
Num. groups: player		11			12	
Var: player constant		0.0717			0.0000	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A8
Detailed regression analysis of dis/entrainment setup for session of games #4.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.4649***	− 0.3236		0.5069***	− 0.0040	
	(0.0795)	(0.4060)		(0.0746)	(0.3438)	
tailored policy$_t$ = 1	− 0.0522	− 0.1407	− 0.0302	− 0.0595	− 0.2676	− 0.0608
	(0.0547)	(0.2580)		(0.0501)	(0.2209)	
deck=1	0.0428	0.1179	0.0253	0.0784	0.3606	0.0803
	(0.0713)	(0.3268)		(0.0651)	(0.2966)	
gofish=1	− 0.3064***	− 1.5281***	− 0.3212	− 0.1950***	− 0.8387***	− 0.1981
	(0.0783)	(0.3936)		(0.0710)	(0.3075)	
Eugenia=1	− 0.0666	− 0.2383	− 0.0512	− 0.0221	− 0.0796	− 0.0180
	(0.0554)	(0.2586)		(0.0503)	(0.2218)	
female=1	0.0052	0.0526	0.0113	0.0744	0.3450	0.0776
	(0.0562)	(0.3790)		(0.0510)	(0.2843)	
turn number	0.0214***	0.1019***	0.0218	0.0115*	0.0512*	0.0116
	(0.0071)	(0.0332)		(0.0065)	(0.0287)	
game number=3	− 0.1434**	− 0.5939*	− 0.1309	0.0605	0.3398	0.0768
	(0.0692)	(0.3326)		(0.0606)	(0.2765)	
game number=4	0.0799	0.4476	0.0970	0.0419	0.2164	0.0495
	(0.0647)	(0.3223)		(0.0604)	(0.2683)	
R^2	0.1180			0.0511		
Adj. R^2	0.0944			0.0305		
Num. obs.	308	308		378	378	
RMSE	0.4761			0.4829		
AIC		403.5155			507.8694	
BIC		440.8165			547.2184	
Log Likelihood		-191.7578			-243.9347	
Num. groups: player		11			11	
Var: player constant		0.1810			0.0772	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A9
Detailed regression analysis of dis/entrainment exposure for session of games #4.

	Mirrored			Against-static		
	OLS	*Generalized linear mixed-effects*	*Average marginal effects*	*OLS*	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.4430***	− 0.4836		0.4890***	− 0.0863	
	(0.1038)	(0.5396)		(0.0749)	(0.3433)	
$exp\_ent_t^{speech\ rate}$	− 0.0029	− 0.0066	− 0.0013			
	(0.0047)	(0.0226)				
$exp\_disent_t^{speech\ rate}$	0.0129***	0.0747***	0.0152	0.0001	0.0010	0.0002
	(0.0048)	(0.0251)		(0.0043)	(0.0189)	
$exp\_ent_t^{pitch}$	− 0.0212*	− 0.1158*	− 0.0236	− 0.0020	− 0.0167	− 0.0038
	(0.0118)	(0.0595)		(0.0104)	(0.0458)	
$exp\_disent_t^{pitch}$	0.0113	0.0196	0.0040			
	(0.0103)	(0.0505)				
$exp\_ent_t^{intensity}$	− 0.0275	− 0.1057	− 0.0216	− 0.0055	− 0.0108	− 0.0024
	(0.0236)	(0.1170)		(0.0198)	(0.0871)	
$exp\_disent_t^{intensity}$	0.0002	0.0446	0.0091			
	(0.0241)	(0.1167)				
deck=1	0.0345	0.0593	0.0121	0.0803	0.3674	0.0821
	(0.0714)	(0.3391)		(0.0658)	(0.2983)	
gofish=1	− 0.2968***	− 1.5809***	− 0.3189	− 0.1965***	− 0.8393***	− 0.1988
	(0.0781)	(0.4065)		(0.0716)	(0.3080)	
Eugenia=1	− 0.0567	− 0.1778	− 0.0364	− 0.0235	− 0.0875	− 0.0198
	(0.0553)	(0.2692)		(0.0505)	(0.2216)	
female=1	− 0.0019	0.0483	0.0099	0.0765	0.3526	0.0795
	(0.0564)	(0.4232)		(0.0520)	(0.2864)	
turn number	0.0196***	0.0985***	0.0201	0.0117*	0.0522*	0.0118
	(0.0070)	(0.0343)		(0.0066)	(0.0288)	
game number=3	− 0.1468**	− 0.6430*	− 0.1352	0.0587	0.3338	0.0757
	(0.0689)	(0.3461)		(0.0618)	(0.2806)	
game number=4	0.0767	0.4210	0.0870	0.0374	0.1929	0.0443
	(0.0649)	(0.3332)		(0.0609)	(0.2684)	
R^2	0.1481			0.0481		
Adj. R^2	0.1105			0.0222		
Num. obs.	308	308		378	378	
RMSE	0.4719			0.4850		
AIC		400.9959			513.0576	
BIC		456.9474			560.2763	
Log Likelihood		-185.4980			-244.5288	
Num. groups: player		11			11	
Var: player constant		0.2497			0.0761	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

