Speech Communication 115 (2019) 78–87

Contents lists available at ScienceDirect

Speech Communication

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

Individual differences in acoustic-prosodic entrainment in spoken dialogue

Andreas Weise[a,*], Sarah Ita Levitan[b], Julia Hirschberg[b], Rivka Levitan[a,c]

[a] Department of Computer Science, The Graduate Center, CUNY, New York, NY 10016, USA
[b] Department of Computer Science, Columbia University, New York, NY 10027, USA
[c] Department of Computer and Information Science, Brooklyn College, CUNY, Brooklyn, NY 11210, USA

ARTICLE INFO

Keywords:
Entrainment
Alignment
Prosody
Gender
English
Chinese

ABSTRACT

The tendency of conversation partners to adjust to each other to become similar, known as entrainment, has been studied for many years. Several studies have linked differences in this behavior to gender, but with inconsistent results. We analyze individual differences in two forms of local, acoustic-prosodic entrainment in two large corpora between English and Chinese native speakers conversing in English. The few previous studies of the effect of non-nativeness on entrainment that exist were based on much smaller numbers of speakers and focused on perceptual rather than acoustic measures. We find considerable variation in both degree and valence of entrainment behavior across speakers with some consistent trends, such as synchronous behavior being mostly positive in direction and somewhat more prevalent than convergence. However, we do *not* find entrainment to vary significantly based on gender, native language, or their combination. Instead, we propose as a hypothesis for further study, that gender mediates more complex interactions between sociocultural norms, conversation context, and other factors.

1. Introduction

Conversation partners tend to adapt their behavior to each other to become more similar. This phenomenon has been studied for many years and is commonly referred to as entrainment. It affects various linguistic dimensions, such as lexical choice (Brennan and Clark, 1996), syntactic structure (Reitter et al., 2006), and acoustic-prosodic features (Levitan and Hirschberg, 2011) and has been found to correlate with desirable conversation outcomes, including task success (Reitter and Moore, 2007), naturalness (Nenkova et al., 2008), and rapport (Lubold and Pon-Barry, 2014).

1.1. State of the art

Studies of entrainment vary greatly both in terms of how data is collected and how similarity is measured. This subsection discusses some of the different methods and their advantages and disadvantages.

The most basic choice regarding data collection is between an interactive and a non-interactive setting. The former is often employed to study social factors impacting entrainment behavior (e.g., Lee et al., 2010; Levitan et al., 2012; Manson et al., 2013) while the latter dampens these factors and allows for greater control to study the link between speech perception and production and how it interacts with remaining social factors (e.g., Goldinger, 1998; Lewandowski and Nygaard, 2018; Namy et al., 2002). Non-interactive settings are usually achieved through a shadowing paradigm in which speakers produce the same words first by reading them and then by repeating them after a previously recorded model talker. Interactive settings vary greatly, from task-oriented conversation (Abel and Babel, 2017; Levitan and Hirschberg, 2011; Pardo, 2006), to tutoring (Ward and Litman, 2007), interviews (Street, 1984), therapy (Lee et al., 2010; Nasir et al., 2018), or spontaneous conversation (Manson et al., 2013; Nasir et al., 2018), illustrating the ubiquity of entrainment in human interaction. We note that Pardo et al. (2018) offers an in-depth review of interactive and non-interactive settings and a study that collected data of both types from the same speakers.

The measurement of entrainment is characterized by a similarly fundamental dichotomy between a subjective but holistic perceptual approach and a more objective but often partial perspective based on acoustic measures such as pitch or speech rate.

Perceived similarity between two speakers is usually determined through AXB tests (e.g., Pardo, 2006; Babel et al., 2014; Lewandowski and Nygaard, 2018; Pardo et al., 2018). In this paradigm, introduced by Goldinger (1998), independent listeners are presented with triplets of samples from a pair of speakers. Sample A is a production of the first speaker before exposure to the partner, sample X is a production by the partner and sample B is a second production by the first speaker after exposure to X. Several listeners are asked to rate whether they

* Corresponding author.
E-mail addresses: aweise@gradcenter.cuny.edu, aweise@gradcenter.cuny.edu (A. Weise), sarahita@cs.columbia.edu, sarahita@cs.columbia.edu (S.I. Levitan), julia@cs.columbia.edu, julia@cs.columbia.edu (J. Hirschberg), rlevitan@brooklyn.cuny.edu, rlevitan@brooklyn.cuny.edu (R. Levitan).

https://doi.org/10.1016/j.specom.2019.10.007
Received 4 April 2019; Received in revised form 15 September 2019; Accepted 30 October 2019
Available online 1 November 2019
0167-6393/

find A or B more similar to X, with balanced presentation as AXB or BXA. Significant preference for B is then interpreted as evidence of entrainment. Some authors use an XAB scheme instead (Kim et al., 2011; Kim, 2012) or an entirely different approach to determine perceived similarity between sample pairs, such as Likert scales (Abel and Babel, 2017).

Acoustic measures of entrainment are far less standardized than perceptual ones. Some treat each acoustic feature individually, for instance through regression (Manson et al., 2013; Ward and Litman, 2007), time series analysis (Pérez et al., 2016), or Pearson correlations and mean comparisons (Levitan and Hirschberg, 2011). Others process many features simultaneously. Lee et al. (2011), for instance, proposed an approach based on principal component analysis to compare 37 features per pair of speech segments. Gravano et al. (2014), meanwhile, worked symbolically with ToBI annotations to determine the similarity of intonational contours. Using speaker recognition techniques based on Gaussian mixture models, Bailly and Martin (2014) assessed how much speakers adjusted their voice overall towards their interlocutor. And in a recent innovation, lastly, Nasir et al. (2018) trained a neural network to process over 200 features per utterance into an encoding, with the L1 norm of the differences between encodings interpreted as a measure of entrainment. While the separate processing of features tends to be easier to automate and has the potential to be used in live settings, it can lead to disparate results which are difficult to interpret. Joint processing, on the other hand, can require training data (Bailly and Martin, 2014; Nasir et al., 2018) or depend on high-quality annotation (Gravano et al., 2014).

Several studies have analyzed perceptual and acoustic measures of entrainment for the same recordings. Some found that individual acoustic features contributed to the perception of entrainment, even if they did not show significant entrainment by themselves (Lewandowski and Nygaard, 2018; Pardo et al., 2013). More commonly, however, no correlation was found between the two types of measures (Abel and Babel, 2017; Babel and Bulatov, 2012; Kim, 2012; Pardo et al., 2010). Note, though, that the methodology for measuring perceived similarity inherently limits the amount of speech that can be analyzed. Therefore, while perceived similarity provides a more holistic assessment of the audio signal of individual utterances (Pardo et al., 2013), automatic acoustic measures can process all of the audio even of long interactions and, thus, have the potential to represent conversations as a whole and the dynamics throughout them.

1.2. Variation in entrainment behavior

Even within the same corpus, applying consistent methodology, some studies have found variation in how pairs of speakers entrain. In a study of entrainment in multiple languages, Levitan et al. (2015a) found evidence of individual differences in entrainment behavior *within* languages. Similarly, Lubold and Pon-Barry (2014) observed variation in local entrainment across pairs of speakers. This variation involves both the number of features entrained on and the valence of the entrainment: that is, whether it is positive, indicating convergent behavior; negative, indicating diverging or complementary behavior; or mixed, indicating convergent behavior for some features and divergent for others. In recent years there has been some indication that both positive and negative entrainment may be beneficial to the conversation (Healey et al., 2014; Pérez et al., 2016), which motivates us to analyze valence.

