Journal of Phonetics 88 (2021) 101074

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue

Rachel Ostrand [a,*], Eleanor Chodroff [b]

[a] IBM Research, Yorktown Heights, NY, USA
[b] Department of Language and Linguistic Science, University of York, Heslington, York, United Kingdom

ARTICLE INFO

Article history:
Received 5 February 2020
Received in revised form 10 May 2021
Accepted 29 May 2021

Keywords:
Alignment
Adaptation
Phonetics
Syntax
Partner-specific
Machine learning

ABSTRACT

During conversation, speakers modulate characteristics of their production to match their interlocutors' characteristics. This behavior is known as *alignment*. Speakers align at many linguistic levels, including the syntactic, lexical, and phonetic levels. As a result, alignment is often treated as a unitary phenomenon, in which evidence of alignment on one feature is cast as alignment of the entire linguistic level. This experiment investigates whether alignment can occur at some levels but not others, and on some features but not others, within a given dialogue. Participants interacted with two experimenters with highly contrasting acoustic-phonetic and syntactic profiles. The experimenters each described sets of pictures using a consistent acoustic-phonetic and syntactic profile; the participants then described new pictures to each experimenter individually. Alignment was measured as the degree to which subjects matched their current listener's speech (vs. their non-listener's) on each of several individual acoustic-phonetic and syntactic features. Additionally, a holistic measure of phonetic alignment was assessed using 323 acoustic-phonetic features analyzed jointly in a machine learning classifier. Although participants did not align on several individual spectral-phonetic or syntactic features, they did align on individual temporal-phonetic features and as measured by the holistic acoustic-phonetic profile. Thus, alignment can simultaneously occur at some levels but not others within a given dialogue, and is not a single phenomenon but rather a constellation of loosely-related effects. These findings suggest that the mechanism underlying alignment is not a primitive, automatic priming mechanism but rather guided by communicative or social factors.

1. Introduction

Language is a flexible medium. Speakers have wide latitude to say what they want to say in many different ways — they can use one grammatical structure or another, use a particular word or its synonym, speak slowly or quickly, at a higher pitch or a lower pitch, or louder or softer. As such, one person's speech production is slightly different from conversation to conversation on many dimensions. Although many of these linguistic "decisions" could be made consciously, most often they are not, and instead reflect subtle tunings performed by the language processing system in response to the speaker's current linguistic environment. Many external factors can influence these production adjustments, and an important one is the identity of, and prior linguistic experience with, the speaker's current listener. When engaged in dialogue, speakers modulate characteristics of their speech to draw closer to the corresponding characteristics of their listener's speech. This process is called *alignment*[1].

Alignment has been demonstrated, to some degree, at every linguistic level. Speakers align their representations of the spatial environment to match their partner's description (Garrod & Anderson, 1987). They align their syntax to match that of an immediately- or recently-preceding sentence (Branigan, Pickering, & Cleland, 2000; Branigan, Pickering, McLean, & Cleland, 2007; Cleland & Pickering, 2003; Cowan, Branigan, Obregón, Bugis, & Beale, 2015; Haywood, Pickering, & Branigan, 2005; Levelt & Kelter, 1982; Reitter & Moore, 2014) and to the overall syntactic bias of the linguistic environment (Kaschak, 2007; Kaschak, Loney, & Borreggine, 2006; Kaschak, Kutta, & Schatschneider, 2011; Ostrand & Ferreira, 2019), although not to their particular listener's

* Corresponding author.
E-mail address: ostrand.rachel@gmail.com (R. Ostrand).

[1] It is also variously referred to as *entrainment*, *accommodation*, or *convergence*.

https://doi.org/10.1016/j.wocn.2021.101074
0095-4470/

syntactic preferences (Ostrand & Ferreira, 2019). They align on lexical choice, producing the same word as they just comprehended (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Horton & Gerrig, 2005; Rosenthal-von der Pütten, Wiering, & Krämer, 2013; Yoon & Brown-Schmidt, 2014), including uniquely to the particular interlocutor who just produced that word. Speakers also align on phonetic features, including their partner's vowel formants (Pardo, Jay, & Krauss, 2010; Pardo, Gibbons, Suppes, & Krauss, 2012; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013) and voice onset time (Nielsen, 2011; Shockley, Sabadini, & Fowler, 2004). They align on temporal characteristics, such as speech and articulation rate (Bell, Gustafson, & Heldner, 2003; Bonin, de Looze, Ghosh, Gilmartin, Vogel, Polychroniou, Salamin, Vinciarelli, & Campbell, 2013; Schultz et al., 2016; Staum Casasanto, Jasmin, & Casasanto, 2010; Street, 1984; Webb, 1969), intra-speaker pause duration (Cappella & Planalp, 1981; de Looze, Oertel, Rauzy, & Campbell, 2011; de Looze & Rauzy, 2011; Edlund, Heldner, & Hirschberg, 2009; Gregory & Hoyt, 1982; ten Bosch, Oostdijk, & de Ruiter, 2004), inter-speaker pause duration (Suzuki & Katagiri, 2007), and turn duration (Matarazzo, Weitman, Saslow, & Wiens, 1963). Speakers also align on low-level acoustic features, including pitch (Babel & Bulatov, 2011; Borrie, Lubold, & Pon-Barry, 2015; Rahimi, Kumar, Litman, Paletz, & Yu, 2017), fundamental frequency (Bonin et al., 2013; Gregory & Webster, 1996), jitter (Borrie et al., 2015; Rahimi et al., 2017), intensity (Bonin et al., 2013; Borrie et al., 2015; Natale, 1975; Rahimi et al., 2017; Suzuki & Katagiri, 2007), and vowel spectra (Gregory, Webster, & Huang, 1993; Gregory, Dagan, & Webster, 1997; Pardo et al., 2013). Speakers even align paralinguistically to their partners' gestures (Bergmann & Kopp, 2012; Holler & Wilkin, 2011; Kimbara, 2008).

Implicit (and sometimes explicit) in much of this prior work, however, is the treatment of alignment as a single, unitary phenomenon. Thus, when alignment is or is not demonstrated on an individual linguistic feature, it is treated as diagnostic of overall, holistic alignment behavior across the linguistic level or even the entire language production system. Although this assumption has substantial implications for the underlying mechanisms which drive alignment, it is largely untested. The present work investigates two related questions regarding the multidimensionality of alignment within the same speech. First, does alignment at one linguistic level necessarily engender alignment at others? Second, does alignment on one feature within a linguistic level entail alignment on other features within that level, and can within-level features be assessed jointly to produce a holistic measure of alignment for that level?

1.1. Alignment as a unitary phenomenon

Most studies of alignment investigate its presence using just a single or a few dependent variables. They treat participants' behavior on those measures as diagnostic of broader alignment behavior of the containing linguistic level, or even more broadly, of alignment of the entire language production system. That is, a study which measures alignment on the individual dimension of pitch may claim that this is evidence of broad "phonetic alignment" or even "alignment as an overall linguistic process", as opposed to evidence for "alignment on pitch." For example, lexical alignment has variously been claimed by demonstrating that participants match their interlocutor's usage of a preferred vs. dispreferred object name (e.g., "couch" vs. "sofa": Branigan, Pickering, Pearson, McLean, & Brown, 2011); a label or expression for an unknown object (e.g., Clark & Wilkes-Gibbs, 1986); a subordinate vs. basic-level object name (e.g., "loafer" vs. "shoe": Brennan & Clark, 1996); and the relative frequency of production of the most frequent words across the entire conversation (e.g., Nenkova, Gravano, & Hirschberg, 2008). Similarly, phonetic alignment has been claimed by demonstrating that talkers shift their production to match their interlocutor's or a model talker's vocal intensity (e.g., Natale, 1975); vowel spectra (e.g., Babel, 2012); fundamental frequency (e.g., Babel & Bulatov, 2011); and voice onset time (e.g., Nielsen, 2011), among many others. Although these studies do all demonstrate the presence of alignment on the particular feature that was studied, it is not necessarily the case that such behavior generalizes to a wider cognitive behavior, either evidencing alignment of all representations at that linguistic level (e.g., alignment on pitch entails alignment on fundamental frequency) or alignment of representations at other linguistic levels (e.g., alignment on pitch entails alignment on usage of an object's dispreferred name).

1.2. Does alignment at one level necessarily lead to alignment at others?

More explicitly, an influential theory predicts that alignment occurs as a unitary phenomenon across linguistic levels as a communicative strategy, with alignment at one level engendering alignment at all other levels. According to the *Interactive Alignment Model* (Pickering & Garrod, 2004), dialogue is made successful on the basis of interlocutors aligning their *situation models*, representations of the linguistic and contextual environment in which they are interacting. If a speaker and a listener have matching representations, the dialogue process can be more efficient because the speaker does not need to model their own representation as well as that of their listener. Instead, for example, to know how the listener expects an object to be referenced, the speaker needs only to query their own model to determine the lexical label that they would use themselves, and that should be the label that the listener expects. This theory posits that alignment of these high-level situation models is achieved via alignment of lower-level linguistic representations, which occurs as a result of a "primitive and resource-free priming mechanism" (p. 172). As a result, because different linguistic levels are interconnected, alignment at one linguistic level should induce alignment at other levels as well. Importantly, because the proposed priming mechanism which causes alignment is resource-free, alignment automatically cascades between linguistic levels by strengthening inter-level links between, for example, the currently-activated lexical item and syntactic structure. The theory therefore makes the prediction that when interlocutors align to each other, they should do so on *all* of their linguistic levels at the same time.

As evidence that linguistic levels align in tandem, the IAM cites effects of the *lexical boost* in dialogue. The lexical boost

is the effect of increased alignment at the syntactic level as a result of alignment at the lexical level. In a typical experimental design, an experimenter or confederate produces a picture description using a particular syntactic structure, and then the participant describes another picture immediately afterward. If the participant's description matches the syntactic structure of the confederate's description, that demonstrates alignment or priming. The lexical boost occurs when the participant is instructed to use a particular verb in their description. When the verb provided to the participant matches the immediately-preceding sentence's verb, the participant is more likely to syntactically align, compared to sentences where the participant's provided verb differs from the preceding verb (Branigan et al., 2000; see also Cleland & Pickering, 2003). According to the IAM, these results demonstrate that speakers align at multiple linguistic levels in tandem. However, although the lexical boost does provide evidence that interlocutors *can* align their representations at *some* different linguistic levels jointly, it does not mean that they *must* do so at *all* levels.

In addition, the lexical boost does not address whether linguistic levels necessarily align together when the participant's production is not fixed by the experiment. In fact, this component of the theory has largely not been tested. Lexical boost effects show that when speakers are forcibly aligned at one level (lexical, by being told which verb to use), then they also align at a different level (syntactic). However, it is largely unknown whether linguistic levels vary together when a speaker can freely select what to produce at each level.

One study did consider this relationship using spontaneous dialogues from two corpora of naturalistic speech (Weise & Levitan, 2018). The authors measured three features at each of the lexical and acoustic-prosodic linguistic levels, and then looked for a correlation in alignment between different features. That is, they investigated whether interlocutors' degree of alignment on one feature was related to those same interlocutors' degree of alignment on a different feature. However, there was no relationship on degree of alignment across features. The authors concluded that alignment — rather than a unitary, behemoth behavior which occurs or does not occur across levels and features in unison — is rather an umbrella of many different loosely-linked behaviors, and perhaps even generated independently by different cognitive mechanisms or for different communicative or social reasons.

