ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Measuring Prosodic Entrainment in Conversation: A Review and Comparison of Different Methods

Joanna Kruyt,[a,b] Dorina de Jong,[c,d] Alessandro D'Ausilio,[c,d] and Štefan Beňuš[a,e]

[a] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia [b] Faculty of Informatics and Information Technologies, Slovak Technical University, Bratislava, Slovakia [c] Istituto Italiano di Tecnologia, Center for Translational Neurophysiology of Speech and Communication, Ferrara, Italy [d] Università di Ferrara, Dipartimento di Neuroscienze e Riabilitazione, Italy [e] Constantine the Philosopher University, Nitra, Slovakia

ARTICLE INFO

Article History:
Received February 8, 2023
Revision received June 30, 2023
Accepted July 28, 2023

Editor-in-Chief: Cara E. Stepp
Editor: Jun Wang

https://doi.org/10.1044/2023_JSLHR-23-00094

ABSTRACT

Purpose: This study aims to further our understanding of prosodic entrainment and its different subtypes by analyzing a single corpus of conversations with 12 different methods and comparing the subsequent results.
Method: Entrainment on three fundamental frequency features was analyzed in a subset of recordings from the LUCID corpus (Baker & Hazan, 2011) using the following methods: global proximity, global convergence, local proximity, local convergence, local synchrony (Levitan & Hirschberg, 2011), prediction using linear mixed-effects models (Schweitzer & Lewandowski, 2013), geometric approach (Lehnert-LeHouillier, Terrazas, & Sandoval, 2020), time-aligned moving average (Kousidis et al., 2008), HYBRID method (De Looze et al., 2014), cross-recurrence quantification analysis (e.g., Fusaroli & Tylén, 2016), and windowed, lagged cross-correlation (Boker et al., 2002). We employed entrainment measures on a local timescale (i.e., on adjacent utterances), a global timescale (i.e., over larger time frames), and a time series–based timescale that is larger than adjacent utterances but smaller than entire conversations.
Results: We observed variance in results of different methods.
Conclusions: Results suggest that each method may measure a slightly different type of entrainment. The complex implications this has for existing and future research are discussed.

One of the hallmarks of being human is not language per se; rather, its use in conversations allows for a fundamental advantage: the development of a dialogically extended mind (Fusaroli et al., 2014). In fact, studying language in isolated individuals, either at the perceiver or producer ends, does not allow for full exploration of the language faculty (Levinson, 2016).

A key phenomenon that appears only in conversations concerns the subtle, dynamic, mutual adjustments that emerge between partners, which contributes to the feeling of being part of a social exchange. Indeed, during conversations, interlocutors tend to become more similar. This has been observed across almost all linguistic levels, including prosody (e.g., Levitan et al., 2012; Natale, 1975; Webb, 1969), lexical choice (e.g., Brennan & Clark, 1996; Garrod & Anderson, 1987), syntax (e.g., Branigan et al., 2000), and dialogue acts (e.g., Mizukami et al., 2016). This work will focus on one particular level of this phenomenon, which has attracted most of the efforts toward an objective and numerical description: the prosodic level.

The phenomenon of increased similarity in interaction has been referred to as "entrainment," "accommodation," "alignment," "convergence," "synchrony," or "imitation," sometimes depending on the theoretical perspective. For instance, Pickering and Garrod's "interactive alignment" account (Pickering & Garrod, 2004, 2013) states that entrainment (henceforth, we will use this label) occurs because of automatic priming mechanisms. On the other hand, according to Giles et al.'s "communication

Correspondence to Joanna Kruyt: joanna.kruyt@savba.sk. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

accommodation theory" (Giles et al., 1991), people alter their communication behaviors to either emphasize or minimize the perceived social distance with their communication partner.

Someone's theoretical understanding of entrainment may not only influence the term they use to describe the phenomenon, but it may also affect the methods they choose or develop to measure entrainment. Multiple analytical strategies have been proposed to quantify the degree of entrainment in an interaction, all informed by different a priori methodological choices or theoretical stances. Assumptions about how, why, and when entrainment occurs can play a role in choosing or developing a method to quantify it: For example, if one assumes that entrainment occurs immediately, one might look for entrainment in short time frames such as adjacent turns. If one assumes that entrainment is a slow process, on the other hand, one may look for entrainment in larger time frames such as entire conversations.

In other words, different methods rely on different assumptions about entrainment and look for entrainment in a different manner. While this plurality of methods is not inherently problematic, as different methods have different goals and therefore have different characteristics, the broad scope of different methods that are used in literature complicates the interpretation and comparison of existing studies.

Comparing methodologies and existing studies on prosodic entrainment is complicated not only by the fact that many different experimental paradigms and many different methods exist, but it is also complicated by the fact that entrainment is highly complex. It occurs on different *levels* of language (e.g., lexical choice, syntax, prosody), can be measured on several different *dimensions* (e.g., synchrony, proximity; see next section for a description of these dimensions), and can be investigated using different *features* (e.g., fundamental frequency [f_o], speech rate).

Understanding exactly what a specific method measures and how this compares to what other methods measure is essential when one considers research that suggests that entrainment on different levels, dimensions, and features may be uncorrelated: Several studies have attempted to find an underlying structure in entrainment on different features, dimensions, and levels (e.g., Ostrand & Chodroff, 2021; Weise & Levitan, 2018) but did not find any meaningful correlations. Ostrand and Chodroff (2021) wrote that "it appears that entrainment, rather than a single behavior or a structured collection of behaviors, is a set of behaviors which are only loosely linked and perhaps independently explained by the competing theories" (Ostrand & Chodroff, 2021, p. 301). If "entrainment" is not a single latent behavior but rather a collection of behaviors without a clear structure or pattern, understanding and specifying the dimension and level of entrainment—and thus the specific type of behavior that is being measured—becomes even more critical.

Considering the complex nature of entrainment and the lack of correlation between entrainment on different features, dimensions, and levels (e.g., Ostrand & Chodroff, 2021; Weise & Levitan, 2018), it is essential that researchers are specific in their definitions of different types of entrainment and the terms used to denote these different types. This article can be considered an initial attempt to empirically investigate the relationship between the many different subtypes of entrainment and the methods that measure them.

Frameworks can help facilitate clarity in prosodic entrainment research in terms of terminology, which can in turn facilitate the interpretation and comparison of existing research. One such framework, developed by Wynn and Borrie (2022) and heavily inspired by the seminal work of Levitan and Hirschberg (2011), classifies entrainment methods according to three variables: entrainment class (which in this article will be referred to as "dimension"), entrainment level (which in this article will be referred to as "timescale"), and entrainment dynamicity. Entrainment "dimension" refers to the type of entrainment that is being measured: "Proximity" refers to the degree of absolute similarity, while "synchrony" describes the relative similarity between two speakers, which is often measured using correlation coefficients. Entrainment "timescale" describes the temporal units in which entrainment is measured: Local entrainment occurs between adjacent utterances, whereas global entrainment can be measured by comparing units that span a larger temporal range. Entrainment "dynamicity" involves the idea that entrainment can change over time. Entrainment may increase as the conversation progresses, or it may fluctuate or remain constant (or "static") throughout the interaction.

In this study, we employed 12 of the most relevant methods to quantify acoustic and prosodic entrainment on the same open-access corpus. The Method section of this article describes the 12 methods, along with their (dis) advantages and analytical procedures. The Results section presents our findings, which suggest relatively little consistency in the results produced by the methods. An interim discussion is presented in the Interim Discussion section, before a brief follow-up study that investigates a potential source of variation in results is presented in the Follow-Up Norming Study: Background section. The General Discussion section contains a general discussion of the results and speculates about additional possible sources of the observed variation, as well as implications for the interpretation of existing research, and practical considerations for future studies. In general, our study aimed to shed light on the relationship between different subtypes of entrainment and the methods used to measure them, to provide an empirical basis for validation of the theoretical framework proposed by Wynn and Borrie (2022), and to provide perspective on why results of entrainment studies report a range of different outcomes.

Method

The methods will be presented grouped by the timescale on which they measure entrainment, that is, locally, globally, or based on time series. For an overview of the 12 methods discussed below and their categorizations in the Wynn and Borrie (2022) framework, see Table 1.

Local Methods

Static Local Proximity: Levitan and Hirschberg's Local Proximity

Levitan and Hirschberg (2011) introduced the notion that there are different types of entrainment, along with a set of methods to determine whether each type of entrainment occurs in a given conversation. We will first focus on the local methods where entrainment is assessed between adjacent interpausal units (IPUs). To determine proximity on a local timescale, that is, local proximity, a "partner difference" and "other difference" were computed. Partner difference was determined by extracting the desired features from a target IPU and its adjacent IPU uttered by the other speaker. The absolute difference between these two feature values was considered the partner difference. Other difference was calculated by extracting desired features from a target IPU and then taking the mean of the differences between the feature in that target IPU and 10 random nonadjacent IPUs uttered by the other speaker. The partner difference and other differences were compared with a paired *t* test. In other words, the "local" difference between a target utterance and its adjacent utterance (partner difference) was compared to the "nonlocal" difference between a target utterance and nonadjacent utterances (other difference) to measure local entrainment.

Static Local Synchrony: Levitan and Hirschberg's Local Synchrony

To calculate turn-level synchrony, or local synchrony, Levitan and Hirschberg (2011) used Pearson's correlation coefficients of features extracted from adjacent IPUs uttered by conversing interlocutors. This method can be considered a measure of local, static synchrony.

Dynamic Local Proximity: Levitan and Hirschberg's Local Convergence

Turn-level convergence, or local convergence, was determined by computing the Pearson correlation between the absolute partner difference (calculated in the same way as for local proximity) and IPU number (as an indication of time. Levitan and Hirschberg's (2011) local convergence can be considered a type of local, dynamic proximity according to Wynn and Borrie (2022) because it measures absolute similarity between speakers and takes into account the time that elapsed in a conversation.

For all methods proposed by Levitan and Hirschberg (2011), the original publication was followed as closely as possible, including the norming procedure: Gender means were calculated by determining the mean per feature per speaker (mean of all IPUs weighted by duration), which were then used to calculate the mean per gender. Each

Table 1. Categorization of methods described and used in this article.

Variable		**Timescale**		
		Local	**Time series**	**Global**
Dimension	Proximity	**Static local proximity** Local proximity, Levitan & Hirschberg (2011)	**Static global proximity** CRQA, Fusaroli & Tylén (2016)	**Static global proximity** Global proximity, Levitan & Hirschberg (2011)
		Dynamic local proximity Local convergence, Levitan & Hirschberg (2011)		**Dynamic global proximity** Global convergence, Levitan & Hirschberg (2011) Geometric approach, Lehnert-LeHouillier, Terrazas, Sandoval, & Boren (2020)
	Synchrony	**Static local synchrony** Local synchrony, Levitan & Hirschberg (2011) Linear mixed-effects models, Schweitzer & Lewandowski (2013)	**Static global synchrony** TAMA, Kousidis et al. (2008) HYBRID, De Looze et al. (2014) WLCC, Boker et al. (2002)	
			Dynamic global synchrony HYBRID, De Looze et al. (2014)	

Note. Please note that following the Wynn and Borrie (2022) framework, all time series–based methods are considered to measure entrainment globally.

feature per IPU was normed using the following formula: feature value in IPU = (original feature value − gender mean) / gender standard deviation.

Significant (dis)entrainment findings for local convergence and local synchrony were only considered valid if no more than one out of 10 correlations performed on surrogate data returned a significant result, following Levitan and Hirschberg. Surrogate data were created by dividing our data into turn-initial IPUs and turn-final IPUs per speaker and then randomly shuffling these. This division was done to ensure that even in the randomly shuffled data, turn-final IPUs from one speaker were always followed by turn-initial IPUs from the other speaker.

Static Local Synchrony: Schweitzer and Lewandowski's Linear Mixed-Effects Models

Schweitzer and Lewandowski (2013) used a method that measures static local synchrony (Wynn & Borrie, 2022). This method is comparable to Levitan and Hirschberg's (2011) method for measuring local synchrony in the sense that it characterizes the relationship between a feature value in a target utterance and the feature value of its adjacent utterance. However, Schweitzer and Lewandowski did not use Pearson correlations as was done by Levitan and Hirschberg but rather used a linear mixed-effects model (LMEM) to see whether the feature value of an utterance could be predicted using the feature value of the preceding utterance of the other speaker, which allowed for the modeling of random intercepts and a clearer distinction between dependent and independent variables. To be more precise, they constructed a model where the fixed effect was the feature value of the preceding utterance and the random effects were the speaker, partner, and dyad. Schweitzer and Lewandowski also included social variables in their formula. The authors assessed the significance of their model using Markov chain Monte Carlo simulations to calculate the highest posterior densities.

The data used in this study differ substantially from the recordings analyzed by Schweitzer and Lewandowski (2013): Their participants were all female, and they obtained ratings of liking. Since our data contained both genders and no liking assessments, our model included gender and excluded liking. Additionally, rather than using whole turns, we extracted features from IPUs at turn exchanges and used these in order to keep the unit of analysis (i.e., IPUs at turn exchanges) the same for all our local methods. We built an LMEM for each feature in R using the lme4 package (Bates et al., 2014). Importantly and in contrast to the original model used by Schweitzer and Lewandowski (2013), conversational partner was removed as a random effect from the LMEM formula due to model convergence failures. The final formulas used to construct the full and the null LMEMs was thus as follows, where gender represents both the gender of the target utterance speaker and their interlocutor, since all dyads in our corpus are gender-matched:

$$\begin{aligned}\text{full}: &\ \text{target utterance} \sim \text{preceding utterance} + \text{gender}\\ &\quad + (1|\text{speaker of target turn}) + (1|\text{dyad})\\ \text{null}: &\ \text{target utterance} \sim \text{gender}\\ &\quad + (1|\text{speaker of target turn}) + (1|\text{dyad}) \qquad (1)\end{aligned}$$

Finally, we used lmerTest (Kuznetsova et al., 2017) to assess the significance of main effects, which approximates the degrees of freedom via t tests using the Satterthwaite approximations. This differs from the method used in the original Schweitzer and Lewandowski (2013) study, where Markov chain Monte Carlo simulations were used to calculate the highest posterior densities, which may cause issues with unreliability, especially in cases where the estimated random effect variances were near zero.[1] Additionally, we constructed a null model for each feature by excluding the feature of preceding utterance as a fixed effect (see above). We then compared this null model to the full model using analysis of variance (ANOVA).

Global Methods

Static Global Proximity: Levitan and Hirschberg's Global Proximity

Following the framework by Wynn and Borrie (2022), Levitan and Hirschberg's (2011) global proximity can be considered to measure global, static proximity. Similarly to proximity on the local timescale, proximity on the global timescale is determined by calculating a "partner difference" and an "other difference." This time, "partner difference" referred to the difference in the mean value of a feature of a speaker and the mean value of a feature of their conversational partner. The "other difference" was calculated by taking the mean of the differences between the mean value of a feature in one speaker and the mean values of that feature of every participant with whom the speaker did not interact. Partner and other differences were compared using a paired t test. Levitan and Hirschberg also outlined an alternative measure to determine global proximity (comparing a speaker's features to their own features in a different conversation). However, this second option only works if participants partake in multiple interactions with different interlocutors, which is not the case for many corpora. Note that the data used in this study do not include multiple conversations per speaker (see Corpus

[1] See explanation for why function *mcmcsamp* is no longer updated to work with newer versions of lme4: https://search.r-project.org/CRAN/refmans/lme4/html/pvalues.html.

section), so only the first method for calculating the other difference mentioned above was used in our study.