Some authors have attempted to identify the sources of these individual differences, focusing on gender, with varying results. In a study of phonetic entrainment measured by perceived similarity, Pardo (2006) found that males in a dependent role entrained more than those in a position of power and males generally entrained more than females. In a larger study with 96 speakers, also using perceptual measures, Pardo et al. (2018) found no difference in the strength of entrainment between the genders or between same- and mixed-gender pairs. They did, however, observe that males were moderately consistent in their entrainment behavior across the two different contexts they analyzed (interactive and shadowing) while females were not. Using acoustic-prosodic measures, Levitan et al. (2012), found that male pairs entrained the least while pairs of mixed gender entrained the most. In a very similar corpus, but with Mandarin speakers, Xia et al. (2014) also found that male pairs entrained the least on intensity but mixed gender pairs entrained the least on speech rate. Reichel et al. (2018), lastly, analyzed another very similar corpus of Slovak speech. They found entrainment for similarly high numbers of acoustic-prosodic features for male and female speakers in positions of power, but with females entraining mostly positively and males mostly negatively. Some of the differences in results between the studies by Pardo and her collaborators and the others might be attributable to the fact that she used perceptual rather than acoustic measures. The other studies, however, analyzed very similar corpora with very similar measures, leaving the language of the speakers as the most notable difference and suggesting that sociocultural norms have an impact on how the genders differ in their entrainment behavior.

A few studies have also addressed the question of whether entrainment varies based on differences between the speakers' native languages and dialects. Kim et al. (2011) examined this in an interactive, task-oriented setting. They found entrainment in conversations in which interlocutors spoke in their shared native language (English or Korean) and dialect but not in English conversations with native language (English, Korean, or Chinese) or dialect differing between interlocutors. Pairs of English native speakers also entrained more than pairs of Korean native speakers. Using a shadowing setting, on the other hand, Kim (2012) obtained virtually the opposite result. In this case, native English speaking shadowers adapted most to model talkers whose native language Korean did *not* match their own, followed by those speaking a different dialect of English. Similarly, Lewandowski and Nygaard (2018) found greater entrainment by shadowers – again all native speakers of English – towards model talkers whose native language Spanish did not match theirs than towards native English speaking model talkers. These differences are despite the fact that all three studies primarily measured entrainment through perceived similarity (Lewandowski and Nygaard (2018) also analyzed three acoustic measures).

Kim et al. (2011) suggest that the lack of entrainment observed in their data among speakers whose native language or dialect does not match could be due to increased cognitive load. Both speech perception and production are more difficult for non-native speakers and at least perception is impeded by dialect mismatches. This interpretation is supported by the fact that Abel and Babel (2017) have since demonstrated decreasing degrees of entrainment with increasing cognitive load. Kim et al. further suggest that native speakers conversing with a non-native speaker may have inadvertently prevented entrainment by adopting "clear speech" (Smiljanic and Bradlow, 2009) in an attempt to increase intelligibility. The fact that the shadowing paradigm largely eliminates this factor might explain the seeming contradiction between the results of Kim et al. (2011) and those of Kim (2012) and Lewandowski and Nygaard (2018). Without the need to be intelligible to an interlocutor, the greater salience of accented speech can become a dominating factor. According to accounts of entrainment as automatic and caused by connections between perception and behavior (Chartrand and Bargh, 1999), this would result in stronger entrainment towards such speech.

In a recent, unpublished report, Loy and Smith (2019) analyzed the influence of non-nativeness on syntactic entrainment. They found that native English-speaking subjects do not differentially align with native or non-native confederates' use of double object (DO) versus prepositional object (PO) constructions. However, if the confederates use only DO phrases, including ungrammatical ones, then subjects entrain more towards non-native than native confederates. If, on the other hand, the non-native confederate merely has a stronger accent but uses both DO and PO constructions, then there is no difference in adaptation based on nativeness. The authors conclude that speakers take their interlocu-

tors' communicative abilities and needs into account when those are prominent in the context of the conversation. This is in line with the previous observation that links between perception and behavior gain importance in contexts where speakers are less constrained by concerns of intelligibility.

In summary, while there is evidence of variation in entrainment behavior, attempts to attribute this to gender or native language of the speakers have led to varying results highlighting the importance of other factors such as the context of exposure to speech, power dynamics, and sociocultural norms. Additionally, due to the relatively small number of speakers used in most of the studies discussed above, (e.g., four model talkers each in Kim, 2012 and Lewandowski and Nygaard, 2018) it is also conceivable that some of their results reflect idiosyncrasies of individual speakers or entrainment targets rather than population differences between male and female or native and non-native speakers, respectively.

1.3. Study overview

We consider two large corpora of dyadic, English speech described in Section 2. Our analysis is based on eight acoustic-prosodic features listed in Section 3 and two local forms of entrainment detailed in Section 4. Throughout Section 5 we analyze our data with regard to differences in entrainment behavior along multiple dimensions. First, Section 5.1 compares the entrainment behavior of speakers in different roles in one of our corpora. Section 5.2 then contributes further evidence of the basic existence of variation across speakers in the same context. In Section 5.3 and 5.4 we attempt to attribute these differences to speaker gender and native language, treating entrainment as a discrete and continuous phenomenon, respectively. The effect of non-nativeness on entrainment has been studied before but only on fewer speakers and in non-interactive settings or with perceptual rather than acoustic measures. The study of such an effect is motivated by the observation that entrainment varies by language (Levitan et al., 2015a) and relies on the speakers' ability to vary their speech, which likely differs between native and non-native speakers. Section 6, finally, discusses our results and plans for future work.

To sum up, this paper offers a systematic analysis of variations in entrainment behavior based on two large corpora and attempts to attribute those variations to gender and native language. In contrast to some prior work, we find that speaker and interlocutor gender are not significant factors in the degree and valence of entrainment behavior, suggesting that a complex interaction between gender and context, rather than gender alone, affects entrainment.

2. Corpora

2.1. Columbia X-Cultural Deception Corpus

The Columbia X-Cultural Deception Corpus (Levitan et al., 2015b) consists of 170 in-person, dyadic conversations in English. All 340 subjects were native speakers of either American English or Chinese. Each conversation consisted of two sessions, with either speaker acting as an interviewer (**ER**) in one of them and as an interviewee (**EE**) in the other. Interviewees would answer 24 biographical questions – 12 randomly chosen ones truthfully, the other half untruthfully, resulting in a combination of deceptive and non-deceptive speech from each participant – while interviewers would try to detect lies. Interviewers read the questions from a printout in the order of their choosing and were encouraged to ask additional, spontaneous follow-up questions to assess the truthfulness of the responses. The authors used Amazon Mechanical Turk to obtain a transcript of the whole corpus and then force-aligned it with the audio. We use inter-pausal units (IPUs) as the basis of our analysis, segments of speech from a single speaker connected by pauses of at most 50 ms each. Maximal sequences of IPUs from one speaker without interruption by the other constitute speaker turns.

2.2. Fisher Corpus

The Fisher Corpus (Cieri et al., 2004) contains over 11,000 dyadic conversations in English, conducted over the phone. We use a subset of 105 of these, selected as described in Suection 2.3. Pairs of subjects, who did not previously know each other, were asked to discuss a given topic for about 10 min. Conversations were transcribed in a semi-automatic process. We use the transcription segments as the smallest units of our analysis. These consist of uninterrupted speech by a single speaker but can include pauses longer than 50 ms, which we remove before feature extraction. We refer to these segments as IPUs and group them into turns in the same way as for the Deception Corpus. The corpus also contains meta-data on the speakers, including their native language and where they were raised.