Although an important first step, this study leaves open some relevant questions. First, as the authors note, they tested just a handful of features which may not cover the full scope of dimensions on which interlocutors might align. Second, due to the design of the speech corpora, the study was not equipped to disentangle partner-specific from context-specific alignment. The authors assessed the similarity between a given participant's speech and their true partner's speech vs. the similarity between a given participant's speech and all of the non-partners with whom that participant never spoke; if similarity to the true partner was greater than that to the non-partners, that was evidence for alignment. However, such a design cannot determine whether the participant modulated their speech specifically to match their *listener* (and would have reverted to their baseline when faced with a new partner), or whether they modulated their speech to match the *recent* linguistic context — all of which was provided by that same partner.

Cohen Priva and Sanker (2018) conducted a similar study, investigating the correlation in convergence between a few acoustic-phonetic features within the same dialogue. This study also found no cross-feature correlation between degree of alignment, except on two highly related measures (descriptors of F0). Similarly, Rahimi et al. (2017) investigated the correlation of alignment on four acoustic and five lexical features, and various statistical descriptors of each. Their results were quite divergent: some correlations were not significant, and of those that reached significance, some had a positive relationship and others had a negative relationship. As there was not a clear mechanism to predict why certain features should be correlated and others not, the degree of alignment shared across linguistic levels remains a question to be further explored.

The present work investigates alignment at different linguistic levels in a conversational interaction where the participant's production is not fixed at any linguistic level. (That is, participants are never instructed to use a particular linguistic feature, unlike lexical boost studies where the participant is told to produce a sentence with a given verb.) We focus on phonetic alignment at the segmental and suprasegmental levels, and syntactic alignment. In particular, we investigate whether speakers align at each level in a partner-specific manner — that is, whether they converge upon the particular linguistic profile of their current listener (as opposed to a linguistic profile made up of their aggregated, across-partner recent exposure), when interacting with two listeners who have different profiles. The Interactive Alignment Model states that the purpose of alignment is to reduce the necessity of modeling the listener's linguistic and situational representation; other theories, such as the Communication Accommodation Theory (Giles, Coupland, & Coupland, 1991), suggest that alignment serves a social purpose and acts as a signal of liking or approval. Under both theories, the starkest situation in which converging on a partner's individual linguistic profile would be necessary is when interacting with multiple partners who have differing linguistic or social profiles; in such a case, it should be important to represent (and thus align to) each partner's linguistic idiosyncrasies individually. Such individually-tailored alignment is referred to as *partner-specific alignment*. An experimental design in which participants interact with two conversational partners who produce differing linguistic distributions allows for the clearest test of whether partner-specific alignment occurs at multiple linguistic levels in tandem, as it allows for the comparison of the participant's speech when speaking to one partner against the participant's speech when speaking to the other partner.

We selected the syntactic and acoustic-phonetic levels to assess alignment within the same interaction because it seemed possible that they could show differing patterns. On the one hand, prior work has shown that speakers do not align their syntax in a partner-specific manner (Ostrand & Ferreira, 2019), but do engage in other forms of syntactic alignment by modulating their syntax to match other types of linguistic context (e.g., Branigan et al., 2000; Gruberg, Ostrand, Momma, & Ferreira, 2019; Kaschak, 2007). On the other hand, many studies (as noted above) have investigated alignment on various phonetic features, including in spontaneous speech tasks, finding positive evidence of alignment. Relatedly, there

are a wide range of acoustic-phonetic dimensions on which speech can be assessed, which raises the possibility of differing alignment behavior among features even within the same linguistic level. This last point informs the second goal of this paper.

1.3. Does alignment on one feature necessarily mean alignment on other features at the same level?

As noted above, a substantial amount of research has investigated alignment in the acoustic domain. However, the acoustics of speech is highly multidimensional, and there are dozens of different individual phonological and phonetic features that have been measured across studies of vocal accommodation, with no real standard as to which feature should be measured when. This leads to some potential problems, as noted in Pardo (2013) and Pardo et al. (2018), among others. First, this inconsistency on which features to measure produces conflicting results from research studying the same linguistic process, as one study might find alignment on one phonetic feature while another finds no alignment on a different feature, even in the same task. In fact, the majority of prior vocal accommodation studies have measured just a single acoustic-phonetic feature; even those studies which have tested multiple features generally measure fewer than five, which does not cover the broad range of possible acoustic-phonetic properties on which speakers might align (cf. Lee et al., 2014; Levitan & Hirschberg, 2011; Mukherjee, D'Ausilio, Nguyen, Fadiga, & Badino, 2017; Pardo et al., 2010, 2013; Rahimi et al., 2017; Weise & Levitan, 2018; for an overview, see Pardo, Urmanche, Wilman, & Wiener, 2017). As a result, the conclusion of the presence or absence of alignment is dependent on the particular feature(s) under investigation; a study which measured alignment on F0 could draw a completely different conclusion as the identical study which measured vowel duration. Second, such variability opens the door to "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) on choosing which features or results to test or report, potentially leading to false positive effects. Roettger (2019) notes that this is a particular concern with phonetic research due to its multidimensionality and high degree of covariance between individual features. Finally, as alignment is a multidimensional and idiosyncratic phenomenon, the haphazard feature selection process can result in missing alignment effects that actually exist in a dialogue, simply because the appropriate feature was not measured. As Pardo (2013) notes, this behavior "yields datasets that are relatively inconsistent and chaotic" (p. 3). There is no single acoustic-phonetic feature which is diagnostic of alignment in all linguistic contexts and tasks, and thus selecting just one or two features is likely to produce misleading or conflicting assessments of alignment. One solution is to measure acoustic alignment in a holistic way, taking into account a wide range of phonetic features which might (or might not) individually be aligned upon, depending on characteristics of the task, the speaker, the social affiliation between interlocutors, and many other, potentially unknown, factors. As Pardo et al. (2018) note, "assessment of phonetic convergence using a holistic measure is preferable to individual acoustic measures because acoustic-phonetic attributes vary inconsistently, there are no standards for selecting particular attributes to examine, and holistic appraisal reflects the multi-dimensionality of the phenomenon."

One solution to this problem of which phonetic feature to choose is to use the human perceptual system as a holistic assessor. After a first group of participants performs the speech production task, a second group of participants listens to the first group's recordings and rates how similar their speech sounds to those of their partner or the model talker using an AXB perceptual similarity task (e.g., Dias & Rosenblum, 2011; Goldinger, 1998; Pardo, 2006; Pardo et al., 2010, 2013, 2017, 2018). A common approach is for raters to select whether the model talker's production of a word sounds more like the participant's pre-exposure (baseline) or post-exposure production; the degree of alignment corresponds to the extent that the post-exposure production is selected. The benefit of this approach is that the human raters are likely taking into account a wide range of phonetic properties in parallel when making this similarity judgement, which therefore does not require the researcher to select an individual feature which may or may not demonstrate alignment. However, this strategy has drawbacks as well. First, it requires a new, large sample of participants to perform the perceptual similarity task for each speaker's productions. Second, there is no way to know (or control) which features the listeners perceive and use to make their judgements, as perception introduces a new source of variability, and makes it difficult to learn which features are the ones being aligned upon.

One recent study investigated the relationship between alignment detected by holistic perceptual similarity judgements and measurements of individual features (Pardo et al., 2017). The authors found that the AXB task detected a numerically small but statistically significant effect of alignment (56%, compared to chance of 50%). However, when alignment was computed for individual acoustic features, only one of five (vocalic duration) showed alignment. The authors additionally found that each of the individual acoustic features significantly predicted the AXB perceptual similarity scores, despite not showing alignment when measured alone. This study lays important groundwork for investigating alignment using both a holistic measure which takes into account many acoustic dimensions simultaneously, and also investigating alignment using a few carefully-chosen individual features.

The present work combines the advantages of a holistic measure of alignment while avoiding the drawbacks of a perceptual similarity task. Here, we measure phonetic accommodation within a dialogue by calculating several hundred common acoustic-phonetic features at both the segmental and suprasegmental levels, and employing a machine learning model to measure the degree of partner-specific acoustic alignment jointly on the set of features. This model uses the full suite of acoustic-phonetic features to determine a global measure of phonetic alignment, and whether at least *some* aspects of the speaker's overall phonetic profile converge upon their listener's, without such convergence being tied to the measurement of an individual feature.

1.4. Present research

The present work investigates whether alignment is a unitary concept; though a prevalent assumption, this has largely not been tested. To do so, we seek to answer two questions: (1) does alignment at one linguistic level necessarily engender alignment at others? and (2) does alignment on one feature within a linguistic level entail alignment on other features within that level? As discussed above, we do so in two ways: by calculating a holistic measure of alignment, and by testing for alignment at multiple levels within the same interaction.

In the current experiment, participants alternately interacted with two conversational partners, playing a picture-matching game. The two partners had distinct speech production profiles on multiple linguistic levels: They had highly differing phonetic profiles (due to differing gender, native language, and accent), and produced contrasting syntactic structures (by consistently producing only one structure of the dative alternation for a given participant). In the game, participants heard picture descriptions from each experimenter, and then the participant described pictures back to each experimenter individually. Degree of alignment was measured at multiple linguistic levels during these dialogues: the acoustic-phonetic features jointly as a holistic measure, and individual features at the temporal-phonetic, spectral-phonetic, and syntactic levels. Alignment was operationalized as the degree to which the participant's speech was more similar to their currently-listening partner's, as opposed to the non-listening partner's, on that linguistic dimension.

This experimental design, where the participant interacts with two partners, allows for a clean test of partner-specific alignment. We test for alignment by comparing the way the participant speaks to one partner against the way the same participant speaks to the other partner. This allows us to determine if the participant modulates their speech to match specific characteristics of their current listener. If participants only interacted with one partner, and we observed modulation on some features (e.g., between a pre- and post- test), we would be unable to determine whether such modulation was alignment specifically to their listener, as compared to alignment to their recent exposure in the aggregate, which incidentally all came from a single partner.

2. Method

De-identified data, scripts for calculating the phonetic features, and predictions from the machine learning model are available at https://osf.io/3hu5r/.

2.1. Participants

The participants were 96 students at the University of California, San Diego, who completed this experiment for course credit. All reported being native, monolingual speakers of English.

2.2. Materials

The stimuli consisted of 96 colored drawings of scenes which could easily be described using a single sentence, each printed on a card approximately 4 1/2″ tall and 3 2/3″ wide. Of these, there were 72 unique dative scenes and 24 unique intransitive scenes. The event actors and actions varied across the pictures to provide variety to the descriptions. The dative pictures could be described using either a prepositional dative (PD) or double object (DO) structure. For example, the scene shown in Fig. 1a could be described either as "The slave is offering grapes to the pharaoh" [PD] or "The slave is offering the pharaoh grapes" [DO]. The intransitive pictures had a simple event structure (e.g., "The woman is sleeping" for Fig. 1b) to make it unlikely that participants would produce a sentence containing a post-verbal object, so as not to prime one of the dative structures. The full set of stimuli sentences is provided in the supplemental materials.

2.3. Procedure

Participants played a conversational picture-matching game with the two experimenters (A and B). The participant and one experimenter sat on opposite sides of a table, separated by an opaque barrier approximately eight inches tall — high enough to block the other person's table space, but low enough to easily see each other's face and upper body. Each partner had a series of pictures in front of them. The task throughout the experiment was for one person (the *Director*) to describe their pictures to the other person (the *Matcher*), who arranged his/her own pictures in the same order as they were described. Over the course of the experiment, the participant alternated between interacting with the two experimenters, one at a time.