Dynamic Global Proximity: Levitan and Hirschberg's Global Convergence

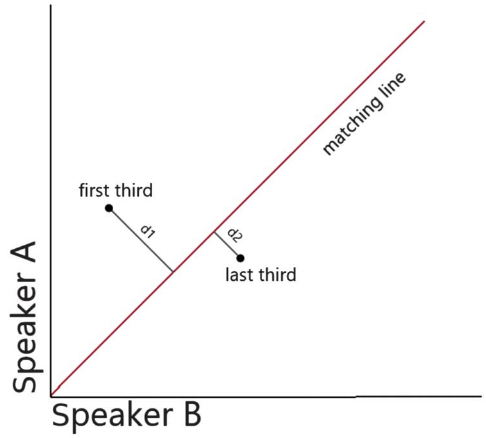
To measure global convergence, Levitan and Hirschberg (2011) divided a conversation in half and compared the difference between speakers' mean feature values between the first and the second half with a paired *t* test. In this study, we calculated the mean of all IPUs per speaker and conversation half weighted by IPU duration. Following Wynn and Borrie (2022), this method measures global, dynamic proximity because it takes into account the change in speakers' features over time. Note that we followed the gender-based *z*-score normalizing procedure described by Levitan and Hirschberg (2011) and above for both the global proximity and convergence measures.

Dynamic Global Proximity: Lehnert-LeHouillier, Terrazas, Sandoval, and Boren's Geometric Approach

The geometric approach by Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020) measures global, dynamic proximity because it compares the absolute similarity between speakers' features at different points in time. They were the first to implement this method to investigate whether interlocutors entrained more during the final third of an interaction than during the first third. To this end, the mean feature value for each speaker was calculated during the first and last third of the interaction and a geometric approach was taken, in which two vectors were drawn: one for each speaker, from their mean feature value during the first third to their mean feature value during the final third of the conversation. Additionally, a "matching line" was drawn such that it represented the scenario in which both speakers' features matched perfectly. The minimum distances between the start and end point of the aforementioned vectors to this matching line were then calculated (see Figure 1). The difference between those minimum distances reflected the overall entrainment. Additionally, each speaker's individual contribution to the overall entrainment was calculated by determining the difference between both thirds along both the *x*- and *y*-axes, respectively. These differences reflected each speaker's contribution to the overall entrainment in percentages.

To quantify entrainment using a geometric approach, we followed the method described in the publications by Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020). Mean features per speaker per third of conversation were calculated by determining the mean over all IPUs by one speaker in each third (based on IPU number), again weighted by duration.

Figure 1. Visual representation of the method used by Lehnert-LeHouillier, Terrazas, and Sandoval (2020). The feature values of both speakers during the first third of the interaction is the start of the line ("first third"), while the feature values during the last third of the interaction form the end of the vector ("last third"). Overall entrainment is determined by calculating the difference between the distances from the start and end point to the matching line, in this case the difference between d1 and d2. To calculate each speaker's individual contribution to the overall entrainment, one can calculate proportion of change along both the *x*-axis (Speaker B) and the *y*-axis (Speaker A). This figure is inspired by the figure in the work of Lehnert-LeHouillier, Terrazas, and Sandoval (2020). f_o = fundamental frequency.



This geometric method is one of the few measures of entrainment that does not indicate whether the entrainment that is measured is statistically significant or not, which may lead to difficulties in interpreting results. One could argue that this method is not an evaluation of entrainment but rather a way to extract entrainment-related features for further statistical analysis.

Time Series–Based Methods

Static Global Synchrony: Kousidis et al.'s Time-Aligned Moving Average

Time-aligned moving average (TAMA) was initially introduced by Kousidis et al. (2008) but has since been adopted and/or modified by others (e.g., De Looze et al., 2014). TAMA is a measure of relative similarity and can thus be considered to detect global, static synchrony. In Kousidis et al., features were extracted by moving a "window" across the speech signal in a stepwise manner. The size of these windows usually lies around 20 s (e.g., De Looze et al., 2014) but can also be larger or smaller, depending on the data and/or research goals, with small windows more closely reflecting local entrainment and large windows reflecting global entrainment (Kousidis et al., 2008). The step size, or increment with which the window is moved across the signal, typically depends on the selected window size but is usually approximately 50% of the window size. Within each window, the desired feature was weighted by its duration: This meant that the duration of an utterance determined how much the utterance contributed to the mean of the window. Therefore, short utterances contributed less to the window's mean than longer utterances. Values were normalized by dividing the raw feature values by the speaker's overall mean of the entire conversation. The resulting time series for each speaker were cross-correlated to provide a measure of entrainment.

As in the original publication, features were extracted from a moving window of 20 s with a step size of 10 s. Features were weighted by duration and normalized following Kousidis et al. (2008). The resulting feature vectors of the two interlocutors were then correlated using a Pearson correlation.

Static Global Synchrony: De Looze et al.'s HYBRID Method

One of the possible drawbacks of TAMA is that utterances may be cut off midway due to the standard window size. To mitigate this issue, De Looze et al. (2014) combined concepts from TAMA and utterance-based methods and created a so-called "HYBRID" method. HYBRID works similarly to TAMA, but the major difference is that in HYBRID, the window size was extended to incorporate the end of utterances that would be cut off by TAMA. Besides that, features were extracted in an identical manner as for TAMA. Pearson correlations between resulting time series were then conducted to quantify the static, global synchrony (following Wynn & Borrie, 2022) between both speakers across the entire conversation.

Dynamic Global Synchrony: De Looze et al.'s HYBRID Method

Additionally, as proposed by De Looze et al. (2014), we extended the analysis to provide a measure of dynamic entrainment. In essence, features were extracted over 10 windows (i.e., 110 s) with a step size of five windows (i.e., 60 s), and the resulting time series were then cross-correlated using a Pearson correlation. This way, we could measure the degree of the entrainment across longer periods of time (De Looze et al., 2014). If the correlation coefficient (rho) of a window was positive, entrainment occurred, and if rho in a window was negative, disentrainment occurred. Furthermore, we assessed whether the strength of the entrainment (measured with Pearson correlations) was stronger in the second half of the conversations (based on the number of windows), as opposed to the first half, with a paired *t* test.

Static Global Synchrony: Boker et al.'s Windowed Lagged Cross-Correlation

Another method that assesses global, static synchrony is windowed lagged cross-correlation (WLCC; Boker et al., 2002). This method has been used to detect synchrony in movement (Duran & Fusaroli, 2017; Schoenherr, Paulick, Worrack, et al., 2019) and physiological processes (Behrens et al., 2020) but has also been adapted to investigate prosodic entrainment (Truong & Heylen, 2012). WLCC estimates the strength of the relationship between two time series, which were extracted by moving a window of predetermined size across the speech signal, similar to TAMA and HYBRID. The cross-correlation between the two resulting time series is then calculated, again similarly to TAMA and HYBRID. However, unlike the time-series methods discussed previously, WLCC also measures entrainment that occurs with a lag or a temporal delay. A so-called "peak-picking" algorithm is then used to find the lag at which the cross-correlation between the speakers is at its highest per individual window (Boker et al., 2002). The (peak) cross-correlations per window can then be combined into a single vector to measure global entrainment.

While WLCC and other time series–based methods can provide rich information into the dynamics of entrainment, stationarity (i.e., the assumption that the statistical properties of the investigated process remain constant) needs to be addressed. When cross-correlation is computed over small windows, it is argued that it does not need to assume stationarity over the entire signal (although the assumption of local stationarity is also debated; Dean & Dunsmuir, 2016). Additionally, with cross-correlation–based methods, one must be mindful of detecting spuriously significant cross-correlations, as a result of autocorrelation within individual time series.

To implement WLCC, we first extracted our features using windows as we did for TAMA and HYBRID. We also used the same norming procedure for the data as we used for TAMA and HYBRID. Then, we used linear interpolation to create new data points in between existing windows so that we had a measure every 5 s instead of every 10 s, following Truong and Heylen (2012). This essentially doubled the amount of extracted features. For WLCC, we used the hyperparameters implemented by Truong and Heylen (2012), namely, a window size of 20 s and a step size of 10 s, with a maximum lag of ± 20 s and a lag step size of 5 s. The absolute peak cross-correlations during real interactions were then compared to the absolute mean values from 38 pseudo-interactions. These pseudo-interactions were created following the methodology described by Truong and Heylen (2012) and were adapted from the studies of Ramseyer and Tschacher (2010) and Bernieri and Rosenthal (1991). First, all interactions were divided into five time frames. Then, a sham interlocutor was created by randomly drawing feature values from speakers who did not interact with the target speaker. The random draw was limited by time frame to maintain the temporal structure of the dialogues to some extent. This meant that the value for the sham interlocutor in the third time frame could only be taken from a real interlocutor's third time frame and not, for example, the second or fourth. Due to violations in the assumption of normality in the data, we used a Wilcoxon signed-ranks test to compare the real and the pseudo-interactions in order to determine whether or not entrainment happened within a dyad.

Static Global Proximity: Fusaroli and Tylén's Cross-Recurrence Quantification Analysis

While all time series–based methods previously discussed in this article relied on cross-correlations, it is important to note that there are also time series–based approaches that do not. An example of such a method is cross-recurrence quantification analysis (CRQA). CRQA has been described as "a non-linear analog to cross-correlation" (Fusaroli & Tylén, 2016, p. 156) because it does not assume that entrainment increases linearly over time and does not assume stationarity, thereby avoiding some of the potential pitfalls of cross-correlation mentioned earlier.

The roots of CRQA lie in the field of dynamical systems, and it was developed to capture how and to what extent two interacting series display recurring properties and patterns in time (Zbilut et al., 1998). The method has been applied to various types of behavioral entrainment, including dance (e.g., Washburn et al., 2014), other types of movement (e.g., Iqbal & Riek, 2015), and prosodic entrainment (e.g., Borrie et al., 2019; Fusaroli & Tylén, 2016). In the case of prosodic entrainment, CRQA quantifies how often and for how long two speakers were speaking similarly during their interaction. Most of this information is extracted from cross-recurrence plots, which are a crucial element of CRQA. For an example of such plots and a more detailed explanation of the various measures that can be extracted from cross-recurrence plots, see Appendix A. In this article, we mostly focused on "recurrence rate" (RR), or the percentage of points that both systems (in this case, speakers) were in similar states (in this case, spoke similarly enough). In order to conduct CRQA, various parameters need to be set: For example, the *radius* is the threshold for when two systems are considered "similar enough," and the *delay* is the maximum delay at which two systems will be compared. Various approaches for setting these parameters exist, depending on research questions (e.g., see the clinically informed approach by Borrie et al., 2019). More information on other parameters can be found in Appendix A.

Global, static proximity was assessed using CRQA, largely following the methodology of Fusaroli and Tylén (2016). While Fusaroli and Tylén (2016) sampled the f_o every 10 ms, we sampled it at 50 ms due to memory issues when computing the CRQA in R. The data were normalized by dividing the raw values by the mean of the speaker in the entire conversation. To determine the value of the delay parameter, we followed a procedure by Abarbanel (1996) that involves calculating the first local minimum of an average mutual information function per conversation. Next, we used the resulting delay parameter to compute the embedding dimension parameter following a procedure described by Kennel et al. (1992) and Abarbanel (1996), which involved calculating the false-nearest neighbor function. For our analysis, we selected both the maximum delay (d = 27) and the maximum embedding dimension (m = 1) as parameters for the CRQA for all conversations. We chose for both parameters the maximal values across all 20 conversations as overembedding is less problematic than underembedding (Webber & Zbilut, 2005). The value for the radius parameter was also kept constant for each conversation at 0.45, so all of the RRs of the real conversations were between 1% and 5% (following recommendations from Wallot & Leonardi, 2018; Webber & Zbilut, 2005). CRQA was executed using the crqa package (Version 2.0.2; Coco & Dale, 2014) in R (R Version 4.0.4; R Core Team, 2018).

The RR of the real conversations was compared with the RR of randomly shuffled conversations created following the suggestions by Fusaroli and Tylén (2016) and Wallot and Leonardi (2018). More specifically, we randomly shuffled the two time series per dyad 50 times and calculated the RR of the surrogate data using the same parameters as those we use for the real conversations. Next, a one-sample t test showed whether the RR of the real conversation was significantly higher than the mean RR of 50 randomly shuffled conversations (i.e., whether the dyad displayed entrainment at an above-chance level in the conversation).

Corpus

This study uses the "LUCID corpus," which was developed and made freely available by Baker and Hazan (2011). In this corpus, 40 native English speakers (19–29 years old, M = 22.6. ± 2.75; 20F, 20 men) were grouped into 20 same-sex dyads that were familiar with one another. Each participant had normal hearing and did not report a history of speech or language disorders. Participants did the Diapix task, a collaborative "spot-the-differences" task in which they were each given a picture that their interlocutor could not see. Participants then had to compare and discuss their pictures to identify their differences. For this study, only the first interactions of the Diapix task were used (i.e., the .wav files and accompanying TextGrids whose filenames end in "cv1"), which had a mean length of 490.87 s (± 159.70 s). More information on the corpus and the participants who participated in the task can be found on Hazan's website.[2]

Preprocessing and Feature Extraction

To facilitate feature extraction, all annotations of laughter, sighs, and other nonspeech sounds were changed to silence annotations. These updated TextGrids have been made available on OSF, as are the Praat scripts that we used for the feature extraction.

For the IPU-based analyses, features were extracted from periods of speech that are surrounded by silences of 50 ms or more (i.e., IPUs). Feature extraction was done using Praat (Boersma, 2006). Median f_o, f_o range, and maximum f_o (max f_o) were extracted in Hertz using pitch floors and ceilings that were adjusted for gender (male: floor = 50, ceiling = 350; female: floor = 75, ceiling = 500) and speaker in order to avoid pitch-tracking issues such as octave jumps. We followed De Looze and Rauzy (2009) in using the following formula to set each speaker's pitch ceiling and floor, respectively: 0.65 quantile of speaker mean × 1.90 and 0.35 quantile of speaker mean × 0.72. We chose to extract median f_o rather than mean f_o, as this is also more robust against pitch-tracking errors. If Praat returned "undefined" for f_o value for an IPU, for example, because a segment of speech was unvoiced, this utterance was excluded from subsequent analyses. The resulting data set was used for the local and global analyses.

The time series–based analyses followed a similar feature extraction process to that described above, except that features were extracted from windows of a larger size than IPUs. For TAMA and WLCC, the window size was 20 s, and the step size was 10 s. The base windows used in the HYBRID method were the same as in TAMA and WLCC (i.e., 20 s) but varied in size to encompass the entire utterance of a speaker within a window (see De Looze et al., 2014). For CRQA, the features were extracted every 50 ms within the pitch floor and ceiling levels previously described.

Results

The results of the analyses will be briefly outlined in this section. For a more detailed insight into the specifics of the results for each method, see Appendices B–K. The results of each method for median f_o entrainment are presented in Table 2.

Local Methods

Levitan and Hirschberg: Local Proximity, Local Synchrony, Local Convergence

Levitan and Hirschberg's (2011) local proximity measures entrainment by comparing two adjacent utterances by different speakers and testing whether the difference between these two is smaller than the difference between one utterance and 10 random, nonadjacent utterances. Results of the *t* tests for local proximity suggested that on median f_o, five out of 20 dyads showed significant entrainment, while no significant (dis)entrainment was observed for any dyads on f_o range and two out of 20 dyads exhibited significant entrainment on max f_o. For a detailed overview of results, see Appendix B.

Levitan and Hirschberg's (2011) local synchrony measures entrainment by correlating features that were extracted from two adjacent utterances by different speakers. Results of the Pearson correlations for local synchrony suggest that three dyads showed significant entrainment while three dyads exhibited significant disentrainment on median f_o. On f_o range, one dyad entrained while one disentrained. Finally, on the dimension of local synchrony and feature of max f_o, two dyads showed significant entrainment and two dyads showed disentrainment. For a detailed overview of results, see Appendix C.

Levitan and Hirschberg's (2011) local convergence measures entrainment by testing whether the difference in features between two adjacent utterances by different speakers decreases over time. Results of Pearson correlations revealed that, on median f_o, five out of 20 dyads exhibited significant entrainment while two out of 20 dyads showed significant disentrainment. One out of 20 dyads significantly entrained on f_o range, and for the feature of max f_o, significant entrainment was found in one dyad while significant disentrainment was found in another one out of 20 conversations. For a detailed overview of results, see Appendix D.

