JSLHR

Research Article

Syncing Up for a Good Conversation: A Clinically Meaningful Methodology for Capturing Conversational Entrainment in the Speech Domain

Stephanie A. Borrie,[a] Tyson S. Barrett,[b] Megan M. Willi,[c] and Visar Berisha[d,e]

Purpose: Conversational entrainment, the phenomenon whereby communication partners synchronize their behavior, is considered essential for productive and fulfilling conversation. Lack of entrainment could, therefore, negatively impact conversational success. Although studied in many disciplines, entrainment has received limited attention in the field of speech-language pathology, where its implications may have direct clinical relevance.

Method: A novel computational methodology, informed by expert clinical assessment of conversation, was developed to investigate conversational entrainment across multiple speech dimensions in a corpus of experimentally elicited conversations involving healthy participants. The predictive relationship between the methodology output and an objective measure of conversational success, communicative efficiency, was then examined.

Results: Using a real versus sham validation procedure, we find evidence of sustained entrainment in rhythmic, articulatory, and phonatory dimensions of speech. We further validate the methodology, showing that models built on speech signal entrainment measures consistently outperform models built on nonentrained speech signal measures in predicting communicative efficiency of the conversations.

Conclusions: A multidimensional, clinically meaningful methodology for capturing conversational entrainment, validated in healthy populations, has implications for disciplines such as speech-language pathology where conversational entrainment represents a critical knowledge gap in the field, as well as a potential target for remediation.

On the surface, conversation is seemingly simple. There are two roles, talking and listening, and conversational partners must alternate between these roles as a message is exchanged. However, successful conversation, it appears, is a much more complex interactional event that requires the coordination or syncing up of behavior (Clark, 1996). This behavioral synchronization, referred here as *conversational entrainment,*[1] describes a pervasive communication phenomenon in which conversational partners subconsciously align their communicative actions with one another. Operationally defined as "spatiotemporal coordination resulting from rhythmic responsiveness to a perceived rhythmic signal" (Phillips-Silver, Aktipis, & Bryant, 2010, p. 5), conversational partners must perceive the behavioral patterns of one another and adjust their own accordingly. This adjusting and aligning of patterned behavior has been observed in verbal (e.g., acoustic prosodic speech features, Lee et al., 2014; syntactic structure, Branigan, Pickering, & Cleland, 2000; lexical use, Kawabata, Berisha, Scaglione, & LaCross, 2016) and nonverbal (e.g., eye movements, Richardson & Dale, 2005; body posture, Shockley, Santana, & Fowler, 2003) aspects of communication.

[a]Department of Communicative Disorders and Deaf Education, Utah State University, Logan
[b]Department of Kinesiology and Health Sciences, Utah State University, Logan
[c]Department of Communication Sciences and Disorders, California State University, Chico
[d]Department of Speech and Hearing Science, Arizona State University, Tempe
[e]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe

Correspondence to Stephanie A. Borrie: stephanie.borrie@usu.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Michelle Ciucci

Received May 26, 2018
Revision received August 22, 2018
Accepted October 31, 2018
https://doi.org/10.1044/2018_JSLHR-S-18-0210

[1] Other terms that have been used to describe this communication coordination phenomenon include *accommodation*, *alignment*, *convergence*, and *synchronization*.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

There is a well-established body of literature correlatively linking entrainment of verbal and nonverbal behavior to productive and fulfilling peer interactions, advancing the idea that conversational entrainment serves cognitive and pragmatic functions essential for successful conversation. A number of studies, for example, show that entrainment of verbal behaviors such as word use (lexical entrainment; e.g., Nenkova, Gravano, & Hirschberg, 2008) or pitch properties (acoustic–prosodic entrainment; e.g., Borrie, Lubold, & Pon-Barry, 2015) track with increased efficiency and performance in tasks that require conversational partners to use verbal communication to achieve goals. Pickering and Garrod (2004) have advanced the idea that aligned or entrained behavior during conversation underlies tightly coupled production and comprehension processes, and that this coupling greatly reduces the computational load of language processing in spoken dialogue. Furthermore, alignment of behavior during conversation supports predictive processes, whereby conversational partners can track and anticipate upcoming realizations of speech (see Pickering & Garrod, 2013, for further details). Entrainment is also considered key for supporting important pragmatic aspects of conversation, including taking turns, interaction smoothness, building rapport, fostering social bonds, and maintaining interpersonal relationships (e.g., Bailenson & Yee, 2005; Chartrand & Bargh, 1999; Lee et al., 2014; Wilson & Wilson, 2005). Furthermore, pragmatic benefits of entrainment may also extend to human–machine communication (Levitan et al., 2016), although the evidence in spoken dialogue systems is, to date, somewhat inconclusive (e.g., Beňuš et al., 2018). In noting some of the key literature in the area of entrainment and its functional utility in human–human communication, it has been concluded that conversational entrainment "… serves as a powerful coordinating device, uniting individuals in time and space to optimize comprehension, establish social presence, and create positive and satisfying relationships" (Borrie & Liss, 2014, p. 816; see also Beňuš, 2014, for a review of social aspects of entrainment). Thus, lack of entrainment or inherent entrainment deficits could impact the success of conversation, contributing to social isolation and diminished quality of life.

Although conversational entrainment has been studied widely across many disciplines, it has received limited attention in the field of speech pathology, where its implications may have direct clinical relevance. Borrie and Liss (2014) recently proposed that any deficit in the ability to produce, perceive, or modify rhythmic behavior will impact entrainment and conversational success. Given that rhythm impairments are pervasive in many populations with communication disorders, entrainment deficits are likely widespread in the field. Indeed, our preliminary work in this area has confirmed that speech entrainment deficits exist in adults with dysarthria (Borrie et al., 2015) and in adults with autism spectrum disorder (Wynn, Borrie, & Sellars, 2018). For example, using a small selection of basic acoustic–prosodic features and simple local, turn-by-turn entrainment measures, Borrie et al. (2015) observed that entrainment of pitch, intensity, and jitter was significantly lower in conversations involving a person with dysarthria conversing with a healthy partner relative to conversations involving two healthy conversational partners. Furthermore, conversations characterized by lower levels of entrainment were correlated with lower levels of communicative efficiency, as measured by task success in goal-oriented dialogues.[2] Thus, it appears that entrainment deficits are both real and of potential consequence in clinical populations with communication disorders. In order to advance entrainment application in speech pathology, we must establish a clinically meaningful methodology for capturing conversational entrainment in the speech domain that, once validated in healthy populations, can be used to learn more about the phenomenon and how it is disrupted in the context of communication disorders.

Critically, a methodology for capturing conversational entrainment in the speech domain for investigating deficit and consequence should be informed by real-world clinical evidence of successful and unsuccessful conversation. The consideration of clinical evidence in objective measurement of entrainment for the purpose of exploring functionality is further motivated by a handful of studies, which have postulated that pragmatic elements of successful conversation may be linked to both similarity and dissimilarity of behavior (De Looze, Scherer, Vaughan, & Campbell, 2014; Reichel, Beňuš, & Màdy, 2018). Speech-language pathologists (SLPs) are professionals trained in the prevention, assessment, and treatment of communication disorders. They have skills and expertise in evaluating social communication and peer interaction, including the cognitive and pragmatic functions of conversation, such as information exchange, turn-taking and interaction cohesiveness, and rapport and connection. Thus, we advance that the expert clinical assessment of conversation afforded by SLPs is key to establishing meaningful measures of conversational entrainment in the speech domain.