In order to create a strong pressure for partner-specific alignment at each linguistic level, the experimenters had highly contrasting linguistic behavior at both the acoustic-phonetic and syntactic levels. To create the phonetic profile contrast, one experimenter was a female native speaker of Mandarin, who spoke English with a strong non-native accent (although she began learning English in primary school, she did not live in an English-speaking country [USA] until age 18, and had a very noticeable non-native accent). The other experimenter was a male native speaker of American English. To create the syntactic contrast, for a given participant, one experimenter (Experimenter A) produced only double object dative sentences (DO) and the other experimenter (Experimenter B) produced only prepositional dative sentences (PD). The assignment of syntactic preference to experimenter was counterbalanced across subjects, so half of participants heard PDs from the non-native experimenter and DOs from the native experimenter, and half heard the reverse pairing.

The experiment began with one experimenter explaining the picture-matching game to the participant and a short practice round using three non-dative pictures. The main experiment consisted of two sequential phases. First, in the *Exposure Phase*, the participant was exposed to each experimenter's acoustic and syntactic production schemas, by listening to a series of picture descriptions from each experimenter. One experimenter entered the participant's testing room and described six pictures using their assigned syntactic structure and their personal acoustic profile; meanwhile, the participant selected each matching picture from a set of 18. The experimenter verified that the participant's picture ordering was correct and corrected any mis-selected pictures (which was extremely rare). When they finished, the experimenter left the

Fig. 1. (a) Sample dative picture and (b) sample intransitive picture used in the present experiment.

room. Then, the second experimenter entered the testing room and described a new set of six pictures while the participant matched. This process, in which each experimenter described six pictures to the participant, constituted *Exposure Round 1*. Then, the first experimenter re-entered the room and described 6 new pictures and left; and the second experimenter re-entered the room and described 6 pictures and left, comprising Exposure Round 2. There were four rounds during the Exposure Phase. All of the pictures that were described by the two experimenters were dative scenes; thus after the Exposure Phase, the participant had heard 24 PD descriptions from one experimenter and 24 DO descriptions from the other experimenter, each using their personal phonetic and syntactic schema.

After the four rounds of the Exposure Phase, the *Test Phase* began. (Note that all Exposure from the two experimenters preceded all Test trials.) The Test Phase was similar except that the participant became the Director and the experimenter the Matcher. The experimenter laid out six dative and six intransitive pictures on the participant's side of the table (without looking so they ostensibly would not know the order), none of which had appeared in the Exposure Phase. The participant described these 12 pictures in order while the experimenter matched. The datives and intransitive pictures were interleaved so as to reduce any effects of trial-to-trial syntactic self-priming. This was *Test Round 1*. The experimenter then left the room. The second experimenter entered and the participant described a new set of six dative and six intransitive pictures, comprising Test Round 2. The first experimenter returned for the participant to describe a third set of 12 pictures (Test Round 3), and then the second experimenter returned for the participant to describe a final set of 12 pictures (Test Round 4). Thus, each participant described a total of 48 pictures across four rounds: 24 pictures to each experimenter, of which 12 were dative scenes and 12 were intransitive scenes.

The participant's productions during the Test Phase — the syntactic structure(s) they used and their vocal acoustic characteristics, as a function of which experimenter they were speaking to — constituted the raw data used in the following analyses.

No picture was repeated across the experiment. All factors, including nuisance factors, were counterbalanced between participants: The order of directing experimenters in the Exposure Phase (Experimenter A vs. Experimenter B directing first); the order of listening experimenters in the Test Phase (Experimenter A vs. Experimenter B matching first); the order of the pictures that the participant described within each round (describing Picture 1, then 2, then 3,... then 12 vs. describing Picture 12, then 11, then 10,... then 1); the structure which the directing experimenter used to describe a particular picture in the Exposure Phase (e.g., Fig. 1a described by the experimenter using a PD vs. DO); and, critically, the assignment of syntactic preference to experimenter identity and thus acoustic profile ([native = DO and non-native = PD] vs. [native = PD and non-native = DO]).

Additional components to the procedure involved each experimenter sharing a "fun fact" about themselves at the start of each round, giving the participant a two-digit math problem when switching experimenters, and a post-experiment questionnaire. Exact details on the timing and reason for including these procedural steps are reported in Experiment 4 and Supplementary Data 2 of Ostrand and Ferreira (2019), which had a similar procedure to this experiment. However, as the facts, math problems, and questionnaire are not relevant to the analyses reported here and were not analyzed for this experiment, they will not be discussed further.

2.4. Phonetic features extraction

For this experiment, three types of analyses were performed: (1) degree of partner-specific holistic acoustic alignment, as calculated using many acoustic-phonetic and temporal features together in a machine learning classifier; (2) degree of partner-specific phonetic alignment as calculated separately on seven *a priori*-selected acoustic-phonetic features (temporal and spectral); and (3) degree of partner-specific syntactic alignment, as measured by participants' rate of prepositional dative production.

The acoustic-phonetic features used in analyses (1) and (2) were calculated using Praat (version 6.0.50; Boersma & Weenink, 2019), Python (version 2.7.17), and R (version 4.0.3). The acoustic-phonetic features covered segmental and suprasegmental aspects of a speaker's speech pattern. Phonetic features were largely tailored to the natural class of

segments (e.g., formants and voice quality measures calculated from sonorants; spectral moments from sibilants), or to suprasegmental features of the utterance- or dialogue-level characteristics of the speech (e.g., rhythm metrics from utterances, or intra-utterance durations from the dialogue). The complete list of acoustic features that were calculated is given in Table 1. The goal of this analysis was to investigate alignment using a suite of features which covers a speaker's full phonetic profile: from individual segmental categories to global temporal properties to more abstract representations of spectral qualities. The purpose of calculating such a range of features is to attempt to capture any phonetic areas on which a speaker might align, and thus be able to detect such alignment within our holistic analysis using the machine learning model. The complete set was used for analysis (1), and a subset (as discussed in the following section) was used for analysis (2).

The goal of the acoustic-phonetic analysis was to create two parallel phonetic profiles for each participant–experimenter pairing. Acoustic-phonetic measurements were extracted from various segmental and suprasegmental properties of the utterances. We attempted to maintain segment-specific representations when relevant; however, token counts varied considerably across segment categories. If the median count of a segment across speakers was low, the category was either fully omitted or combined with a related category for later averaging for each participant–experimenter pairing.

For all phonetic analyses, the four Test Phase rounds were segmented individually and each analyzed as separate recordings. Thus, each participant had four recordings, two directed to the native experimenter and two to the non-native experimenter. Each recording was approximately 70 s long.

Recordings were first manually segmented at the utterance level (i.e., each picture description), and then force aligned using the Montreal Forced Aligner (MFA; McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017). This resulted in phone-level alignments of canonical pronunciations for each utterance. Measurements were organized by domain. These were: sonorant-specific measures (formants and vowel quality), obstruent-specific measures (spectral moments), and general temporal measures (segment-specific durations and rhythm).

Sonorant-specific measures were extracted from all vowels, glides, liquids, and nasals. These measures were: F0, F1, F2, F3, and F4 at the first quartile, midpoint, and third quartile of the segment. Voice quality measures of local jitter, ppq5, local shimmer, apq5, mean period duration, standard deviation of the period duration, harmonic-to-noise ratio (HNR), and the standard deviation of HNR were also extracted.

All sonorant-specific measures were obtained using Praat. Vowel formants were measured using the Burg estimation method over the full recording from one round with a maximum formant of 5500 Hz for female speakers and 5000 Hz for male speakers. F0 and HNR measures were extracted with F0 intervals of 100–500 Hz for female speakers and 50–300 Hz for male speakers. Local jitter, ppq5, local shimmer, apq5, period duration estimates were measured using the default duration interval and a maximum period factor of 1.3.

Due to low token counts, the glides [w j] and diphthongs [aʊ aɪ oɪ], were removed, and the following broad vowel categories were created: I ([i ɪ]), E ([eɪ ɛ]), A ([ɑ ɔ oʊ])[2], U ([u ʊ]), ER ([ɝ ɚ]), and AH ([ʌ ə]). Segment-specific categories were [æ], [l], [ɹ], [n], [m] and [ŋ]. Following profile-averaging, the vowel triangle area was also computed from the midpoint F1 and F2 averages for I, A, and U.

Obstruent-specific measures were retained for only [s] and [z], largely due to low token counts for other fricatives. In addition, stop consonants were omitted from analysis based on their highly dynamic realizations (closure, burst, and possible aspiration phases), and the difficulty that arises from isolating these events. The four spectral moments — center of gravity, variance, skewness, and kurtosis — were calculated from the middle 20 ms of each sibilant following high-pass filtering at 300 Hz and multitaper spectral analysis in R (using the *multitaper* package, version 1.0–14). All instances of [s] and [z] met this minimal duration threshold. The multitaper spectral analysis had 8 tapers and the time-bandwidth parameter set to 4 (Iskarous, Shadle, & Proctor, 2011; Reidy, 2015).

Mel Frequency Cepstral Coefficients (MFCCs) were calculated using the *Python Speech Features* package in Python (Lyons, Wang, & Gianluca, 2020). Thirteen bands were calculated on the individual utterances (picture descriptions) from each recording (round), and then the median was taken across all utterances from the same recording for each coefficient.

Duration was also calculated from each MFA-aligned segment and represented in the profile for each of the above-mentioned segmental categories. A set of temporal features was also calculated for each utterance (picture description), and metrics of intra- and inter-utterance pause durations were also extracted. For the rhythm analysis, phone-level alignments were converted to C and V intervals. For each utterance (picture description), the rhythm metrics of rateCV, %V, ΔC (Ln), ΔV (Ln), ΔPeak (Ln), and mean Peak (Ln) were calculated using the *DurationAnalyzer* Praat script (Dellwo, 2019). These respectively correspond to the number of segments per second (rate CV), percent vocalic (%V), variance of the logged consonant and vocalic durations (ΔC, ΔV), and mean and variance of the vocalic logged peak-to-peak durations (ΔPeak, mean Peak). These features exhibit relatively high between-speaker variability, potentially allowing for increased speaker-specific alignment (Dellwo, Leemann, & Kolly, 2015). In addition, speech rate was calculated as the number of syllables per second across the entire recording, as estimated from the automatic detection of syllable nuclei using the *Syllable Nuclei* Praat script (de Jong & Wempe, 2009).

Two types of pause duration were computed. Intra-utterance pause duration, capturing another measure of speech rate, was calculated as the average pause duration within each utterance, and then averaged across the utterances within each recording (round). Pauses were detected automatically using the Montreal Forced Aligner. Inter-utterance pause duration was calculated across the entire recording (round) for each participant, and was calculated as the duration from the offset of one utterance (picture description) to the onset of the next, capturing how long participants

[2] The A category is best characterized as the set of non-high back vowels. The decision to group [oʊ] with the low back vowels instead of the high back vowels was a judgment call motivated primarily by its traditional featural description. All vowels in this group share the [-high] feature and both [ɔ] and [oʊ] share the [-low] feature.

Table 1
Acoustic-phonetic features calculated on the participants' and experimenters' recordings.