Schweitzer and Lewandowski: Mixed-Effects Models

Schweitzer and Lewandowski (2013) used mixed-effects models to test whether the feature of an utterance by one speaker could be predicted by the feature of the preceding utterance, produced by the other speaker. For median f_o, results of the lmerTest ANOVA comparing the full model to the null model (excluding the fixed effect of preceding utterance feature) suggested that the full model was a significantly better fit, $\chi^2(1) = 88.64$, $p < .001$. Furthermore, results from lmerTest suggest that gender is a significant main effect in the full model, $b = -84.42$, $t(23.75) = -19.41$, $p < .001$. This suggests that the median f_o difference between males and females is

[2] https://valeriehazan.com/wp/index.php/lucid-corpus-london-ucl-clear-speech-in-interaction/.

Table 2. Results of the 12 methods for quantifying entrainment on median f_o. Entrainment is indicated with a "+" while disentrainment is indicated with a "–".

	Local[a]				Global[a]				Time series[a]			
Dyad	L&H prox	L&H conv	L&H sync	LMEM[b]	L&H prox	L&H conv [b]	geom.	HYBRID	TAMA	HYBRID	WLCC	CRQA
F03F04	+			+		+	+		+		+	+
F11F12	+	–	+	+		+	+		+		–	+
F13F14		+		+		+	–				+	+
F15F16				+		+	–		+	+	–	+
F21F22	+	+	+	+		+	–		+		+	+
F25F26	+		+	+		+	+			+	+	+
F31F32				+		+	–				+	+
F37F38			–	+		+	+				–	+
F41F42				+		+	–		+	+	–	+
F47F48				+		+	–		+	+	+	+
M07M08		+	–	+		+	–				–	+
M09M10				+		+	+		+		+	+
M11M12				+		+	–				–	+
M13M14				+		+	–		+	+	–	+
M15M16				+		+	–				+	+
M17M18		–	–	+		+	+				+	+
M25M26				+		+	+				+	+
M33M34				+		+	–					+
M35M36	+	+		+		+	–				+	+
M41M42				+		+	+				+	+

Note. If a method suggested that no significant (dis)entrainment occurred in an interaction, the corresponding cell in the table is left empty. Note that some methods (e.g., Lehnert-LeHouillier, Terrazas, Sandoval, and Boren's geometric approach) did not rely on measures of significance. All results from this method are thus provided.

[a] Legend: L&H = Levitan and Hirschberg (2011); prox = proximity; conv = convergence; sync = synchrony; LMEM = linear mixed-effects models; geom. = Lehnert-LeHouillier, Terrazas, Sandoval, and Boren's geometric approach; TAMA = time-aligned moving average; WLCC = windowed lagged cross-correlation; CRQA = cross-recurrence quantification analysis. [b] Note that these methods measure entrainment over the entire corpus, rather than per conversation.

approximately 84 Hz, which is not surprising. Additionally, lmerTest suggested that the median f_o of the preceding utterance was a significant main effect, b = 0.16, t(4145.76) = 9.52, p < .001. According to the model, the target utterance increases by 0.16 Hz when the preceding utterance increases by 1 Hz. In other words, if the preceding utterance is higher in f_o, so is the utterance that directly follows it, thereby suggesting that entrainment occurred on median f_o.

On the contrary, the lmerTest ANOVA conducted on the models for f_o range did not return a significant result, $\chi^2(1)$ = 2.65, p = .10, suggesting that the full model is not a significantly better fit for the data. In the full model, gender was a significant main effect according to the output provided by lmerTest, b = −37.6, t(18.20) = −6.87, p < .001, such that men had a larger f_o range than women by 37.6 Hz, on average. f_o range of the preceding utterance was not found to be a significant main effect, b = 0.02, t(4150.10) = 1.61, p = .107, suggesting that no significant entrainment occurred on f_o range.

For max f_o, the lmerTest ANOVA results suggested that the full model is a significantly better fit than the null model, χ^2 (1) = 39.08, p < .001. In the full model, lmerTest results suggested that gender was a significant fixed effect, b = −114.82, t(21.61) = −17.78, p < .001, such that there was, on average, approximately a 114-Hz difference between the max f_o of men and women, which again is not surprising. Additionally, lmerTest results of the full model suggest that the f_o max of the preceding utterance is a significant main effect, b = 0.10, t(4148.92) = 6.28, p < .001, which suggests that entrainment occurred on max f_o. For a more detailed overview of results, see Appendix E.

Global Methods

Static Global Proximity: Levitan and Hirschberg's Global Proximity

Levitan and Hirschberg's (2011) global proximity tests whether speakers spoke more similarly to their conversation partner than to the speakers in the corpus with whom they did not interact. No significant entrainment nor disentrainment was found on the dimension of global proximity for median f_o, $t(19) = -1.87$, $p = .077$, while significant entrainment was found on f_o range, $t(19) = -3.07$, $p = .006$, and max f_o, $t(19) = -2.44$, $p = .024$. For details on the results of this analysis, see Appendix F.

Dynamic Global Proximity: Levitan and Hirschberg's Global Convergence

Levitan and Hirschberg's (2011) global convergence measures whether two speakers spoke more similarly during the second half of an interaction than during the first half. Significant global convergence was found on median f_o, $t(19) = 2.45$, $p = .024$, but no significant results were obtained regarding (dis)entrainment on f_o range, $t(19) = -0.38$, $p = .705$, or max f_o, $t(19) = -0.83$, $p = .415$. For more details on the results of this analysis, see Appendix F.

Lehnert-LeHouillier, Terrazas, Sandoval, and Boren: Geometric Approach

The geometric approach by Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020) was used to investigate whether two speakers spoke more similarly in the final third of their conversation than during the first third. For median f_o, entrainment was found in eight out of 20 dyads, while disentrainment was observed in the remaining 12 dyads. Out of 20 dyads, six exhibited entrainment on f_o range, while 14 showed disentrainment. For max f_o, eight of the dyads showed entrainment, while the other 12 showed disentrainment. Note that this method does not provide information regarding the statistical significance of the results. For more details on the results of this analysis, such as the contributions of each individual to the observed entrainment, see Appendix G.

Time Series–Based Methods

Kousidis et al.: TAMA

We used TAMA by Kousidis et al. (2008) to measure the cross-correlation between the features of two speakers. Pearson correlations over the entire conversation suggest that seven out of 20 dyads showed statistically significant entrainment (i.e., $p < .05$) on solely median f_o, while another dyad featured significant entrainment on f_o range and max f_o. Moreover, one dyad displayed significant entrainment on all measures (i.e., median f_o, f_o range, and max f_o). Details on these analyses can be found in Appendix H.

De Looze et al.: HYBRID Method

The HYBRID method introduced by De Looze et al. (2014) tests, just like TAMA, the cross-correlation between the features of two speakers. One dyad displayed entrainment for all measures (i.e., median f_o, f_o range, and max f_o) when looking at the entire conversation. Another four dyads only showed significant entrainment in median f_o. The correlation for the other dyads did not achieve significance (i.e., $p < .05$).

Furthermore, the HYBRID method was also used to assess how the degree of entrainment changes throughout a conversation. We did this by calculating a Pearson correlation across just one period of the conversation (i.e., stretch of 10 windows) and repeating the same process every five windows along. Out of 20 dyads, seven displayed at least one period of significant entrainment on median f_o, and only one of these dyads showcased two periods of entrainment. Four dyads exhibited periods of significant disentrainment on median f_o, with one dyad showing two periods of disentrainment. Additionally, eight dyads showed periods of entrainment on f_o range, with six dyads showing such behavior during more than one period. Seven dyads displayed periods of disentrainment on f_o range, with one dyad showing two periods of disentrainment. Five out of the eight dyads that showed at least one period of entrainment on f_o range also showed periods of disentrainment on this same feature. For max f_o, nine dyads showed periods of significant entrainment, with seven dyads exhibiting more than one such period. Three dyads exhibited periods of disentrainment, with only one dyad showcasing more than one such period. Here, only one of the nine dyads showing entrainment also showed disentrainment during the conversation. More detailed information regarding the results can be found in Appendix I.

Finally, the HYBRID method sees whether the cross-correlation between the features of two speakers is stronger during the last half of the conversation compared to the first half. No significant differences in Pearson correlation were found between the first and second half of the conversation for median f_o, $t(19) = -0.93$, $p = .36$; f_o range, $t(19) = 1.26$, $p = .22$; and f_o max, $t(19) = 1.03$, $p = .32$.

Boker et al.: WLCC

Similarly to TAMA and the HYBRID method, WLCC by Boker et al. (2002) uses cross-correlations between speakers to assess entrainment. WLCC takes into account that entrainment can happen with a temporal delay. Results of the WLCC analyses suggest that 19 out of the 20 dyads displayed significantly higher absolute peak cross-correlations than the pseudo-interactions in median f_o. Using the mean values of the real conversations, we inferred that 12 conversations showed overall entrainment, while the seven other conversations displayed overall disentrainment. The same number of dyads (i.e., 19) showed higher absolute peak cross-correlations than the pseudo-interactions in f_o range, of which 11 exhibited overall entrainment. Eighteen of the 20 dyads reached significance for the measure of max f_o entrainment, with 11 showing overall entrainment and seven disentrainment.

The overall difference between the real and pseudo-interactions was significant for all measures, namely, median f_o (z = 13.03, $p < .01$), max f_o (z = 13.84, $p < .01$), and f_o range (z = 12.48, $p < .01$), suggesting that significant differences between the real conversations and the pseudo-interactions were observed on this feature across all conversations. Please refer to Appendix J for more details on the results of these analyses.

Fusaroli and Tylén: CRQA

CRQA (e.g., Fusaroli & Tylén, 2016) measures how many times the two speakers spoke similarly across the entire conversation and whether this amount is above chance level or not. RR, or the amount of time speakers spoke similarly, was 3.55% ± 0.76%. All dyads showed significantly higher amounts of RRs than surrogate samples, indicating that every dyad showed entrainment above chance level. Please refer to Appendix K for a more detailed breakdown of the analyses.

Interim Discussion

As can be observed in the various results presented in the previous section and in Table 2, little consistency was observed between the results of these different methods used to measure prosodic entrainment in the same corpus. One possible source of this variation in results could be that different methods rely on different norming procedures. For all the analyses we conducted, we aimed to stay as true as possible to the original methods as they were described in their original publications, which is why we sometimes implemented different norming procedures. To investigate whether different norming procedures account for some of the variation in results, a brief follow-up study will be presented in which we aimed to investigate whether different norming procedures can lead to varying results when entrainment is measured in the same corpus and using the same local entrainment measurement methods. Specifically, we used the local methods by Levitan and Hirschberg (2011), since these are some of the most commonly used entrainment measurements in the literature.

Follow-Up Norming Study

In the speech sciences in general, norming is done for a variety of reasons, such as accounting for sex-based differences between speakers (e.g., Adank et al., 2004), to ensure that any measured differences between speakers are due to variables of interest rather than differences such as differences in vocal tract length or body size. Many different norming procedures exist for several different purposes. For example, some norming procedures may focus exclusively on norming features of vowels (Adank et al., 2004).

Since this follow-up study is focused on entrainment research, we will only include norming methods that have been used in previously published studies that measure entrainment. Specifically, we will focus on f_o entrainment, since the majority of recently published papers include measures of entrainment on this feature.

In such studies, norming seems to be used mainly to minimize the effects of either sex-based differences or individual differences between speakers. A gender-based norming procedure was implemented by Levitan and Hirschberg (2011), who used gender means and standard deviations to z score raw values (described in Dynamic Local Proximity: Levitan and Hirschberg's (2011) Local Convergence section in more detail). When it comes to norming for individual differences, various approaches have been taken. In several entrainment studies that employ cross-correlations between extracted time series, such as studies of Kousidis et al. (2008) or De Looze et al. (2014), raw features were extracted from within windows of a set number of seconds, and then these raw features were divided by a speaker's mean over all utterances, weighted by duration. Other approaches have also been taken to account for individual variability. In the study of Schweitzer and Lewandowski (2013), for example, raw speech features were entered into LMEMs, where individual speaker was added as a random effect. This can be considered a type of norming, since individual variability is somehow controlled for. In yet other studies, entrainment measurements are derived from raw, unnormalized features (e.g., Lehnert-LeHouillier, Terrazas, & Sandoval, 2020; Lehnert-LeHouillier, Terrazas, Sandoval, & Boren, 2020).

Follow-Up Norming Study: Methods

In this follow-up study, we used the same speech corpus as we used for the other analyses (see Corpus section for more details). We extracted features from it and normed these in several different ways before measuring entrainment using three methods. We employed two different norming methods: speaker- and gender-based norming. We followed Levitan and Hirschberg (2011) and used a gender-based z-scoring procedure to norm our data by gender:

$$\text{feature value in IPU} = (\text{raw feature value} - \text{gender mean}) / \text{gender standard deviation}. \tag{2}$$

Because we are not extracting features from windows but from IPUs, it does not make sense to implement a norming procedure that involves weighting features by window duration as was used by Kousidis et al. (2008)

and De Looze et al. (2014). Instead, we normed our data by speaker by using a z-scoring procedure similar to our gender-based norming:

$$\text{feature value in IPU} = (\text{raw feature value} - \text{speaker mean}) / \text{speaker standard deviation}. \qquad (3)$$

Finally, we ran the entrainment analyses using the raw extracted f_o values as was done, for example, by Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020). We measured local entrainment using methods by Levitan and Hirschberg (2011), that is, proximity, convergence, and synchrony, as these are some of the most commonly used methods in the field. For more details on these methods, see Static Local Proximity: Levitan and Hirschberg's Local Proximity, Static Local Synchrony: Levitan and Hirschberg's Local Synchrony, and Dynamic Local Proximity: Levitan and Hirschberg's Local Convergence sections, respectively, where the methods are presented in more detail.

Follow-Up Norming Study: Results

Table 3 presents the results of the local proximity analyses. Differences in results based on norming can be seen in two of 20 conversations, so norming procedure seemed to influence results in 10% of cases. In both cases, no significant (dis)entrainment was found in gender-normed data, while significant entrainment was measured in the raw and speaker-normed data. Results of the raw and speaker-normed data are always the same.

Results for the local convergence analyses are presented in Table 4, which shows that the same results are found regardless of norming in 15 of 20 conversations, so differences are found in 25% of cases. The results of gender-normed and raw data are the same in 17 of 20 conversations, so most differences can be observed between these methods and speaker-normed data. Additionally, fewer significant results are obtained from the speaker-normed data (three) than from the other two data sets (five each).

Table 5 presents the measurements of local synchrony. Results differ in six of 20 conversations, so in 30% of cases. Again, few differences can be seen between raw and gender-normed data, and most differences are between these sets and speaker-normed data. Interestingly, no significant disentrainment was measured in the speaker-normed data, while it was found in both the raw and gender-normed data.

In total, when all three analyses are taken together, results suggest that different norming procedures lead to

Table 3. Results of the paired *t* tests conducted to measure local proximity in differently normed datasets.