Conversational entrainment has been evidenced in acoustic–prosodic features such as speaking rate (e.g., Local, 2007), fundamental frequency measures (e.g., Borrie et al., 2015; Duran & Fusaroli, 2017; Levitan & Hirschberg, 2011), vocal intensity (e.g., Local, 2007; Natale, 1975), voice quality (e.g., Borrie & Delfino, 2017; Levitan & Hirschberg, 2011), vowel spectra (e.g., Babel, 2012; Vallabha & Tuller, 2004), voice onset time (e.g., Fowler, Brown, Sabadini, & Weihing, 2003; Nielsen, 2011), and latency and utterance durations (e.g., Matarazzo, Weitman, Saslow, & Wiens, 1963; Matarazzo & Wiens, 1967). These studies using a single or small number of self-selected features afford important insight into the nature of entrainment in the speech domain. However, fewer studies have examined entrainment of speech behaviors in large feature sets that capture multiple dimensions of the speech signal (although

[2] The dialogue elicitation task required conversational partners to work together, as quickly and accurately as possible, to identify the differences between pairs of pictures. Task success was defined as the number of differences identified in 12 min of spoken dialogue.

for recent work involving larger feature sets, see Lee et al., 2014; Nasir, Baucom, Georgiou, & Narayanan, 2017). In a review of research on phonetic convergence (analogous to speech entrainment), Pardo (2013) comments that, “there is currently no compelling rationale or standard for choosing one acoustic attribute over another” (p. 559) and that characterization of the phenomenon must ultimately include multiple dimensions. Thus, a methodology for capturing conversational entrainment in the speech domain should characterize the communication phenomenon across a broad range of acoustic–prosodic features, spanning rhythmic (e.g., signal envelope), articulatory (e.g., spectral features), and phonatory (e.g., pitch properties) dimensions of speech.

This Study

The purpose of this study was to build and validate a clinically meaningful methodology to capture conversational entrainment in multiple dimensions of speech, using a novel computational approach involving expert clinical assessment of conversation and a corpus of experimentally elicited, goal-oriented conversations involving healthy participants. Our methodology begins with extracting large acoustic–prosodic feature sets that represent rhythmic, phonatory, and articulatory dimensions of speech. These feature sets are then reduced, retaining shared information among the individual acoustic–prosodic behaviors. We then use the expert conversation assessments from five SLPs and cross-recurrence quantification analysis (CRQA), a nonlinear technique that allows us to quantify shared organization of behavior over time (Coco & Dale, 2014; Zbilut, Giuliani, & Webber, 1998), to capture global entrainment in the speech domain. Using this methodology, our first research question asked: Which dimensions of the speech signal are entrained during spoken dialogue? To address this question, we compare real conversations with a sham conversational corpus, constructed of randomly generated dialogs between not-in-conversation partners. If acoustic–prosodic features of the speech signal are really entrained during conversations, then entrainment values should be higher in the real conversational corpus as compared to the sham conversational corpus. This analysis also serves as verification that our computational methodology involving automatic acoustic–prosodic feature extraction methods, feature reduction techniques, recurrence quantification, and expert clinical assessment of conversation is sufficiently robust to capture conversational entrainment in speech signal dimensions. Our second research question asked: Do measures of speech signal entrainment predict an objective measure of conversational success, communicative efficiency? To address this question, we use a series of machine learning approaches to detail the predictive relationship between measures of speech signal entrainment and a measure of communicative efficiency derived from task success in the goal-oriented conversations. We provide additional support for this predictive relationship by modeling the nonentrained speech signal measures (i.e., measures that do not differentiate between real and sham conversations). If it really is the alignment of speech signal behavior that facilitates communicative efficiency in conversation, then models built on speech signal entrainment measures should outperform models built on nonentrained measures. A validated methodology for characterizing conversational entrainment in the speech domain lays the groundwork for entrainment application in clinical settings.

Method

Participants

This study is based on a corpus of 57 experimentally elicited conversations, involving 114 participants (99 women and 15 men) aged 19–28 years old (M = 22.41) engaged in university-level education. All participants were native speakers of American English with no self-reported history of speech, language, hearing, or cognitive impairment. Participants were paired up, at random, to form a dyad and partake in a conversational task. Note that gender was not controlled for when forming dyads, so some dyads were female–female (n = 43), other dyads were female–male (n = 13), and one dyad was male–male (n = 1).

Conversational Task

Each dyad participated in a single recording session. Conversational partners were seated facing one another and fitted with wireless CVL Lavalier microphones, synced with a Shure BLX188 DUAL Lavalier System connected to a Zoom H4N Portable Digital Recorder. Separate audio channels for each conversational partner and standard settings (48 kHz; 16-bit sampling rate) were employed for audio recording of the conversational task.

The conversation task was based on the Diapix task,[3] a collaborative “spot-the-difference” task whereby dyads must work together, verbally comparing scenes, to identify differences between sets of pictures (Van Engen et al., 2010). Each partner in the dyad was given one of a pair of pictures and instructed to hold their picture at an angle at which it was not visible to their partner sitting across the table from them. The pair of pictures depict virtually identical scenes (e.g., yard, beach), differing from one another by 10 small details (e.g., number of people, color of t-shirt). The dyad was told that their goal was to work together, simply by speaking to one another, to identify the 10 differences between the pair of pictures as accurately and as quickly as possible. When all the differences were identified, the dyad was given another pair of pictures to work through. Dyads were tasked with working through as many pairs of pictures as possible in a 10-min time frame. Total recording time from each dyad was, therefore, 10 min. No additional rules (i.e., who could talk when) or roles (i.e., giver, receiver) were given so dyads were free to verbally interact in any way they saw fit to problem-solve

[3] For more details on the Diapix, see http://groups.linguistics.northwestern.edu/speech_comm_group/diapix/

the task. The task is not considered cognitively demanding; however, conversational partners must work together to be successful.

Expert Clinical Assessment of Conversation

SLPs work with people with communication disorders and are often called upon to make judgments about what constitutes a conversational breakdown (Garcia & Orange, 1996). Using a 7-point Likert-type rating scale (1 = *strongly disagree*, 4 = *neutral*, 7 = *strongly agree*), five SLPs assessed each of the 57 conversational recordings according to the extent that they agreed with the following statement: The conversation pair sound like they are in-sync or aligned with one another, with high ratings (scores above 4) indicative of a natural cohesiveness to the interaction, smooth turn-taking and conversational flow, and a sense of rapport and connection between conversational participants and low ratings (scores below 4) indicative of an awkward, disconnected, and disengaged interaction (see the Appendix). Thus, this score reflects expert clinical assessment of conversational success, also indicative of a holistic impression of conversational entrainment. The clinicians were required to listen to the first 2 min of a conversation before making their assessment rating.[4] An expert clinical assessment score was calculated for each of the 57 recordings by averaging the individual ratings across the five SLPs.

Measure of Communicative Efficiency

The dialogue elicitation tool, the Diapix task, grants us an objective measure of an aspect of conversational success, communicative efficiency.[5] This objective measure has been previously observed to correlate with measures of acoustic–prosodic entrainment (Borrie & Delfino, 2017; Borrie et al., 2015; Willi, Borrie, Barrett, Tu, & Berisha, 2018). Recall that the Diapix task required the conversational partners to work together as accurately and quickly as possible to identify the differences between pairs of pictures. Total number of differences identified in the 10-min recording was then used as a simple, gross measure of communicative efficiency: Relatively low and high numbers of identified differences indicate relatively low and high communicative efficiency, respectively. The measure of communicative efficiency is, therefore, an objective evaluation of how proficiently the dyad used verbal communication to collaboratively work through the demands of the goal-oriented dialogue task.