Feature type	Phonetic class	Number of features	Description	Reference
F0 at Q1, Midpoint, Q3 [Hz]	Sonorants	36	Approximate frequency of the period	Boersma & Weenink, 2019
F1, F2, F3, F4 at Q1, **Midpoint**, Q3 [Hz]	Sonorants	144	Concentrations of acoustic energy predominantly reflecting resonant cavities of the vocal tract	Boersma & Weenink, 2019
Jitter (local, ppq5) [%]	Sonorants	24	Frequency variation between periods	Boersma & Weenink, 2019; Teixeira, Oliveira, & Lopes, 2013
Shimmer (local, apq5) [%]	Sonorants	24	Amplitude variation between periods	Boersma & Weenink, 2019; Teixeira et al., 2013
Period (mean [s], standard deviation [s])	Sonorants	24	Mean and standard deviation of the period duration	Boersma & Weenink, 2019
Harmonics-to-noise ratio (HNR [dB], standard deviation [dB])	Sonorants	24	Log (base 10) ratio of periodic to non-periodic components in the signal multiplied by 10	Boersma & Weenink, 2019; Fernandes, Teixeira, Guedes, Junior, & Teixeira, 2018
Vowel triangle area [Hz^2]	I, A, U	1	Area between I, A, and U in the midpoint F1×F2 space Vowel triangle area $=\lvert 0.5 * (F1_I * (F2_A - F2_U) + F1_A * (F2_U - F2_I) + F1_U * (F2_I - F2_A)) \rvert$	Skodda, Grönheit, & Schlegel, 2012
Center of gravity [Hz]	Sibilants	2	Energy-weighted mean frequency	Forrest, Weismer, Milenkovic, & Dougall, 1988; Iskarous et al., 2011
Variance [Hz^2]	Sibilants	2	Variance of energy from the mean across frequency bins	Forrest et al., 1988; Iskarous et al., 2011
Skewness [dimensionless]	Sibilants	2	Skewness of the energy distribution across frequency bins	Forrest et al., 1988; Iskarous et al., 2011
Kurtosis [dimensionless]	Sibilants	2	Kurtosis (peakiness) of the energy distribution across frequency bins	Forrest et al., 1988; Iskarous et al., 2011
Duration [s]	Sonorants, sibilants	14	Duration of the segment	McAuliffe et al., 2017
Mel Frequency Cepstral Coefficients (MFCCs)	Utterance-level	13	Median of the power spectrum in 13 frequency bands	Lyons et al., 2020
rateCV [segments / s]	Utterance-level	1	Number of segments divided by the duration of the utterance (sec), excluding silent periods	Dellwo et al., 2015
% V [%]	Utterance-level	1	Percent vocalic across consonant and vocalic intervals	Dellwo et al., 2015
ΔC, ΔV [ln s^2]	Utterance-level	2	Variance of the log consonant or vowel durations within an utterance	Dellwo et al., 2015
Mean Peak [ln s], ΔPeak [ln s^2]	Utterance-level	2	Mean and variance of log peak-to-peak durations within an utterance, where the peak is defined by the location of peak energy in a vocalic interval	Dellwo et al., 2015
Speech rate [syllables / s]	Utterance-level	1	Number of detected syllable nuclei divided by the total duration	de Jong & Wempe, 2009
Intra-utterance pause duration (median, 95th percentile) [s]	Utterance-level	2	Average duration of silent intervals within an utterance (picture description)	McAuliffe et al., 2017
Inter-utterance pause duration (median, **95th percentile**) **[s]**	Recording-level	2	Duration from the offset of one utterance (picture description) to the onset of the next	McAuliffe et al., 2017

Note. Unless otherwise indicated, the acoustic-phonetic profile comprised the means for each measure–category pairing. Sonorants consisted of the following 12 segment categories: I, E, A, U, ER, AH, AE, L, R, N, M, NG. Sibilants consisted of the following two segments: S, Z. Features shown in bold were used in the individual-feature analysis.

waited for their partner to comprehend their sentence and select the matching picture after each picture description.

For each segmental measure, outliers beyond 2.5 standard deviations of the speaker- and category-specific mean were excluded, with the exception of the vowel triangle area, which was calculated directly from resulting midpoint F1 and F2 means for I, A, and U following outlier exclusion. For the suprasegmental measures of rhythm and pause durations, outliers beyond 2.5 standard deviations of the speaker-specific mean were excluded. For all measures except pause durations and MFCCs, means were then obtained for each measure and category (local segmental category or global recording category) to form each speaker's profile. The median was used for MFCCs as these talker-specific distributions were somewhat more skewed. For the intra- and inter-utterance pause durations, median was calculated to estimate central tendency, as well as the 95th percentile of the distribution to give an estimate of the maximal pause length without being susceptible to extreme outliers.

In total, 323 acoustic-phonetic and temporal measures were calculated for each experimenter and for each of four Test Phase rounds for each participant (see Table 1).

2.5. Analysis strategy

There are a few ways that conversational partners might align their speech to each other during a dialogue. For the three analyses presented here, we employ *global proximity* (cf. Levitan & Hirschberg, 2011; Weise & Levitan, 2018) as our alignment measure. This measure tests how close overall the participant's feature values are to those of their current listener, as compared with those of their non-listener. This measure of global proximity contrasts with *convergence*, which assesses the degree to which a participant's speech *becomes more* similar to that of their listener's over time. Convergence is more suited to a true back-and-forth dialogue task, in which the two interlocutors alternate speaking. In contrast, proximity is more suited to a task like the present one, in which the participant gets ample exposure to their partner's speech upfront, and then either does or does not modulate features of their own speech to match those of their listener's.

2.5.1. Holistic measure of alignment

Analysis (1) tested for the presence of partner-specific alignment across many acoustic-phonetic features in the same

model. As discussed above, dozens of different measures are used across vocal accommodation studies, but most individual studies measure at most a small handful. As a result, that study's conclusion as to presence or absence of alignment is determined by its presence or absence on that individual feature. The goal of this analysis was to test for the overall presence of phonetic alignment without relying on only one metric or creating a massive multiple comparisons problem. To that end, a suite of 323 acoustic-phonetic features (see Table 1) was calculated on the participants' and experimenters' sound files, and jointly used as features in a machine learning classifier to predict the degree of partner-specific alignment.

To create the input features for the classifier, the set of 323 acoustic-phonetic measures were computed on all participant and experimenter recordings. The two experimenters were recorded individually, speaking their sentences from the Exposure Phase as they did during the experiment. However, the experimenter recordings were not collected during an interaction with any particular participant, and the same experimenter recording was compared to all participant recordings. This works *against* finding evidence of phonetic alignment, as there may have been conversation-specific idiosyncrasies of how the experimenter spoke to a particular participant (e.g., suffering from a cold) which affected the participants' acoustic production but is not reflected in the generic experimenter recording employed as the baseline here. In such a situation, the participant's speech may actually have been closer to the experimenter's speech on the day of the experiment and thus may have aligned more than they appear to when compared against just the experimenter's generic recording. Additionally, the experimenters may themselves have somewhat converged to the participants over the course of the experiment, which is not captured by the generic recording.

For each participant's recordings, the array of 323 acoustic-phonetic feature values was subtracted from each experimenter's array of feature values, and the by-feature difference squared, to capture the similarity between the participant's and each experimenter's speech while ignoring the direction[3]. Thus, for each participant recording, two difference score arrays were created (following Levitan & Hirschberg, 2011). The *Listener difference* array came from subtracting the participant's feature values from the feature values of the experimenter to whom they were speaking. The *Non-listener difference* array came from subtracting the participant's feature values from the feature values of the other experimenter, to whom they were *not* currently speaking. For example, Participant #1 described pictures to Experimenter A in Test Phase Round 1. The feature array from the Test Round 1 participant recording was subtracted from Experimenter A's to form the Listener difference, and from Experimenter B's to form the Non-listener difference. (See Fig. 2 for a graphical depiction of this process.)

These difference score arrays became the input features to a machine learning binary classifier using logistic regression with elastic net regularization. The classifier used leave-one-subject-out cross-validation to determine prediction accuracy and thus degree of partner-specific alignment. Thus the classifier was trained on (i.e., given as input) the difference score arrays for the four rounds of 95 participants and told which ones were Listener difference scores and which were Non-listener difference scores, and then predicted, for the 96th (left-out) subject, which difference score arrays were Listener and which were Non-listener. This process iterated, with each participant left out in turn; therefore, a prediction of Listener/Non-listener was made for each difference score for each participant. Each prediction was either correct or incorrect; averaging across all predictions produces an overall accuracy score. This accuracy indexes the degree to which the model could differentiate a Listener array from a Non-listener array, and thus the degree to which participants sounded more like Experimenter A than B when addressing A, and more like Experimenter B than A when addressing B.

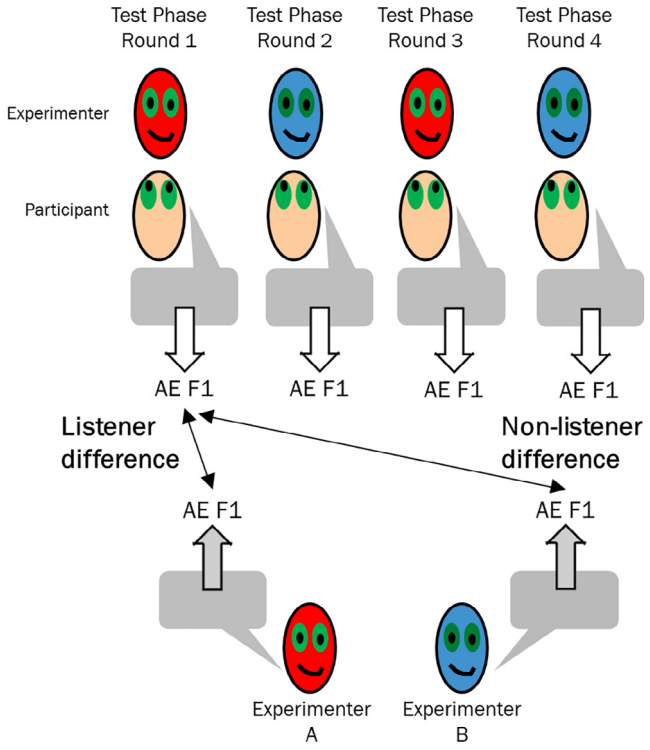

Fig. 2. A demonstration of the process for computing Listener and Non-listener difference arrays as inputs to the machine learning classifier. Each participant produced four recordings (Test Phase Rounds 1–4), two spoken to each experimenter. First, a particular feature (here, F1 of [æ], referred to as AE F1) is calculated on each of the participant's four recordings. Second, AE F1 is calculated on each experimenter's recording. Third, the participant's Test Round 1 AE F1 is subtracted from each experimenter's AE F1, and the difference is squared, producing two difference scores as shown by the solid arrows. This subtraction is repeated for each of the four participant recordings. When the participant described to Experimenter A during that round (as in Test Round 1 in this figure), then $(\text{Experimenter A} - \text{Participant})^2$ is the Listener difference and $(\text{Experimenter B} - \text{Participant})^2$ is the Non-listener difference.

2.5.2. Individual measures of alignment

Phonetic Alignment. For analysis (2), the individual-feature phonetic analysis, Listener and Non-listener difference scores were calculated for each tested feature, and then aggregated across the participant's four rounds to produce a single Listener score and a single Non-listener score for each participant.

The formulas used to compute the two difference scores for a particular feature across the four rounds are as follows,

[3] Using the difference squared (rather than absolute value) as the measure of distance magnitude allows this metric to robustly capture differences between the speakers both when the participant's feature value is between that of the two experimenters' as well as when the participant's feature value is either less than or greater than both experimenters'.

where **x** refers to the participant's value on a particular feature, with the subscript indicating whether the participant was speaking to Experimenter A or B, and in which round. **a** and **b** represent the feature values of Experimenter A and B, respectively. Thus, $\mathbf{x}_{a,\text{round1}}$ denotes the participant's value on a particular feature calculated when the participant was speaking to Experimenter A during Round 1.

