Intrapersonal dependencies in multimodal behavior

Blomsma, Pieter A ; Linders, Guido M ; Vaitonyté, Julija ; Louwerse, Max M

DOI: https://doi.org/10.1145/3383652.3423872

Intrapersonal dependencies in multimodal behavior

Pieter A. Blomsma
Dept. of Cognitive Science & Artificial Intelligence,
Tilburg University
Tilburg
p.a.blomsma@uvt.nl

Guido M. Linders
Dept. of Cognitive Science & Artificial Intelligence,
Tilburg University
Tilburg
g.m.linders@uvt.nl

Julija Vaitonyte
Dept. of Cognitive Science & Artificial Intelligence,
Tilburg University
Tilburg
j.vaitonyte@uvt.nl

Max M. Louwerse
Dept. of Cognitive Science & Artificial Intelligence,
Tilburg University
Tilburg
m.m.louwerse@uvt.nl

ABSTRACT

Human interlocutors automatically adapt verbal and non-verbal signals so that different behaviors become synchronized over time. Multimodal communication comes naturally to humans, while this is not the case for Embodied Conversational Agents (ECAs). Knowing which behavioral channels synchronize within and across speakers and how they align seems critical in the development of ECAs. Yet, there exists little data-driven research that provides guidelines for the synchronization of different channels within an interlocutor. This study focuses on intrapersonal dependencies of multimodal behavior by using cross-recurrence analysis on a multimodal communication dataset to better understand the temporal relationships between language and gestural behavior channels. By shedding light on the intrapersonal synchronization of communicative channels in humans, we provide an initial manual for modality synchronisation in ECAs.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Embodied Conversational Agents, Verbal and nonverbal coordination, Spoken Dialog Systems, Conversational Behavior Models, Cross-recurrence Analysis

ACM Reference Format:
Pieter A. Blomsma, Guido M. Linders, Julija Vaitonyte, and Max M. Louwerse. 2020. Intrapersonal dependencies in multimodal behavior. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423872

1 INTRODUCTION

Natural communication is multimodal [14]. It includes tightly interwoven verbal (i.e., speech) and non-verbal (i.e., facial expressions, eye gaze, body posture, and gestures) channels. The use of multiple channels facilitates communication in allowing speakers to express messages more clearly. For instance, the advantage of multimodality is apparent in co-speech gestures, hand and arm movements that spontaneously occur with spoken language [37]. Gestures can help disambiguate information in the spoken modality [21]. One can gesture while saying "The cup was *this* big" or point to a certain glass on the table while uttering "Can you pass me *that* one?". While gestures can convey complementary information to speech, they can also be redundant [3]. Gestures can thus vary in their semantic relationship to speech. Spoken and gestural modalities are not only related on a semantic level, they are also coupled on a temporal level [18, 47], such that the most meaningful part of gesture slightly precedes speech content [19]. Thus, on the production side, speech and gestures are related and on the comprehension side, listeners draw on both linguistic utterances and gestures to understand a speaker's message [23].

Studying the interplay of speech and gestures is important for understanding human-human interaction, but even more so for human-machine interaction. That is, when humans interact with Embodied Conversational Agents (ECAs), these agents display similar verbal and non-verbal competences to humans [9]. However, generating ECAs' non-verbal behavior, including gestures that need to be aligned with the generated speech, is not trivial. In order to do so, several aspects need to be considered. For example, should gestures be generated on the basis of speech, or should speech and gestures be two independent systems? Are there any contingencies and constraints as to which types of gestures occur with speech? If so, how are they aligned across time?

In human interlocutors, temporal dependencies between different multimodal behaviors have been shown to exist in face-to-face settings [33]. In the current study, we ask the question whether similar dependencies between spoken and gestural modalities also exist within a single speaker. Thus, to what extent are verbal and non-verbal behaviors aligned within an individual? The answer to this question is crucial in the development of ECAs. Without knowing which types of behavior go together, the generation of naturally-looking multimodal cues is comparable to looking for a needle in a haystack. First, this is because the proverbial haystack of possible behaviors is the combinatorial explosion of all the degrees of freedom of each modality. Without guidance, the number of possible behaviors is semi-infinite. Another difficulty involves selecting

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk
2020. ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423872

behaviors that come across as natural. Without guidance, such selection is not feasible. Therefore, we suggest that, to limit the number of possible behaviors, it is worthwhile to look at the relations between the spoken and gestural modalities at the intrapersonal level. It is plausible that certain speech-gesture combinations are not possible, meaning they do not occur in human communication. Such information would help diminish the size of the haystack and, in turn, simplify the process of finding the needle. Thus, a better understanding of the mechanisms that govern multimodal behavior within one individual could directly inform the development of ECAs.