Feature Extraction

Trained research assistants manually coded each 10-min audio file, annotating individual spoken utterances by speaker using the Praat textgrid function (Boersma & Weenink, 2017).[6] A spoken utterance is defined as a pause-free unit of speech, where pauses are greater than 50 ms, from a single speaker. Thus, pauses less than 50 ms are included in the spoken utterance. This definition of spoken utterance is the same as interpausal units (Levitan & Hirschberg, 2011). Simultaneous illustrations of the associated spectrograms were used to aid coding accuracy. All audio files were normalized using a reference level and down-sampled to 16 kHz prior to feature extraction.

Five speech feature subsets, spanning rhythmic (envelope modulation spectrum [EMS], rhythm metrics), articulatory (long-term average spectrum [LTAS], mel-frequency cepstral coefficients [MFCCs]), and phonatory (voice report) dimensions of the speech signal were extracted from each spoken utterance. Each feature subset included a number of acoustic–prosodic features, which resulted in a 429-feature vector for each utterance. Similar speech feature extraction methods have been reported previously (Berisha, Liss, Sandoval, Utianski, & Spanias, 2014; Tu, Jiao, Berisha, & Liss, 2016; Tu, Berisha, & Liss, 2017; Willi et al., 2018). Specific feature subsets are described briefly below, but please refer to the following link for comprehensive calculation details, http://www.public.asu.edu/~visar/IS2018Supp.pdf.

EMS

The EMS feature subset is made up of 60 features related to rhythmic dimensions of speech. Specifically, EMS is a spectral analysis of the low-rate amplitude modulations in the speech signal, with measures that capture modulations within the entire speech signal envelope and within specific frequency bands. These modulation measures provide information related to temporal regularities in speech and have been shown to significantly correlate with acoustic vocalic and consonantal segmental rhythm metrics (Liss, LeGendre, & Lotto, 2010).[7]

As per Liss et al. (2010), the EMS features are calculated by obtaining the amplitude envelopes for the original speech signal and nine octave bands with center frequencies of approximately 30, 60, 120, 480, 960, 1920, 3840, and 7680 Hz using eight-order Butterworth filters. Then, the mean of each amplitude envelope is removed, and the power spectra for each signal were calculated. Finally, six EMS metrics were computed from each power spectra (i.e., the nine octave bands and full signal), resulting in a 60-dimensional speech feature vector.

[4] Two minutes was selected to enable clinicians to evaluate 57 conversations within a reasonable time frame. Although all SLPs agreed that 2 min was ample time to evaluate a conversation, we acknowledge that the evaluation may change over the course of the conversation as partners become familiar with one another.

[5] Communicative efficiency is operationally defined as "increasing the rate of communication without sacrificing intelligibility or comprehensibility" (Duffy, 2015, p. 386).

[6] Each speaker channel was coded for all spoken utterances, regardless of whether the utterance was "talked over."

[7] Additional research is necessary to determine whether functional or perceptual significance can be assigned to amplitude modulation within specific frequency bands.

Rhythm Metrics

The rhythm metrics feature subset is made up of 12 features related to rhythmic dimensions of speech. Specifically, rhythm metrics is an analysis of voice timing based on voiced and voiceless interval durations. These duration-based measures provide information related to syllable structure and stress patterns and have been shown to capture speech rhythm differences between and within languages (e.g., Dellwo & Fourcin, 2013; White & Mattys, 2007) and populations with rhythmic speech disorders (Liss et al., 2009).[8]

The rhythm metric features were calculated using a Praat script, based on the periodicity detection algorithm outlined in Boersma (1993). Voiced and unvoiced intervals were extracted using a 5-ms time step and default parameters for pitch floor, pitch ceiling, silence threshold, and voicing threshold. The script used the pitch track function to assess voicing on a frame-by-frame basis by using an autocorrelation-based method to estimate the periodicity and assuming that the speech signal is unvoiced when the pitch track is undefined. Details of the pitch estimation algorithm used to partition the speech signal into voiced and unvoiced segments are described fully in Boersma (1993). The duration of the vocalic and intervocalic segments and a series of related metrics were extracted, resulting in a 12-dimensional speech feature vector.

LTAS

The LTAS feature subset is made up of 99 features broadly related to articulatory dimensions of speech. Specifically, LTAS is an analysis of the average energy distribution across frequency over an utterance. These measures have been shown to capture differences in speaker gender and age, as well as professional and dysphonic voices (e.g., Cleveland, Sundberg, & Stone, 2001; Linville, 2002; Mendoza, Valencia, Muñoz, & Trujillo, 1996).

The LTAS features were calculated by obtaining the average spectral information for the original speech signal and nine octave bands with center frequencies described in the EMS feature extraction. The 10 band signals (the original full-band signal and nine octave band signals) were then framed using a 20-ms nonoverlapping rectangular window, and the root-mean-square of each frame is estimated. Ten features were extracted for each of the signals, resulting in a 99-dimensional speech feature vector.

MFCCs

The MFCCs feature subset is made up of 234 features broadly related to articulatory dimensions of speech. Specifically, MFCC is an analysis of coefficients that represent the short-term power spectrum of a speech segment. These coefficient measures, introduced by Davis and Mermelstein (1990), have been widely used in automatic speech recognition (e.g., Martin & Jurafsky, 2000). The lower order cepstral coefficients relate to the frequency response of the vocal tract, and the higher order cepstral coefficients relate to the frequency spectrum of the source signal. Speaker-dependent characteristics can be suppressed by only processing the lower order cepstral coefficients.

Based on a standard power spectrum estimate, the MFCC features were first subjected to a log-based transform of the frequency axis (mel-frequency scale) and then decorrelated by using an inverse discrete cosine transform. We then calculated the coefficients from the 13th-order MFCCs (including 0th order) and their first- and second-order derivatives using a 20-ms window with 10-ms frame increment. Then, the mean, standard deviation, range, skewness, kurtosis, and mean absolute deviations were calculated for each coefficient or derivative feature, resulting in a 234-dimensional speech feature vector.

Voice Report

The voice report feature subset is made up of 24 individual features related to phonatory dimensions of speech. The features, extracted using a custom Praat script, included fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio, affording information about pitch, cycle-to-cycle pitch variation, cycle-to-cycle amplitude variation, and an estimate of the noise level in the human voice, respectively. The phonatory features were extracted using a 5-ms time step and default parameters for pitch floor, pitch ceiling, silence threshold, and voicing threshold. Additional phonatory features and measures of central tendency and variation were also included in the voice report feature set, resulting in a 24-dimensional acoustic–prosodic feature vector.

Feature Reduction

As described above, each speech feature set is made up of a number of features. We employed independent components analysis (ICA) to reduce the dimensionality of the feature sets (Comon, 1992; Marchini, Heaton, & Ripley, 2017) for the entrainment analysis. As a close relative of principle components analysis, ICA aims to establish a variable that represents the highest amount of the shared variance across the individual features in a set. Each feature set had a high amount of shared variability (all Cronbach's αs > .70), suggesting that the ICA captured a high degree of the original variability across the features. We used ICA to perform feature reduction for the five feature sets: EMS (using the five features—of the total 60—relating to a center frequency of 480; this is likely to capture the rhythmic patterns associated with changes in vowel energy), rhythm metrics (using all 12 features), LTAS (using all 99 features), MFCC (using all 234 features), and voice report (using all 24 features). Thus, this produced five variables, representing rhythmic (EMS, rhythm metrics), articulatory (LTAS, MFCC), and phonatory (voice report) speech dimensions, to be used in the entrainment analyses.