Listener difference =

$$\sqrt{(\mathbf{a}-\mathbf{x}_{a,\text{round1}})^2+(\mathbf{b}-\mathbf{x}_{b,\text{round2}})^2+(\mathbf{a}-\mathbf{x}_{a,\text{round3}})^2+(\mathbf{b}-\mathbf{x}_{b,\text{round4}})^2}$$

Non-listener difference =

$$\sqrt{(\mathbf{b}-\mathbf{x}_{a,\text{round1}})^2+(\mathbf{a}-\mathbf{x}_{b,\text{round2}})^2+(\mathbf{b}-\mathbf{x}_{a,\text{round3}})^2+(\mathbf{a}-\mathbf{x}_{b,\text{round4}})^2}$$

A numerical example may be illustrative. For example, Experimenter A's raw value on some particular feature is 5 and Experimenter B's raw value is 7, and this participant is speaking to A on rounds 1 and 3 and to B on rounds 2 and 4. If the participant always speaks exactly like A (i.e., always produces a raw value of 5) and does not modulate her speech when speaking to either experimenter, then the two difference scores would be as follows:

Listener difference =

$$\sqrt{(5-5_{a,\text{round1}})^2+(7-5_{b,\text{round2}})^2+(5-5_{a,\text{round3}})^2+(7-5_{b,\text{round4}})^2}$$
$$=2.8$$

In rounds 1 and 3, when Experimenter A is the listener, the participant's value (5) is subtracted from A's value (5); in rounds 2 and 4, when Experimenter B is the listener, the participant's value (5) is subtracted from B's value (7).

Non-listener difference =

$$\sqrt{(7-5_{a,\text{round1}})^2+(5-5_{b,\text{round2}})^2+(7-5_{a,\text{round3}})^2+(5-5_{b,\text{round4}})^2}$$
$$=2.8$$

Here, the participant's value is subtracted from B's in rounds 1 and 3 (when the participant was addressing A), and the participant's value is subtracted from A's in rounds 2 and 4 (when the participant was addressing B).

The measurement of alignment comes about by comparing the Listener difference with the Non-listener difference. In this example, we know the participant did not align because she always produced a value of 5. The comparison of Listener and Non-listener difference scores bear this non-alignment out, as they are identical (both 2.8).

In contrast, a participant who *does* align to their current partner should have a *smaller* Listener difference than Non-listener difference, as in the following example, when the participant's value is smaller when talking to a partner with a smaller feature value, and larger when talking to a partner with a larger feature value:

Experimenter A $= 5$

Experimenter B $= 7$

$\mathbf{x}_{a,\text{round1}} = 6$

$\mathbf{x}_{b,\text{round2}} = 7$

$\mathbf{x}_{a,\text{round3}} = 6$

$\mathbf{x}_{b,\text{round4}} = 7$

Listener difference =

$$\sqrt{(5-6_{a,\text{round1}})^2+(7-7_{b,\text{round2}})^2+(5-6_{a,\text{round3}})^2+(7-7_{b,\text{round4}})^2}$$
$$=1.4$$

Non-listener difference =

$$\sqrt{(7-6_{a,\text{round1}})^2+(5-7_{b,\text{round2}})^2+(7-6_{a,\text{round3}})^2+(5-7_{b,\text{round4}})^2}$$
$$=3.2$$

In this case, the participant does align in a partner-specific manner, and the Listener difference (1.4) is less than the Non-listener difference (3.2).

The present experimental design of two partners that the participant alternates between and this difference score technique captures the participant's relative degree of alignment regardless of how close their baseline phonetic properties are to either experimenter's. For example, a female participant is likely to be closer to the female experimenter on a number of phonetic features, independent of degree of alignment. However, even if the participant spoke exactly like Experimenter A for all four rounds (as in the first example above) on every acoustic-phonetic dimension, this analysis technique will result in no alignment. For the two rounds in which the participant was speaking to Experimenter A, the Listener difference will be 0 — thus showing very high alignment. In contrast, for those same rounds, the Non-listener difference will be high, because the participant's phonetic properties will be very different from Experimenter B's. The reverse will be true for the two rounds in which the participant was speaking to Experimenter B — the Listener difference will be high and the Non-listener difference will be 0. Combined across the four rounds (two spoken to each experimenter), the Listener and Non-listener difference scores will be identical. This holds true for any participant who speaks identically to the two experimenters and never aligns to their listener — regardless of whether the participant's speech is close to Experimenter A's, close to Experimenter B's, somewhere in between the two, or far above or below both experimenters. Similarly, if the participant's value on some feature is extreme (whether due to natural tendency, a respiratory infection, measurement error, etc.), this experimental design and difference score calculation remains robust, as the participant's extreme value goes into both the Listener and Non-listener difference scores. Thus that participant's difference from *both* partners may be particularly large, but the *comparative difference* to the Listener vs. the Non-listener — namely, the degree of partner-specific alignment — will be unaffected by the participant's extreme raw value.

In keeping with the analysis strategy of the majority of phonetic alignment research, this analysis employs seven phonetic features to individually measure the degree of partner-specific alignment: two measuring temporal characteristics (speech rate and inter-utterance pause duration) and five measuring spectral characteristics (midpoint F1 at each of the four corner vowel categories of A, AE, I, U, and the area of the

vowel triangle defined in midpoint F1 × F2 space for A, I, and U). These seven features to be analyzed individually were selected *a priori* (with no other features analyzed individually) because they are commonly used measures in research on phonetic alignment. For each feature, the by-participant Listener and Non-listener scores were statistically compared with a paired t-test to determine degree of alignment on that feature.

Speech rate was calculated as the number of syllables per second, across the entire recording (cf. Bell et al., 2003; Bonin et al., 2013; Levitan & Hirschberg, 2011; Matarazzo, Wiens, Saslow, Dunham, & Voas, 1964; Pardo et al., 2010; Schultz et al., 2016; Staum Casasanto et al., 2010; Street, 1984; Uther, Knoll, & Burnham, 2007; Webb, 1969).

Inter-utterance pause duration was calculated as the length of time that participants paused after completing one picture description sentence and before beginning the next, capturing how long participants gave their partner to comprehend their sentence and locate the matching picture before moving on to the next description. This produced a distribution of durations across the recording. For this feature, the 95th percentile of the distribution was used, to give an estimate of the maximal pause length without being susceptible to extreme outliers. The 95th percentile of inter-utterance duration is less likely to be correlated with speech rate than the average or median of intra-utterance pause duration would be (cf. Cappella & Planalp, 1981; de Looze et al., 2011; de Looze & Rauzy, 2011; Edlund et al., 2009; Gregory & Hoyt, 1982; ten Bosch et al., 2004).

F1, the first vowel formant, was calculated at the midpoint of the vowel for each of the corner vowel categories of A ([ɑ ɔ oʊ]); AE ([æ]); I ([i ɪ]); and U ([u ʊ]). The means of these four measures were estimated from the distribution of category-specific values across the sound file (cf. Hwang, Brennan, & Huffman, 2015; Pardo et al., 2010, 2012, 2013).

The area of the *vowel triangle* was derived by calculating the area between the vertices I, A, and U in F1 × F2 space (cf. Babel, 2010, 2012; Pardo et al., 2012).

Syntactic Alignment. Participants' descriptions were transcribed and coded offline (blind to listener identity and syntactic preference) as one of two dative alternations (PD or DO) or neither. Descriptions that could not be categorized as one of the alternations, either because they were not a dative structure (17.9%) or were missing one of the objects (6.7%), were excluded. One picture was excluded from the syntactic analyses because it did not elicit an acceptable description from at least 30% of participants.

Syntactic production was analyzed with generalized logit mixed-effect models (GLMM) in R (version 3.4.0; R Core Team, 2019) using the *lme4* package (version 1.1.13; Bates, Mächler, Bolker, & Walker, 2015). All factors were fully within-subject and within-item. The two independent variables each had two factor levels coded as −0.5 and + 0.5. To determine the statistical significance for each fixed effect, the omnibus model (with the two main effects and the interaction term included) was compared with a reduced model that had the effect in question removed. When a model failed to converge, first, correlations between random effects were removed, and then the random effect accounting for the least variance was iteratively removed until the model converged. When subset models failed to converge, an omnibus model was created which matched the subset model's random effect structure, and was used as the comparison in the statistical test.

3. Results

Data for this experiment are available at: https://osf.io/4y6wn/. This repository includes raw acoustic-phonetic values of each of the features calculated for each of the participants' and experimenters' recordings, as well as the syntactic structure produced by each participant for each picture description. The repository also includes the difference scores, showing the degree of alignment from each participant to each experimenter on each feature. The machine learning predictions for each participant round and difference are posted as well.

3.1. Holistic measure of alignment

The 323 acoustic-phonetic features noted above (see Table 1) were used as features in a machine learning binary classifier, classifying each of the left-out subject's difference scores as either a Listener difference or Non-listener difference. The collective set of acoustic-phonetic features revealed positive evidence of partner-specific alignment, as the classifier performed significantly above chance (accuracy = 55.0%, chance = 50%, $p = .003$), correctly predicting on average 4.396 (maximum = 8; chance = 4) of the participants' difference scores as either the Listener or the Non-listener difference. This accuracy score is on par with estimates of convergence derived from listeners rating perceptual similarity in an AXB task (generally around 56%; Pardo et al., 2018)[4].

3.2. Individual measures of alignment

3.2.1. Phonetic alignment: temporal

As noted above, for the temporal-phonetic features of speech rate and inter-utterance pause duration, a Listener difference score and a Non-listener difference score was calculated for each participant, aggregating across their production on the four Test rounds. These by-subject difference scores were compared using a paired, two-tailed *t*-test for each feature (following Levitan & Hirschberg, 2011). This measure indicates how closely a participant's production on the given feature matched their listener's, as compared to their non-listener's, production. A smaller difference score to one experimenter than the other means that the participant's speech was closer (more similar) to that experimenter's speech.

Participants aligned their speech rate in a partner-specific manner. Participants' speech rate was closer to their listener's speech rate (listener difference score [LDS] = 0.94 syllables/second) than to their non-listener's speech rate (non-listener difference score [NDS] = 1.07 syllables/second; $t(95) = 4.502$, $p < .0001$). See Fig. 3a.

Participants also aligned their pause duration in a partner-specific manner. Their longest pause was closer to the length

[4] It is important to note that we do not claim that the machine learning model is doing something akin to the human perceptual system during an AXB task, or that the features which the machine learning model uses to make its predictions are the same as those which the human perceptual system does (although this is an intriguing question for future research). The point here is just that both holistic measures of alignment produce very similar accuracy scores, and the AXB results provide a useful baseline metric for prediction performance for the machine learning model.