In this study, we use a cross-recurrence quantification analysis to investigate speech-gesture dependencies within a speaker during face-to-face communication. The cross-recurrence analysis is applied to a multimodal communication corpus of about 25 hours of dialog that is encoded for speech (thirteen categories of dialog acts) and gestures (five types of gestures) at 250 ms intervals. The main goal of this study is to show that cross-recurrence quantification analysis is useful to unravel potential underlying patterns in intrapersonal multimodal behavior. We focus specifically on the correlations between different dialog acts and gesture behaviors. These results can be used by agent developers to increase the multimodal realism of their dialogue systems.

2 BACKGROUND

2.1 Embodied Conversational Agents

An important part of ECAs' multimodal behavior generation is the temporal coordination of speech and gestures. Gesture plays a prominent role in conveying information in human communication [40] and humans are able to notice whether speech and gesture of ECAs are consistent or not [17]. In general, ECAs use rule-based approaches to produce gestures that are based on the speech that has been uttered [46]. However, it is difficult to construct precise rules about how the different modalities interact in communication as the relationships between different modalities on the production side remain unclear. Furthermore, large richly-annotated multimodal corpora are scarce due to the time-intensive and difficult labor required to create such corpora [27]. Also for this reason, most available corpora are specific to a domain or purpose. Hence, there are very few general-purpose multimodal corpora.

Currently, rules lack the precision on a temporal level that data-driven techniques could provide, as rules are usually manually extracted from these multimodal corpora and implemented in ECAs [41]. BEAT was one of the first rule-based systems and it uses syntactic and semantic information to generate gestures and eye gaze behavior during speech [11]. Similar systems have been designed subsequently, such as the NUMACK system [28], the NVBGenerator [30] and more recently, the Cerebella system [31].

With the increase in multimodal corpora and computing power, researchers slowly started to investigate extracting patterns from data automatically. The first data-driven approaches focused on data from individual speakers [39]. These approaches were quickly extended to include data from multiple speakers. For example, [48] collected a dataset consisting of TED videos with subtitles and used an end-to-end learning approach to learn the relation between gesture and speech. Others have used prosody to automatically learn mappings between prosody and gestures [12] or a deep learning approach that learns gesture behavior from prosody, syntax and semantics [13]. Some researchers have used hybrid approaches that combine a rule-based approach with a data-driven one [4, 43]. Focusing only on iconic gestures, [4] uses a Bayesian network to decide which gesture to use, after which it is further specified by a set of rules. Finally, in a recent approach gestures were generated in an adaptive way through optimizations, based on feedback from human participants [6].

2.2 Dialog acts and gestures

Dialog acts are used to represent the pragmatic (contextual) meaning of user utterances in dialog. Although the communicative intention of a speaker is not marked explicitly in speech, dialog acts capture such intentions [1, 44].

Dialog acts and gestures could potentially be coupled in ECAs to generate realistic multimodal behavior. Just like prosodic, syntactic and lexical units, dialog acts are linguistic units that may relate to gestures, however wether such relationship exists, and if so, to what extent is an open question.

There are many theoretical arguments for the existence of such a relationship. [36] was the first to note that gestures also have a pragmatic function, next to their propositional content. [2] introduced a new class of hand gestures, called *interactive gestures*, stating the importance of (hand) gestures for the organization of discourse and turn-taking. On a pragmatic level, next to a contribution to the organization of the discourse, [24] distinguishes three other pragmatic functions. Gestures can provide the way speech should be interpreted, add to the meaning of speech by being an operator and also elucidate the speech act being used [24].