[8] Voicing intervals provide a recurring, suprasegmental temporal measure, which for simplicity reasons are referred to as *rhythm metrics*. It is, however, acknowledged that these measures may not provide a comprehensive model of speech rhythm.

Speech Signal Entrainment Analysis

We used CRQA to objectively quantify speech signal entrainment in conversations. CRQA is a nonlinear time-series technique that evaluates instances or points in time in which two different streams of the same type of information (i.e., rhythmic speech features of conversational partners' spoken utterances) visit similar states, termed *recurrence*. The approach quantifies how and to what extent recurrence of behavior occurs over time (Coco, Dale, & Keller, 2017; see Duran & Fusaroli, 2017, for application of CRQA to speech behavior in a conversational paradigm). To do this, CRQA produces a "recurrence plot" that marks all points wherein the two systems, the conversational partners, were aligned at each possible time point. Using this information, several measures were quantified:

1. *Recurrence rate* is defined as the number of single instances of alignment between conversational partners accounting for the number of turns taken over the entire conversation. Higher recurrence rate values indicate higher amounts of single-instance entrainment.
2. *Sustained recurrence* is defined as the amount of alignment between conversational partners that is maintained for longer than a single instance (also referred to as *nline*). Higher sustained recurrence values indicate higher amounts of sustained entrainment.
3. *Length* is defined as the average length/time that conversational partners stay aligned with one another. Higher length values indicate that the dyad maintained entrainment for longer stretches, on average.
4. *Max length* is defined as the longest length/uninterrupted time that conversational partners stay aligned with one another. Higher max length values indicate longer periods of entrainment.
5. *Entropy* is defined as the variability in length/time that conversational partners are entrained with one another. Higher entropy values indicate that the time in which the dyad entrained vary more widely across the conversation.

Several aspects of CRQA make it particularly useful for understanding conversational dynamics. First, it is a nonlinear approach. As opposed to other measures traditionally used to measure entrainment (e.g., correlation between adjacent speaking turns, aggregation across the conversation), this allows for far greater flexibility in the types and complexity of repeated patterns that the model can capture. Second, it analyzes the entire conversation simultaneously. That is, all possible lags (i.e., all possible time delays between the conversational partners) are considered in the quantification. In this way, we make no assumptions about who is "leading" the conversation and at what temporal scale entrainment occurs. Third, CRQA produces several interpretable measures that summarize not only the amount of entrainment but also stability and complexity of the entrained behavior (as described previously). This approach has been validated as a robust method to capture conversational entrainment or alignment in a wide variety of verbal and nonverbal communicative behaviors, including speech rate (Duran & Fusaroli, 2017) and eye gaze (Coco et al., 2017; see Fusaroli, Konvalinka, & Wallot, 2014, for a review of CRQA application in social interaction). We implement the CRQA analysis using the "crqa" package in the R statistical environment (crqa package Version 1.0.6 and R Version 3.5.1; Coco & Dale, 2014; R Core Team, 2018) and compare the results using the "furniture" package (Version 1.7.13; Barrett & Brignone, 2017).

Parameter Settings: Integrating Expert Clinical Assessment

A key aspect of using CRQA regards selecting a number of parameters it uses to calculate the measures. Many of the parameters are intuitive (e.g., the normalization parameter specifies whether the CRQA analysis is performed on the original acoustic features or a *z*-scored version of the features). However, some of the parameters do not have intuitive values as they require prior knowledge regarding the temporal scale at which entrainment occurs or how similar features between two individuals should be to be considered entrained. Specifically, the delay and radius parameters are often unknown and can impact the resulting CRQA measures. Delay refers to the interval necessary for the conversational partners to be optimally aligned. Changes to delay often do not have a large impact on the resulting CRQA measures. Radius, on the other hand, refers to the threshold at which the two states are considered "similar enough" and has a meaningful impact on the resulting measures.

Some previous work has reported on an algorithm that seeks for the parameters that produce a recurrence rate between 3% and 5% (Coco & Dale, 2014). However, this approach makes it difficult to compare real and sham conversations (described below) on recurrence rate (any differences are due to chance given the algorithm is arbitrarily stopping between 3% and 5%). To find meaningful parameters that are not driven by obtaining an arbitrary, prespecified recurrence rate, we sought to use a data-driven approach wherein the selected parameters would be based on having high predictive power of the expert clinical assessment of conversation, thus affording the optimal CRQA parameters that align with real-world evidence of successful and unsuccessful communication.

Using 10-fold cross-validated *k*-nearest neighbor models, the parameters (delay and radius) that produced the highest prediction accuracy of expert clinical assessment were selected across a wide range of possible values. The tested values were based on the algorithm used by Duran and Fusaroli (2017) with the values tested being both above and below the algorithmically produced values. Notably, we did not try multiple embedding values given the algorithm consistently produced a single value across all the conversations and dimensions. This data-driven parameter selection approach is built on approaches common in machine learning (Hastie, Tibshirani, & Friedman, 2009) and allows the CRQA to have parameters that can be used across all the conversations (both real and sham). The delay–radius combinations that resulted in the highest cross-validated accuracies are reported in Table 1. These parameters were then used to produce the CRQA measures for each conversation and feature set.

Sham Conversations

To test which of the various CRQA measures based on the optimal parameters are considered entrained, we created a corpus of sham conversations to disrupt interdependent behaviors that should transpire between in-conversation partners. The sham corpus was created by randomly generating conversations between participants who did not converse with one another, maintaining the same female–male ratio of the original dyads. If our measures of entrainment really capture the interdependent coordinated behavior that presumably occurs during real conversation, then recurrence output should be significantly greater in conversations between two in-conversation partners (real) versus two out-of-conversation partners (sham). The use of sham conversations (also termed *virtual pairs*) to validate measures of conversational entrainment is not new (e.g., Bernieri, Reznick, & Rosenthal, 1988; Duran & Fusaroli, 2017; Lee et al., 2014).

A total of 500 sham conversations, as opposed to 57 (to match the number of real conversations), were generated to reduce uncertainty in our estimates and more confidently test for differences across real and sham conversations. By increasing the number of shams, the Type II error rates are reduced without any inflation of Type I error rates. However, because of the 30 *t* tests (i.e., testing overall differences across the five measures and testing the differences across the 25 feature–measure combinations [e.g., Envelope modulation spectrum-Sustained ecurrence]), we used the Bonferoni adjustment (i.e., keeping the Type I error rate across all the comparisons at .05). Herein, that resulted in an alpha level of .0017. Thus, if the *p* value was less than .0017, the measure/feature–measure was deemed a verified measure of speech signal entrainment.

Table 1. Parameters (delay and radius) selected for each feature set based on the highest cross-validated prediction of clinical assessment of conversational entrainment.

Variable	Accuracy	Kappa	Delay–radius
Rhythm			
EMS	75.2	0.345	4–9
Rhythm metrics	76.0	0.411	16–9
Articulation			
MFCC	76.5	0.393	16–8
LTAS	69.6	0.122	5–9
Phonation			
Voice report	76.6	0.323	5–8

Note. EMS = envelope modulation spectrum; MFCC = mel-frequency cepstral coefficient; LTAS = long-term average spectrum.