	Gender						Raw						Speaker z					
Dyad	entr.	*t*	*df*	*p*	PD	OD	entr.	*t*	*df*	*p*	PD	OD	entr.	*t*	*df*	*p*	PD	OD
F03F04		−1.87	182	.063	1.52	1.68	+	−2.09	182	.038	21.65	24.42	+	−2.41	182	.017	0.97	1.10
F11F12	+	−2.45	299	.015	1.82	2.02	+	−2.73	299	.007	25.87	28.98	+	−2.11	299	.036	0.96	1.06
F13F14		−1.31	186	.192	1.82	2.00		−1.73	186	.086	25.89	29.33		−0.85	186	.395	1.01	1.09
F15F16		0.33	165	.743	1.76	1.76		0.11	165	.909	25.01	25.23		−1.89	165	.061	0.90	1.06
F21F22	+	−2.88	207	.004	1.69	2.04	+	−2.44	207	.016	24.00	28.34	+	−3.57	207	< .001	0.82	1.05
F25F26	+	−2.89	148	.004	1.00	1.19	+	−2.62	148	.010	14.18	16.68	+	−3.25	148	.001	0.86	1.04
F31F32		−0.63	252	.532	1.60	1.65		0.12	252	.902	22.71	22.62		0.32	252	.749	1.13	1.11
F37F38		0.11	161	.910	2.74	2.73		0.67	161	.503	39.05	37.77		−1.67	161	.096	0.93	1.03
F41F42		−0.22	144	.825	1.36	1.39		0.13	144	.897	19.38	19.39		−0.98	144	.330	1.15	1.22
F47F48		−0.79	140	.429	1.83	1.94		−0.13	140	.896	26.00	26.58		−0.84	140	.404	0.97	1.05
M07M08		−0.21	188	.831	1.97	1.99		−0.21	188	.837	21.42	21.54		−1.45	188	.148	0.98	1.06
M09M10		−1.04	317	.298	0.92	0.96		−0.73	317	.464	9.95	10.28		−1.18	317	.240	0.94	1.00
M11M12		−1.67	265	.095	1.08	1.17		−0.64	265	.523	11.69	12.08		−1.57	265	.118	0.99	1.08
M13M14		−0.51	248	.613	1.40	1.44		−0.60	248	.548	15.24	15.63		−1.70	248	.091	0.91	1.00
M15M16		−0.59	200	.554	0.99	1.01		−1.09	200	.278	10.70	11.25		−1.50	200	.134	1.04	1.12
M17M18		−0.49	298	.622	2.28	2.32		0.18	298	.855	24.70	24.59		−0.79	298	.431	1.01	1.04
M25M26		−1.39	208	.165	1.08	1.15		−1.59	208	.114	11.76	12.68		−1.18	208	.240	0.90	0.95
M33M34		0.15	154	.880	0.95	0.94		−0.19	154	.848	10.27	10.38		−0.68	154	.499	1.01	1.04
M35M36		−1.67	250	.097	1.00	1.09	+	−2.33	250	.021	10.85	12.28	+	−2.39	250	.018	1.00	1.13
M41M42		−1.21	124	.230	0.90	0.97		−0.41	124	.680	9.77	10.02		−0.80	124	.424	1.07	1.12

Note. Mean partner and other differences are presented in the "PD" and "OD" columns, respectively. The "entr." column indicates whether significant entrainment occurred ("+") or no significant results were found (cell left blank).

Table 4. Results of the local convergence analyses conducted on differently normed datasets.

Dyad	Gender				Raw				Speaker z			
	entr.	*r*	*p*	sig./10	entr.	*r*	*p*	sig./10	entr	*r*	*p*	sig./10
F03F04		–.09	.250	2		–.09	.250	0		–.03	.677	0
F11F12	–	–.17	.003	0	–	–.17	.003	0	–	–.18	.002	0
F13F14	+	.21	.004	1	+	.21	.004	1	+	.23	.001	0
F15F16		.11	.147	0		.11	.147	0		.03	.703	0
F21F22	+	.17	.012	1	x	.17	.012	2		.12	.086	0
F25F26		–.05	.536	0		–.05	.536	0		–.01	.869	0
F31F32		.10	.127	0		.10	.127	2		.05	.391	0
F37F38		.03	.660	0		.03	.660	0		.05	.535	1
F41F42		.02	.790	0		.02	.790	0		–.09	.299	0
F47F48		.05	.560	1		.05	.560	1		.15	.080	0
M07M08	x	.28	< .001	8	x	.28	< .001	7		–.06	.446	0
M09M10		.02	.734	0		.02	.734	0		.06	.255	0
M11M12		.00	.981	0		.00	.981	0		.06	.328	0
M13M14		–.03	.674	0		–.03	.674	0		–.10	.119	0
M15M16		.13	.067	1		.13	.067	0		–.02	.728	0
M17M18	–	–.17	.003	1	–	–.17	.003	0	–	–.15	.011	1
M25M26	x	–.21	.002	3	x	–.21	.002	4		–.07	.323	2
M33M34		.00	.989	0		.00	.989	0		–.06	.491	0
M35M36	x	.15	.017	2		.15	.017	0		.10	.121	0
M41M42	+	.20	.025	0		.20	.025	1	x	.20	.028	2

Note. The “sig./10” column represents how many of the correlations performed on randomly shuffled data returned a significant result. If 1 > such correlation was significant, any significant results are considered invalid. The “entr.10” column indicates whether significant entrainment (+) or disentrainment occurred (–), results are invalid (×), or no significant results were found (cell left blank).

Table 5. Results of the local synchrony analyses conducted on differently normed datasets.

Dyad	Gender				Raw				Speaker z			
	entr.	*r*	*p*	sig./10	entr.	*r*	*p*	sig./10	entr	*r*	*p*	sig./10
F03F04	x	.16	.027	2	+	.16	.027	0	+	.20	.006	0
F11F12	+	.19	.001	1	+	.19	.001	0	+	.16	.004	0
F13F14		.03	.709	0		.03	.709	1		.11	.152	0
F15F16		–.12	.137	0		–.12	.137	1		.06	.445	0
F21F22	+	.27	< .001	0	+	.27	< .001	0	+	.38	< .001	0
F25F26	+	.27	.001	0	+	.27	.001	1	+	.27	.001	0
F31F32		–.01	.861	0		–.01	.861	0		.01	.847	1
F37F38	–	–.28	< .001	0	x	–.28	< .001	3		.11	.171	1
F41F42		–.06	.439	1		–.06	.439	0		.01	.916	1
F47F48		–.05	.568	0		–.05	.568	0		.11	.205	1
M07M08	–	–.50	< .001	1	–	–.50	< .001	1		.10	.189	0
M09M10		.00	.931	1		.00	.931	1		.09	.091	0
M11M12		.01	.847	0		.01	.847	1	+	.12	.046	1
M13M14		–.05	.423	0		–.05	.423	0		.06	.357	1
M15M16		–.01	.912	1		–.01	.912	0		.14	.052	0
M17M18	–	–.34	< .001	0	–	–.34	< .001	0		.03	.661	0
M25M26		.03	.655	0		.03	.655	1	+	.14	.048	1
M33M34		–.03	.747	0		–.03	.747	0		.04	.629	0
M35M36		.09	.137	2		.09	.137	0		.12	.055	1
M41M42		.07	.433	0		.07	.433	1		.09	.334	1

Note. The “sig./10” column represents how many of the correlations performed on randomly shuffled data returned a significant result. If 1 > such correlation was significant, any significant results are considered invalid. The “entr.” column indicates whether significant entrainment (+) or disentrainment occurred (–), results are invalid (×), or no significant results were found (cell left blank).

differing results in 12 of 60 conversations, or in 20% of all cases.

Follow-Up Norming Study: Discussion

This follow-up study aimed to investigate the effect of norming procedures on the outcome of entrainment measurements. In total, results from the different analyses were the same regardless of norming procedure for 48 out of 60 tests. In other words, in 80% of the cases, norming did not seem to affect entrainment measurements.

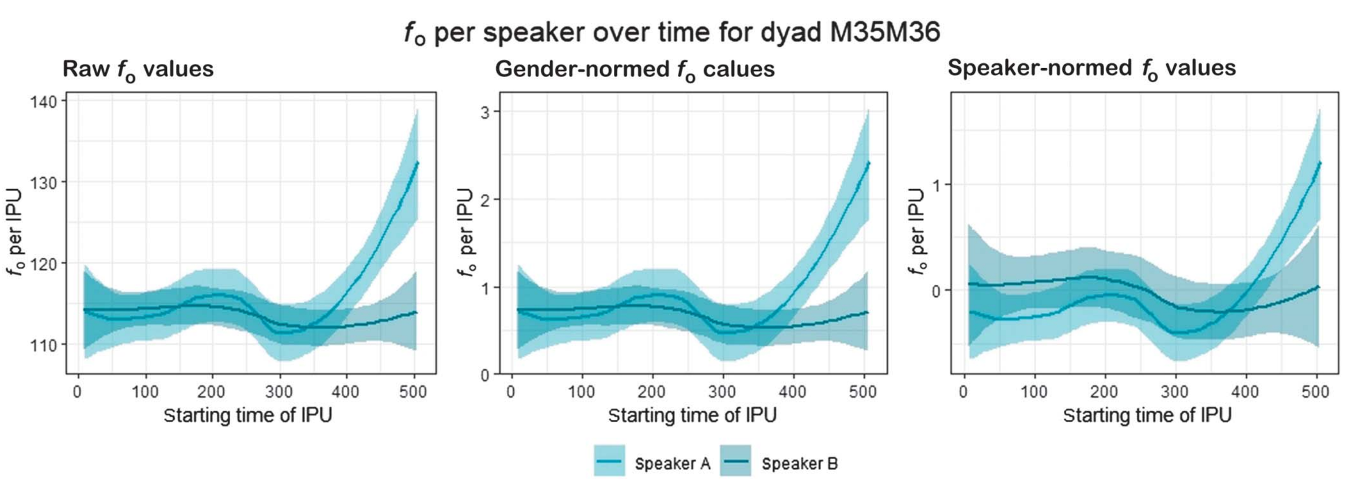
The results raise several questions. First of all, they suggest that other methodological factors may be influencing results in entrainment research, as differences in norming seem to explain some, but not all, of the variance in results found in the previously presented results in Results section. Future research may also elucidate which other decisions during the analysis process, such as, for example, the setting of parameters related to pitch-tracking issues and outlier removal, can further explain some of the discrepancies.

Norming procedure seems to influence entrainment measurements in 20% of all cases. Specifically, most differences were found between speaker-normed data versus the raw and gender-based normed data. This suggests that gender-based norming may not be as reliable a norming procedure as speaker-based norming. After all, the purpose of gender-based norming is to account for gender-based differences between speakers, so if results of entrainment measurements on gender-normed data closely mimic the results of the same analyses on raw data, it may not be as effective at controlling for gender-based differences as it should be. Indeed, in later works by the same authors as Levitan and Hirschberg (2011; e.g., Levitan et al., 2012), they use a speaker-based, rather than a gender-based, norming procedure.

Something that stands out in Tables 4 and 5 is that the test statistics for the raw and gender-normed data are identical. This is because both measures rely on a Pearson correlation, and the gender-norming procedure was a linear function. In the corpus used in this study, all dyads were gender-matched, meaning that the raw data for each speaker underwent the exact same linear transformation with the same gender mean and standard deviation as their partner; in other words, the correlation between the two sets of values did not change. This can be seen when the raw and normed values of each IPU are plotted against their starting time (see Figure 2): Although the scale on the *y*-axis is different for the raw and gender-normed values, the relationship between the features of both speakers does not differ. This highlights that implementing gender-based norming when using correlation-based methods to calculate entrainment may not be a suitable choice if dyads are matched on gender.

Interestingly, despite the fact that the relationship between speakers in the raw and gender-normed data sets remains unchanged after norming, the methods did occasionally lead to different results. For example, in conversation M35M36 (see Table 4 and Figure 2), the local convergence assessment found entrainment in the raw data, while results were considered invalid for the gender-normed data. This can be traced back to the validation step suggested by Levitan and Hirschberg (2011): All correlations are repeated 10 times on randomly shuffled data, and results are only considered valid if no more than one of these correlations returns a significant result. The fact that our analysis on raw and gender-normed data can return the exact same test statistics but lead to different results (i.e.,

Figure 2. Plots of raw or normed median fundamental frequency (f_o) values per interpausal unit (IPU) in dyad M35M36.



significant disentrainment or invalid results) begs a question about yet another potential source of variation: the generation of surrogate or baseline data. This will be discussed in more detail in the Possible Sources of Variation section.

To summarize the results of this follow-up study, different norming procedures may influence the outcomes of entrainment analyses. In the entrainment measurements conducted in this follow-up study, using raw values and implementing a gender-based norming procedure seem to lead to very similar results when using correlation-based methods, likely because our corpus consisted of gender-matched dyads. Importantly, differences in norming procedures accounted for some, but not all, of the variance in results observed in the previously presented analyses. Other possible sources of this variance, as well as the practical and theoretical implications of a lack of agreement between results, will be discussed in the next section.

General Discussion

In the main study of this article, we set out to compare the results of 12 methods that quantify or assess prosodic entrainment in the same corpus. The methods were divided in three broad groups based on the timescale in which entrainment was measured to happen: local, global, and time series–based methods. Three main patterns stand out from the results of our analyses: There seems to be little correlation

1. between entrainment on different features (see Appendices B–K),
2. between the results of methods in different groups (see Table 2), and
3. between the results of different methods within a group.

A follow-up study was conducted to investigate a potential source of this variation in results, namely, the effect of different norming procedures (see Follow-Up Norming Study section). Results of this follow-up study suggest that norming can explain some, but not all, of the variance observed in the results. Since differences in norming procedures cannot explain all the variance in results, each of the three main findings listed above will be discussed in the following sections. Possible explanations for these findings will be discussed, and practical and theoretical implications of the findings will be outlined.

Interpretation of Results

Entrainment on median f_o, max f_o, and f_o range do not seem to correlate with one another, even when entrainment on these features is measured by the same method (see the Appendices B–K for more details). This finding is in line with existing research that suggests that if an individual entrains on one feature, they may not necessarily entrain on a different feature, and that entrainment on different features appears to be unrelated (e.g., Priva & Sanker, 2018; Sanker, 2015; Weise & Levitan, 2018). This finding highlights the notion that entrainment may not be a single construct but rather a set of behaviors that may be governed by different mechanisms.

Entrainment results for max f_o and f_o range are more similar than entrainment results between these features and median f_o, which makes sense considering the fact that max f_o and f_o range are closely linked. Methods in which results for max f_o and f_o range are relatively similar may be more reliable at capturing entrainment than methods for which the results of these two features differ drastically.

The main purpose of this article is to compare the results of different methods for measuring entrainment to investigate the relationship between different subtypes of entrainment and the methods used to measure these different subtypes. To facilitate comparison between results of different methods, the remainder of this discussion will focus on the results pertaining to median f_o entrainment. Results of the different methods we implemented for median f_o entrainment are presented in Table 2. This table illustrates minimal agreement between the methods used in this study: No clear patterns can be observed in the results. This could be expected to some extent, since different methods were developed for different goals and therefore possibly measure different subtypes of entrainment. Results of the methodologies used in this article could thus also be interpreted as suggesting that there is no clear pattern between different subtypes of entrainment in conversations: If one subtype of entrainment is measured, that does not mean other subtypes of entrainment are also present in the same conversation.

It is worth noting that some methods used in this study seem to detect a lot of entrainment, while other methods return no significant results. For example, HYBRID returned significant results in five out of 20 conversations, while no significant (dis)entrainment was found on median f_o by Levitan and Hirschberg's global proximity (2011)—though it should be noted that the latter produces one measurement for the entire corpus rather than per conversation. Additionally, some methods indicated significant disentrainment in some conversations (e.g., Levitan and Hirschberg's local synchrony; Levitan & Hirschberg, 2011), while others only detected entrainment (e.g., TAMA). It is possible that different methods are more or less conservative in their detection of entrainment, though it is difficult to substantiate this claim as there is no "gold standard" for measuring entrainment.

To further compare the results of different methods, we will focus on one conversation as an example. Figure 3 shows a smoothed plot of median f_o per IPU against the starting time of each IPU of one dyad (F37F38). Upon visual inspection, the conversation seems to display a high level of entrainment: Both speakers seem to covary their median f_o across the conversation, where one speaker's rise in median f_o goes along with the rise in median f_o for the other speaker. One may thus expect that most methods would suggest that entrainment had occurred in this conversation, especially the methods that measure synchrony. However, this is not the case: Only four out of 12 methods implemented in this study suggest that entrainment occurred in this conversation, though two of these (Levitan and Hirschberg's global convergence and global proximity [Levitan & Hirschberg, 2011] and Schweitzer and Lewandowski's LMEM [Schweitzer & Lewandowski, 2013]) produce a single entrainment measurement for the entire corpus, so it is difficult to draw any conclusions about (dis)entrainment in individual conversations such as F37F38. Moreover, two methods suggest that disentrainment was observed (see Table 6).