Table A10
Detailed regression analysis results for all against-static sessions taken as a whole.

	OLS	*Generalized linear mixed-effects*	*Average marginal effects*
constant	0.4893***	− 0.0383	
	(0.0446)	(0.2025)	
$exp\_ent_t^{speech\ rate}$	− 0.0028	− 0.0123	− 0.0029
	(0.0028)	(0.0118)	
$exp\_disent_t^{speech\ rate}$	− 0.0019	− 0.0076	− 0.0018
	(0.0036)	(0.0155)	
$exp\_ent_t^{pitch}$	− 0.0090*	− 0.0382*	− 0.0090
	(0.0053)	(0.0226)	
$exp\_ent_t^{intensity}$	0.0248**	0.1033**	0.0243
	(0.0113)	(0.0495)	
deck=1	0.0736**	0.3317**	0.0773
	(0.0348)	(0.1525)	
gofish=1	− 0. 1566***	− 0.6580***	− 0.1580
	(0.0372)	(0.1569)	
Eugenia=1	0.0184	0.0680	0.0160
	(0.0258)	(0.1102)	
female=1	− 0.0239	− 0.1011	− 0.0238
	(0.0271)	(0.1365)	
turn number	0.0124***	0.0529***	0.0124
	(0.0034)	(0.0146)	
game number=3	− 0.0140	− 0.0347	− 0.0082
	(0.0305)	(0.1315)	
game number=4	0.0075	0.0727	0.0170
	(0.0322)	(0.1413)	
session=2	0.0299	0.1226	0.0287
	(0.0399)	(0.1972)	
session=3	− 0.0388	− 0.1724	− 0.0411
	(0.0424)	(0.2058)	
session=4	0.0328	0.1334	0.0312
	(0.0439)	(0.2138)	
R^2	0.0357		
Adj. R^2	0.0264		
Num. obs.	1456	1456	
RMSE	0.4885		
AIC		1965.2753	
BIC		2049.8105	
Log Likelihood		-966.6377	
Num. groups: player		45	
Var: player constant		0.0496	

Notes: Standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01

References

Acosta, J.C., Ward, N.G., 2011. Achieving rapport with turn-by-turn, user-responsive emotional coloring. Speech Commun. 53 (9–10), 1137–1148.

Anderson, C.M., 2012. Ambiguity aversion in multi-armed bandit problems. Theory Decis. 72 (1), 15–33. https://doi.org/10.1007/s11238-011-9259-2.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Beňuš, Š., 2014. Social aspects of entrainment in spoken interaction. Cognit. Comput. 6 (4), 802–813. https://doi.org/10.1007/s12559-014-9261-4.

Beňuš, Š., Gravano, A., Levitan, R., Levitan, S.I., Willson, L., Hirschberg, J., 2014. Entrainment, dominance and alliance in supreme court hearings. Knowl. Based Syst. 71, 3–14.

Boersma, P., Weenink, D., 2018. Praat: doing phonetics by computer [computer program].Version 6.0.42, retrieved 15 August 2018 from http://www.praat.org.

Bourhis, R.Y., Giles, H., 1977. The Language of Intergroup Distinctiveness. In: Giles, H. (Ed.), Language, ethnicity and intergroup relations, 13. European Association of Experimental Social Psychology, The address of the publisher, pp. 119–135.

Brennan, S.E., Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation. J. Exp. Psychol. 22 (6), 1482.

Brockmann, C., Isard, A., Oberlander, J., White, M., 2005. Modelling alignment for affective dialogue. In Proc. of the UM'05 Workshop on Adapting the Interaction Style to Affective Factors.