2.3. Selection of balanced subsets

We note that while neither corpus was specifically designed to study entrainment, evidence of local entrainment has been found both in the Deception Corpus (Levitan et al., 2018) and the Fisher Corpus (Nasir et al., 2018). We use these two corpora because they are larger than those underlying most previous studies of entrainment while allowing us to analyze the effects not just of gender but also of native language on entrainment behavior. To do so, we select subsets of conversations that are balanced with regard to these characteristics.

Since our entrainment measures are asymmetric (see Section 4), they each yield one value per speaker. We group speakers by the combination of their native language, their gender, their interlocutor's native language, and the gender of their interlocutor and refer to these combinations as **speaker types**. One speaker type, for instance, is that of "male English native speakers responding to female Chinese native speakers", which we label by the abbreviated characteristics as "ME-FC". This results in the following 16 speaker types: FC-FC, FC-FE, FC-MC, FC-ME, FE-FC, FE-FE, FE-MC, FE-ME, MC-FC, MC-FE, MC-MC, MC-ME, ME-FC, ME-FE, ME-MC, and ME-ME. Note that the Fisher Corpus does not contain any conversations between pairs of Chinese native speakers, so four speaker types do not occur in that corpus: FC-FC, FC-MC, MC-FC, and MC-MC.

The smallest number of instances for any speaker type in either corpus is 15. Therefore, we choose 15 conversations per speaker type to generate balanced subsets from our corpora. Each conversation between speakers that differ in gender, native language, or both serves as an instance for two different speaker types. Each conversation between speakers of the same native language and gender, on the other hand, could serve as two instances of the same speaker type. Instead, we choose to use 15 different conversations for those speaker types as well and ignore one speaker in each of them. In doing so for the Deception Corpus, we balance the number of EEs and ERs and the number of speakers who are EE first or ER first. Note that for the rest of the paper we mean these balanced subsets whenever we refer to our corpora.

2.4. IPU statistics

Our analysis focuses on turn exchanges (see Section 4), using turn-initial and turn-final IPUs that do not overlap. In total, the Deception Corpus contains 88,363 such IPUs with an average of 5.10 syllables ($\sigma = 4.67$) and a duration of 1.19 seconds ($\sigma = 0.91$) per IPU, for a total of over 29 hours of speech. On average, there are 294.5 relevant IPUs ($\sigma = 143.32$) per speaker, with a minimum of 70 and a maximum of 751. Our analysis of the Fisher Corpus is based on 13,576 IPUs with an average of 11.31 syllables ($\sigma = 12.53$) and a duration of 1.88 seconds ($\sigma = 1.85$) per IPU, about 7 hours of speech overall. For this corpus, the average number of IPUs per speaker is 56.57 ($\sigma = 15.86$) with a minimum of 19 and a maximum of 100. We note that it is not uncommon for research on acoustic entrainment to be based on short segments of speech. For instance, Kim et al. (2011) and Abel and Babel (2017) both

used samples with lengths between 0.5 and 1.5 seconds for their perceptual measure of similarity in conversational speech. Also, while we only use up to two IPUs per turn, we note that the average number of IPUs per turn is 2.45 ($\sigma = 2.82$) in the Deception Corpus ($\mu = 2.83, \sigma = 3.58$ for interviewees; $\mu = 2.08, \sigma = 1.66$ for interviewers) and 1.61 ($\sigma = 1.31$) in the Fisher Corpus.

Further analysis of the number and length of IPUs reveals differences between speaker groups (details and statistical tests in Appendix A). First, we observe that the Chinese native speakers in our corpora use fewer syllables per IPU, that their IPUs are shorter in duration (in the Fisher Corpus only), and that they speak more slowly than the English native speakers. Conversations involving English native speakers, on the other hand, contain fewer turn exchanges. All this can be attributed to the cognitive load of conversing in a nonnative language, allowing native speakers to communicate faster and more efficiently. The latter matches the results of van Engen et al. (2010). Next, we find that female subjects in our data speak more slowly, in longer utterances than males, and that their conversations involve fewer turn exchanges. Lastly, interviewees in the Deception Corpus use fewer syllables per IPU but their IPUs last longer, i.e., they speak more slowly. This suggests that responding to the questions – and trying to lie convincingly half the time – resulted in greater cognitive load than asking them, coming up with follow-ups, and trying to discern truthfulness.

2.5. Speaker demographics

All English native speakers in our corpora either specified that they were raised in the US or we informally confirmed their accent to be American. Most of the Chinese native speakers were raised in China or Taiwan. For those raised in the US we informally confirmed the presence of a non-native accent. Most of the non-native speakers listed their native language as “Mandarin”, the others as “Chinese”, with no specific variety or dialect.

The average and standard deviation of the age of speakers in the Deception Corpus ($\mu = 23.2, \sigma = 4.6$) are lower than in the Fisher Corpus ($\mu = 34.2, \sigma = 11.7$). This is due to the fact that its participants were recruited largely from the Columbia University student body whereas recruiting for the Fisher Corpus was based on broader online and print advertising.

English proficiency among the non-native speakers varies greatly, from limited fluency to only subtle non-native accents. For the Fisher Corpus we have no data on language proficiency but the Deception Corpus lists the age at which each speaker first started learning English ($\mu = 9.8, \sigma = 3.4$). There is no significant correlation between the number of years that speakers have been learning English and either of our entrainment measures on any feature, both for the raw values and their magnitude. Therefore, in the rest of the paper we do not differentiate non-native speakers beyond their gender.

3. Features

To study entrainment, we extract eight acoustic-prosodic features from each IPU using Praat (Boersma and Weenink, 2018), a free speech analysis software. *Pitch*, the fundamental frequency of voiced speech segments, describes the tone of an utterance while its loudness is represented by *intensity*, the energy of the acoustic signal. We consider the mean and maximum values for both. *Speaking rate*, the utterance speed, is estimated using syllables per second. *Jitter* and *shimmer* are measures of small variations in pitch and intensity, respectively, which are perceived as vocal harshness. The *noise-to-harmonics ratio* (NHR), lastly, is associated with hoarseness. We z-score normalize each feature per speaker. That is, we use the normalized value $z = (x - \mu)/\sigma$, where x denotes the raw feature value while μ and σ are the speaker's mean and standard deviation for the respective feature over all IPUs.

Table 1
Significant differences in the entrainment behavior of the same speakers in the role of EE and ER, respectively. All entries refer to synchrony.

Feature	Cohen's d	p	
max. intensity	0.41	8.1e-07	*
speech rate	-0.41	1.4e-06	*
NHR	0.26	0.00212	*
shimmer	0.19	0.01352	.
jitter	0.18	0.01824	.

4. Entrainment measures

In this work we focus on local measures of entrainment which are based on similarity at the IPU level rather than aggregates over longer segments of conversation. We apply two of the local measures defined by Levitan and Hirschberg (2011). Local convergence determines to what extent the similarity at turn exchanges increases or decreases over the course of a conversation. Synchrony, on the other hand, measures the degree of coordination at turn exchanges, whether feature values for both speakers tend to rise and fall together. To compute them, we first determine the initial IPU of each turn (*target IPU*) and pair it with the last IPU of the partner's most recent turn (*partner IPU*), excluding pairs that overlap. We collect target IPUs separately per speaker, allowing us to attribute similarity to the responding speaker who has a more active role in facilitating it. This yields two asymmetric values per speaker pair and entrainment measure.

Specifically, both measures are defined using Pearson correlation coefficients. Convergence is the correlation between the negated absolute differences between target IPUs and their partner IPUs and time, represented by the number of turn exchanges. Synchrony is the correlation between the feature values for target IPUs and those for partner IPUs. To ensure that results are significant, we also compute each correlation for the same data in ten random permutations. We only consider a correlation for the real data to be significant if at most one correlation for a random permutation is significant.