Interestingly, visual inspection of conversation F37F38 (see Figure 3) suggests high levels of synchrony, but only one method that measures synchrony seemed to find significant entrainment in F37F38 (see Table 6), namely, Schweitzer and Lewandowski's (2013) LMEM approach. Importantly, this method measures entrainment over a whole corpus, so it is difficult to say whether this method measured significant synchrony in F37F38, specifically. Methods that measure synchrony in each individual conversation suggest that there is either no significant entrainment (TAMA; HYBRID) or even measure significant disentrainment (Levitan and Hirschberg's measure for local synchrony [Levitan & Hirschberg, 2011]; WLCC).

On the contrary, most methods that seemed to detect significant entrainment in conversation F37F38 measured proximity (i.e., Levitan and Hirschberg's global convergence, Lehnert-LeHouillier et al.'s geometric approach, and CRQA). Again, it should be noted that some of the methods that produced significant results for conversation F37F38, such as Schweitzer and Lewandowski's (2013) LMEM and Levitan and Hirschberg's (2011) global convergence, produced a measure of entrainment over all conversations pooled together. It is thus difficult to conclude whether (dis)entrainment occurred in a specific conversation in the corpus. In summary, visual inspection suggests high synchrony, but methods that measure synchrony do not. Results suggest that there may be a high degree of proximity in F37F38, though not all methods that assess proximity

Figure 3. Median fundamental frequency (f_o) of conversation F37F38 per speaker against the starting time of the utterance.

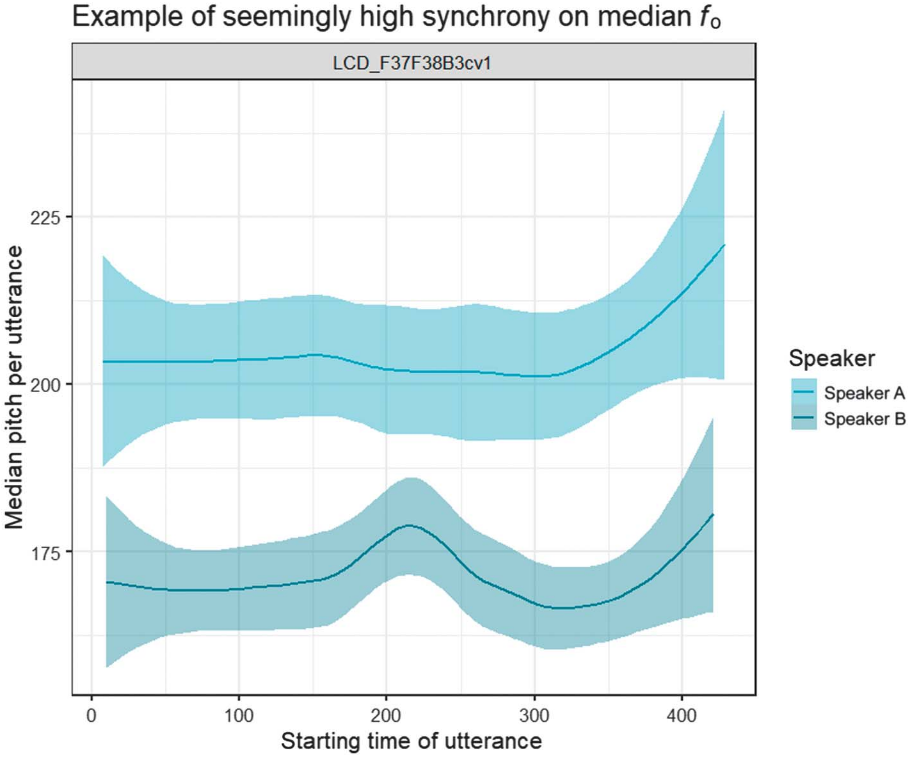


Table 6. Results of the analyses for conversation F37F38, where + indicates entrainment and – indicates disentrainment.

Variable		Time-scale		
		Local	**Time series**	**Global**
Dimension	Proximity	**Static local proximity** Local proximity, Levitan & Hirschberg (2011)	**Static global proximity** +CRQA, Fusaroli & Tylén (2016)	**Static global proximity** Global proximity, Levitan & Hirschberg (2011)
		Dynamic local proximity Local convergence, Levitan & Hirschberg (2011)		**Dynamic global proximity** +*Global convergence, Levitan & Hirschberg (2011) +Geometric approach, Lehnert-LeHouillier, Terrazas, & Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, & Boren (2020)
	Synchrony	**Static local synchrony** –Local synchrony, Levitan & Hirschberg (2011) +*Linear mixed-effects models, Schweitzer & Lewandowski (2013)	**Static global synchrony** TAMA, Kousidis et al. (2008) HYBRID, De Looze et al. (2014) –WLCC, Boker et al. (2002)	
			Dynamic global synchrony HYBRID, De Looze et al. (2014)	

Note. An absence of either a + or – means that this method did not find entrainment. The asterisk (*) indicates that a method measures entrainment in all analyzed conversations at once, rather than per conversation. Please note that following the Wynn and Borrie (2022) framework, all time series–based methods are considered to measure entrainment globally.

returned significant results. In other words, results between methods may be considered inconsistent with expectations based on visual inspection, and with one another.

To sum up, findings from different methods are inconsistent with one another, which is in part to be expected as different methods were developed for different purposes, and may measure different subtypes of entrainment. Importantly, even the methods used in this article that measure the same type of entrainment according to the framework developed by Wynn and Borrie (2022) sometimes have diverging results (see Table 6). While these findings could be interpreted as questioning reproducibility in prosodic entrainment research, similarly to how different methods for measuring interpersonal synchrony in movement have been shown to lack convergent findings (Schoenherr, Paulick, Strauss, et al., 2019), alternative interpretations of these findings are also possible. For example, results may suggest that there are no strong, systematic relationships between different subtypes of entrainment, and the variance in results raises questions about the interpretation of entrainment measurements. Possible sources of this variation in results are discussed in the next section.

Possible Sources of Variation

There are multiple potential explanations for the variation in the results of the different methods. One possible methodological source of variation was addressed in the follow-up study presented in Follow-Up Norming Study section, namely, different norming procedures. Results suggest that differences in norming account for some, but not all, of the variation in results.

Results of the follow-up study also highlighted an additional potential source of variation: Different methods rely on different techniques to conclude whether entrainment occurred. For instance, some methods compare real conversations to surrogate data, while others rely on a significance limit. Surrogate data can, in turn, be generated via different methods such as randomly shuffling IPUs (see Levitan & Hirschberg, 2011) or windows of a larger size (see CRQA) within a dyad. In other studies, "real" conversations are compared to surrogate conversations that are created by combining values of speakers who never interacted and analyzing these as one conversation (e.g., WLCC implemented by Truong & Heylen, 2012; global proximity by Levitan & Hirschberg, 2011). In the WLCC, surrogate data were created by randomly selecting values from speakers who were not part of the target dyad but from similar points in time so the temporal structure resembled that of a real conversation. Such methodological differences may (partially) account for the lack of similarity in results of different methods. Future research may focus on the different methods of generating surrogate data and how this may impact results, though it must be kept in mind that methods for generating surrogate data are heavily dependent on feature extraction methods (e.g., one cannot randomly shuffle windows if features were extracted from IPUs rather than windows).

An additional potential source of variance is the "resolution" at which a method measures entrainment: For example, in conversation F37F38 (see Figure 3 and Table 6), two different methods were used to measure static, local synchrony, and both methods found opposing results (entrainment vs. disentrainment). Importantly, one of these methods (i.e., Levitan and Hirschberg's local synchrony; Levitan & Hirschberg, 2011) measured entrainment in each conversation, whereas the other methods (i.e., using LMEMs as was done in Schweitzer & Lewandowski, 2013) measured entrainment in all conversations at once. It is possible that other conversations in the corpus skewed the results of the LMEM method and that the final result thus more closely reflects those conversations than it reflects F37F38. This difference is one explanation for why two methods that both measure local, static synchrony seemingly produce different outcomes. The difference in "resolution" of these methods may mean that results of such methods should not be directly compared.

As was mentioned before, setting the parameters, such as window size, step size, maximum lag, lag step size, and radius, for the time series–based methods is crucial as these parameters can influence results. In this study, we used parameters that have been previously used in entrainment research (e.g., window size of 20 s as was done by De Looze et al., 2014) or used preestablished procedures for the setting of parameters (see Method section on CRQA). Nonetheless, these parameters may be a source of variation in our results. We used the same window and step sizes for TAMA, HYBRID, and WLCC, except that the windows in HYBRID were extended to encompass the end of an utterance that would be cut off in TAMA and WLCC. Any differences between HYBRID and the other two methods are likely due in part to the different window sizes these methods employ.

While this difference in window size was purposefully introduced in the HYBRID method to ensure that utterances were not cut off midway, it may be a source of variation in the results, despite the fact that HYBRID and TAMA otherwise measure the same subtype of entrainment, namely, static, global synchrony. While one may thus expect slightly different results from these two methods, as they were designed differently, this begs questions such as whether both methods truly measure the same type of entrainment, whether one method is perhaps more sensitive or accurate than the other method, and how any diverging findings between these two methods should be interpreted.

Similarly, we used different maximum lag variables for WLCC and CRQA. While these methods work quite differently, it is nonetheless possible that the difference in maximum lag accounts for part of the variation in results. Importantly, these two methods were the only ones that explicitly investigated entrainment with a time lag, and results from the two methods are quite different. While CRQA found significant entrainment in every conversation, perhaps suggesting that also looking for entrainment with a time lag may make a method more sensitive, WLCC did not produce such uniform results, though significant entrainment or disentrainment was found in most conversations using this method. Again, WLCC and CRQA work quite differently and measure different subtypes of entrainment, but the finding that both methods produced a large number of significant results could suggest that including a time lag makes a method more sensitive to detecting entrainment. The effects of different time lag parameters on the observation of significant (dis)entrainment can be investigated in further research.

Future research could investigate how different parameters' values influence how much (dis)entrainment is measured. This could shed light on how different parameters can affect the outcome of studies and can inform the parameter setting process in future studies to further our understanding of entrainment.

Besides methodological explanations, there may also be more theoretical accounts of the variation found in the results. In this study, we used the framework described by Wynn and Borrie (2022) to classify our methods in different groups according to three variables: entrainment timescale (i.e., local, global, and a separate time series–based group), entrainment dimension (i.e., proximity vs. synchrony), and entrainment dynamicity (i.e., static vs. dynamic). However, this framework does not capture all differences between the methods. Other frameworks, such as the one developed by Rasenberg et al. (2020) in the context of multimodal entrainment research, may highlight additional differences between methods. For example, according to the Wynn and Borrie (2022) framework, CRQA and Levitan and Hirschberg's (2011) global proximity both measure static, global proximity (see Table 1). Applying the Rasenberg et al. (2020) framework to these two methods highlights a key difference between the two: While in Levitan and Hirschberg's method, utterances are grouped by time or sequence (e.g., halves of conversations) and the similarity in form between these two groups is measured, in CRQA, utterances are grouped by their form, and the amount of time that two speakers are similar is measured. In this way, the Rasenberg et al. (2020) framework can highlight conceptual differences between methods, which may explain part of the variance in results that we observed in our study.

Additionally, it can be argued that different distinctions made between methods reflect differences in the theoretical understanding of entrainment. For example, CRQA measures the elapsed time between two instances of shared behavior, suggesting the underlying assumptions that entrainment will occur, that disentrainment is not as interesting to measure as entrainment, and that the elapsed time between two instances of shared behavior is more informative than the degree of similarity between the two instances.

Practical Considerations for Future Research

Besides giving rise to methodological and theoretical questions, this study also highlights some practical considerations for future research. The vast range of methods for measuring entrainment may pose challenges for researchers: Selecting the most appropriate method for a given study may seem difficult. Several methodological considerations should be made before the selection of a method. For example, as mentioned before, one may want to pay attention to the "resolution" of the method. Some methods, such as Schweitzer and Lewandowski's (2013) LMEM approach or Levitan and Hirschberg's (2011) global measures reveal whether entrainment occurred in a group of conversations. Such methods may be most useful when one simply wants to assess whether entrainment occurs in a group of conversations, for example, after attempting to implement entrainment in a spoken dialogue system. Other methods, such as Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020) geometric approach and Levitan and Hirschberg's (2011) local methods indicate whether entrainment occurred in each analyzed dialogue. This type of method may be particularly useful if one aims to compare entrainment in different groups (e.g., clinical vs. control groups) or different conditions (e.g., face-to-face vs. separated by a curtain or screen).

Finally, in some methods, it is easier to include other variables or covariates than in others: For example, the output from the Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020) method can be entered as a dependent variable in statistical tests, or additional fixed or random effects can be added to the mixed-effects model in methods based on the study of Schweitzer and Lewandowski (2013), whereas Levitan and Hirschberg's (2011) methods already involve specific statistical tests and the inclusion of a covariate may be more complicated. Especially considering the high complexity of entrainment as a phenomenon, the possibility of including additional variables or covariates may be favorable as it allows one to investigate and control for multiple factors that influence entrainment. This may prove crucial to further our understanding of entrainment.

Importantly, researchers may also want to consider whether and how a method considers entrainment "significant": Some methods, such as Lehnert-LeHouillier, Terrazas, and Sandoval's (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren's (2020) geometric approach, do not include a measure of significance, though most methods rely on statistical tests to compare adjacent to nonadjacent utterances, first conversation halves to second conversation halves, or real conversations to surrogate data. Regardless of statistical test, typically the standard *p*-value cutoff of < .05 is used, but as in many other fields, there are discussions about whether arbitrary *p* values should be relied upon as heavily as they have been in past research. Researchers may want to include additional analyses to assess whether findings are significant and whether this significance is meaningful. For example, investigating whether any significant findings of entrainment are large enough to actually be perceivable in conversation may prove useful in studies that focus on the social role of entrainment.

An additional characteristic of methods that is not captured by any of the described frameworks, but that one may want to take into account, is the notion of "directionality." Some methods, such as the geometric approach by Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020), provide information about who is entraining to whom. Other methods, such as many of the time series–based approaches including WLCC, provide insights into any lag/lead relationships in entrainment, that is, whether one person follows the patterns produced by the other speaker. Insight into who is entraining to whom may be especially relevant for researchers who study entrainment in clinical populations (e.g., Borrie et al., 2015, 2019; Lehnert-LeHouillier, Terrazas, & Sandoval, 2020; Lehnert-LeHouillier, Terrazas, Sandoval, & Boren, 2020) or in human–machine interaction and from a theoretical perspective may be informative regarding predictions based on Giles et al.'s (1991) communication accommodation theory.

Similarly, researchers may want to ensure that they are using a method that captures dynamic entrainment if they are interested in the social functions of entrainment, as it has been hypothesized that fluctuations in entrainment may reflect changes in social behavior, intentions, or the speakers' mutual involvement (De Looze et al., 2014). Dynamic methods may thus be more suitable for research that is focused on the understanding or modeling of the relationship between entrainment and interpersonal aspects such as speaker involvement.

Additionally, researchers should decide whether they want to focus on entrainment or also want to be able to capture disentrainment: Though research has suggested that entrainment is associated with various positive social measures such as more effective communication, building of rapport, and feelings of closeness (Borrie et al., 2015; Chartrand & Bargh, 1999; Levitan et al., 2012), entrainment is not always beneficial. For example, too much entrainment could be perceived as imitation or mockery (Giles, 1979). Furthermore, when one person raises their voice during an argument, it is unlikely that entrainment on intensity, that is, the other person also starts shouting, would be socially beneficial. Indeed, research has shown that disentrainment may also positively influence the development of a conversation (Pérez et al., 2016). One may thus want to use a method that can not only measure entrainment but also disentrainment, such as those by Levitan and Hirschberg (2011), Lehnert-LeHouillier, Terrazas, and Sandoval (2020), and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020), and TAMA, HYBRID, or WLCC, rather than measures that can only measure entrainment, such as CRQA.