From an empirical point of view, however, there is little evidence for such relationship. [22] tried to establish an empirical relationship between dialog acts and gestures by analyzing the distribution of gestures across dialog acts in a small multimodal corpus of non-task-based conversations between three participants and found that turn-keeping utterances and fillers contain relatively the most hand gestures, while backchannels contain the least [22]. Moreover, self-touch motions occurred most with backchannels and laughter. This research helps to understand the different distributions of gestures among dialog acts, however the results do not necessarily imply a direct relationship between gesture and dialog acts. The data could simply reflect the fact that gestures help in dividing the information into small parts and thus help in conceptualizing information [25]. More concretely, gestures are more frequent during high cognitive load [26] and turn-keeping phrases intuitively have a higher cognitive load than backchannels. These results are interesting from a conversational agent point of view, since an ECA needs to know when to gesture. A similar study, looking at instances of hand gestures, observed relationships between dialog acts and certain interactive gestures [42].

So, while theoretically a relationship between gestures and dialog acts is not unlikely, current empirical evidence is for their relationship is unclear. In this study, we hope to shed more light on this relationship by investigating it with cross-recurrence quantification analysis.

2.3 Cross-recurrence Quantification Analysis

No default framework exists to scrutinize the temporal relationships between the different behavioral channels. We looked at frameworks utilized in studies regarding behavioral coordination between speakers and selected cross-recurrence quantification analysis (CRQA) [15]. Many other analytic frameworks disregard the temporal organization of the relationships and primarily aggregate data over the temporal dimensions of analysis. Such frameworks calculate event frequencies, rates or magnitudes [20]. Although such calculations have produced many insights, in the current research, we are primarily interested in the time-related patterns across the intention and gesture channels. If two events, i.e., a gesture activity and a dialog activity, often happen simultaneously, we could quantify this relationship by calculating the correlation between the two channels representing the activities. However, since the multimodal production system is more complex, this view would be too simplistic and a simple correlation would not be able to capture potentially complex relationships between the two channels. For example, it might be the case that events on two channels never happen at the exact same time, but follow or precede each other. This means that, next to the strength of the correlation, we also do not know the timing of this correlation.

In order to find out how two events relate temporally to each other, we need a measure that can quantify the relationship between them at different time shifts. One way to do this is to look at the co-occurrence of the events in time. For this, we use cross-recurrence analysis. This measure looks at the correlations in time, the recurrences of discrete events. The main idea behind cross-recurrence analysis is to determine how often events of one time series are succeeded (or preceded) by events of another time series, which is expressed in a proportional measure, called recurrence rate. This measure is obtained for a specific delay as follows: one of the time series is delayed, i.e., shifting all values of that time series a number of steps into the past. This means that the result of delaying a time series with one time step is that all values shift one step into history, such that the original value of the first time step is removed and the last element of time series is deleted. Thus, the length of the time series is one step shorter. In order to compare the shifted time series with a non-shifted time series, the last element of the non-shifted time series is also removed, such that both time series have the same length. The recurrence rate is obtained by creating a new time series that contains a 1 for each time point if both time series contain 1 for that time point as well. The recurrence rate is equal to the sum of the resulting vector divided by the length of that vector. The recurrence rate for a specific delay indicates how often an event on one channel co-occurs with an event on the other channel. Analyzing the recurrence rate for multiple delays informs us on how often events co-occur and within which time-frames [35].

3 METHOD

3.1 Dataset

A multimodal communication corpus of about 25 hours of dialog is used. 48 students from the University of Memphis participated (30 female, 18 male; 1 Asian, 19 African-American, 28 Caucasian) in a Map Task scenario [33]. This scenario requires two persons: an instruction giver, who has a map containing a route, and instruction follower, with a slightly different map without a route. The aim of the task is to reproduce the route of the instruction giver's map onto the instruction follower's map. The interlocutors can freely interact. However, they cannot see each other's map.

The corpus consisted of 13 encoded verbal and 10 encoded non-verbal behavior channels per interlocutor at 250 ms time intervals, with a total of 731,824 encoded intervals. The non-verbal behavior of each interval was encoded for five types of gestures according to the gesture coding scheme proposed by [37]: beat, deictic, iconic, metaphoric and symbolic gestures (emblems). Some gesture types were subdivided over multiple channels. The coders had a high inter-rater agreement score, quantified by Cohen's κ (.82).