We use these measures because variation in convergence and synchrony has been observed in prior research. For both measures, correlations can be positive or negative. Positive synchrony and convergence constitute accommodating behavior, speakers adjusting their speech to become more similar to partners. Negative synchrony can be viewed as complementary behavior which correlates with positive speaker perception (Pérez et al., 2016). It is doubtful whether negative convergence can be viewed favorably as well, as it indicates speakers becoming less and less similar over time. Nonetheless, we include negative convergence in our analysis as our focus in this paper is primarily on the occurrence and variation of behaviors rather than their positive or negative connotations.

5. Individual differences

5.1. Variation by role

We first explore whether speakers in the Deception Corpus vary their entrainment behavior based on the role they perform in the interaction (Interview**ER** or Interview**EE**). This is done with a series of 16 repeated measures t-tests, one for each of eight features and either entrainment measure. To reduce the probability of Type I error, we control for false discovery rate (FDR) using the procedure of Benjamini and Hochberg (1995). That is, for a given significance level α, we determine the largest integer k such that $p_k < k^*\alpha/n$, where p_k is the kth smallest p value and n is the number of tests. We then consider the k smallest p values significant. Table 1 lists those differences that reach significance ($\alpha = 0.05$, marked with “*”) or approach it ($\alpha = 0.1$, marked with “.”).

Table 2
Percentages of speakers entraining on at least one feature and details on their entrainment behavior, per corpus and entrainment measure.

	Deception (EE)		Deception (ER)		Fisher	
	conv.	synch.	conv.	synch.	conv.	synch.
Entraining speakers	47%	53%	42%	47%	39%	46%
Valence						
positive	42%	68%	40%	65%	37%	69%
negative	52%	18%	51%	26%	52%	22%
mixed	6%	14%	9%	9%	11%	9%
#Features						
1	68%	55%	73%	64%	74%	69%
2	25%	30%	19%	28%	23%	25%
3+	7%	15%	8%	8%	3%	6%
max.	4	5	4	4	4	4

Note that all of these results are for synchrony. None for convergence even approaches significance.

The table also lists effect sizes, measured by Cohen's *d*, with positive values indicating relatively stronger entrainment in the role of EE compared to ER, and negative values vice versa. That is, speakers change their speech rate more in synchrony with their interlocutor when they are interviewers than when they are interviewees and do the opposite for maximum intensity and NHR. It is unclear at this time what causes this behavior. All effects are small ($|d| < 0.5$) or very small ($|d| < 0.2$). Despite this, the differences motivate us to analyze the roles separately throughout the remainder of the paper.

5.2. Variation across speakers

There is considerable variation in convergence and synchrony behavior across the speakers in our corpora. Table 2 lists the percentages of speakers that exhibit significant convergence and synchrony, respectively, for at least one feature. For each measure and corpus, two fifths to one half of all speakers entrain. Synchrony, in each corpus, is slightly more prevalent than convergence. However, this difference is not significant according to χ^2-tests for either subcorpus of the Deception Corpus (EE: $\chi^2(1) = 1.9, p = 0.17$; ER: $\chi^2(1) = 0.84, p = 0.36$) or for the Fisher Corpus ($\chi^2(1) = 0.69, p = 0.41$). Table 2 also provides details on the valence and number of features entrained on, which are discussed below.

Looking at valence in Table 2, we note that in all corpora, many more speakers exhibit positive than negative synchrony. We again use χ^2-tests to assess significance. The differences are highly significant for all our corpora (EE: $\chi^2(1) = 63.3, p = 1.8\text{e-}15$; ER: $\chi^2(1) = 31.5, p = 2.0\text{e-}08$; Fisher $\chi^2(1) = 35.1, p = 3.1\text{e-}09$). That is, those speakers who significantly adapt their voice in immediate response to a change in their partner's voice tend to do so in the same rather than the opposite direction as the partner. Convergence, on the other hand, is more balanced between positive and negative entrainment, with slight trends towards negative convergence that are not significant (EE: $\chi^2(1) = 1.8, p = 0.18$; ER: $\chi^2(1) = 2.0, p = 0.16$; Fisher $\chi^2(1) = 2.3, p = 0.13$).

Between half and three quarters of the speakers who entrain at all do so on only one of the eight features we investigate here. Between 19 and 30% entrain on two features. The remaining speakers, between 3 and 15%, entrain on three or more features, up to a maximum of five. For instance, while 47% of speakers do not exhibit significant synchrony for any feature in the EE subcorpus, others entrain on five out of eight, illustrating the wide range of individual differences. Lastly, we note a tendency for speakers to entrain on more features for synchrony than for convergence. Repeated measures *t*-tests, comparing the number of features with significant synchrony and convergence, respectively, for each speaker, show that this result is significant for the EE subcorpus ($t(239) = 2.76, p = 0.006$) but not for the other corpora (ER: $t(239) = 1.5, p = 0.13$; Fisher: $t(179) = 1.6, p = 0.1$).

It is worth noting the similarity of results between the Deception subcorpora and the Fisher Corpus. We conduct a series of χ^2-tests – treating the corpora as three different categories – to check the nominal differences that do exist for significance. The number of speakers that exhibit entrainment on at least one feature does not differ across the corpora, neither for convergence ($\chi^2(2) = 2.58, p = 0.28$) nor for synchrony ($\chi^2(2) = 2.75, p = 0.25$). The differences between EE and ER in this regard are also not significant (both $p > 0.2$). Furthermore, there is no difference in the valence distribution across corpora. This is true whether a "0" valence for no entrainment is included in the test (conv.: $\chi^2(6) = 4.21, p = 0.65$; sync.: $\chi^2(6) = 6.69, p = 0.35$) or not (conv.: $\chi^2(4) = 1.65, p = 0.80$; sync.: $\chi^2(4) = 3.85, p = 0.43$). The same holds for the differences between EE and ER (all four $p > 0.14$). Lastly, there is no significant difference between the number of entrained features. As in Table 2, we group "3 and above" to avoid data sparsity issues. Again, we run tests for all corpora – including "0" (conv.: $\chi^2(6) = 5.29, p = 0.51$; sync.: $\chi^2(6) = 10.52, p = 0.10$) and excluding it (conv.: $\chi^2(4) = 2.81, p = 0.59$; sync.: $\chi^2(4) = 7.58, p = 0.11$) – as well as for EE and ER only (all four $p > 0.12$).

5.3. Discrete variation across speaker types

In this Subsection we continue to consider entrainment behavior in the aggregate and treat it as discrete, but analyze it by speaker type (see Section 2.3) to begin to explore the influence of gender and native language. Figs. 1–3 show the percentages of speakers of each type who entrain only positively, only negatively, mixed, or not at all, per corpus and measure. Substantial variation both in the percentages of entraining speakers and the valence is evident.

At the most basic level, we observe that even speakers of the same type exhibit different behaviors. Among FC-ME speakers in the EE subcorpus, for instance, about 25% of speakers converge only positively and only negatively, respectively, while 50% do not converge at all. Other speakers vary their behavior for different features, entraining positively for some and negatively for others. Over 30% of FC-ME speakers in the EE subcorpus do this for synchrony, for instance.

The overall percentage of entraining speakers also varies widely across speaker types, even within the same corpus and for the same measure. For instance, while only about 20% of FE-FE speakers in the EE subcorpus show significant positive or negative synchrony, almost 90% of FC-FE speakers do. Similarly, only 20% of FE-ME speakers in the ER subcorpus converge or diverge, compared to 60% of FC-FE speakers.