Predictive Models of Communicative Efficiency

We examined if the speech signal measures (CRQA measures with parameters set by the expert clinical assessments) classified as entrained (i.e., measures significantly different between real and sham conversations) are predictive of an objective measure of conversational success, communicative efficiency. For validation, we examined the predictive value of speech signal measures classified as not entrained (i.e., measures not significantly different between real and sham conversations). For the predictive models, we used three dissimilar, yet common, machine learning techniques: Lasso (or elastic net) regression, support vector machines, and *k*-nearest neighbors (Hastie et al., 2009). Lasso regression is a linear approach built on linear regression that, given certain parameters, can both select important variables and handle high multicollinearity naturally. It has been used in a wide variety of situations, including predicting clinical judgments of the presence of speech disorders (Ballard et al., 2016). Support vector machines and *k*-nearest neighbors, on the other hand, are nonlinear approaches to the prediction task. Support vector machines project the data on a much larger subspace and identify a separating hyperplane in the new subspace; they have been recently used in classification tasks involving measures of acoustic–prosodic entrainment and functional conversational outcomes (Nasir et al., 2017; Willi et al., 2018). The *k*-nearest neighbors' classifier is a simpler approach in which classification is based on the closest *k* neighboring points (closeness based on Euclidean distances). Nonparametric classifiers, such as *k*-nearest neighbors, have the benefit of making no assumptions on the underlying distribution of the data (i.e., data need not follow a Gaussian distribution; Berisha, Wisler, Hero, & Spanias, 2016). The use of three different approaches to the classification task was motivated by the desire for confidence in not only model predictions but in the measures that are considered to drive these predictions.

The model-specific parameters of the machine learning approaches were selected based on 10-fold cross-validation, wherein many combinations of parameters were tested. Both the accuracy of 10-fold cross-validated prediction and the relative importance of each feature–measure were evaluated. Each model was assessed in R Version 3.5.1 using the "caret" package (Version 6.0-78; Kuhn, 2017).

Results

Expert Clinical Assessment of Conversation

Expert clinical assessment of conversation, according to SLP judgments of conversational success, in line with a holistic impression of conversational entrainment, ranged from 1 to 7, with 4 = *neutral conversation* (i.e., not successful enough to be considered "in sync" but not unsuccessful enough to be considered "not in sync"). Interrater reliability between the five clinician judgments across the conversational corpus was high (Cronbach's α = .92). This provides important validation that the SLPs in this study are reliable, with one another, at assessing conversational success as it relates to a holistic view of conversational entrainment.

Ratings from five SLPs were averaged to achieve a mean expert clinical assessment score for each of the 57 conversations. Using the average ratings, conversations could be classified as unsuccessful (scores below 4) and successful (scores above 4). This categorization resulted in 19 conversations classified as unsuccessful and 38 conversations classified as successful. As discussed previously, this clinical evidence was then used to set the delay and radius parameters of the CRQA measures for each feature set.

Entrained CRQA Measures

A summary of the CRQA output measures for both real and sham conversations can be found in Table 2 (results pooling over the feature sets) and Table 3 (results by individual feature set). The tables both report the *p* values from independent-samples *t* tests—adjusting for any instances of unequal variances—comparing the measures from the two types of conversations. Table 2 shows that, pooling the five feature sets, both sustained recurrence and max length were significantly different at the .0017 level (Bonferoni adjustment).

Table 3 shows objective evidence of entrainment in all feature sets studied (EMS, rhythm metrics, LTAS, MFCC, voice report). Across all five feature sets, measures of sustained recurrence were significantly higher in the real conversations than the sham conversations. These significant differences between real and sham conversations present evidence of two important concepts: (a) a distinction between measures that can be classified as measures of speech signal entrainment and measures that are not entrained and (b) validation that our computational approach involving automatic acoustic–prosodic feature extraction, feature reduction, recurrence quantification, and expert clinical assessment is able to capture evidence of conversational entrainment in multiple dimensions of speech behavior. Both aspects provide information necessary to investigate relationships between measures of speech signal entrainment and an objective measure of communicative efficiency. Notably, several features were approaching the conservative alpha level of .0017. Of these, max length for MFCC, rhythm metrics, and voice report was approaching significance ($p = .015$, $p = .038$, and $p = .048$, respectively). Furthermore, length and entropy in LTAS were approaching significance as well ($p = .016$ and $p = .023$, respectively). Thus, with more conversational samples, these feature–measures may also be considered measures of speech signal entrainment.

Table 2. Comparison of real conversations with sham conversations across the feature sets.

	Conversation		
	Real (*n* = 57)	**Sham (*n* = 500)**	
Measure	***M* (*SD*)**	***M* (*SD***	***p***
Entropy	0.234 (0.054)	0.218 (0.044)	.035
Length	2.072 (0.026)	2.061 (0.024)	.003
Max length	3.272 (0.324)	3.208 (0.285)	< .001
Sustained recurrence	127.884 (55.484)	83.312 (30.193)	< .001
Recurrence rate	6.096 (0.746)	5.953 (0.562)	.166

Note. *p* Value is based on independent *t* test adjusting for any unequal variances.

Predictive Models of Communicative Efficiency

Both the entrained and the nonentrained models predicted communicative efficiency of the goal-oriented conversations. Across the 57 conversations collected, communicative efficiency scores ranged from 10 to 30 ($M = 19.2$, $SD = 5.4$). Each model of the three types of machine learning techniques (Lasso, support vector machines, and k-nearest neighbors) predicted the continuous efficiency score without any arbitrary categorization of the measure. The entrained models included the five feature–measures that were entrained (i.e., measures significantly different between real and sham conversations), whereas the nonentrained models included the 20 feature–measures that were not entrained (i.e., measures that were not significantly different between real and sham conversations). The amount of variance accounted for R^2 and error rates (root-mean-square error) for the three entrained predictive models, and the three nonentrained models are shown in Table 4. The entrained models consistently explained more of the variance and had better predictive accuracies (lower error) than the nonentrained models in predicting communicative efficiency. For an estimate of effect size of the difference between entrained and nonentrained models, we calculated the standardized mean difference (a form of Cohen's *d*). The effect sizes between the Lasso and SVM models were moderate to large effect sizes.

Figure 1 shows the relative importance (*x*-axis) of each verified entrainment measure (sustained recurrence) by feature (*y*-axis) for the three models. Both the linear approach (Lasso) and the nonlinear approaches (support vector machines and *k*-nearest neighbors) revealed similar patterns of importance where several measures were consistently important drivers of the high predictive accuracies. Among these, sustained recurrence of articulatory (MFCC, LTAS) and rhythmic (EMS, rhythm metrics) speech signal dimensions emerged as important in predicting communicative efficiency of the conversations in at least two of the three models.

Discussion

Here, we developed and validated a novel computational methodology to investigate the communication phenomenon of conversational entrainment in the speech

Table 3. Comparison of real conversations with sham conversations by feature set.