Beat gestures are rhythmic hand movements that do not directly convey meaning but help marking discourse and organizing speech. Beat gestures were encoded over two channels: *beat single* contained the isolated, standalone beat events and *beat multiple* channel contained the sequences with multiple connected beat gestures. Deictic gestures referred to locations in space (i.e., pointing at something or someone). This space can be real (pointing at present objects and people) as encoded in the *deictic concrete* channel, or methaporical (pointing at abstract ideas, located in space), as encoded in the *deictic abstract* channel. Iconic gestures conveyed information about actions and object attributes by bearing partial resemblance to them, e.g., gestures depicting the path of the movement or the shape of the object. The behavior was encoded for iconic gestures, related to landmarks (*iconic landmark*) and those related to route information (*iconic route*). Metaphoric gestures were similar to iconic ones, in that they are also pictorial, but instead of bearing a resemblance to concrete entities, metaphoric gestures depict abstract ideas. Metaphoric gestures were captured in three different channels: *metaphoric level action*, gestures related to actions; *metaphoric meta-action*, metaphoric gestures related to meta-actions and *metaphoric metaphor*, metaphoric gestures that do not belong to the action and meta-action categories. Symbolic gestures (also called emblems) were conventionalized gestures (e.g., "thumbs up") and were least dependent on the speech content, as they do not require speech to be disambiguated. All symbolic gestures were encoded in one channel: *symbolic*.

The communicative intention was encoded by thirteen different dialog act types that are typically used for Map Task scenarios [7]. The types included dialog acts that communicate new information (*instruct*, *explain*, *check* and *align*), responses to previous dialog (*reply-yes*, *reply-no*, *reply-what*, *acknowledgement*, *clarification*), dialog acts related to preparations in the experiment (*ready*) and an *unknown* category for unclassifiable dialog acts. Dialog acts were encoded at utterance level, thus each conversational turn could potentially consist of a sequence of dialog acts. The coders had a good inter-rater agreement score, quantified by Cohen's κ (.67).

The 23 behavior channels were encoded as binary time series. If a behavior channel contained an event at a certain interval (e.g., the person was making an iconic gesture related to the route), the value at that time point was encoded as 1. If no event was measured at that interval (e.g., person was not making an iconic gesture), the value was encoded as 0. The dataset was also used in [33] and [5].

3.2 Analysis

The dataset was loaded into the statistical computing software R as a matrix, such that each row represented a specific interval and each column, a specific behavior channel. Hence, the matrix comprised 731,824 rows and 23 columns.

Subsequently, the recurrence rate was calculated for every possible combination of behavior channels. Thus, 23 behavior channels times 22 behavior channels resulted in 506 recurrence rate calculations. Each recurrence rate calculation involved calculating the recurrence rate per experiment (session) for every delay between 0 and 240 intervals (as each interval is 250 ms, this range corresponds with 0 and 60 seconds). The final result for each of the 506 recurrence rate calculations was the mean of the recurrence rates for each interval over all sessions for that specific combination. The final results were then plotted as a recurrence plot. As an example, the recurrence plots for *deictic concrete* with *beat multiple* and *deictic concrete* with *beat single* are shown in Figure 1.

For the calculation of the recurrence rates, we used a custom-made script [1]. This script produced exactly the same results as the widely-used CRQA analysis package for R [15]. The only difference with the CRQA analysis package is that, due to some minor optimizations in the custom made script, it allowed us to use longer time series as input and to shorten the calculation times significantly. Finally, because the recurrence rate is a proportional value that is difficult to interpret on its own, we also calculated the random (chance level) recurrence rate for each combination to facilitate a contrast and to determine the significance of the results at a later stage. Unlike other works that have facilitated such contrast by randomized baseline patterns [33], which can be wobbly and difficult to interpret by the human eye, this work uses the random recurrence rate, which produces straight baselines that make interpretation more accessible for human raters. The random recurrence rate (RRR) was calculated by dividing the number of events of the second behavior channel by the length (number of intervals) of the first behavior channel, as in Equation (1), where:

- $T1_s$ is the number of events of the first behavior channel.
- $T1_l$ is the length of the first behavior channel.
- $T2_s$ is the number of events of the behavior channel.
- $T2_l$ is the length of the second behavior channel.

$$RRR = \frac{T1_s}{T2_l}\frac{T2_s}{T1_l} \tag{1}$$

3.3 Results selection

To minimize the risk of both Type 1 error, i.e., classifying the recurrence plot as having the effect where there is none, and Type 2 error, i.e., selecting a wrong time window, such that the plot contains a real effect within a different time range, the results were selected in three stages, both automatically and manually. Each stage analyzed only the results selected in the previous stage.