Furthermore, we continue to note a trend for synchrony to be more positive than negative for most speaker types, as observed in Section 5.2. We test for significance of this observation per corpus by treating the number of speakers of each type with only positive and only negative synchrony, respectively, as paired samples. The difference is, in fact, significant for both subcorpora of the Deception Corpus (EE: $t(15) = 6.0, p = 2.4\text{e-}05$; ER: $t(15) = 3.3, p = 0.005$) and for the Fisher Corpus ($t(11) = 5.9, p = 0.0001$). This result matches the one from Section 5.2, suggesting that the differences found there are distributed more or less evenly

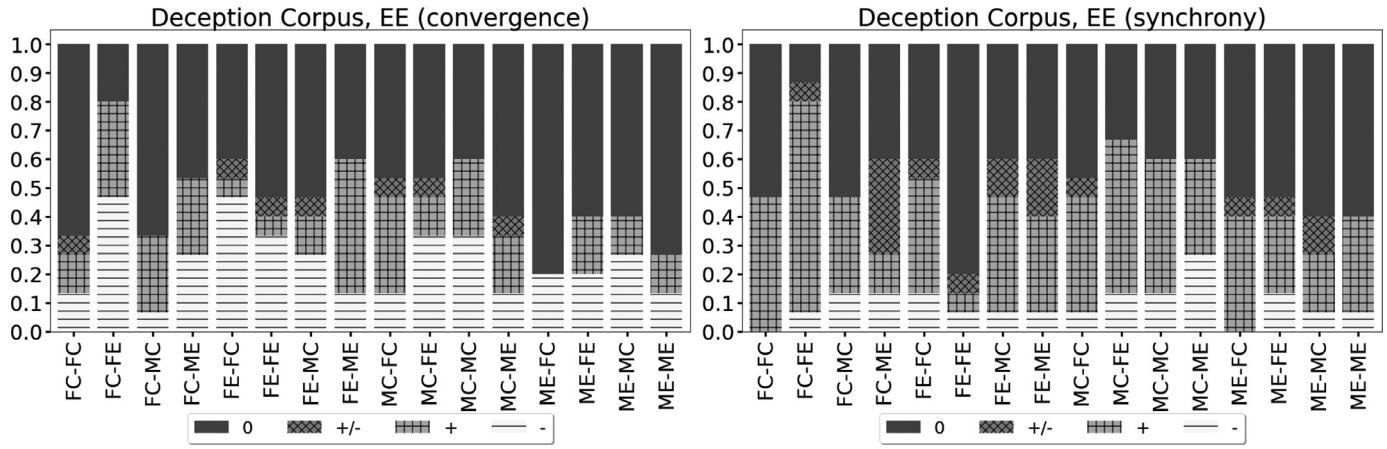

Fig. 1. Percentages of speakers who entrain negatively (-), positively (+), mixed (+/-) or not at all (0), per measure and speaker type, for the Deception Corpus (EE).

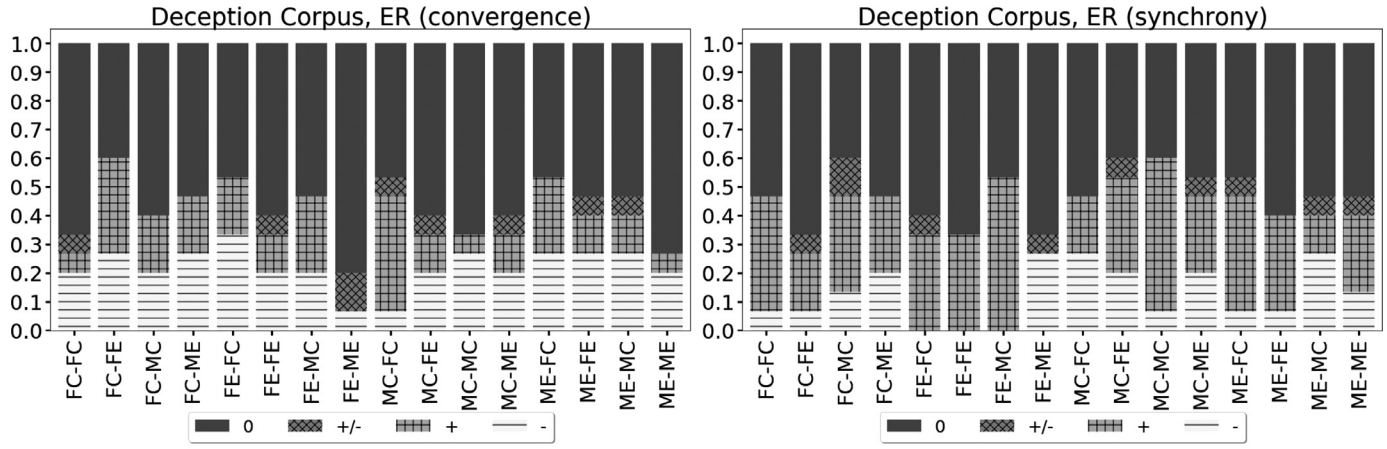

Fig. 2. Percentages of entraining speakers for the Deception Corpus (ER).

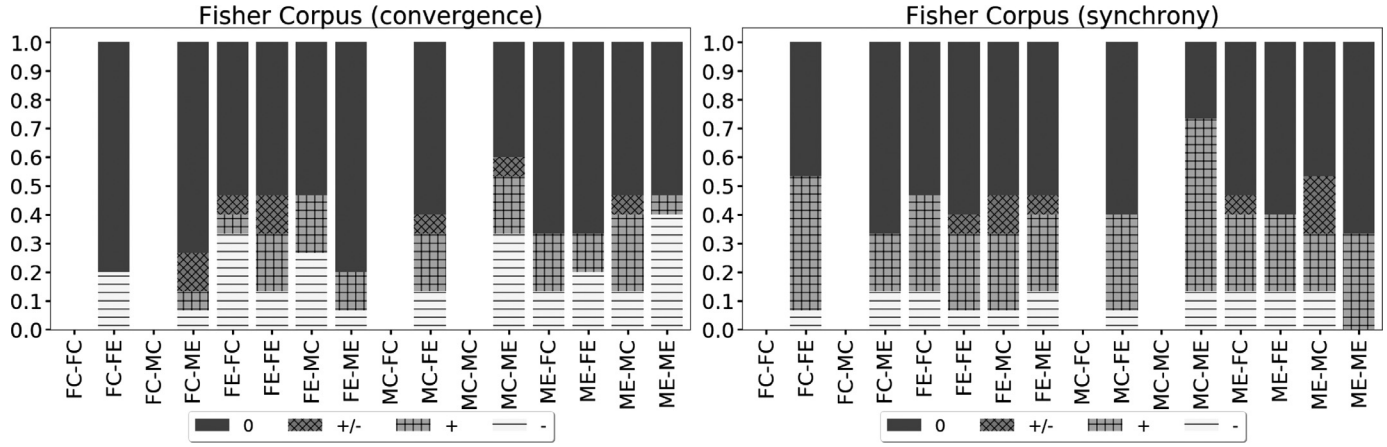

Fig. 3. Percentages of entraining speakers for the Fisher Corpus. Missing speaker types are left blank.

across speaker types rather than being caused by idiosyncratic behavior of individual speaker types.

The slight tendency of convergence to be more negative than positive, on the other hand, does not reach the level of significance for any corpus, with the lowest $p = 0.17$. We also use paired t-tests to compare the number of speakers exhibiting significant synchrony and convergence, respectively, for each speaker type. The tendency for synchrony to be more common than convergence is *not* significant when controlling for FDR (EE: $t(15) = 2.2, p = 0.04$; ER: $t(15) = 1.3, p = 0.23$; Fisher: $t(11) = 1.9, p = 0.08$).