	Conversation		
	Real (*n* = 57)	Sham (*n* = 500)	
Feature-Measure	***M* (*SD*)**	***M* (*SD***	***p***
Rhythm			
EMS			
Entropy	0.229 (0.118)	0.228 (0.087)	.924
Length	2.067 (0.047)	2.060 (0.098)	.332
Max length	3.281 (0.491)	3.238 (0.542)	.540
Sustained recurrence	123.404 (53.996)	87.390 (33.083)	< .001
Recurrence rate	6.332 (0.641)	6.209 (0.484)	.163
Rhythm metrics			
Entropy	0.236 (0.164)	0.213 (0.111)	.310
Length	2.075 (0.090)	2.060 (0.039)	.233
Max length	3.439 (0.907)	3.178 (0.592)	.038
Sustained recurrence	131.018 (80.683)	88.838 (58.434)	< .001
Recurrence rate	6.268 (1.561)	6.145 (1.558)	.577
Articulation			
MFCC			
Entropy	0.205 (0.115)	0.191 (0.097)	.377
Length	2.060 (0.060)	2.052 (0.033)	.360
Max length	3.211 (0.411)	3.064 (0.510)	.015
Sustained recurrence	77.456 (28.873)	59.598 (21.320)	< .001
Recurrence rate	4.960 (0.452)	5.081 (0.451)	.059
LTAS			
Entropy	0.293 (0.148)	0.246 (0.105)	.023
Length	2.096 (0.074)	2.072 (0.038)	.016
Max length	3.579 (0.801)	3.388 (0.644)	.087
Sustained recurrence	181.000 (160.925)	104.822 (71.407)	< .001
Recurrence rate	6.975 (2.546)	6.665 (1.873)	.375
Phonation			
Voice report			
Entropy	0.205 (0.140)	0.211 (0.103)	.772
Length	2.061 (0.060)	2.059 (0.035)	.810
Max length	3.351 (0.641)	3.172 (0.589)	.048
Sustained recurrence	126.544 (73.830)	75.912 (42.826)	< .001
Recurrence rate	5.945 (1.166)	5.664 (1.289)	.093

Note. *p* Value is based on independent *t* test adjusting for any unequal variances. EMS = envelope modulation spectrum; MFCC = mel-frequency cepstral coefficient; LTAS = long-term average spectrum.

domain. Key was the use of expert clinical assessment of conversation, as judged by five SLPs, to inform measurement of conversational entrainment across large feature sets that capture a broad representation of the speech signal. Based on the operational statement that if speech signal behaviors are really entrained during conversations, then alignment values should be higher in real conversations as compared to artificially generated sham conversations involving two out-of-conversation participants, we afford empirical evidence of conversational entrainment across rhythmic, articulatory, and phonatory dimensions of speech.

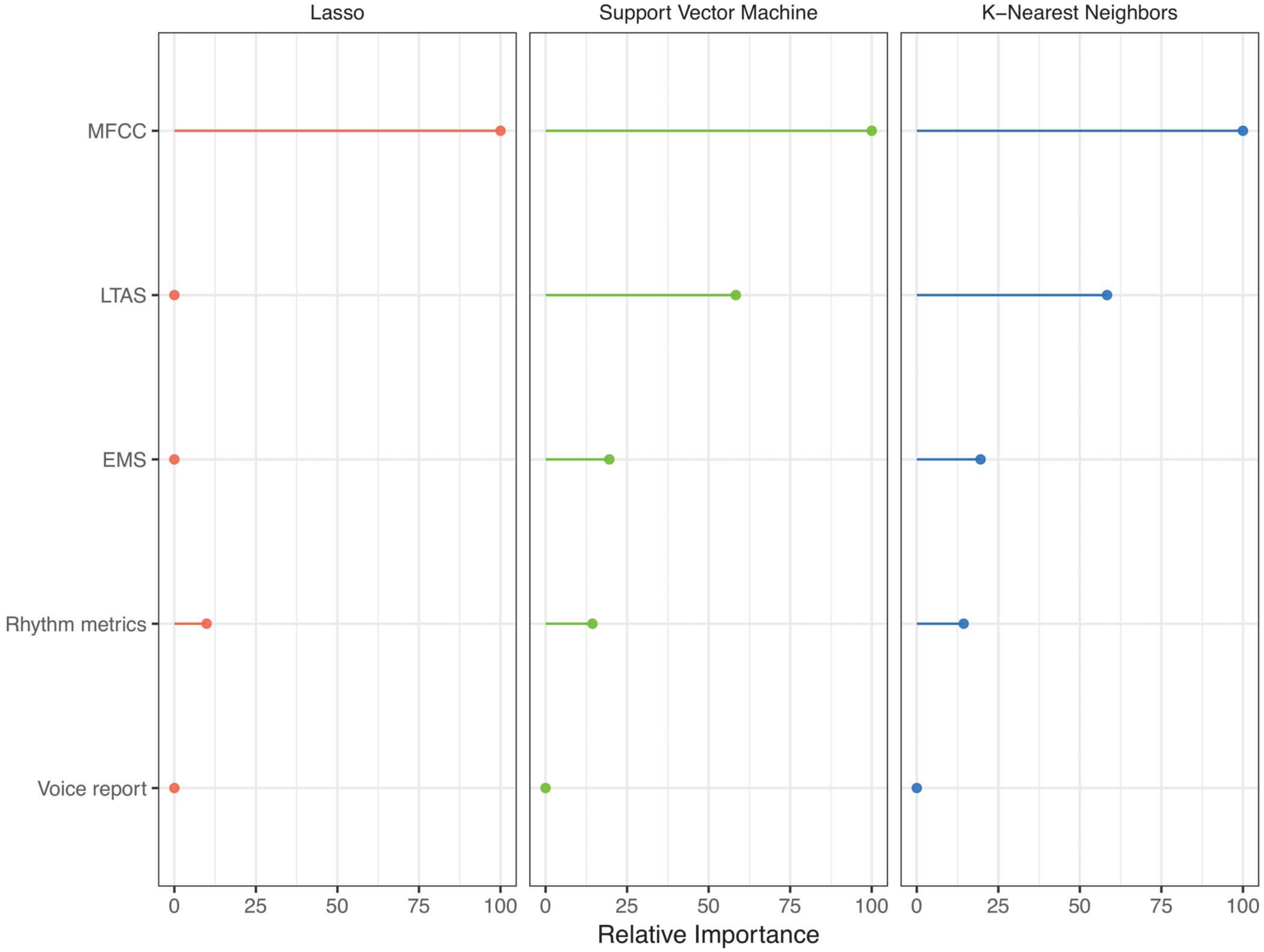
Table 4. Prediction accuracies for the entrained models and the nonentrained models and the effect sizes for the difference between the models.

	Entrained		Nonentrained		Cohen's *d*	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
Lasso	.347	4.854	.223	5.390	.628	0.629
SVM	.325	5.060	.296	5.442	.338	0.321
KNN	.304	4.899	.262	5.235	.148	0.274

Note. Cohen's *d* is a standardized mean difference of the entrained and nonentrained groups. RMSE = root-mean-square error; SVM = support vector machine; KNN = k-nearest neighbors.

Speech signal entrainment has typically been quantified using synchrony measures (e.g., Pearson correlation) on single acoustic–prosodic features, computed across conversational partners' speaking turn change (e.g., Borrie et al., 2015; Levitan & Hirschberg, 2011). Recently, more sophisticated projection approaches involving principal components analysis (Lee et al., 2014) or linear discriminant analysis (Willi et al., 2018), applied to larger feature sets, have been used to evaluate the communication phenomenon. Although these approaches provide valuable insight into the degree of acoustic–prosodic entrainment in a conversation, the use of CRQA allows us to examine time-evolving interdependent behavior, capturing not only the amount of entrainment but also organization that

Figure 1. Results of the communicative efficiency prediction models, highlighting the relative importance of each entrained (sustained recurrence) feature set. MFCC = melfrequency cepstral coefficient; LTAS = long-term average spectrum; EMS = envelope modulation spectrum.



reflects stability and complexity of behavioral dependency (see Fusaroli et al., 2014, for complete details). To briefly recount the CRQA measures, recurrence rate quantifies the amount of entrainment on the basis of a single instance of alignment. Sustained recurrence also quantifies the amount of entrainment, but this measure is based on alignment that is maintained for longer than a single instance. Length and max length quantify the average length/time and longest length/time of aligned behavior, respectively, and are thus considered indicative of entrainment stability. Finally, entropy captures the variability in the length of aligned behavior, reflecting entrainment complexity.