The first selection automatically discarded recurrence plots without any significant difference between the recurrence rate and the random recurrence rate. To calculate if such significant difference exists for a combination, the delay with the largest difference between the recurrence rate and the random recurrence rate was taken as input for a significance analysis. The significance for this delay was calculated with a mixed-regression analysis with actual recurrence rate for that delay against the random cross-recurrence rate (see Equation 1), with role, session and dialog as fixed factors. The second selection automatically discarded recurrence plots where 20% or more of the time delays had a zero cross-recurrence rate, as this indicates data sparsity. The final selection was done by four raters, on an individual basis. For each recurrence plot, each rater judged whether the recurrence plots contained an effect or not. Plots that were not unanimously classified as having an effect were discussed in a group meeting. Plots for which at least three raters agreed upon a certain classification were then reclassified, while the others were identified as conflicted and removed from the analysis. In sum, all selected results had a significant peak or valley, contained less than 20% zero cross-recurrence rates and the cross-recurrence plot of those results were classified by at least three raters as having an effect.

[1] Code publicly available via: https://github.com/pblomsma/CRQA-turbo

4 RESULTS AND DISCUSSION

Cross-recurrence rates were analyzed for 506 combinations (23 times 22 behavior channels). Of the 506 combinations, 133 have been found significant by automatic analysis (described in Section 3.3), i.e., the largest distance between the random recurrence rate and the actual recurrence rate was significant for those 133 combinations. Four human raters classified 130 of those results unanimously as having an effect. An example of a recurrence plot with a significant effect is shown in Figure 1. In this figure, the temporal relationship between deictic gestures and beat gestures is examined from the perspective that beat gesture(s) happen at the same time or after the deictic gesture has started. Both types of gesture show the largest effect at time point 0, which means that both gestures most often occur simultaneously. It is also possible that beat gestures most often occur before deictic gestures, but that cannot be derived from Figure 1 as it only shows one side of the recurrence plot.

The results of the remaining 130 combinations were then further analyzed. For every recurrence plot, we identified the delay that corresponded with largest difference between random recurrence rate and actual recurrence rate between 0 and 1250ms. This time window was chosen to maximize the possibility that two behaviors are somehow related to each other. If the largest difference resulted from the actual recurrence rate being below the random recurrence rate, then such relation was classified as a mutual exclusive relation ("Mutex"), as this indicates that an event on the first behavior channel is not followed by an event on the second behavior channel. In other words, those events do not go together for that specific delay. On the other hand, recurrence plots where the largest difference resulted from the actual recurrence rate being higher than the random recurrence rate, were classified as "synchronized", as an event of the first behavior channel was followed above average by an event on the second behavior channel for that specific delay. These results are summarized in Figure 2. Due to the low number of participants, we deviated from good statistical practice by not splitting our dataset in two sets (i.e., one for the selection of effects, and one for the significance calculation), but used the same data for both the selection and the selective analysis. A potential consequence of this circular dependency is that the reported results could show