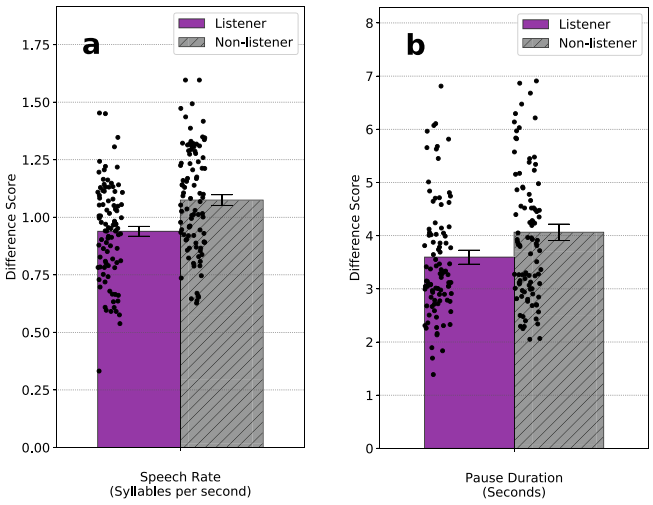

Fig. 3. Degree of temporal alignment on two features, as measured by the participant's difference from their listener as compared to their non-listener. A smaller value means that the participant's speech was more similar to the experimenter's speech on that feature. Error bars show standard error of the mean; points show each participant's difference scores. (a) Partner-specific alignment on speech rate. (b) Partner-specific alignment on pause duration. One participant's points were excluded from both the listener and non-listener scatter plot for display purposes (but were retained for all analyses and displaying the means and error bars) in (b).

of their listener's longest pause (LDS = 3.60 seconds) compared to their non-listener's longest pause (NDS = 4.06 seconds; $t(95) = 4.183$, $p < .0001$). See Fig. 3b.

3.2.2. Phonetic alignment: spectral

Similarly, for the spectral features of A, AE, I, and U midpoint F1 and the vowel triangle area, the Listener and Non-listener overall difference scores were calculated for each participant and then compared using a two-tailed, paired t-test for each feature to investigate partner-specific alignment.

Participants did not align the midpoint F1 at any of the corner vowels. Participants' average F1 values of the A category (non-high back vowels: [ɑ ɔ oʊ]) were equivalently different from their listener's (LDS = 116.44 Hz) and their non-listener's (NDS = 115.77 Hz); and the numerical difference is in the direction of anti-alignment ($t(95) = 1.825$, $p = .071$). Similarly, participants' average F1 values of the AE category ([æ]), of the I category (high front vowels: [i ɪ]), and of the U category (high back bowels: [u ʊ]) were equivalently different from their listener's and their non-listener's ([æ]: LDS = 193.53, NDS = 192.82; [i] and [ɪ]: LDS = 129.64, NDS = 130.45; [u] and [ʊ]: LDS = 187.71, NDS = 187.66; all $t(95) < 1$, all $p > .7$). See Fig. 4a-d.

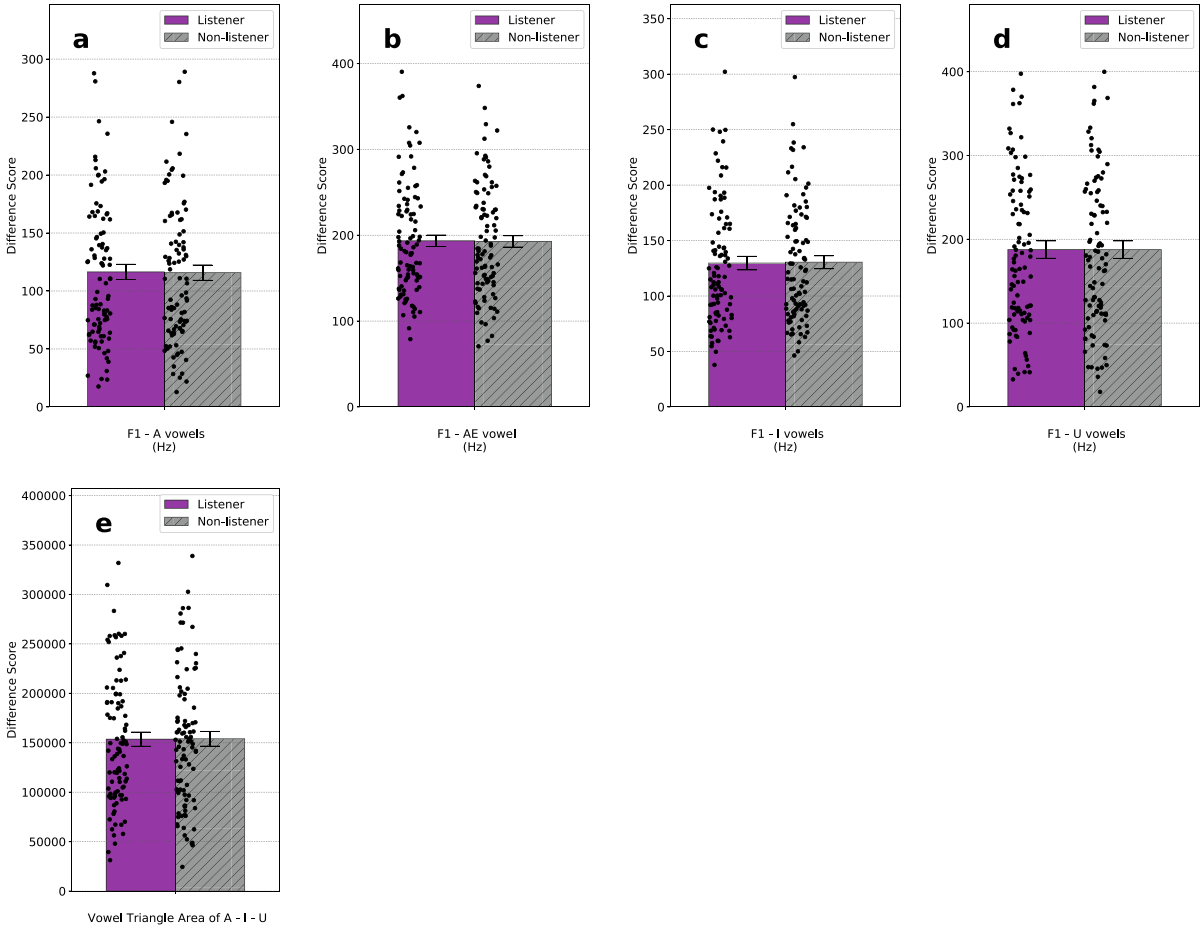

Fig. 4. Degree of spectral alignment on five features, as measured by the participant's difference from their listener as compared to their non-listener. A smaller value means that the participant's speech was more similar to the experimenter's speech on that feature. Error bars show standard error of the mean; points show each participant's difference scores. Figures show the lack of partner-specific alignment on F1 at the midpoint of (a) A vowels ([ɑ ɔ oʊ]); (b) AE vowel ([æ]); (c) I vowels ([i ɪ]); (d) U vowels ([u ʊ]). (e) Lack of partner-specific alignment on the area of the vowel triangle defined by the F1 and F2 averages for A – I – U vowels. One participant's points were excluded from both the listener and non-listener scatter plot for display purposes (but were retained for all analyses and displaying the means and error bars) in (a), (c), (d), and (e).

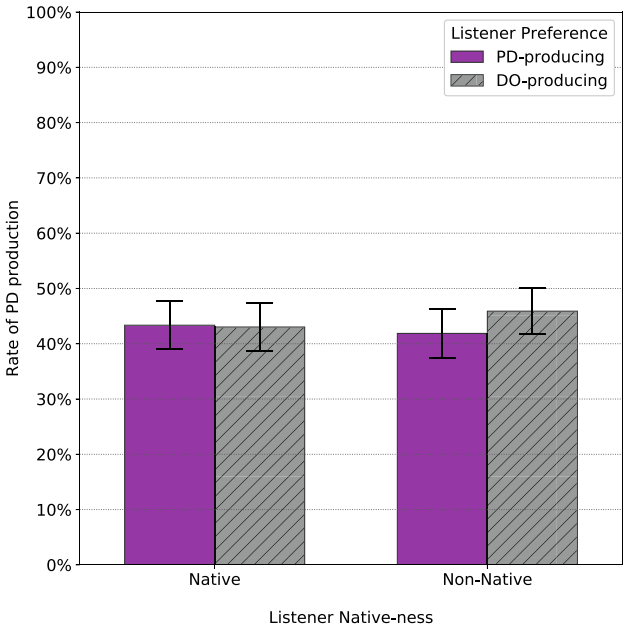

Fig. 5. Degree of syntactic alignment. Percentage of trials on which participants produced a prepositional dative, as a function of their listener's native-ness and syntactic preference. Error bars show standard error of the mean. PD = prepositional dative; DO = double object dative.

Finally, the area of participants' vowel triangle was equivalently different from their listener's (LDS = 153613.35) as compared to their non-listener's (NDS = 154000.06) vowel triangle ($t(95) < 1$, $p = .930$). See Fig. 4e.

3.2.3. Syntactic alignment

Participants' frequency of PD productions were submitted to a 2 (Listener Preference: PD, DO) × 2 (Listener Native-ness: Native, Non-native) GLMM. Participant means are reported below and shown in Fig. 5.

Participants did not engage in partner-specific syntactic alignment. There was no main effect of Listener Preference ($\chi^2(1) < 1$, $p = .941$), meaning that participants produced the same number of PDs regardless of whether they were speaking to the experimenter who had produced only PDs (43.2%) or to the experimenter who had produced only DOs (43.9%). Participants also did not modulate their syntactic production as a function of their listener's identity (no main effect of Listener Native-ness: $\chi^2(1) = 1.402$, $p = .236$), meaning that participants produced PDs at the same rate regardless of whether they were speaking to the native experimenter (42.6%) or the non-native experimenter (44.5%). There was similarly no interaction ($\chi^2(1) < 1$, $p = .605$), meaning that participants did not align syntactically to one listener as compared to both.

4. General discussion

The current experiment is among the first to investigate alignment at multiple linguistic levels within a given interaction, and demonstrates that alignment can simultaneously occur for some linguistic features and levels, and not occur for others. Participants *did not* align on several pre-selected spectral-phonetic or syntactic characteristics. However, participants *did* align on a few pre-selected temporal-phonetic characteristics, and, in a small but robust effect, a holistic measure of partner-specific phonetic alignment was observed when measuring a large number of acoustic-phonetic features jointly using a machine learning model.

Unlike in a standard trial-to-trial priming design, the present findings of alignment cannot be explained by recency priming, as participants received all of their experimental linguistic exposure from both experimenters first, and only spoke back to the two experimenters afterwards. As a result, the effects of alignment that are observed here must be driven by participants learning a partner-specific linguistic profile over the course of the initial Exposure Phase. This makes detecting the presence of alignment more striking, as it relies on participants engaging in partner-specific statistical learning over time rather than just increased activation from the immediately preceding trial.

One possible mechanism that could account for the observed partner-specific alignment on temporal-phonetic features could be due to the "foreign-talk effect," in which native speakers tend to speak more slowly to non-native compared to native interlocutors (e.g., Scarborough, Brenier, Zhao, Hall-Lew, & Dmitrieva, 2007). If the non-native experimenter spoke more slowly, paused for longer, and had longer vowel durations compared to the native experimenter, then the observed alignment on temporal measures could be due not to partner-specific alignment, but rather to participants generally slowing their speech when talking to a non-native listener. However, an examination of the raw feature values for the two experimenters shows this is not the case. In fact, the non-native experimenter spoke more *quickly* and produced *shorter* pauses compared to the native experimenter. Most vowel durations were similar between the two experimenters; of those that differed, some were longer for the non-native experimenter and some were longer for the native experimenter. Overall, the raw temporal features show that the non-native experimenter generally spoke more quickly than did the native experimenter. Table 2 shows the raw feature values for the two experimenters for a demonstrative subset of temporal measures; the complete list of raw features values for the two experimenters is available at the noted OSF repository.