Buschmeier, H., Bergmann, K., Kopp, S., 2009. An alignment-capable microplanner for natural language generation. Proceedings of the 12th European Workshop on Natural Language Generation. Association for Computational Linguistics, pp. 82–89.

Chartrand, T.L., Bargh, J.A., 1999. The chameleon effect: the perception–behavior link and social interaction. J. Pers. Soc. Psychol. 76 (6), 893.

Crumpton, J., Bethel, C.L., 2016. A survey of using vocal prosody to convey emotion in robot speech. Int. J. Soc. Robot.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., Kleinberg, J., 2012. Echoes of power: Language effects and power differences in social interaction. Proceedings of the 21st international conference on World Wide Web. ACM, pp. 699–708.

De Jong, M., Theune, M., Hofs, D., 2008. Politeness and alignment in dialogues with a virtual guide. Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems, pp. 207–214.

De Looze, C., Scherer, S., Vaughan, B., Campbell, N., 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. Speech Commun 58, 11–34.

Fandrianto, A., Eskenazi, M., 2012. Prosodic entrainment in an information-driven dialog system. Interspeech 2012.

Gauder, L., Reartes, M., Gálvez, R.H., Štefan, B., Gravano, A., 2018. Testing the effects of acoustic/prosodic entrainment on user behavior at the dialog-act level. Proc. 9th International Conference on Speech Prosody 2018, pp. 374–378. https://doi.org/10.21437/SpeechProsody.2018-76.

Giles, H., 1979. En p. smith. 1979. accommodation theory: optimal levels of convergence. H. Giles and St. Clair (eds.), Language and Social Psychology 45–87.

Giles, H., Coupland, N., Coupland, I., 1991. 1. Accommodation theory: communication, context, and. Contexts of accommodation: Developments in applied sociolinguistics 1.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., Perkins, W.H., 1988. Relationship between changes in voice pitch and loudness. J. Voice 2 (2), 118–126. https://doi.org/10.1016/S0892-1997(88)80067-5.

Gravano, A., Beňuš, Š., Levitan, R., Hirschberg, J., 2015. Backward mimicry and forward influence in prosodic contour choice in standard american english. Sixteenth Annual Conference of the International Speech Communication Association.

Healey, P.G.T., Purver, M., Howes, C., 2014. Divergence in dialogue. PLoS ONE 9 (6), 1–6. https://doi.org/10.1371/journal.pone.0098598.

Hu, Z., Halberg, G., Jimenez, C.R., Walker, M.A., 2016. Entrainment in Pedestrian Direction Giving: How Many Kinds of Entrainment? Situated Dialog in Speech-Based Human-Computer Interaction. Springer, pp. 151–164.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I., 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, 1. IEEE.I–I

Hung, V., Elvir, M., Gonzalez, A., DeMara, R., 2009. Towards a method for evaluating naturalness in conversational dialog systems. Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on. IEEE, pp. 1236–1241.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. Lmertest package: tests in linear mixed effects models. J. Stat. Softw. 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lee, M.D., Zhang, S., Munro, M., Steyvers, M., 2011. Psychological models of human and optimal performance in bandit problems. Cognitive Systems Research 12 (2), 164–174. https://doi.org/10.1016/j.cogsys.2010.07.007.The 9th International Conference on Cognitive Modeling. Manchester, UK, July 2009

Leeper, T. J., 2017. Interpreting regression results using average marginal effects with R's margins.

Levitan, R., 2014. Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue. Columbia University.

Levitan, R., Beňuš, Š., Gálvez, R.H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., Hirschberg, J., 2016. Implementing acoustic-prosodic entrainment in a conversational avatar. Interspeech 2016 1166–1170.

Levitan, R., Beňuš, Š., Gravano, A., Hirschberg, J., 2015. Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison. Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 325–334.

Levitan, R., Benus, S., Gravano, A., Hirschberg, J., 2015. Entrainment and turn-taking in human-human dialogue. AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction.

Levitan, R., Gravano, A., Hirschberg, J., 2011. Entrainment in speech preceding backchannels. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2. Association for Computational Linguistics, pp. 113–117.

Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. Twelfth Annual Conference of the International Speech Communication Association.

Looze, C.D., Rauzy, S., 2011. Measuring speakers' similarity in speech by means of prosodic cues: Methods and potential. INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, 2011, pp. 1393–1396.