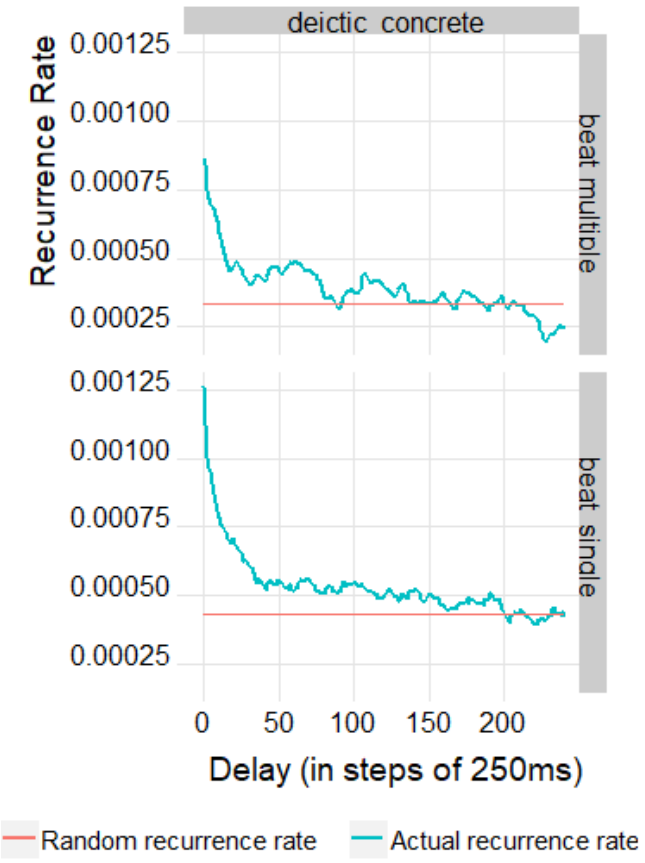

Figure 1: Recurrence plots for the behavior channel *beat multiple* and *deictic concrete*, and *beat single* and *deictic concrete*. The graph can be interpreted as follows: the actual recurrence rate for e.g., the delay 50 equals 0.00045 in the upper graph. This signifies that in 0.045% of the total analyzed time an event occurred at the behavior channel *beat multiple*, 50 intervals (of 250 ms) after an event occurred at the *deictic concrete* behavior channel. As the recurrence rate does not take into account the number of total events of both behavior channels, the rate is difficult to interpret on its own, therefore the random recurrence rate (see Equation 1) is also plotted to contrast the recurrence rate and facilitate its interpretation.

inflated correlations[29]. The next paragraphs discuss the notable aspects of Figure 2, grouped by details regarding the gesture and dialog act modality, and the relations between both modalities.

4.1 Within gesture modality

One would have expected that (1) a person could only produce one gesture at a time and (2) that gesture events in general would not occur often. Thus, one would not expect to find any temporal synchronizations between gestures, and expect only mutual exclusive relations. However, because of the second conjecture it is likely that the number of events would not result in any significant relations, and we would therefore find only non-significant mutual exclusive relations between the gestures (resulting in grey squares Figure 2).

Indeed, the results do not show any mutually exclusive relationship within the gesture modality. This could sprout from a data sparsity problem: i.e., not enough gesture events were available in the data to produce significant cross-recurrence valleys around time point zero. However, this non-exclusivity could also be explained by the fact that the boundaries between different gesture types are less clearly delineated than they are for the spoken modality. While it is possible to encode and classify gestures, they are not discrete and categorical the way words are [34]. Except for emblems that are conventional gestures, gestures are not "frozen" and can take on many forms. In consequence, being spontaneously produced, gestures are representative of thought processes or mental representations of an event [8], suggesting that gestures are not stored in the mental lexicon and happen on the fly.

However, against what one would expect, deictic and beat gestures seem to synchronize with a peak at 0 ms. Thus, speakers often start a deictic and beat gesture (both the *beats single* and *beats multiple* version) at the same time. Beat gestures are often found superimposed over other types of gestures [10], which could explain this finding. It could also be an encoding artifact. The starting phase of a beat gesture and a deictic gesture are quite similar, this similarity could have elicited uncertainty on the part of the encoders. Indeed, an inspection of the data reveals that out of the total 1562 intervals that were encoded as *beat single*, 987 (63.4%) were also encoded as deictic gestures. 705 (61.8%) of the 1141 intervals, that were encoded as *beat multiple*, were also encoded as deictic gestures. Thus, further research is needed to investigate if gestures in general do not happen together (with the exception of the aforementioned superimposed beat gestures).

4.2 Within dialog act modality

Regarding the relations among different dialog acts, one would expect to only find significant mutual exclusive relations between dialog acts with a peak delay at 0 ms. First of all, a speaker can only express one dialog act at a time, which implies that no dialog act combination can be synchronized and thus dialog acts should be mutually exclusive. Secondly, every utterance in the in the corpus (that contained about 25 hours of dialog) was encoded with a specific dialog act, the corpus should have enough dialog act data points to reap significant results. Finally, as a speaker most probably does not pause between dialog acts, one would expect to see 0 ms peaks.