Our data is too sparse to apply χ^2-tests to identify the influence of full speaker types consisting of all combinations of gender and native language. The use of χ^2 is discouraged unless the average expected count is at least 5.0 (Moore et al., 2009, p.532), which in our case would require at least 20 instances per speaker type while we only have 15. Instead, we test for the influence of gender and native language separately. For

Table 3
Results of one- and two-way ANOVAs with $p < 0.05$ for all measures, features, and corpora.

Corpus	Interaction	Measure	Feature	df	F	p	Tukey
Deception (EE)	gender	synchrony	max. pitch	3	2.82	0.040	MM < FF
Deception (EE)	language	synchrony	speech rate	3	3.84	0.010	CE < EC
Deception (EE)	language	synchrony	NHR	3	3.11	0.027	—
Deception (EE)	language	convergence	mean pitch	3	3.25	0.023	CE > EC
Deception (EE)	gender	convergence	shimmer	3	2.84	0.039	MM > FF
Deception (ER)	language	synchrony	jitter	3	3.00	0.031	EC > EE
Deception (ER)	gender	convergence	mean intensity	3	2.91	0.035	MM < MF
Deception (ER)	gender	convergence	shimmer	3	2.98	0.032	MM < FF
Deception (ER)	language	convergence	NHR	3	2.76	0.043	CE < EE
Fisher	gender	synchrony	mean intensity	3	3.06	0.030	MM > MF
Fisher	language:gender	synchrony	speech rate	6	2.35	0.034	—

Table 4
Statistics for the turn-final and turn-initial IPUs included in our analysis of the Deception and Fisher corpora, overall as well as per gender and native language. Duration is in seconds and numbers in parentheses are standard deviations.

	Deception			Fisher		
	syllables	duration	number of IPUs	syllables	duration	number of IPUs
All	5.26 (1.39)	1.21 (0.25)	147.27 (79.54)	11.91 (5.27)	1.97 (0.76)	56.57 (15.86)
Gender						
Female	5.28 (1.38)	1.25 (0.26)	138.22 (76.03)	12.26 (5.28)	2.07 (0.76)	55.42 (15.60)
Male	5.23 (1.40)	1.17 (0.24)	156.32 (82.04)	11.57 (5.25)	1.88 (0.75)	57.71 (16.11)
Native Lang.						
Chinese	4.77 (1.16)	1.21 (0.26)	167.23 (82.95)	8.60 (3.02)	1.70 (0.57)	61.80 (15.80)
English	5.75 (1.43)	1.22 (0.25)	127.31 (70.68)	13.02 (5.40)	2.06 (0.79)	54.82 (15.54)

Table 5
Statistics on the relevant IPUs of the Deception subcorpora.

	Deception (EE)			Deception (ER)		
	syllables	duration	number of IPUs	syllables	duration	number of IPUs
All	5.07 (1.56)	1.24 (0.28)	145.47 (79.56)	5.44 (1.16)	1.19 (0.23)	149.07 (79.62)
Gender						
Female	5.09 (1.59)	1.28 (0.29)	134.34 (73.95)	5.48 (1.10)	1.23 (0.23)	142.11 (78.11)
Male	5.06 (1.55)	1.20 (0.26)	156.60 (83.57)	5.40 (1.22)	1.14 (0.22)	156.04 (80.75)
Native Lang.						
Chinese	4.54 (1.26)	1.21 (0.27)	164.22 (84.87)	4.99 (1.00)	1.20 (0.24)	170.25 (81.15)
English	5.61 (1.66)	1.26 (0.28)	126.72 (69.22)	5.89 (1.14)	1.17 (0.21)	127.90 (72.33)

each gender type (FF, FM, MF, MM) and each native language type (EE, EC, CE, CC; the last one only for the Deception Corpus) we analyze the number of speakers exhibiting each type of valence (+, -, +/-, 0). Note that the overall number of speakers per type is 45 for gender pairs in the Fisher Corpus and 60 for all others. None of the tests shows significance, with the lowest $p = 0.11$. That is, we do not find any influence of gender or native language here on the valence of synchrony or convergence.

5.4. Continuous variation across speaker types

To detect more subtle variations in the strength and valence of the entrainment behavior of different speakers, we now treat our entrainment measures as continuous rather than discrete and analyze them for each feature individually instead of in the aggregate. To do so, we conduct three analyses of variance (ANOVAs) for each combination of corpus, measure, and feature. Gender type, native language type (both one-way ANOVAs), and full speaker type (two-way ANOVA), respectively, are the independent variables, the values of the entrainment measures are the dependent variables.

Table 3 lists all results with $p < 0.05$. None of them reach the level of significance when controlling for FDR to account for the high number of tests (144). Nonetheless, we also apply Tukey's test post-hoc for each of these ANOVAs. The last column of Table 3 contains the pairwise differences with $p < 0.05$, at most one and in two cases none.

Keeping in mind that the results are not significant, we note that they are also not consistent, either for gender or native language type. For instance, male pairs tend to entrain more than female and mixed pairs on some features but less on others, even within the same corpus (Deception (EE)). This suggests that trends in entrainment behavior, when they are found, should not be assumed to be consistent for different features.

Following the work of Pérez et al. (2016), we also run ANOVAs for the absolute values of the synchrony measure for each feature. Only five of these additional ANOVAs yield $p < 0.05$, 4 of them with $p > 0.025$, the lowest $p = 0.004$. This is far from significant when correcting for 72 tests. We conclude that gender and native language cannot directly explain the variation in entrainment behavior which we observe.

6. Discussion and conclusion

We present a systematic analysis of variation in two types of local, acoustic-prosodic entrainment based on two large corpora. Our work shows that, while entrainment behavior varies greatly, this variation cannot be directly attributed to gender, contrary to the conclusions drawn by previous studies. We also investigate the influence of native language on entrainment and find that it, too, does not explain differences in behavior, either on its own or in combination with gender. In fact, the only speaker characteristic that we do find to predict some differences in the behavior of the *same* speakers is whether they act as interviewee or interviewer.

Regarding overall trends in our data, we find that about half of all speakers exhibit a form of synchrony and a similar number converge or diverge on at least one feature. This is roughly comparable with the

Table 6
Results of *t*-tests for various differences between speaker groups in our corpora. Positive *t*-statistics indicate a higher average with regard to the criterion for group 1 than group 2, and vice versa. *p* values up to 0.044 are significant after accounting for false discovery rate (Benjamini and Hochberg, 1995). Non-significant *p* values are marked in the rightmost column.