It is worth noting that several of the methods that we employed in this study were developed over a decade ago (e.g., Boker et al., 2002; Kousidis et al., 2008; Levitan & Hirschberg, 2011), when the use of advanced statistical procedures was not as common in speech science as it is now. It is possible that more frequent and widespread use of more advanced statistical modeling in the field of speech sciences over the next few years will advance our understanding of entrainment. An increased focus on tools such as nonlinear statistical models and improved acoustic modeling may help us paint a clearer picture of entrainment.

In the meantime, the results presented in this article have substantial implications for future entrainment research. Results may suggest that the different methods each measure different subtypes of entrainment, or quantify entrainment on different dimensions, and that these different subtypes of entrainment show no particularly strong relationship to each other. This study thus highlights the importance of specifying which subtype or dimension of entrainment is being measured and emphasizes the importance of frameworks such as those introduced by Wynn and Borrie (2022) and Rasenberg et al. (2020).

However, one must also wonder how useful it is to develop even more detailed frameworks: If each method is viewed as measuring a different dimension or subtype of entrainment, can those subtypes still be considered subtypes? In other words, if entrainment on different features, levels, and dimensions is unrelated (Ostrand & Chodroff, 2021; Weise & Levitan, 2018) and different methods for measuring entrainment capture slightly different types of entrainment, should we continue viewing and researching entrainment as one phenomenon? The results presented in this study support the notion that entrainment is not one behavior, but rather a set of behaviors, that may not be as strongly associated with one another as was once thought. This raises additional questions. For example, what distinctions should be made, and on what basis?

Future research may want to further focus on investigation whether each method truly captures a different subtype of entrainment, for example, by using simulated data in which each subtype of entrainment can be clearly mimicked. Different sham conversations could be simulated, in which different (combinations of) entrainment subtype(s) are present. These simulated conversations could be analyzed with a range of methods to test how each of these methods assessed the intended entrainment subtype. Such a study could also elucidate whether other practical or methodological differences, such as the generation of surrogate data, explain more of the variance between results.

"Entrainment" is highly complex and seems to be influenced by many different factors, including social (Levitan et al., 2012; Reichel et al. 2018), individual (Menshikova et al., 2020; Weise et al., 2019), and methodological (Wynn & Borrie, 2020) ones, among others, and understanding the way these different factors influence different types of entrainment and interact with other factors is a monumental task. Using frameworks to categorize the set of behaviors referred to as "entrainment" may elucidate whether different subtypes or dimensions of entrainment are influenced by similar factors, governed by shared mechanisms, or serve similar social or communicative functions, and may thus further our understanding of this set of complex interpersonal phenomena.

Conclusions

This study aimed to further our understanding of prosodic entrainment by comparing the results of 12 different methods for measuring entrainment in the same corpus. The main finding of this study is that there is little correlation between the results of methods that measure different subtypes of entrainment, but also occasionally between results of methods that measure the same subtype of entrainment. The article investigated, outlined, and discussed several potential sources of the observed variability of the results and discussed the implications of this variability for the future of entrainment research on both a practical level, related to experimental design and planning of analyses, as well as a theoretical level: After all, the findings raise several important questions, such as whether entrainment should be viewed and researched as one phenomenon, to what extent specifying different subtypes of entrainment is helpful, and how findings of existing studies should be compared. Answers to these questions can ultimately facilitate better understanding of this complex behavior.

Data Availability Statement

The corpus we used can be accessed on Valerie Hazan's website.[3] The subset of files that we used for this study, along with the updated TextGrids, can be found on this OSF respository.[4] All Praat and R scripts used for the

[3] https://valeriehazan.com/wp/index.php/lucid-corpus-london-ucl-clear-speech-in-interaction/.

[4] https://osf.io/yd3cg/files/.

analyses presented in this article have also been uploaded to the same OSF repository[3].

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant Agreement No. 859588 to all authors and the Slovak Granting Agency Grant VEGA 2/0165/21 and the Slovak Research and Development Agency Grant APVV-21-0373 to Štefan Beňuš and Joanna Kruyt. The authors would like to extend thanks to Johannah O'Mahony and Carol Figuera for their detailed and thoughtful comments on an early draft of this article.

References

Abarbanel, H. D. I. (1996). *Analysis of observed chaotic data*. Springer. https://doi.org/10.1007/978-1-4612-0763-4

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America, 116*(5), 3099–3107. https://doi.org/10.1121/1.1795335

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods, 43*(3), 761–770. https://doi.org/10.3758/s13428-011-0075-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. ArXiv. https://doi.org/10.48550/arXiv.1406.5823

Behrens, F., Moulder, R. G., Boker, S. M., & Kret, M. E. (2020). Quantifying physiological synchrony through windowed cross-correlation analysis: Statistical and theoretical considerations. *BioRxiv*.

Bernieri, F., & Rosenthal, R. (1991). Interpersonal coordination: Behavioral matching and interactional synchrony. In R. S. Feldman & B. Rime (Eds.), *Foundations of nonverbal behavior* (pp. 401–432). Cambridge University Press.

Boersma, P. (2006). *Praat: Doing phonetics by computer*. http://www.praat.org/

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods, 7*(3), 338–355. https://doi.org/10.1037/1082-989X.7.3.338

Borrie, S. A., Barrett, T. S., Willi, M. M., & Berisha, V. (2019). Syncing up for a good conversation: A clinically meaningful methodology for capturing conversational entrainment in the speech domain. *Journal of Speech, Language, and Hearing Research, 62*(2), 283–296. https://doi.org/10.1044/2018_JSLHR-S-18-0210

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,* Article 1187. 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

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.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: An R package. *Frontiers in Psychology, 5,* Article 510. https://doi.org/10.3389/fpsyg.2014.00510

De Looze, C., & Rauzy, S. (2009). Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration. *Proceedings of Interspeech 2009*, 2919–2922. https://doi.org10.21437/Interspeech.2009-739

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. https://doi.org/10.1016/j.specom.2013.10.002

Dean, R. T., & Dunsmuir, W. T. M. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods, 48*(2), 783–802. https://doi.org/10.3758/s13428-015-0611-2

Duran, N. D., & Fusaroli, R. (2017). Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement. *PLOS ONE, 12*(6), Article e0178140. https://doi.org/10.1371/journal.pone.0178140

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology, 32,* 147–157. https://doi.org/10.1016/j.newideapsych.2013.03.005

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science, 40*(1), 145–171. https://doi.org/10.1111/cogs.12251

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. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. N. St. Clair (Eds.), *Language and social psychology* (pp. 45–65). Blackwell.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Context of Accommodation: Developments in Applied Sociolinguistics,* 1. https://doi.org/10.1017/CBO9780511663673

Iqbal, T., & Riek, L. D. (2015). A method for automatic detection of psychomotor entrainment. *IEEE Transactions on Affective Computing, 7*(1), 3–16.

Kennel, M. B., Brown, R., & Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A, 45*(6), 3403–3411. https://doi.org/10.1103/PhysRevA.45.3403

Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., McDonnell, C., & Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. In *Ninth Annual Conference of the International Speech Communication Association* (pp. 1692–1695).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lehnert-LeHouillier, H., Terrazas, S., & Sandoval, S. (2020). Prosodic entrainment in conversations of verbal children and teens on the autism spectrum. *Frontiers in Psychology, 11,* Article 582221. https://doi.org/10.3389/fpsyg.2020.582221

Lehnert-LeHouillier, H., Terrazas, S., Sandoval, S., & Boren, R. (2020). The relationship between prosodic ability and conversational prosodic entrainment. *Speech Prosody, 2020,* 769–773.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic–prosodic entrainment and social behavior. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 11–19.

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic–prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of Interspeech 2011*, 3081–3084. https://doi.org/10.21437/Interspeech.2011-771

Menshikova, A., Kocharov, D., & Kachkovskaia, T. (2020). Phonetic entrainment in cooperative dialogues: A case of Russian. *Proceedings of Interspeech 2020*, 4148–4152. https://doi.org/10.21437/Interspeech.2020-2696

Mizukami, M., Yoshino, K., Neubig, G., Traum, D. R., & Nakamura, S. (2016). Analyzing the effect of entrainment on dialogue acts. *Proceedings of the SIGDIAL 2016 Conference*, 310–318.

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

Ostrand, R., & Chodroff, E. (2021). It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. *Journal of Phonetics, 88,* Article 101074. https://doi.org/10.1016/j.wocn.2021.101074

Pérez, J. M., Gálvez, R. H., & Gravano, A. (2016). Disentrainment may be a positive thing: A novel measure of unsigned acoustic–prosodic synchrony, and its relation to speaker engagement. *Proceedings of Interspeech 2016*, 1270–1274. https://doi.org/10.21437/Interspeech.2016-587

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(02), 169–190. https://doi.org/10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Priva, U. C., & Sanker, C. (2018). Distinct behaviors in convergence across measures. *Proceedings of the Annual Conference of the Cognitive Science Society,* 1518–1523.

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

Ramseyer, F., & Tschacher, W. (2010). Nonverbal synchrony or random coincidence? How to tell the difference. In Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (Eds.), *Development of multimodal interfaces: Active listening and synchrony* (pp. 182–196). Springer. https://doi.org/10.1007/978-3-642-12397-9_15

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science, 44*(11), Article e12911. https://doi.org/10.1111/cogs.12911

Reichel, U. D., Beňuš, Š., & Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication, 100,* 46–57. https://doi.org/10.1016/j.specom.2018.04.009

Sanker, C. (2015). Comparison of phonetic convergence in multiple measures. *Cornell Working Papers in Phonetics and Phonology,* 60–75.

Schoenherr, D., Paulick, J., Strauss, B. M., Deisenhofer, A. K., Schwartz, B., Rubel, J. A., Lutz, W., Stangier, U., & Altmann, U. (2019). Identification of movement synchrony: Validation of windowed cross-lagged correlation and -regression with peak-picking algorithm. *PLOS ONE, 14*(2), Article e0211494. https://doi.org/10.1371/journal.pone.0211494

Schoenherr, D., Paulick, J., Worrack, S., Strauss, B. M., Rubel, J. A., Schwartz, B., Deisenhofer A.K., Lutz W., Stangier U., & Altmann U. (2019). Quantification of nonverbal synchrony using linear time series analysis methods: Lack of convergent validity and evidence for facets of synchrony. *Behavior Research Methods, 51*(1), 361–383. https://doi.org/10.3758/s13428-018-1139-z

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. *Proceedings of Interspeech 2013*, 525–529. https://doi.org/10.21437/Interspeech.2013-148

Truong, K. P., & Heylen, D. (2012). Measuring prosodic alignment in cooperative task-based conversations. *Proceedings of Interspeech 2012*, 843–846. https://doi.org/10.21437/Interspeech.2012-190

Wallot, S., & Leonardi, G. (2018). Analyzing multivariate dynamics using cross-recurrence quantification analysis (CRQA), diagonal-cross-recurrence profiles (DCRP), and multidimensional recurrence quantification analysis (MDRQA)—A tutorial in R. *Frontiers in Psychology, 9,* Article 2232. https://doi.org/10.3389/fpsyg.2018.02232

Washburn, A., DeMarco, M., de Vries, S., Ariyabuddhiphongs, K., Schmidt, R. C., Richardson, M. J., & Riley, M. A. (2014). Dancers entrain more effectively than non-dancers to another actor's movements. *Frontiers in Human Neuroscience, 8,* Article 800. https://doi.org/10.3389/fnhum.2014.00800

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

Webber, C. L., Jr., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 142–177). NSF.

Weise, A., & Levitan, R. (2018, June). 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*, 297–302. http://dx.doi.org/10.18653/v1/N18-2048

Weise, A., Levitan, S. I., Hirschberg, J., & Levitan, R. (2019). Individual differences in acoustic–prosodic entrainment in spoken dialogue. *Speech Communication, 115,* 78–87. https://doi.org/10.1016/j.specom.2019.10.007

Wynn, C. J., & Borrie, S. A. (2020). Methodology matters: The impact of research design on conversational entrainment outcomes. *Journal of Speech, Language, and Hearing Research, 63*(5), 1352–1360. https://doi.org/10.1044/2020_JSLHR-19-00243

Wynn, C. J., & Borrie, S. A. (2022). Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics, 94,* Article 101173. https://doi.org/10.1016/j.wocn.2022.101173

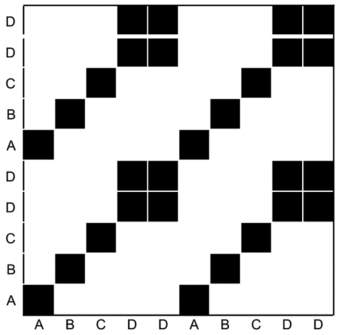
Zbilut, J., Giuliani, A., & Webber, C. (1998). Recurrence quantification analysis and principal components in the detection of short complex signals. *Physics Letters A 237*(3). 131–135. https://doi.org/10.1016/S0375-9601(97)00843-8

Appendix A (p. 1 of 3)

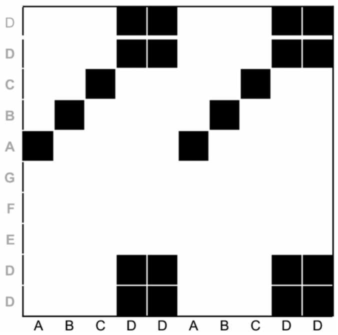
Cross-Recurrence Quantification Analysis Intro

The roots of cross-recurrence quantification analysis (CRQA) lie in the field of dynamical systems. CRQA has been applied to measure various types of behavioral entrainment, including dance (e.g., Washburn et al., 2014) and other types of movement (e.g., Iqbal & Riek, 2015). A crucial element of CRQA is the cross-recurrence plot, from which a wealth of information can be extracted. For an example of a recurrence plot (recurrence in one signal) and a cross-recurrence plot (recurrence between two signals), see Figures A and B, respectively (taken from Wallot & Leonardi, 2018). Figure A shows the recurrence plot of one time series with categorical data. Figure B shows the cross-recurrence plot of two time series that each contains categorical data. Various measures that can be extracted from such cross-recurrence plots are explained in the table below. In our analysis, we focused on recurrence rate (RR).