Indeed, there is no synchronization between any of the dialog act combinations. However, contrary to the expectations, not all combinations have significant mutually exclusive relations either. Furthermore, not all significant mutually exclusive relations have a peak delay at 0 ms,

Why is the peak of all the significant mutually exclusive relations not at the 0 ms delay? There are several possible explanations. First, small interruptions by the interlocutor with e.g., a backchannel or a short acknowledgement could cause the speaker to stop speaking for a moment to subsequently continue with a different dialog act. Secondly, it could be an artifact of the data, since the start of a new dialog act, after a previous one ends is often within 250 ms [45]. The

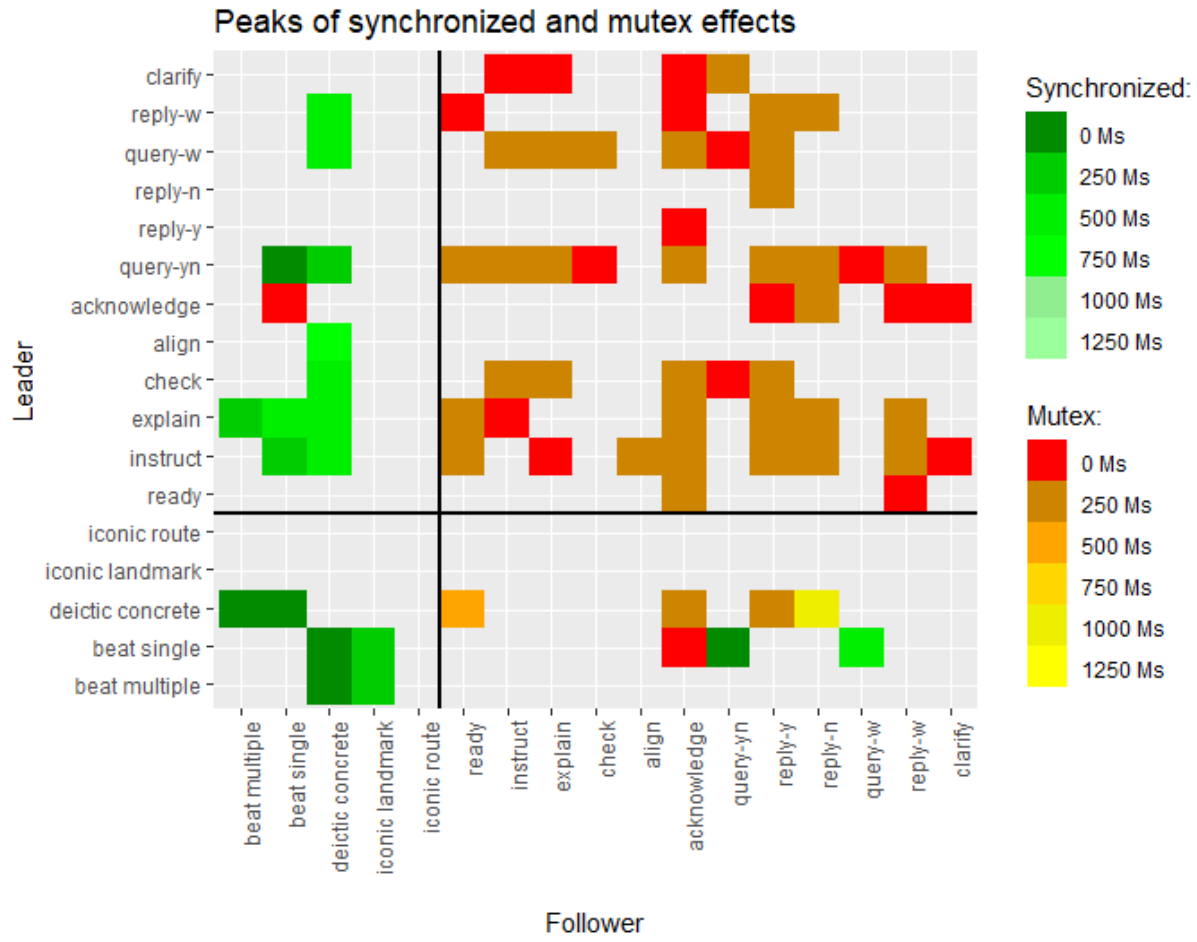

Figure 2: Overview of combinations of channels which had their most significant peak or valley within 0 ms and 1250 ms. Combinations that had their most significant peak or valley outside of this timeframe, or had no significant difference at all, were left out of this overview.

reason that some combinations have non-significant results may, in hindsight, sprout from a data sparsity problem. Not all dialog acts have a high frequency in the data. As a result, some combinations of dialog acts do not occur often enough to reap significant results.