The current results reveal that, at least for goal-oriented conversations and clinically meaningful parameter settings, entrainment is largely a sustained and stable phenomenon. When examined collectively across all feature sets, measures of sustained recurrence and max length were significantly greater in real versus sham conversations, whereas recurrence rate, length, and entropy were not. When broken down by specific feature set, we see that the measure of sustained recurrence is consistently informative for differentiating between real and sham conversations across rhythmic, articulatory, and phonatory dimensions of speech. Measures of length and max length are approaching informative on particular feature sets (i.e., articulatory and phonatory dimensions) but, due to the conservative *p* value cutoff (using the Bonferoni adjustment) used in the current study, are not considered significant. Although the cutoff value was selected to mitigate the chance of Type I error in the 30 comparisons, it does raise the possibility that additional measures may also capture speech signal entrainment. Future studies using larger numbers of conversations may shed more light on this. From the current findings alone, we surmise that single instances of shared coordination (i.e., recurrence rate) between conversational participants may not be an optimal measure to capture clinically meaningful alignment in the speech domain. Rather, sustained entrainment, where alignment transpires for longer than a single instance, appears to maximally characterize the interdependency of acoustic–prosodic behavior that occurs in face-to-face, goal-oriented communicative interaction. Indeed, others have noted that the type of interaction (e.g., level of shared information, agreement vs. disagreement, deception vs. truth) influences the organization and structure of aligned behavior (Coco et al., 2017; Duran & Fusaroli, 2017).

Based on results from three different machine learning approaches, the current study demonstrates that, as a group, measures of speech signal entrainment (i.e., those measures that are significantly different between real and sham conversations) predict an objective measure of successful conversation, communicative efficiency. Furthermore, the models built on entrained speech signal measures outperform models built on nonentrained speech signal measures, highlighting the role of conversational entrainment in communicative efficiency. Although other studies have shown that measures of speech signal entrainment track with communicative efficiency/task success (e.g., Borrie et al., 2015; Nenkova et al., 2008), this is the first study of its kind to explicate a predictive relationship with speech signal entrainment measures informed by expert clinical assessment of conversation. Furthermore, as illustrated in Figure 1, sustained entrainment of features that represents articulatory behavior was the key driver for predicting communicative efficiency. Entrainment of rhythmic features played a smaller role in the prediction task, whereas phonatory feature entrainment played virtually no role. Much of the existing work in the area of acoustic–prosodic entrainment, including our own previous work, has targeted a single feature or speech signal dimension. The current findings, however, provide evidence that a singular focus is likely insufficient when evaluating conversational entrainment and functional outcomes, advancing the idea that characterization of conversational entrainment, at least in the speech domain, necessitates a multidimensional framework.

Setting CRQA Parameters With Expert Clinical Assessment

It is important to acknowledge how the parameters were set for the application of CRQA to the speech data in the current study. Given that the parameters are critical for the methodology to distinguish between instances of alignment versus instances that are not aligned, selecting appropriate parameters is no trivial task. Existing literature has used an algorithmic approach to select the parameters (Coco & Dale, 2014; Duran & Fusaroli, 2017). This approach individualizes parameters for each conversation by selecting a recurrence rate between certain percentage point values (e.g., 3%–5%). Herein, we used novel approach to set CRQA parameters, using real-world evidence from expert clinical assessment of conversation. This approach selected the parameters that produced the highest cross-validated predictive accuracy of expert clinical assessment and was used across all conversations, real and sham.

This data-driven parameter selection approach, although novel in the application of CRQA, is widely used to select the parameters in machine learning techniques (Hastie et al., 2009). It removes potential researcher bias from the selection of the parameters and generally allows a much better model fit. This is also true of CRQA. In this case, this approach ensures that the measures produced are meaningfully related to clinical evidence of successful conversation. Notably, the parameters that would have been chosen in the next highest predictive accuracy situations were not much different than those from the highest accuracy. This suggests that this approach provides a close range of parameters that perform similarly across the conversations. As such, this approach was not heavily dependent on some particularly well-performing parameters that would likely not generalize to other data. Future work, when coupled with clinical evidence about the conversations, can use this parameter selection approach with CRQA to find high-performing parameters that can be used across the conversations. In addition, work assessing the generalizability of the chosen parameters would shed light on similarities across conversations.

Clinical Implications and Future Directions

The current study affords a clinically meaningful methodology, developed and validated in healthy populations, for learning more about the communication phenomenon of entrainment and how it is disrupted in the context of communication disorders. Much more research is undoubtedly required to comprehend the utility of entrainment measures in speech-language pathology, but we advance that such investigations could have high yield. There presently exist no theory-driven automated tools to assess conversation in terms of productive and fulfilling interactions. Although expert clinical assessment of conversation should always be gold standard, we speculate that automated measures of entrained speech signal behavior may provide predictive and decision-making support for conversational assessment. Furthermore, in theory, these objective measures will have the capacity to go beyond what can be detected by the trained clinical ear, computing the degree of speech signal entrainment, revealing the locus of impairment (i.e., rhythm, articulation, or phonation), identifying potential intervention targets, and monitoring treatment progress and outcomes. Thus, this methodology, developed and validated in conversations involving healthy participants, lays the groundwork for ensuing entrainment investigations in clinical populations.

Conclusion

We weaved together automatic acoustic–prosodic feature extraction methods, feature reduction techniques, recurrence quantification, and expert clinical assessment to characterize conversational entrainment within a multidimensional framework. Using a real versus sham validation procedure, we find evidence of sustained entrainment in rhythmic, articulatory, and phonatory dimensions of speech. We provide additional validation for the methodology, showing that key output measures, verified measures of speech signal entrainment, predicted an objective measure of conversational success, communicative efficiency. This clinically meaningful methodology for capturing conversational entrainment, validated in goal-oriented conversations involving healthy participants, has potential application for clinical disciplines such as speech-language pathology where conversational entrainment represents a critical knowledge gap in the field, as well as a potential target for remediation.

Acknowledgments

This research was supported by the National Institute on Deafness and Other Communication Disorders Grant R21DC016084 awarded to Stephanie Borrie and Visar Berisha. We gratefully acknowledge research assistants in the Human Interaction Laboratory at Utah State University for the laborious task of manually coding all speech utterances in the conversational corpus.

References

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40,* 177–189.

Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science, 16,* 814–819.

Ballard, K. J., Azizi, L., Duffy, J. R., McNeil, M. R., Halaki, M., O'Dwyer, N., … Robin, D. H. (2016). A predictive model for diagnosing stroke-related apraxia of speech. *Neuropsychologia, 81,* 129–139.

Barrett, T. S., & Brignone, E. (2017). Furniture for quantitative scientists. *The R Journal, 9,* 142–148.

Beňuš, Š. (2014). Social aspects of entrainment in spoken interaction. *Cognitive Computation, 6*(4), 802–813.

Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., & Levitan, R. (2018). Prosodic entrainment and trust in human–computer interaction. In *Proceedings of the 9th International Conference on Speech Prosody* (pp. 220–224). Baixas, France: International Speech Communication Association.