The present results have important implications: They demonstrate that treating alignment as a unitary concept with a single generating mechanism, both between and within linguistic levels, is insufficient. Here, we showed that speakers can align on some individual features at a given level while simultaneously not aligning at a different linguistic level or even on other related features at the same level. As a component of this, these results may provide evidence against a central claim of the Interactive Alignment Model (Pickering & Garrod, 2004), that alignment occurs automatically and that alignment

Table 2
Raw feature values for the two experimenters on a selection of temporal features. Bolded values demonstrate (numerically) faster speech.

Temporal Feature	Native Experimenter	Non-native Experimenter
Speech rate [syllables/s]	1.37	**2.19**
Inter-utterance pause duration 95th percentile [s]	4.25	**1.59**
Intra-utterance pause duration 95th percentile [s]	0.57	**0.34**
Duration A [s]	**0.11**	0.12
Duration AH [s]	0.18	**0.16**
Duration ER [s]	**0.15**	0.18

Table 3
Within-subject correlation in degree of alignment between features at different linguistic levels.

Feature	R^2			p		
	Speech rate	Pause duration	Syntax	Speech rate	Pause duration	Syntax
F1 – A	.002	.000	.014	.648	.861	.249
F1 – AE	.000	.000	.004	.984	.901	.550
F1 – I	.027	.000	.003	.110	.852	.596
F1 – U	.000	.000	.005	.867	.943	.507
Vowel triangle area	.004	.010	.001	.541	.342	.736
Syntax	.000	.001	–	.942	.794	–

at one linguistic level necessarily leads to alignment at other levels. Therefore, alignment connections across levels may not be as robust as was thought, and testing for alignment using only one or two features can miss the presence of alignment as a whole.

4.1. Does alignment at one level necessarily lead to alignment at others?

In contrast to some prior studies of alignment, in which participants are told to shadow given words or use a particular verb in their picture description, in the current experiment, participants' production was not fixed so they were free to speak however they wanted at each linguistic level. This allows for testing the underlying theoretical contention that different linguistic levels naturally vary together, and thus whether more alignment at one level is associated with more alignment at another. The results of the current experiment suggest that this is not the case — speakers may align on some linguistic levels, and on some features of a particular linguistic level, while not aligning on others.

However, the reported results demonstrate this result in the aggregate, across participants, but it is not necessary that all speakers must align in the same situations. It is therefore possible that a subset of participants aligned at multiple levels during their interaction, and other participants did not align at *any* level during their interaction, leading to aggregate differences in alignment across levels. To investigate this possibility, a series of within-subject pairwise correlations were conducted between the degree of alignment on each of the individual features and those from the other linguistic levels tested in analysis (2). That is, we calculated how much each subject aligned on each of the individually-tested features, by subtracting their Listener difference score from their Non-listener difference score, yielding a positive number if that participant aligned on that feature (i.e., their Listener difference was smaller than their Non-listener difference) and a zero or negative number if that participant did not align or anti-aligned on that feature. Then, we ran a series of within-subject pairwise correlations, matching the features at each linguistic level with those at the other levels (e.g., correlation between speech rate and syntax), to assess whether a particular subject who aligned at one level (e.g., temporal-phonetic) also aligned at a different level (e.g., syntactic)[5]. There was no correlation between any pair of cross-level features: Across the 17 pairwise correlations, the most robust correlation (even without correction for multiple comparisons) was $R^2 = 0.027$ with $p = .110$, with most R^2 values below 0.01. (The full correlation matrix is presented in Table 3). This demonstrates that an individual speaker's degree of alignment at one linguistic level is not predictive of their degree of alignment at a different level.

This observation raises questions about the proposed mechanism and trajectory of alignment within a conversation. Alignment behavior could result from a mechanism which applies automatically, tracking incoming linguistic variability and causing the speaker to modulate to small degrees as each new piece of evidence comes in, using transient activation or longer-lasting implicit learning. Alternatively, alignment behavior could result from a mechanism which is first sensitive to communicative or social factors of the language context, inducing alignment only on those features for which there would be a contextual benefit. In the former case, where alignment is driven by an automatic priming mechanism, alignment should occur at all levels of representation simultaneously. However, the present work demonstrates that it is possible for linguistic levels to be out of sync. These results suggest that the mechanism producing alignment is likely not entirely automatic and primitive, and that simultaneous multi-level alignment may not have as strong a cognitive or communicative role as was posited.

In contrast, the current evidence supports an alignment mechanism which is sensitive to contextual or communicative features of the dialogue, rather than a truly resource-free mechanism. One important component of conversational adaptation is communicative utility (cf. Ostrand & Ferreira, 2019). In this case, rather than alignment occurring by default across linguistic levels, the alignment mechanism is sensitive to whether doing so is likely to provide a communicative benefit or improve the likelihood of communicative success during the dialogue. If so, alignment is more likely to occur. If not, there is less pressure for a speaker to align to their conversational partner. In the present experiment, speakers aligned temporal aspects of their speech to their current partner; speech rate and pause duration could reasonably affect intelligibility and likelihood of communicative success, especially when speaking to a non-native listener (Conrad, 1989; Derwing, 1990), although we note that in the present study, participants generally spoke *faster* to the non-native listener to align to her speech rate. In contrast, speakers did not align their syntax to their current partner. The present experiment only tested for syntactic alignment on the dative alternation, which is very common in English and structurally simple; therefore it is unlikely that participants inferred that mismatching their listener's syntax would impair successful dialogue. However, although

[5] Each feature was correlated against all individually-tested features from the other two levels, but not the other feature(s) from the same level. As there is likely high correlation in raw values between closely-related measures, it does not make sense to test for the correlation in degree of alignment between features at the same linguistic level. For example, a speaker's F1 on the A vowels will be strongly related to their F1 on the AE vowel. To reduce the number of correlations presented, and because the goal of this analysis is to assess between-level alignment consistency, correlations were only conducted between features at different levels.

there was no syntactic alignment in the current experiment, speakers might syntactically align if their partners produced other, more complex syntactic structures, where doing so might provide some communicative utility. Similarly, speakers did not align their vowel formants to those of their listeners. Although accent and pronunciation can certainly affect communicative success, specific modulations to F1 may not have been seen as a communicative benefit. Such a mechanism, in which alignment does not automatically activate but rather is modulated by social or communicative factors, is supported by prior work which found that a given speaker varies greatly on how much they align based on the task or conversational setting they are in (Pardo et al., 2018). If alignment occurred automatically and without evaluation of these extra-linguistic factors, then different speakers might align to different degrees compared to each other, but an individual speaker should be fairly consistent in their degree of alignment across conversational settings.

Indeed, the results of the current experiment may shed light on why and when certain linguistic features are entrained upon and perhaps speak to conflicting results across different tasks in different studies. As noted above, one possibility is that the features which are aligned are those which are more appropriate to be learned as part of a partner-specific linguistic profile, because they convey some communicative or social benefit. In contrast, the features which did not evidence alignment in the current experiment (but which have been demonstrated in prior experiments of varying types), might be ones which are less relevant as part of a partner-specific linguistic profile but are more susceptible to transient priming activation and recency effects. It is difficult to draw conclusive evidence across studies with many different paradigms (separated exposure and test vs. trial-to-trial priming, live interlocutor vs. recorded model talker, relatively free speech vs. heavily constrained speech, etc.) but this is a question that should be explored in future research.

Another important difference from much prior work is that the present experiment employed interaction with multiple, alternating interlocutors, and a separated exposure and test phase. The Interactive Alignment Model proposes that alignment occurs to enhance communicative success by removing the necessity of separately maintaining both your own representation of the linguistic situation and also your partner's, because if the two representations match, there is no need to model your listener. Most prior work on alignment and the lexical boost, however, has investigated trial-to-trial priming, finding that the current trial's syntactic structure is influenced by the immediately preceding trial's syntactic structure and verb choice. An experimental design with only one interlocutor, in which alignment is assessed as whether the participant's production matches the immediately preceding production, cannot test whether a speaker learns and maintains representations of each of their listeners individually, and aligns their own representation to that of their *current* listener's. This is because a single-interlocutor, trial-to-trial priming design conflates the representation of the current listener with that of the current overall linguistic context, and thus any alignment observed could be due not to representing the listener's situation model, but merely increased activation of that linguistic feature due to recency. In contrast, the present work investigates partner-specific alignment *per se*: whether speakers modulate their speech so as to match the linguistic properties produced by their current listener *in particular*, irrespective of any intervening exposure. It does so by deconfounding the linguistic experience received from a particular listener and the linguistic experience from the overall context. This is because each participant interacts with two partners alternately throughout the experiment, and alignment is measured as the degree to which the participant spoke more like Experimenter A than B when addressing A, and more like Experimenter B than A when addressing B. This experimental design directly tests whether speakers learn a separate linguistic profile of each of their partners and then reflect that profile in their subsequent interactions with that partner uniquely.

4.2. Does alignment on one feature necessarily mean alignment on other features at the same level?

The second important implication of the current work is that speakers engaged in partner-specific alignment on some features at the acoustic level, but not other features at the same level. In particular, at the acoustic level, speakers did not show partner-specific alignment on five common spectral-phonetic features, but they did show partner-specific alignment on two common temporal features, and a holistic suite of acoustic features was jointly able to differentiate between speech to the two partners. As discussed in the Introduction, most research on vocal accommodation measures just one or a handful of acoustic features. The present work suggests that, when the goal is to investigate whether speakers align acoustically overall (as opposed to the much more specific question of whether speakers align on, e.g., F1 values of the vowel [ɑ]), this strategy may produce misleading results. There are substantial discrepancies in the results between different studies of acoustic alignment, likely due to the fact that different studies measure different features. The present work suggests that a valuable strategy for future research may be to use a wider range of acoustic features to jointly measure speakers' degree of alignment, to characterize at a more general level what types of features induce alignment in which linguistic contexts, and what types do not.

5. Conclusion

Within the same dialogue, speakers may align partner-specifically at some linguistic levels, and simultaneously not align at other levels. In the present experiment, participants aligned on temporal-phonetic measures, aligned on some spectral-phonetic measures but not others, and did not align on a syntactic measure. This suggests that, contrary to a common assumption across alignment research and a core claim of the Interactive Alignment Theory, alignment is not a unitary phenomenon which shifts in sync across the language production system, as speakers do not necessarily align different linguistic levels together. In addition, speakers may align on some features at a given level, but not on other features at the same level, and thus testing for alignment using only one or two features may miss the presence of alignment in a dialogue. Therefore, although alignment *can* occur at multiple

levels simultaneously, communicative success does not *require* alignment within and across levels in tandem.

CRediT authorship contribution statement

Rachel Ostrand: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Eleanor Chodroff:** Software, Formal analysis, Data curation, Writing - review & editing.

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.

Acknowledgements

Thanks to Coy Shaffstall for asking a great question which started this project, and to Coy Shaffstall and Kexin Chen for testing the participants, transcribing speech, and pretending to love math, kayaking, and purple. Thanks also to Vic Ferreira for helpful comments and the use of his lab, Matt Goldrick for good discussions and introducing the authors to each other, and Carla Agurto, Sara Berger, David Piorkowski, and Pablo Polosecki for assistance with aspects of the code. An early version of this work was presented at the Architectures and Mechanisms for Language Processing (AMLaP) conference in 2018. Running of the experiment was supported in part by a grant to Vic Ferreira from the National Institutes of Health [R01 HD051030].