A



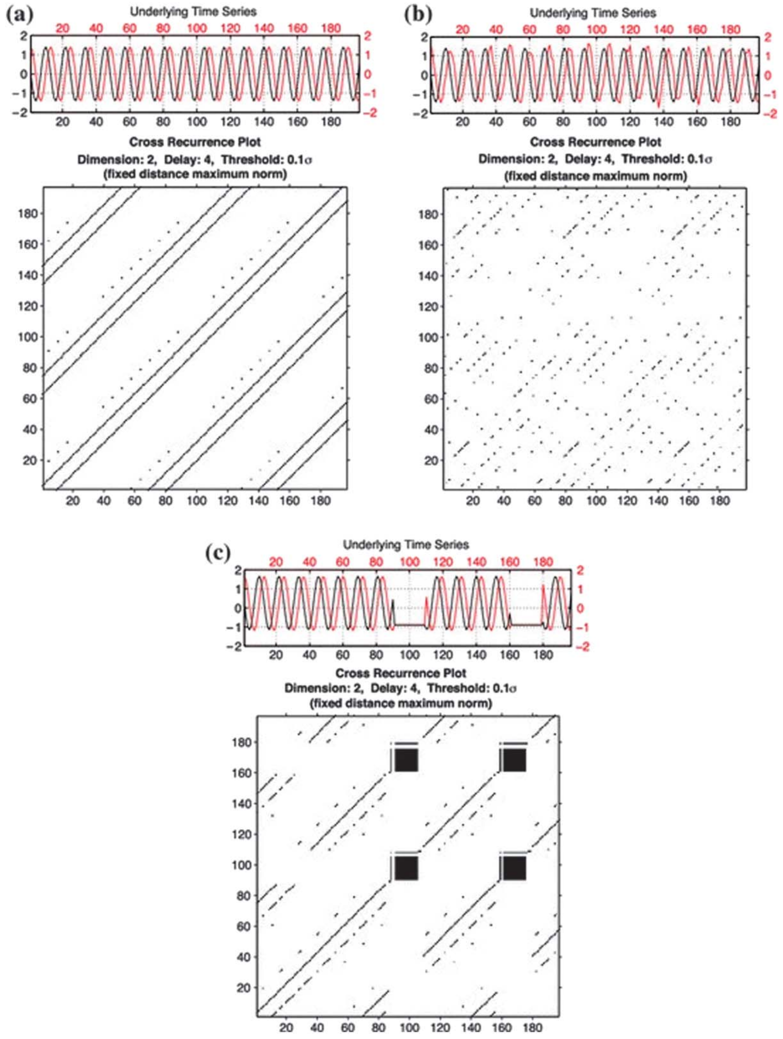
B



Term	Description
Recurrence rate (RR)	The number of points in which both systems are in a similar state, divided by the total number of points. Sometimes denoted as %REC, or percentage recurrence. This is a measure of how similar two systems are, or how often they visit similar states.
Length (L)	The average length of recurring trajectories, i.e., the average amount of time that both systems are in a similar state. Sometimes denoted as ADL or average diagonal length.
Maximum length (LMAX)	Researchers can also choose to extract the maximum length, or the longest uninterrupted recurring trajectories. Also denoted as MDL or maximum diagonal length. This is a measure of how stable the coordination between two systems is: If the alignment between two systems is unstable or sensitive to noise, the LMAX will be lower than if the alignment is more stable.
Entropy (ENTR)	The amount of variability in the length of recurring trajectories. This is a measure of the complexity of the alignment between two systems: If entropy is high, there is a lot of variability in the length of the recurring trajectories and the alignment is thus not very regular.

For more examples of cross-recurrence plots, see the plot below taken from Fusaroli et al. (2014). In the plots, both diagonal and vertical structures can be observed. Diagonal structures provide information about the recurring trajectories or the way in which two systems share similar structures over time. Examples of such measures are explained in the table above. For example, if one compares Figures 1a and 1b from Fusaroli et al. (2014), Figure 1a shows more and longer diagonal lines, reflecting more recurring trajectories than Figure 1b (which was made from the same time series as Figure 1a, but with the inclusion of Gaussian noise, which resulted in a less strong coupling between the two systems).

Cross-recurrence plots also allow for the extraction of measures related to vertical structures, which can provide information regarding the amount of time two systems spend in similar states. For example, in Figure 1c by Fusaroli et al. (2014, see above), the short vertical lines in the image result from the flat lines in the time-series plot, or a time during which both systems were in similar states. Measures derived from vertical structures are not used often in prosodic entrainment research.



Note that in the plots by Fusaroli et al. (2014), some parameters are listed, such as dimension, delay, and threshold. These reflect important parameters that have to be set to conduct CRQA. The "**delay**" (*d*) refers to the maximum lag at which two systems will be compared. Since CRQA measures the amount of time and the number of times that both time series visit similar states, the threshold for when two systems are considered "similar" has to be set. This parameter is referred to as "threshold" in the Fusaroli et al. (2014) plots but is typically called the "**radius**" (*r*). Selecting the radius is relatively easy for categorical variables such as pauses or lexical units, but far more complex for continuous variables such as acoustic–prosodic features.

Various procedures exist for estimating the different parameters. For example, Borrie et al. (2019) used a clinically informed approach to determine the radius: Speech and language pathologists were asked to rate how "in sync" two interlocutors were. A *k*-nearest neighbor method was used to determine which delay and radius best predicted the clinicians' assessments, and these parameters were used in CRQA to quantify entrainment. Other researchers use different methods of setting parameters: A commonly used procedure is to set all parameters such that the recurrence rate is within a set range of percentages in every conversation, though this does not always facilitate the comparison between real and surrogate conversation, which is often used to assess significance of entrainment.

Another variable that has to be set during CRQA is the "embedding dimension" (*m*, listed as dimension in the Fusaroli et al., 2014, plots), which in our analyses was set to 1. This parameter is related to the number of latent variables that may govern complex systems. More information on embedding dimensions and how to set them can be found in the study of Wallot and Leonardi (2018), who provide a detailed overview of all the required steps for setting parameters and conducting CRQA.

Further Reading

The vast majority of the information presented in this Appendix was taken from the following papers, each of which provides an overview of CRQA and its applications for analyzing social interaction in varying amounts of detail. For further information on the method and its potential for analyzing entrainment, the reader is recommended to read the following articles:

Borrie, S. A., Barrett, T. S., Willi, M. M., & Berisha, V. (2019). Syncing up for a good conversation: A clinically meaningful methodology for capturing conversational entrainment in the speech domain. *Journal of Speech, Language, and Hearing Research*, *62*(2), 283–296.

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis. In *Translational recurrences: From mathematical theory to real-world applications* (pp. 137–155). Springer.

Wallot, S., & Leonardi, G. (2018). Analyzing multivariate dynamics using cross-recurrence quantification analysis (CRQA), diagonal-cross-recurrence profiles (DCRP), and multidimensional recurrence quantification analysis (MDRQA)–A tutorial in R. *Frontiers in Psychology*, *9*, 2232.

Appendix B

Levitan and Hirschberg Local Proximity Results

Table B1. Results for the paired *t* tests for local proximity (following Levitan & Hirschberg, 2011). The “entr.” column indicates whether significant entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

	f_o						Range						Max					
Dyad	*t*	*df*	*p*	**entr.**	**Partner diff**	**Other diff**	*t*	*df*	*p*	**entr.**	**Partner diff**	**Other diff**	*t*	*df*	*p*	**entr.**	**Partner diff**	**Other diff**
F03F04	–1.99	182	.049	+	1.52	1.70	–0.43	182	.669		3.68	3.75	–0.92	182	.356		2.98	3.12
F11F12	–2.06	299	.040	+	1.82	1.99	–1.93	299	.055		2.49	2.69	–3.94	299	< .001	+	1.90	2.23
F13F14	–1.27	186	.207		1.82	2.00	–0.84	186	.403		3.39	3.53	–0.24	186	.808		3.06	3.10
F15F16	–0.18	165	.859		1.76	1.80	0.93	165	.354		2.65	2.51	0.31	165	.758		2.03	2.00
F21F22	–2.52	207	.012	+	1.69	1.99	–0.14	207	.889		3.29	3.32	–1.56	207	.121		2.49	2.73
F25F26	–2.97	148	.004	+	1.00	1.19	0.74	148	.457		2.62	2.52	–1.03	148	.306		1.75	1.87
F31F32	–0.50	252	.614		1.60	1.64	1.97	252	.050		2.75	2.52	–0.51	252	.611		2.01	2.05
F37F38	–0.22	161	.828		2.74	2.77	0.07	161	.944		3.35	3.35	0.06	161	.951		3.02	3.02
F41F42	–0.13	144	.894		1.36	1.38	1.57	144	.118		2.24	2.04	–0.34	144	.735		1.57	1.60
F47F48	–0.30	140	.767		1.83	1.88	–0.74	140	.461		2.60	2.72	–0.72	140	.470		2.05	2.13
M07M08	0.05	188	.964		1.97	1.97	0.55	188	.586		2.07	2.02	–0.62	188	.538		1.97	2.01
M09M10	–0.77	317	.444		0.92	0.95	–0.80	317	.425		1.82	1.88	–0.76	317	.450		1.16	1.19
M11M12	–0.64	265	.522		1.08	1.12	0.05	265	.961		2.12	2.12	–0.37	265	.708		1.34	1.36
M13M14	–0.48	248	.635		1.40	1.44	–1.56	248	.119		2.31	2.49	–2.80	248	.006	+	1.43	1.64
M15M16	–0.85	200	.396		0.99	1.03	–0.02	200	.982		1.79	1.79	–0.55	200	.583		1.13	1.17
M17M18	–0.21	298	.834		2.28	2.30	0.51	298	.613		3.37	3.31	0.60	298	.552		2.36	2.30
M25M26	–1.41	208	.161		1.08	1.16	0.34	208	.731		2.64	2.58	–0.18	208	.859		1.78	1.79
M33M34	–0.43	154	.665		0.95	0.97	–0.06	154	.949		2.20	2.21	–0.04	154	.972		1.39	1.39
M35M36	–2.39	250	.017	+	1.00	1.13	–1.18	250	.238		2.05	2.15	–1.36	250	.175		1.31	1.39
M41M42	–0.54	124	.588		0.90	0.93	–0.78	124	.434		1.49	1.58	–1.49	124	.139		0.86	0.97

Appendix C

Levitan and Hirschberg Local Synchrony Results

Table C1. Results of the Pearson correlation for local synchrony following Levitan and Hirschberg (2011). The “sig./10” column indicates how many of the correlations performed with randomly shuffled data returned significant results. Note that for a result to be considered significant and valid, p < .05 and < 1 of the random correlations must have been significant. The “entr.” column indicates whether significant entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

Dyad	f_o				Range				Max			
	r	***p***	**sig./10**	**entr.**	***r***	***p***	**sig./10**	**sig.**	***r***	***p***	**sig./10**	**entr.**
F03F04	0.16	.027	0		–0.04	.631	0		–0.05	.510	1	
F11F12	0.19	.001	0	+	0.13	.022	1	+	0.28	< .001	0	+
F13F14	0.03	.709	1		0.03	.659	1		0.10	.166	2	
F15F16	–0.12	.137	0		–0.11	.177	0		–0.05	.484	1	
F21F22	0.27	< .001	0	+	–0.08	.245	0		0.04	.579	1	
F25F26	0.27	.001	0	+	–0.09	.278	1		0.07	.403	1	
F31F32	–0.01	.861	0		–0.02	.797	1	–	0.12	.056	0	
F37F38	–0.28	< .001	0	–	0.03	.668	2		–0.15	.065	2	
F41F42	–0.06	.439	0		–0.12	.140	0		–0.11	.186	1	
F47F48	–0.05	.568	2		0.03	.746	0		0.07	.381	0	
M07M08	–0.50	< .001	0	–	–0.04	.558	0		–0.34	< .001	0	–
M09M10	0.00	.931	0		0.02	.745	0		0.08	.163	0	
M11M12	0.01	.847	1		0.01	.830	0		0.04	.564	0	
M13M14	–0.05	.423	0		0.12	.053	0		0.16	.014	0	+
M15M16	–0.01	.912	0		–0.07	.316	2		0.01	.914	0	
M17M18	–0.34	< .001	1	–	–0.03	.564	0		–0.15	.012	0	–
M25M26	0.03	.655	1		0.01	.857	0		0.05	.501	0	
M33M34	–0.03	.747	0		–0.10	.200	0		–0.12	.132	0	
M35M36	0.09	.137	1		0.03	.602	0		0.05	.393	1	
M41M42	0.07	.433	3		0.13	.158	0		0.11	.219	1	

Appendix D

Levitan and Hirschberg Local Convergence Results

Table D1. Results of the Pearson correlation for local convergence following Levitan and Hirschberg (2011). The “sig./10” column indicates how many of the correlations performed with randomly shuffled data returned significant results. Note that for a result to be considered significant and valid, $p < .05$ and no more than 1 of the random correlations must have returned a significant result. The “entr.” column indicates whether significant and valid entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

	f_o				Range				Max			
Dyad	***r***	***p***	**sig./10**	**entr.**	***r***	***p***	**sig./10**	**entr.**	***r***	***p***	**sig./10**	**entr.**
F03F04	–0.09	.250	0		–0.02	.761	0		0.01	.846	0	
F11F12	–0.17	.003	0	–	0.00	.964	0		–0.09	.103	0	
F13F14	0.21	.004	0	+	0.00	.976	0		0.04	.555	0	
F15F16	0.11	.147	0		0.00	.973	0		0.10	.194	0	
F21F22	0.17	.012	0	+	–0.02	.772	0		0.00	.966	0	
F25F26	–0.05	.536	1		0.00	.958	2		–0.09	.259	1	
F31F32	0.10	.127	0		–0.03	.677	0		–0.06	.349	0	
F37F38	0.03	.660	0		0.00	.952	0		0.03	.724	0	
F41F42	0.02	.790	0		–0.07	.411	0		0.02	.793	0	
F47F48	0.05	.560	0		0.01	.900	0		0.04	.639	0	
M07M08	0.28	< .001	0	+	0.02	.836	0		0.00	.958	0	
M09M10	0.02	.734	0		0.00	.988	0		0.04	.479	0	
M11M12	0.00	.981	0		0.05	.423	1		0.02	.746	0	
M13M14	–0.03	.674	0		0.21	.001	1	+	0.15	.016	0	+
M15M16	0.13	.067	3		0.00	.991	0		0.06	.406	0	
M17M18	–0.17	.003	0	–	–0.09	.124	3		–0.20	.001	0	–
M25M26	–0.21	.002	3		0.03	.672	0		0.06	.423	0	
M33M34	0.00	.989	0		–0.10	.223	0		–0.07	.407	0	
M35M36	0.15	.017	0	+	0.12	.056	2		0.06	.343	2	
M41M42	0.20	.025	0	+	0.13	.156	0		0.14	.125	0	

Appendix E

Schweitzer and Lewandowski Linear Mixed-Effects Models Results

Table E1. Results from the linear mixed-effects model for median f_o, loosely following Schweitzer and Lewandowski (2013). *p* values were estimated via *t* tests using the Satterthwaite approximations to degrees of freedom (using lmerTest package).

Effect	Estimate	*SE*	*df*	*t*	*p*
Intercept	173.23	4.41	97.74	39.3	< .001
f_o preceding utterance	0.16	0.02	4145.76	9.52	< .001
Gender (male)	−84.42	4.35	23.76	−19.41	< .001

Table E2. Results from the model comparison between the full and null model for median f_o using lmerTest ANOVA, where the null model includes gender as a fixed effect and target speaker and dyad as random effects, whereas the full model also includes the f_o of the preceding utterance as a fixed effect.

Model	*df*	AIC	BIC	logLik	Deviance	χ^2	χ^2 *df*	*p*
Null	5	35878	35910	−17934	35868			
Full	6	35792	35830	−17890	35780	88.64	1	< .001

Table E3. Results from the linear mixed-effects model for f_o range, loosely following Schweitzer and Lewandowski (2013). *p* values were estimated via *t* tests using the Satterthwaite approximations to degrees of freedom (using lmerTest package).

Effect	Estimate	*SE*	*df*	*t*	*p*
Intercept	83.42	3.69	25.44	22.59	< .001
Range preceding utterance	0.02	0.02	4170.71	1.31	< .001
Gender (male)	−46	4.85	18.95	−9.48	< .001

Table E4. Results from the model comparison between the full and null model for f_o range using ANOVA, where the null model includes gender as a fixed effect and target speaker and dyad as random effects, whereas the full model also includes the f_o range of the preceding utterance as a fixed effect.

Model	*df*	AIC	BIC	logLik	Deviance	χ^2	χ^2 *df*	*p*
Null	5	41771	41803	−20880	41761			
Full	6	41770	41808	−20879	41758	2.66	1	.103

Table E5. Results from the linear mixed-effects model for max f_o. loosely following Schweitzer and Lewandowski (2013). *p* values were estimated via *t* tests using the Satterthwaite approximations to degrees of freedom (using lmerTest package).

Effect	Estimate	*SE*	*df*	*t*	*p*
Intercept	228.62	5.88	58.78	38.88	< .001
Max preceding utterance	0.10	0.02	4148.92	6.28	< .001
Gender (male)	−114.82	6.56	21.61	−17.78	< .001

Table E6. Results from the model comparison between the full and null model for max f_o using ANOVA, where the null model includes gender as a fixed effect and target speaker and dyad as random effects, whereas the full model also includes the max f_o of the preceding utterance as a fixed effect.