Corpus	Group 1	Group 2	Criterion	df	*t*	*p*	n.s.
Fisher	Chinese	English	syllables	238	−6.03	6.4e-09	
Deception	Chinese	English	syllables	598	−9.25	3.8e-19	
Deception (EE)	Chinese	English	syllables	298	−6.27	1.3e-09	
Deception (ER)	Chinese	English	syllables	298	−7.26	3.5e-12	
Fisher	Chinese	English	number of IPUs	238	3.00	0.0030	
Deception	Chinese	English	number of IPUs	598	6.35	4.4e-10	
Deception (EE)	Chinese	English	number of IPUs	298	4.19	3.6e-05	
Deception (ER)	Chinese	English	number of IPUs	298	4.77	2.9e-06	
Fisher	Chinese	English	duration	238	−3.25	0.0013	
Deception (EE)	Chinese	English	duration	298	−1.66	0.098	x
Fisher	Chinese	English	speech rate	238	−6.07	5.1e-09	
Deception	Chinese	English	speech rate	598	−11.1	3.2e-26	
Deception (EE)	Chinese	English	speech rate	298	−7.31	2.5e-12	
Deception (ER)	Chinese	English	speech rate	298	−10.7	1.1e-22	
Fisher	Female	Male	number of IPUs	238	−1.12	0.27	x
Deception	Female	Male	number of IPUs	598	−2.80	0.0052	
Deception (EE)	Female	Male	number of IPUs	298	−2.44	0.0015	
Deception (ER)	Female	Male	number of IPUs	298	−1.52	0.12	x
Fisher	Female	Male	duration	238	2.02	0.044	
Deception	Female	Male	duration	598	3.89	1.1e-04	
Deception (EE)	Female	Male	duration	298	2.34	0.020	
Deception (ER)	Female	Male	duration	298	3.31	0.0011	
Fisher	Female	Male	speech rate	238	−2.56	0.011	
Deception	Female	Male	speech rate	598	−4.57	6.0e-06	
Deception (EE)	Female	Male	speech rate	298	−3.94	1.0e-04	
Deception (ER)	Female	Male	speech rate	298	−3.37	8.6e-04	
Deception	EE	ER	syllables	598	−3.27	0.0011	
Deception	EE	ER	duration	598	2.55	0.011	
Deception	EE	ER	speech rate	598	−12.1	2.2e-30	

findings of Levitan et al. (2015a) for English. However, while they found synchrony to be mostly negative, it is predominantly positive in our corpora. They also found only positive convergence while in our data convergence and divergence are about equally common. These differences in findings suggest that the conversation context – collaborative, task-oriented dialogues versus deceptive interviews and spontaneous speech, respectively – influences the valence of entrainment. In addition, we find that synchrony occurs for more features than convergence, significantly so for interviewees in the Deception Corpus, and that the number of features entrained on varies widely between speakers.

Gender alone does not explain the differences in entrainment we find in our data, neither for its rate of occurrence, nor its strength, nor its valence. This finding is unlike those from many previous studies which did report gender differences. It does, however, accord with the results of Weise and Levitan (2018), who found that the overall entrainment behavior of speakers does not form clusters based on gender. Their work was based on the Switchboard Corpus, which is very similar to the Fisher Corpus analyzed here.

We also find no significant differences between native and non-native English speakers. This is despite the signs of greater cognitive load we find among non-native speakers (see Section 2.4) and the decrease in entrainment this predicts (Abel and Babel, 2017). In particular, our results neither match those of Kim et al. (2011) nor those of (Kim, 2012) and Lewandowski and Nygaard (2018). The most notable difference between those studies and ours is that their analyses were based primarily on perceptual rather than acoustic measures. Only Lewandowski and Nygaard (2018) considered acoustic measures at all and found no consistent difference for them based on model talker accent, unlike for perceived similarity. Another potential explanation for the lack of differences based on native language in our data is dialect. Kim et al. (2011) found that mismatches in regional dialect among pairs of native speakers of English were enough to eliminate differences in entrainment compared to pairs with a non-native speaker. The Fisher Corpus, by design, contains a wide variety of dialects and many of the speaker pairs in our selection were mismatched with regard to dialect. For the Deception Corpus this information was not tracked. However, the Columbia University student body is geographically diverse so that many of those speaker pairs may have had a different dialect. On the other hand, we found substantial evidence of entrainment among all speaker groups while Kim et al. found none among speakers mismatched in dialect or native language. That is, even if dialects were mismatched in our data, this may have had less impact than in their data and thus might not explain the difference in findings. Finally, we note that language proficiency of non-native speakers also does not influence entrainment in our data.

We conclude that entrainment behavior is not generally influenced by gender, native language, or their interaction alone. Previous results have detected an influence of other factors such as liking (Lee et al., 2010; Lubold and Pon-Barry, 2014) or power (Danescu-Niculescu-Mizil et al., 2012) on entrainment, which are also predicted by theoretical accounts of the phenomenon (Giles et al., 1991). In light of this, we propose as a hypothesis for further study that gender merely mediates more complex interactions between power, sociocultural norms, liking, personality, and conversation context, and that this influence may vary between linguistic features.

Lastly, it is worth noting the remarkable similarity between our results for the Deception subcorpora and the Fisher Corpus (Table 2 and Section 5.2). We find no significant differences in the rate of occurrence or the valence of entrainment, nor for the number of features entrained on. So while there are slight differences in individual features' local similarity based on the truthfulness of the responses (Levitan et al., 2018), and while we find differences between the speaker roles for individual features' synchrony (Table 1), in the aggregate and with regard to synchrony and convergence, speakers entrain very similarly in the context of deceptive interviews and spontaneous speech.

In our future work, we intend to analyze additional corpora and meta-data, e.g., for speaker personality, for the influence of gender under various circumstances to clarify the interaction with other factors.

Since one of the most statistically significant results in this paper is the difference between behavior of the same speakers in different roles, it would also be fascinating to have the same speakers interact in at least two different settings, such as spontaneous versus task-oriented speech, to investigate our hypothesis that gender has varying influence on entrainment depending on conversation context. Pardo et al. (2018) did analyze entrainment for the same speakers in two contexts and found that the correlation of the degrees of entrainment per speaker across settings was stronger for males than for females. However, this was for interactive and non-interactive settings with the model talkers in the shadowing part being different from the interlocutors in the interactive part. Experiments with the exact same pairs in different contexts should also be conducted in the future.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by the Air Force Office of Scientific Research (no. FA9550-11-1-0120) and by the National Science Foundation (no. 1845710).

Appendix A. Statistical analysis of length and number of IPUs

This section analyzes the average length – number of syllables and duration in seconds – and number of IPUs per speaker in our data, overall and by speaker group, i.e., based on gender and native language. All numbers refer only to those IPUs included in the analysis, i.e., turn-final and turn-initial IPUs without overlaps. We compute averages for each speaker and then average those values across all speakers in a speaker group. Table 4 lists these statistics for the Deception and Fisher corpora, Table 5 for the Deception subcorpora of interview**ER**s and interview**EE**s.

We note that the relatively high standard deviations for the number of IPUs in the Deception Corpus are not due to an imbalance in the number of conversations that were used per speaker (see also Section 2.3). For each speaker included in the analysis, we used all relevant IPUs from both parts of the conversation. The number of exchanges needed to answer all biographical questions simply varied across subject pairs. The lower standard deviations in the Fisher Corpus result from the fact that those conversations were timed to all be roughly the same length of 10 min.

There are numerous apparent differences between speakers of different groups in our data. We run *t*-tests to compare the speaker averages for many of these differences and list the results in Table 6. Chinese native speakers conversing in English use fewer syllables per IPU and more IPUs per conversation than English native speakers in all of our corpora. In the Fisher corpus, they also speak in IPUs of shorter duration, while this difference is not significant in the EE Deception subcorpus. Non-native speech rate is significantly lower in all of our corpora. Female speakers in the Deception Corpus overall and in the EE subcorpus use fewer IPUs per conversation, while that same tendency is not significant in the other corpora. Females also speak in longer IPUs (by duration) and more slowly in all corpora. Lastly, interviewees use fewer syllables per IPU but those IPUs last longer, resulting in lower speech rate than that of the interviewers.

References

Abel, J., Babel, M., 2017. Cognitive load reduces perceived linguistic convergence between dyads. Lang. Speech 60 (3), 479–502. doi:10.1177/0023830916665652.