References

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*, 437–456. https://doi.org/10.1017/s0047404510000400.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*(1), 177–189. https://doi.org/10.1016/j.wocn.2011.09.001.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer (6.0.50) [Computer software]. http://www.praat.org/.

Babel, M., & Bulatov, D. (2011). The role of fundamental frequency in phonetic accommodation. *Language and Speech*. https://doi.org/10.1177/0023830911417695.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human–computer interaction, ICPhS-15, 2453–2456. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_2453.pdf.

Bergmann, K., & Kopp, S. (2012). Gestural alignment in natural dialogue. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (pp. 1326–1331). Cognitive Science Society. https://escholarship.org/uc/item/73z0q063.

Bonin, F., de Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., & Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. INTERSPEECH-2013, 539–543. https://www.isca-speech.org/archive/interspeech_2013/i13_0539.html.

Borrie, S. A., Lubold, N., & Pon-Barry, H. (2015). Disordered speech disrupts conversational entrainment: A study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01187.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13–B25. https://doi.org/10.1016/s0010-0277(99)00081-5.

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition, 104*(2), 163–197. https://doi.org/10.1016/j.cognition.2006.05.006.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition, 121*(1), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482.

Cappella, J. N., & Planalp, S. (1981). Talk and silence sequences in informal conversations III: Interspeaker influence. *Human Communication Research, 7*(2), 117–132. https://doi.org/10.1111/j.1468-2958.1981.tb00564.x.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language, 49*(2), 214–230. https://doi.org/10.1016/S0749-596X(03)00060-3.

Cohen Priva, U., & Sanker, C. (2018). Distinct behaviors in convergence across measures. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 1518–1523). Cognitive Science Society. https://cogsci.mindmodeling.org/2018/papers/0294/index.html.

Conrad, L. (1989). The effects of time-compressed speech on Native and EFL listening comprehension. *Studies in Second Language Acquisition, 11*(1), 1–16. https://doi.org/10.1017/S0272263100007804.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human−computer dialogue. *International Journal of Human-Computer Studies, 83*, 27–42. https://doi.org/10.1016/j.ijhcs.2015.05.008.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385.

de Looze, C., Oertel, C., Rauzy, S., & Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. *ICPhS, 2011*, 1294–1297 https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/de%20Looze/de%20Looze.pdf.

de Looze, C., & Rauzy, S. (2011). Measuring Speakers' similarity in speech by means of prosodic cues: Methods and potential. *INTERSPEECH, 2011*, 1393–1396 https://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_1393.pdf.

Dellwo, V. (2019). Praat script: Duration Analyzer (0.03) [Computer software]. https://www.pholab.uzh.ch/static/volker/software/plugin_duratio nAnalyzer.zip.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America, 137*(3), 1513–1528. https://doi.org/10.1121/1.4906837.

Derwing, T. M. (1990). Speech rate is no simple matter: Rate adjustment and NS–NNS communicative success. *Studies in Second Language Acquisition, 12*(3), 303–313. https://doi.org/10.1017/S0272263100009189.

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception, 40*(12), 1457–1466. https://doi.org/10.1068/p7071.

Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and Gap Length in Face-to-Face Interaction. INTERSPEECH 2009, 2779–2782. https://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_2779.pdf.

Fernandes, J., Teixeira, F., Guedes, V., Junior, A., & Teixeira, J. P. (2018). Harmonic to noise ratio measurement—Selection of window and length. *Procedia Computer Science, 138*, 280–285. https://doi.org/10.1016/j.procs.2018.10.040.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America, 84*(1), 115–123. https://doi.org/10.1121/1.396977.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*(2), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, J., & Coupland, N. (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics* (Studies in Emotion and Social Interaction (pp. 1–68). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511663673.001.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. https://doi.org/10.1037/0033-295x.105.2.251.

Gregory, S. W., Dagan, K., & Webster, S. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior, 21*(1), 23–43. https://doi.org/10.1023/A:1024995717773.

Gregory, S. W., & Hoyt, B. R. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psycholinguistic Research, 11*(1), 35–46. https://doi.org/10.1007/BF01067500.

Gregory, S. W., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology, 70*(6), 1231–1240. https://doi.org/10.1037/0022-3514.70.6.1231.

Gregory, S. W., Webster, S., & Huang, G. (1993). Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language & Communication, 13*(3), 195–217. https://doi.org/10.1016/0271-5309(93)90026-j.

Gruberg, N., Ostrand, R., Momma, S., & Ferreira, V. S. (2019). Syntactic entrainment: The repetition of syntactic structures in event descriptions. *Journal of Memory and Language, 107*, 216–232. https://doi.org/10.1016/j.jml.2019.04.005.

Haywood, S. L., Pickering, M. J., & Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue?. *Psychological Science, 16*(5), 362–366. https://doi.org/10.1111/j.0956-7976.2005.01541.x.

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior, 35*(2), 133–153. https://doi.org/10.1007/s10919-011-0105-6.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition, 96*(2), 127–142. https://doi.org/10.1016/j.cognition.2004.07.001.

Hwang, J., Brennan, S. E., & Huffman, M. K. (2015). Phonetic adaptation in non-native spoken dialogue: Effects of priming and audience design. *Journal of Memory and Language, 81*, 72–90. https://doi.org/10.1016/j.jml.2015.01.001.

Iskarous, K., Shadle, C. H., & Proctor, M. I. (2011). Articulatory–acoustic kinematics: The production of American English /s/. *The Journal of the Acoustical Society of America, 129*(2), 944–954. https://doi.org/10.1121/1.3514537.

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition, 35*(5), 925–937. https://doi.org/10.3758/BF03193466.

Kaschak, M. P., Kutta, T. J., & Schatschneider, C. (2011). Long-term cumulative structural priming persists for (at least) one week. *Memory & Cognition, 39*(3), 381–388. https://doi.org/10.3758/s13421-010-0042-3.

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition, 99*(3), B73–B82. https://doi.org/10.1016/j.cognition.2005.07.002.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior, 32*(2), 123–131. https://doi.org/10.1007/s10919-007-0044-4.

Lee, C.-C., Katsamanis, A., Black, M. P., Baucom, B. R., Christensen, A., Georgiou, P. G., & Narayanan, S. S. (2014). Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech & Language, 28*(2), 518–539. https://doi.org/10.1016/j.csl.2012.06.006.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*(1), 78–106. https://doi.org/10.1016/0010-0285(82)90005-6.

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Interspeech, 2011*. https://doi.org/10.7916/D8V12D8F.

Lyons, J., Wang, D. Y.-B., Gianluca, Shteingart, H., Mavrinac, E., Gaurkar, Y., Watcharawisetkul, W., Birch, S., Zhihe, L., Hölzl, J., Lesinskis, J., Almér, H., Lord, C., & Stark, A. (2020). jameslyons/python_speech_features: Release v0.6.1 (0.6.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3607820.

Matarazzo, J. D., Weitman, M., Saslow, G., & Wiens, A. N. (1963). Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior, 1*(6), 451–458. https://doi.org/10.1016/s0022-5371(63)80031-6.

Matarazzo, J. D., Wiens, A. N., Saslow, G., Dunham, R. M., & Voas, R. B. (1964). Speech durations of astronaut and ground communicator. *Science, 143*(3602), 148–150. https://doi.org/10.1126/science.143.3602.148.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. INTERSPEECH 2017, 498–502. https://doi.org/10.21437/Interspeech.2017-1386.

Mukherjee, S., D'Ausilio, A., Nguyen, N., Fadiga, L., & Badino, L. (2017). The relationship between F0 synchrony and speech convergence in dyadic interaction. INTERSPEECH 2017, 2341–2345. https://doi.org/10.21437/Interspeech.2017-795.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology, 32*(5), 790–804. https://doi.org/10.1037/0022-3514.32.5.790.

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High Frequency Word Entrainment in Spoken Dialogue. Proceedings of ACL-08: HLT, Short Papers, 169–172. https://www.aclweb.org/anthology/P08-2043.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*(2), 132–142. https://doi.org/10.1016/j.wocn.2010.12.007.

Ostrand, R., & Ferreira, V. S. (2019). Repeat after us: Syntactic alignment is not partner-specific. *Journal of Memory and Language, 108*. https://doi.org/10.1016/j.jml.2019.104037.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*(4), 2382–2393. https://doi.org/10.1121/1.2178720.

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00559.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics, 40*(1), 190–197. https://doi.org/10.1016/j.wocn.2011.10.001.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics, 72*(8), 2254–2264. https://doi.org/10.3758/BF03196699.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language, 69*(3), 183–195. https://doi.org/10.1016/j.jml.2013.06.002.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics, 79*(2), 637–659. https://doi.org/10.3758/s13414-016-1226-0.

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. *Journal of Phonetics, 69*, 1–11. https://doi.org/10.1016/j.wocn.2018.04.001.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(02), 169–190. https://doi.org/10.1017/s0140525x04000056.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., & Yu, M. (2017). Entrainment in multi-party spoken dialogues at multiple linguistic levels. Interspeech 2017, 1696–1700. https://doi.org/10.21437/Interspeech.2017-1568.

Reidy, P. F. (2015). A comparison of spectral estimation methods for the analysis of sibilant fricatives. *The Journal of the Acoustical Society of America, 137*(4), EL248–EL254. https://doi.org/10.1121/1.4915064.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language, 76*, 29–46. https://doi.org/10.1016/j.jml.2014.05.008.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 10*(1), 1. https://doi.org/10.5334/labphon.147.

Rosenthal-von der Pütten, A. M., Wiering, L., & Krämer, N. (2013). Great minds think alike. Experimental study on lexical alignment in human-agent interaction. *I-Com, 12*(1), 32–38. https://doi.org/10.1524/icom.2013.0005.

Scarborough, R., Brenier, J., Zhao, Y., Hall-Lew, L., & Dmitrieva, O. (2007). An acoustic study of real and imagined foreigner-directed speech. *The International Congress of Phonetic Sciences*, 2165–2168. https://doi.org/10.1121/1.4781735.

Schultz, B. G., O'Brien, I., Phillips, N., McFarland, D. H., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics, 37*(5), 1201–1220. https://doi.org/10.1017/S0142716415000545.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66*(3), 422–429. https://doi.org/10.3758/BF03194890.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Skodda, S., Grönheit, W., & Schlegel, U. (2012). Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease. *PLoS ONE, 7*(2). https://doi.org/10.1371/journal.pone.0032132.

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In Ohlsson, S., & Catrambone, R. (Eds.), Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (pp. 127–132). Austin, TX: Cognitive Science Society, https://escholarship.org/uc/item/3vg3g1ds.

Street, R. L. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research, 11*(2), 139–169. https://doi.org/10.1111/j.1468-2958.1984.tb00043.x.

Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science, 19*(2), 131–141. https://doi.org/10.1080/09540090701369125.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis – Jitter, Shimmer and HNR parameters. *Procedia Technology, 9*, 1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124.

ten Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In Sojka, P., Kopeček, I., & Pala, K. (Eds.), Text, speech and dialogue (pp. 563–570). Springer. https://doi.org/10.1007/978-3-540-30120-2_71.

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication, 49*(1), 2–7. https://doi.org/10.1016/j.specom.2006.10.003.

Webb, J. T. (1969). Subject speech rates as a function of interviewer behaviour. *Language and Speech, 12*(1), 54–67. https://doi.org/10.1177/002383096901200105.

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), pp. 297–302. https://www.aclweb.org/anthology/N18-2048.

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40*(4), 919–937. https://doi.org/10.1037/a0036161.