Model	*df*	AIC	BIC	logLik	Deviance	χ^2	χ^2 *df*	*p*
Null	5	41421	41453	−20706	41411			
Full	6	41384	41422	−20686	41372	39.08	1	< .001

Appendix F

Levitan and Hirschberg Global Proximity and Convergence Results

Table F1. Results of the paired *t* tests for global convergence and global proximity following Levitan and Hirschberg (2011). “First half” and “second half” in the global convergence table refer to the mean difference between speakers in the first and second halves of the conversation. The “entr.” column indicates whether significant entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

Global proximity																	
f_o						**Range**						**Max**					
t	***df***	***p***	**Partner diff**	**Other diff**	**entr.**	***t***	***df***	***p***	**Partner diff**	**Other diff**	**entr.**	***t***	***df***	***p***	**Partner diff**	**Other diff**	**entr.**
–1.87	19	.077	0.89	1.14		–3.07	19	.006	0.76	1.16	+	–2.44	19	.024	0.78	1.15	+
Global convergence																	
f_o						**Range**						**Max**					
t	***df***	***p***	**First half**	**Second half**	**entr.**	***t***	***df***	***p***	**First half**	**Second half**	**entr.**	***t***	***df***	***p***	**First half**	**Second half**	**entr.**
2.45	19	.024	0.21	–0.31	+	–0.388	19	.705	–0.18	–0.08		–0.83	19	.415	0.04	0.18	

Appendix G

Lehnert-LeHouillier et al. Geometric Approach Results

Table G1. Results of the geometric analysis to assess entrainment following Lehnert-LeHouillier, Terrazas, and Sandoval (2020) and Lehnert-LeHouillier, Terrazas, Sandoval, and Boren (2020). The "first" and "last" columns indicate the difference between speakers in the first and last third of the conversation, while "diff." represents the difference between the two. The "entr." column indicates whether entrainment (+) or disentrainment (–) occurred. The columns with "cont. A" and "cont. B" indicate each speaker's relative contribution to the overall observed entrainment (in percentages).

	f_o						Range						Max					
Dyad	**First**	**Last**	**diff.**	**entr.**	**cont. A**	**cont. B**	**First**	**Last**	**diff.**	**entr.**	**cont. A**	**cont. B**	**First**	**Last**	**diff.**	**entr.**	**cont. A**	**cont. B**
F03F04	8.16	4.12	4.04	+	64%	36%	14.62	19.69	–5.07	–	83%	17%	26.35	25.55	0.80	+	42%	58%
F11F12	6.97	2.89	4.09	+	41%	59%	12.80	13.63	–0.83	–	46%	54%	4.33	8.71	–4.38	–	28%	72%
F13F14	6.52	7.68	–1.16	–	69%	31%	24.18	5.56	18.63	+	81%	19%	5.59	6.09	–0.50	–	88%	12%
F15F16	10.62	13.40	–2.77	–	93%	7%	6.18	13.05	–6.87	–	93%	7%	15.93	25.42	–9.50	–	49%	51%
F21F22	6.54	18.48	–11.95	–	13%	87%	21.98	33.55	–11.58	–	92%	8%	21.01	34.03	–13.03	–	83%	17%
F25F26	6.78	1.89	4.89	+	95%	5%	1.30	16.12	–14.82	–	15%	85%	7.73	7.26	0.47	+	15%	85%
F31F32	10.74	12.50	–1.75	–	11%	89%	6.56	6.60	–0.05	–	60%	40%	17.28	5.75	11.52	+	82%	18%
F37F38	24.91	23.73	1.19	+	85%	15%	3.32	14.75	–11.43	–	38%	62%	19.21	31.42	–12.21	–	26%	74%
F41F42	8.48	9.29	–0.81	–	57%	43%	9.68	14.18	–4.49	–	40%	60%	18.33	20.27	–1.94	–	42%	58%
F47F48	10.29	10.30	0.00	–	50%	50%	0.89	3.07	–2.18	–	41%	59%	1.16	5.62	–4.46	–	12%	88%
M07M08	13.41	17.65	–4.24	–	51%	49%	12.15	5.42	6.74	+	55%	45%	24.48	19.32	5.17	+	44%	56%
M09M10	4.66	1.83	2.83	+	17%	83%	0.14	7.77	–7.63	–	15%	85%	4.98	1.99	2.99	+	5%	95%
M11M12	4.44	5.90	–1.47	–	41%	59%	6.94	3.06	3.88	+	76%	24%	2.10	4.27	–2.16	–	73%	27%
M13M14	7.64	12.21	–4.57	–	25%	75%	2.02	5.76	–3.74	–	41%	59%	7.74	9.73	–1.99	–	45%	55%
M15M16	3.52	6.41	–2.88	–	73%	27%	0.60	0.27	0.33	+	62%	38%	5.10	6.35	–1.25	–	5%	95%
M17M18	18.42	12.20	6.22	+	27%	73%	14.70	1.76	12.94	+	10%	90%	24.70	11.25	13.45	+	36%	64%
M25M26	6.24	1.92	4.32	+	67%	33%	1.69	11.67	–9.99	–	21%	79%	7.61	5.71	1.90	+	50%	50%
M33M34	3.06	5.61	–2.55	–	78%	22%	14.15	10.35	3.80	+	31%	69%	12.00	11.12	0.87	+	38%	62%
M35M36	0.85	6.55	–5.70	–	85%	15%	2.50	16.35	–13.85	–	52%	48%	5.94	14.38	–8.45	–	62%	38%
M41M42	2.04	1.42	0.62	+	63%	37%	0.07	5.85	–5.78	–	64%	36%	0.17	4.61	–4.44	–	85%	15%

Appendix H

Kousidis et al. Time-Aligned Moving Average Results

Table H1. Results of the Pearson correlation on TAMA across the entire conversation following Kousidis et al. (2008) and De Looze et al. (2014). The estimate (*t*), *p* value, and estimate (rho) are shown per dyad and feature. The "entr." column indicates whether significant entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

	f_o				Range				Max			
Dyad	***t***	***p***	**rho**	**entr.**	***t***	***p***	**rho**	**entr.**	***t***	***p***	**rho**	**entr.**
F03F04	2.24	.03	0.36	+	1.02	.31	0.18		1.17	.25	0.20	
F11F12	2.03	.05	0.24	+	0.38	.70	0.05		1.04	.30	0.12	
F13F14	1.02	.24	0.18		2.59	.01	0.37	+	2.94	.01	0.42	+
F15F16	2.22	.03	0.35	+	–0.02	.98	0.00		1.03	.31	0.17	
F21F22	2.06	.04	0.29	+	–0.40	.69	–0.06		0.25	.80	0.04	
F25F26	1.74	.09	0.27		–0.64	.53	–0.10		0.04	.97	0.01	
F31F32	0.55	.59	0.07		–0.48	.63	–0.06		0.91	.37	0.11	
F37F38	0.83	.41	0.13		–0.21	.84	–0.03		–0.3-	.76	–0.05	
F41F42	2.12	.04	0.33	+	0.93	.36	0.15		0.97	.34	0.16	
F47F48	2.12	.04	0.37	+	0.63	.54	0.12		1.09	.29	0.20	
M07M08	–1.31	.20	–0.20		–0.54	.59	–0.08		0.22	.82	0.03	
M09M10	2.03	.05	0.28	+	–0.81	.42	–0.11		0.30	.77	0.04	
M11M12	0.33	.74	0.05		1.60	.12	0.22		0.60	.55	0.08	
M13M14	2.05	.04	0.23	+	3.95	< .01	0.42	+	4.44	< .01	0.46	+
M15M16	–0.03	.98	0.00		0.02	.98	0.00		–0.53	.60	–0.08	
M17M18	–0.36	.72	–0.04		–0.96	.34	–0.10		–1.21	.23	–0.13	
M25M26	0.36	.72	0.06		0.59	.56	0.09		0.57	.57	0.09	
M33M34	–0.65	.52	–0.12		–1.17	.25	–0.22		–0.70	.49	–0.13	
M35M36	0.55	.58	0.08		0.85	.40	0.12		0.64	.52	0.09	
M41M42	1.92	.07	0.34		1	.33	0.19		0.72	.48	0.13	

Appendix I

De Looze et al. HYBRID Results

Table I1. Results of the Pearson correlation on HYBRID across the entire conversation following De Looze et al. (2014). The estimate (*t*), *p* value, and estimate (rho) are shown per dyad and feature. The S/A/N column indicates the ratio of the periods in which people showed synchrony (S), asynchrony (A), or no change (N) across the entire conversation (assessed by calculating Pearson correlation on windows of data). In other words, the S/A/N column indicates the number of periods during which a dyad showed significant entrainment (S), significant disentrainment (A), or no significant (dis)entrainment (N). The "entr." column indicates whether significant entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

	f_o					Range					Max				
Dyad	***t***	***p***	**rho**	**entr.**	**S/A/N**	***t***	***p***	**rho**	**entr.**	**S/A/N**	***t***	***p***	**rho**	**entr.**	**S/A/N**
F03F04	1.57	.13	0.26		0/0/6	1.06	.30	0.18		0/0/6	0.93	.36	0.16		0/0/6
F11F12	1.66	.10	0.19		0/0/14	–0.50	.62	–0.06		0/1/13	0.31	.76	0.04		2/0/12
F13F14	0.91	.37	0.14		0/0/8	1.04	.31	0.16		3/0/5	1.52	.14	0.23		2/0/6
F15F16	2.77	.01	0.42	+	1/0/6	0.20	.84	0.03		2/1/4	1.40	.17	0.23		3/0/4
F21F22	1.97	.05	0.28		1/0/8	0.31	.76	0.05		0/0/9	0.68	.50	0.10		0/0/9
F25F26	3.64	< .01	0.50	+	0/0/8	–1.48	.15	–0.23		0/0/8	–0.29	.77	–0.05		0/1/7
F31F32	0.37	.72	0.05		1/1/11	0.34	.74	0.04		2/0/11	1.36	.18	0.16		3/0/10
F37F38	1.05	.30	0.16		1/0/7	0.44	.67	0.07		0/0/8	0.28	.78	0.04		0/0/8
F41F42	2.51	.02	0.38	+	2/0/5	0.95	.35	0.15		0/0/7	0.87	.39	0.14		1/0/6
F47F48	2.95	.01	0.49	+	0/0/5	1.16	.26	0.21		0/1/4	2.03	.05	0.36		0/0/5
M07M08	–1.75	.09	–0.26		0/1/7	–0.97	.34	–0.15		0/0/8	–0.17	.86	–0.03		0/0/8
M09M10	1.43	.16	0.20		0/0/10	–0.95	.34	–0.13		2/0/8	0.76	.45	0.11		2/0/8
M11M12	0.06	.95	0.01		0/0/10	0.84	.41	0.12		0/0/10	0.03	.98	0.00		0/0/10
M13M14	2.00	.05	0.23	+	0/0/15	3.17	< .01	0.35	+	3/1/11	3.88	< .01	0.41	+	2/0/13
M15M16	0.26	.80	0.04		0/0/8	0.08	.93	0.01		0/0/8	–0.18	.86	–0.03		0/0/8
M17M18	–0.37	.71	–0.04		1/2/14	–1.44	.15	–0.16		1/2/14	–1.78	.08	–0.19		0/2/15
M25M26	0.27	.79	0.04		0/0/8	0.78	.44	0.12		0/0/8	0.49	.62	0.08		0/0/8
M33M34	–1.00	.32	–0.19		0/1/4	–0.31	.76	–0.06		1/1/3	–0.04	.97	–0.01		1/1/3
M35M36	–0.25	.80	–0.04		0/0/9	0.30	.77	0.04		2/1/6	–0.16	.87	–0.02		2/0/7
M41M42	1.75	.09	0.31		1/0/4	1.11	.28	0.21		0/0/5	0.74	.46	0.14		0/0/5

Appendix J

Windowed Lagged Cross-Correlation Results

Table J1. Results of the Wilcoxon signed-rank test across the entire conversation for WLCC (Boker et al., 2002) to determine whether the real dyads showed higher cross-correlations than the pseudodyads. The mean correlation coefficient (rho), estimate (*z*), and *p* value are shown per dyad and feature. The “entr.” column indicates whether significant entrainment (+) or disentrainment (–) occurred and is left blank if results were not significant.

Dyad	f_o				Range				Max			
	rho	*z*	*p*	entr.	rho	*z*	*p*	entr.	rho	*z*	*p*	entr.
F03F04	0.10	2.20	.01	+	0.22	3.03	< .01	+	–0.23	2.69	< .01	–
F11F12	–0.29	3.66	< .01	–	–0.03	3.60	< .01	–	0.03	3.52	< .01	+
F13F14	0.12	3.06	< .01	+	0.12	3.19	< .01	+	0.02	3.37	< .01	+
F15F16	–0.02	2.84	< .01	–	–0.03	2.79	< .01	–	0.09	1.00	.16	
F21F22	0.23	2.80	< .01	+	0.00	2.32	.01	+	0.17	2.37	.01	+
F25F26	0.30	2.52	.01	+	–0.08	2.78	< .01	–	–0.08	2.72	< .01	–
F31F32	0.03	3.45	< .01	+	0.05	4.76	< .01	+	0.20	3.23	< .01	+
F37F38	–0.17	4.79	< .01	–	–0.08	3.01	< .01	–	0.17	4.44	< .01	+
F41F42	–0.06	2.00	.02	–	–0.15	1.56	.06		–0.23	2.36	.01	–
F47F48	0.15	2.21	.01	+	0.08	2.64	< .01	+	0.00	1.68	.05	+
M07M08	–0.06	3.11	< .01	–	–0.29	2.06	.02	–	–0.29	1.31	.10	
M09M10	0.04	4.26	< .01	+	–0.09	4.84	< .01	–	0.08	2.89	< .01	+
M11M12	–0.01	3.19	< .01	–	0.09	1.75	.04	+	–0.12	4.64	< .01	–
M13M14	–0.02	3.93	< .01	–	–0.07	3.16	< .01	–	–0.23	1.88	.03	–
M15M16	0.03	2.69	< .01	+	0.08	4.21	< .01	+	0.08	3.58	< .01	+
M17M18	0.06	2.88	< .01	+	0.18	4.27	< .01	+	0.06	3.67	< .01	+
M25M26	0.02	1.65	.05	+	0.23	2.46	.01	+	0.08	2.07	.02	+
M33M34	0.37	0.71	.24		–0.23	2.44	.01	–	–0.15	1.80	.04	–
M35M36	0.16	2.35	.01	+	0.26	4.21	< .01	+	–0.04	4.14	< .01	–
M41M42	0.30	2.90	< .01	+	0.45	2.62	< .01	+	0.01	2.77	< .01	+

Appendix K

Cross-Recurrence Quantification Analysis Results

Table K1. Results of the one-sample *t* test for CRQA (Fusaroli & Tylén, 2016) to test whether the recurrence rates of the real dyads were higher than the recurrence rates of random dyads for f_o. The estimate (*t*), *p* value, and effect size (*d*) are shown per dyad.

Dyad	***t***	***p***	***d***	**entr.**
F03F04	–78.37	< .001	11.08	+
F11F12	–59.71	< .001	8.44	+
F13F14	–44.29	< .001	6.26	+
F15F16	–65.93	< .001	9.32	+
F21F22	–42.81	< .001	6.05	+
F25F26	–49.85	< .001	7.05	+
F31F32	–68.21	< .001	9.65	+
F37F38	–58.71	< .001	8.30	+
F41F42	–47.49	< .001	6.72	+
F47F48	–58.03	< .001	8.21	+
M07M08	–87.65	< .001	12.40	+
M09M10	–72.49	< .001	10.25	+
M11M12	–68.11	< .001	9.63	+
M13M14	–64.68	< .001	9.15	+
M15M16	–72.49	< .001	10.25	+
M17M18	–79.16	< .001	11.19	+
M25M26	–84.25	< .001	11.92	+
M33M34	–69.26	< .001	9.79	+
M35M36	–77.82	< .001	11.01	+
M41M42	–56.95	< .001	8.05	+