Lopes, J., Eskenazi, M., Trancoso, I., 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. Computer Speech & Language 31 (1), 87–112.

Lubold, N., Pon-Barry, H., Walker, E., 2015. Naturalness and rapport in a pitch adaptive learning companion. IEEE Automatic Speech Recognition and Understanding Workshop.

Lubold, N., Walker, E., Pon-Barry, H., Ogan, A., 2018. Automated pitch convergence improves learning in a social, teachable robot for middle school mathematics.

International Conference on Artificial Intelligence in Education. Springer, pp. 282–296.
Marge, M., Miranda, J., Black, A.W., Rudnicky, A.I., 2010. Towards improving the naturalness of social conversations with dialogue systems. Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, pp. 91–94.
Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An integrative model of organizational trust. Academy of management review 20 (3), 709–734.
Michalsky, J., Schoormann, H., 2017. Pitch convergence as an effect of perceived attractiveness and likability. Proc. Interspeech 2017, pp. 2253–2256. https://doi.org/10.21437/Interspeech.2017-1520.
Natale, M., 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. .J Pers. Soc. Psychol. 32 (5), 790.
Nenkova, A., Gravano, A., Hirschberg, J., 2008. High frequency word entrainment in spoken dialogue. Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers. Association for Computational Linguistics, pp. 169–172.
Pardo, J.S., 2006. On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119 (4), 2382–2393.
Pérez, J.M., Gálvez, R.H., Gravano, A., 2016. Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement. Interspeech 2016, pp. 1270–1274. https://doi.org/10.21437/Interspeech.2016-587.
Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and brain sciences 27 (2), 169–190.
Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behavioral and Brain Sciences 36 (4), 329–347.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.
Racey, D., Young, M.E., Garlick, D., Ngoc-Minh Pham, J., Blaisdell, A.P., 2011. Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules. Learning & Behavior 39 (3), 245–258. https://doi.org/10.3758/s13420-011-0025-7.
Reichel, U.D., Štefan, B., Mády, K., 2018. Entrainment profiles: comparison by gender, role, and feature set. Speech Commun. 100, 46–57. https://doi.org/10.1016/j.specom.2018.04.009.
Reichel, U.D., Mády, K., Cole, J., 2018. Prosodic entrainment in dialog acts. arXiv preprint arXiv:1810.12646.
Reitter, D., Keller, F., Moore, J.D., 2011. A computational cognitive model of syntactic priming. Cogn Sci 35 (4), 587–637.
Reitter, D., Moore, J.D., 2014. Alignment and task success in spoken dialogue. J Mem Lang 76, 29–46.
Robbins, H., et al., 1952. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society 58 (5), 527–535.
Sadoughi, N., Pereira, A., Jain, R., Leite, I., Lehman, J.F., 2017. Creating prosodic synchrony for a robot co-player in a speech-controlled game for children. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp. 91–99.
Schulz, E., Konstantinidis, E., Speekenbrink, M., 2018. Putting bandits into context: how function learning supports decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition 44 (6), 927–943.
Schweitzer, A., Lewandowski, N., 2014. Social factors in convergence of f1 and f2 in spontaneous speech. Proc. of the 10th International Seminar on Speech Production, Cologne..
Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., Clark, R., Saurous, R.A., 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. International Conference on Machine Learning, pp. 4700–4709.
Steyvers, M., Lee, M.D., Wagenmakers, E.-J., 2009. A bayesian analysis of human decision-making on bandit problems. J. Math. Psychol. 53 (3), 168–179.
Street Jr, R.L., 1984. Speech convergence and speech evaluation in fact-finding interviews. Hum. Commun. Res. 11 (2), 139–169.
Sutton, R.S., Barto, A.G., et al., 1998. Reinforcement learning: An introduction. MIT press.
Violante, L., Zivic, P.R., Gravano, A., 2013. Improving speech synthesis quality by reducing pitch peaks in the source recordings. HLT-NAACL, pp. 502–506.
Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., Saurous, R.A., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. International Conference on Machine Learning, pp. 5167–5176.
Ward, A., Litman, D., 2007. Measuring convergence and priming in tutorial dialog. University of Pittsburgh.
Weise, A., Levitan, R., 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 297–302. https://doi.org/10.18653/v1/N18-2048.
West, B.T., Welch, K.B., Galecki, A.T., 2014. Linear mixed models: A practical guide using statistical software. Chapman and Hall/CRC.