Berisha, V., Liss, J., Sandoval, S., Utianski, R., & Spanias, A. (2014). Modeling pathological speech perception from data with similarity labels. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014* (pp. 915–919). Piscataway, NJ: IEEE.

Berisha, V., Wisler, A., Hero, A. O., & Spanias, A. (2016). Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing, 64,* 580–591.

Bernieri, F. J., Reznick, J. S., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology, 54,* 243–253.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences* (Vol. 17, No. 1193, pp. 97–110). Amsterdam, the Netherlands: Institute of Phonetic Sciences, University of Amsterdam.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer(Version 6.0.39) [Computer program]. Retrieved from http://www.praat.org/

Borrie, S. A., & Delfino, C. (2017). Conversational entrainment of vocal fry in young adult female American English speakers. *Journal of Voice, 31,* 513.e25–513.e32.

Borrie, S. A., & Liss, J. M. (2014). Rhythm as a coordinating device: Entrainment with disordered speech. *Journal of Speech, Language, and Hearing Research, 57,* 815–824.

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,* 1187.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 74,* B13–B25.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76,* 893–910.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Cleveland, T. F., Sundberg, J., & Stone, R. E. (2001). Long-term-average spectrum characteristics of country singers during speaking and singing. *Journal of Voice, 15,* 54–60.

Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: An R package. *Frontiers in Psychology, 5,* 510.

Coco, M. I., Dale, R., & Keller, F. (2017). Performance in a collaborative search task: The role of feedback and alignment. *Topics in Cognitive Science, 10,* 55–79.

Comon, P. (1992). *Independent components analysis. In higher-order statistics* (pp. 29–38). Oxford, United Kingdom: Elsevier. Retrieved from https://hal.archives-ouvertes.fr/hal-00346684

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In A. Waibel & K.-F. Lee (Eds.), *Readings in speech recognition* (pp. 65–74). San Mateo, CA: Morgan Kaufmann.

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

Dellwo, V., & Fourcin, A. (2013). Rhythmic characteristics of voice between and within languages. *Revue Tranel (Travaux neuchâtelois de linguistique), 59,* 87–107.

Duffy, J. R. (2015). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). St. Louis, MO: Elsevier Mosby.

Duran, N. D., & Fusaroli, R. (2017). Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement. *PLoS One, 12*(6), e0178140.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49,* 396–413.

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: The promises and challenges of using cross recurrence quantification analysis. In N. Marwan, M. Riley, A. Giuliani, & C. Webber, Jr. (Eds.), *Translational recurrences. Springer proceedings in mathematics & statistics* (Vol. 103). New York, NY: Springer, Cham.

Garcia, L. J., & Orange, J. B. (1996). The analysis of conversational skills of older adults: Current research and clinical approaches. *Canadian Journal of Speech-Language Pathology and Audiology, 20,* 123–135.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Elements* (vol. 1, 2nd ed.). New York, NY: Springer. https://doi.org/10.1007/b94608.

Kawabata, K., Berisha, V., Scaglione, A., & LaCross, A. (2016). A convex model for linguistic influence in group conversations. In *Proceedings of the Interspeech 2016* (pp. 1442–1446).

Kuhn, M. (2017). *Caret: Classification and regression training* (R package Version 6.0-78). Retrieved from https://CRAN.R-project.org/package=caret

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,* 518–539.

Levitan, R., Beňuš, Š., Gálvez, R. H., Gravano, A., Savoretti, F., Trnka, M., . . . Hirschberg, J. (2016). Implementing acoustic-prosodic entrainment in a conversational avatar. In *Proceedings of Interspeech* (pp. 1166–1170). San Francisco, United States: International Speech Communication Association.

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic–prosodic entrainment with respect to multiple levels and dimensions. In R. Pieraccini & A. Colombo (Eds.), *Proceedings of Interspeech 2011*. Brisbane, Australia, International Speech Communications Association.

Linville, S. E. (2002). Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice, 16,* 472–479.

Liss, J. M., LeGendre, S., & Lotto, A. J. (2010). Discriminating dysarthria type from envelope modulation spectra. *Journal of Speech, Language, and Hearing Research, 53,* 1246–1255.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., & Caviness, J. N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research, 52,* 1334–1352.

Local, J. (2007). Phonetic detail and the organization of talk-interaction. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI)* (pp. 6–10). Saarbrucken, Germany: International Phonetic Association.

Marchini, J. L., Heaton, C., & Ripley, B. D. (2017). *FastICA: FastICA algorithms to perform ICA and projection pursuit* (R package Version 1.2-1). Retrieved from https://CRAN.R-project.org/package=fastICA

Martin, J. H., & Jurafsky, D. (2000). *Speech and language processing: An introduction to natural language processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Matarazzo, J. D., Weitman, M., Saslow, G., & Wiens, A. W. (1963). Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior, 6,* 451–458.

Matarazzo, J. D., & Wiens, A. N. (1967). Interviewer influence on durations of interviewee silence. *Journal of Experimental Research in Personality, 2,* 56–69.

Mendoza, E., Valencia, N., Muñoz, J., & Trujillo, H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of Voice, 10,* 59–66.

Nasir, M., Baucom, B. R., Georgiou, P., & Narayanan, S. (2017). Predicting couple therapy outcomes based on speech acoustic features. *PLoS One, 12,* e0185123.

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.

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of ACL/HLT 2008* (pp. 169–172). Columbus: The Ohio State University.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39,* 132–142.

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology, 4,* 559.

Phillips-Silver, J., Aktipis, C. A., & Bryant, G. A. (2010). The ecology of entrainment: Foundations of coordinated rhythmic movement. *Music Perception, 28,* 3–14.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 169–225.

Pickering, M. J., & Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences, 36,* 377–392.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reichel, U. D., Beňuš, Š., & Màdy, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication, 100,* 46–57.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29,* 1045–1060.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 326–332.

Tu, M., Berisha, V., & Liss, J. (2017). Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Proceedings of Interspeech* (pp. 1849–1853). Baixas, France: International Speech Communication Association.

Tu, M., Jiao, Y., Berisha, V., & Liss, J. M. (2016). Models for objective evaluation of dysarthric speech from data annotated by multiple listeners. In *50th Asilomar Conference on Signals, Systems and Computers, 2016* (pp. 827–830). Piscataway, NJ:IEEE.

Vallabha, G. K., & Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. *The Journal of Acoustical Society of America, 116,* 1184–1197.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech, 53,* 510–540.

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35,* 501–522.

Willi, M. M., Borrie, S. A., Barrett, T. S., Tu, M., & Berisha, V. (2018). A discriminative acoustic–prosodic approach for measuring local entrainment. *Proceedings of Interspeech*. Baixas, France: International Speech Communication Association.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*(6), 957–968.

Wynn, C. J., Borrie, S. A., & Sellars, T. (2018). Speech rate entrainment in children and adults with and without autism spectrum disorder. *American Journal of Speech-Language Pathology, 27*(3), 965–974.

Zbilut, J. P., Giuliani, A., & Webber, C. L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A, 246,* 122–128.

Appendix

Clinical Rating Scale

To what extent do you agree with this statement:

The conversational pair sound like they are in-sync or aligned with one another.

Note that high ratings (scores above 4) are indicative of a natural cohesiveness to the interaction, smooth turn-taking and conversational flow, and a sense of rapport and connection between conversational participants and low ratings (scores below 4) are indicative of an awkward, disconnected, and disengaged interaction.

1. Strongly Disagree
2. Disagree
3. Slightly Disagree
4. Neutral
5. Slightly Agree
6. Agree
7. Strongly Agree