Babel, M., Bulatov, D., 2012. The role of fundamental frequency in phonetic accommodation. Lang. Speech 55 (2), 231–248. doi:10.1177/0023830911417695.

Babel, M., McGuire, G., Walters, S., Nicholls, A., 2014. Novelty and social preference in phonetic accommodation. Lab. Phonol. 5 (1), 123–150. doi:10.1515/lp-2014-0006.

Bailly, G., Martin, A., 2014. Assessing objective characterizations of phonetic convergence. In: Interspeech 2014, pp. P–19.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57 (1), 289–300.

Boersma, P., Weenink, D., 2018. PRAAT, a system for doing phonetics by computer. http://www.fon.hum.uva.nl/praat/.

Brennan, S.E., Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation.. J. Exp. Psychol.: Learn.MemoryCognit. 22 (6), 1482–1493. doi:10.1037/0278-7393.22.6.1482.

Chartrand, T.L., Bargh, J.A., 1999. The chameleon effect: the perception-Behavior link and social interaction. J. Pers. Soc. Psychol. 76 (6), 893–910.

Cieri, C., Miller, D., Walker, K., 2004. The Fisher corpus: a resource for the next generations of speech-to-Text. LREC 4, 69–71. 10.1.1.61.8327

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., Kleinberg, J.M., 2012. Echoes of power: language effects and power differences in social interaction. In: 21st International Conference on World Wide Web, pp. 699–708. doi:10.1145/2187836.2187931.

van Engen, K.J., Baese-Berk, M., Baker, R.E., Choi, A., Kim, M., Bradlow, A.R., 2010. The wildcat corpus of native- and foreign-accented english: communicative efficiency across conversational dyads with varying language alignment profiles. Lang. Speech 53 (4), 510–540. doi:10.1177/0023830910372495.

Giles, H., Coupland, N., Coupland, J., 1991. Accommodation Theory: Communication, Context, and Consequence. In: Contexts of accommodation: Developments in applied sociolinguistics. Cambridge University Press, pp. 1–68. doi:10.1017/CBO9780511663673.001.

Goldinger, S.D., 1998. Echoes of echoes? an episodic theory of lexical access. Psychol. Rev. 105 (2), 251–279. doi:10.1037/0033-295X.105.2.251.

Gravano, A., Beňuš, Š., Levitan, R., Hirschberg, J., 2014. Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In: Spoken Language Technology (SLT), 2014 IEEE Workshop on, pp. 578–583.

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

Kim, M., 2012. Phonetic accommodation after passive exposure to native and nonnative speech. Northwestern University.

Kim, M., Horton, W.S., Bradlow, A.R., 2011. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. Lab. Phonol. 2 (1), 125–156. doi:10.1515/labphon.2011.004.

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P.G., Narayanan, S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, pp. 793–796.

Lee, C.C., Katsamanis, A., Black, M.P., Baucom, B.R., Georgiou, P.G., Narayanan, S.S., 2011. An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3101–3104.

Levitan, R., Beňuš, Š., Gravano, A., Hirschberg, J., 2015. Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese : A cross- linguistic comparison. In: SIGdial, pp. 325–334. doi:10.1016/j.knosys.2014.05.020.

Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, pp. 3081–3084.

Levitan, R., Willson, L., Gravano, A., Beňuš, Š., Hirschberg, J., Nenkova, A., 2012. Acoustic-prosodic entrainment and social behavior. In: NAACL HLT, pp. 11–19.

Levitan, S.I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., Rosenberg, A., 2015. Cross-Cultural production and detection of deception from speech. Proceedings of the 2015 ACM Workshop on Multimodal Deception Detection - WMDD '15 1–8. doi:10.1145/2823465.2823468.

Levitan, S.I., Xiang, J., Hirschberg, J., 2018. Acoustic-prosodic and lexical entrainment in deceptive dialogue. In: Speech Prosody, pp. 532–536. doi:10.21437/SpeechProsody.2018-108.

Lewandowski, E.M., Nygaard, L.C., 2018. Vocal alignment to native and non-native speakers of english. J. Acoust. Soc. Am. 144 (2), 620–633. doi:10.1121/1.5038567.

Loy, J. E., Smith, K., 2019. Syntactic adaptation may depend on perceived linguistic knowledge: native english speakers differentially adapt to native and nonnative confederates in dialogue.

Lubold, N., Pon-Barry, H., 2014. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In: Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge, pp. 5–12. doi:10.1145/2666633.2666635.

Manson, J.H., Bryant, G.A., Gervais, M.M., Kline, M.A., 2013. Convergence of speech rate in conversation predicts cooperation. Evol. Human Behav. 34 (6), 419–426. doi:10.1016/j.evolhumbehav.2013.08.001.

Moore, D.S., McCabe, G.P., Craig, B.A., 2009. Introduction to the Practice of Statistics, 6th W.H. Freeman and Company.

Namy, L.L., Nygaard, L.C., Sauerteig, D., 2002. Gender differences in vocal accommodation: the role of perception. J. Lang. Soc. Psychol. 21 (4), 422–432. doi:10.1177/026192702237958.

Nasir, M., Baucom, B., Narayanan, S., Georgiou, P., 2018. Towards an Unsupervised Entrainment Distance in Conversational Speech using Deep Neural Networks. In: Interspeech, pp. 3423–3427. doi:10.21437/Interspeech.2018-1395.

Nenkova, A., Gravano, A., Hirschberg, J., 2008. High frequency word entrainment in spoken dialogue. In: ACL HLT, pp. 169–172.

Pardo, J.S., 2006. On phonetic convergence during conversational interaction.. J. Acoust. Soc. Am. 119 (4), 2382–2393. doi:10.1121/1.2178720.

Pardo, J.S., Cajori Jay, I., Krauss, R.M., 2010. Conversational role influences speech imitation. Atten. Percept. Psychophys. 72 (8), 2254–2264. doi:10.3758/APP.72.8.2254.

Pardo, J.S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E., 2013. Phonetic convergence in shadowed speech: the relation between acoustic and perceptual measures. J. Mem. Lang. 69 (3), 195–198. doi:10.1016/j.jml.2013.06.002.

Pardo, J.S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., Ward, M., 2018. A comparison of phonetic convergence in conversational interaction and speech shadowing. J. Phon. 69, 1–11. doi:10.1016/j.wocn.2018.04.001.

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. In: INTERSPEECH, pp. 1270–1274. doi:10.21437/Interspeech.2016-587.

Reichel, U.D., Beňuš, Š., Mády, K., 2018. Entrainment profiles: comparison by gender, role, and feature set. Speech Commun. 100, 46–57. doi:10.1016/j.specom.2018.04.009.

Reitter, D., Moore, J.D., 2007. Predicting success in dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 808–815.

Reitter, D., Moore, J.D., Keller, F., 2006. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In: CogSci, pp. 685–690.

Smiljanic, R., Bradlow, A.R., 2009. Speaking and hearing clearly: talker and listener factors in speaking style changes.. Linguist. Lang. Compass 3 (1), 236–264.

Street, R.L., 1984. Speech convergence and speech evaluation in fact-finding interviews. Hum. Commun. Res. 11 (2), 139–169. doi:10.1111/j.1468-2958.1984.tb00043.x.

Ward, A., Litman, D., 2007. Automatically measuring lexical and acoustic / prosodic convergence in tutorial dialog corpora. In: Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE), pp. 57–60.

Weise, A., Levitan, R., 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In: NAACL HLT, pp. 297–302.

Xia, Z., Levitan, R., Hirschberg, J., 2014. Prosodic entrainment in mandarin and english : a cross-linguistic comparison. In: Speech Prosody, pp. 65–69.