4.3 Between modalities

As indicated in the background section, several researchers have given theoretical arguments as to why a relationship between gestures and pragmatic functions (which are represented by dialog acts) should exist. However, empirically, little of this relationship has been verified yet. Figure 2 shows multiple significant mutually exclusive and synchronized relations. Most synchronized relations show that dialog acts lead gestures. In other words, a gesture starts after a dialog act has started. For example, the *deictic abstractness concrete* gesture starts after an *instruct* or *explain* dialog acts has started (among others). This is contrary to what has been argued in the literature [32]. However, the *beat single* gesture often starts before a *yes-no* or *what query*. Most mutually exclusive relations are led by the gesture, or in case of the *acknowledge* dialog act and the *beat single* gesture, are at the same time (0 ms). Deictic gestures do not go together with the dialog acts *ready*, *acknowledge*, and *reply-yes* and *reply-no*. Interestingly, those dialog acts are also often short and, according to literature, do not involve high cognitive effort [26]. However, further research is needed to claim with more certainty to what extent deictic gestures and cognitive effort are related. Why are *acknowledge* and *beat single* gestures mutually exclusive? Some studies show that beat gestures are used by the speaker to emphasize certain cues [16]. Thus, the likely explanation for why *acknowledge* and *beats* are mutually exclusive is that the former is mainly used in listening while the latter has a role in speaking – for emphasis and discourse structuring [38]. Thus, it seems logical that neither go together, since one is mainly used while speaking and the other while listening.

Table 1: Spread indications

Channel A	Channel B	Start	End	Peak	
Beat multiple	Deictic	-21.00	21.75	0.00	+++
	Iconic landmark	-8.50	11.00	-0.25	+++
	Explain	-3.25	12.50	0.25	+++
Beat single	Beat multiple*	-0.50	0.00	0.00	- - -
	Deictic concrete	-50.00	60.00	0.00	+++
	Iconic landmark	-32.00	13.25	-0.25	+++
	Instruct	-37.75	8.00	0.25	+++
	Explain	-5.50	11.00	0.50	+++
	Acknowledge	-2.00	3.25	0.00	- - -
	Query-YN	-1.50	27.00	0.00	+++
	Query-W*	-2.75	0.00	-0.50	+++
	Clarify*	0.00	9.00	2.25	+++
Deictic concrete	Ready*	-2.25	0.00	-0.75	- - -
	Instruct	-37.50	60.00	0.50	+++
	Explain	-1.00	1.75	0.00	+++
	Check*	-18.50	0.00	0.00	+++
	Acknowledge	-19.00	2.75	0.00	- - -
	Query-YN	-2.00	2.50	0.25	+++
	Reply-Y	-6.50	6.00	-0.25	- - -
	Reply-N*	-4.00	0.00	-1.00	- - -
	Query-W*	-1.75	0.00	0.00	+++
	Check*	0.00	13.00	0.50	+++
	Align*	0.00	11.25	1.00	+++
	Query-W*	0.00	2.00	0.50	+++
	Reply-W*	0.00	9.75	0.50	+++
Iconic landmark	Reply-W	-0.75	4.25	1.25	+
	Reply-Y*	0.00	24.50	1.75	- - -
Iconic route	Iconic landmark*	-11.25	0.00	0.00	- - -
	Acknowledge*	-4.25	0.00	0.00	- -

Note. Pluses and minuses mark positive and negative regression coefficients. Positive coefficients correspond with peaks, negative coefficients with valleys (mutex). The number of symbols indicates p-level: +++ < 0.001, ++< 0.01, + < 0.05. Combinations marked with a * had only direction that was taken into consideration for this table, as the other direction was filtered out by the selection process.

4.4 Precision of event generation

The results have shown the existence of several temporal dependencies between combinations of dialog acts and gestures. How can those dependencies, both mutual exclusive and synchronized, be translated to exact multimodal behavior generation rules for ECAs? Figure 2 can be used to draw up a first set of rules about which dialog acts and gestures do or do not go together. However, if we specifically want to know within which time interval both events should take place, then Figure 2 does not provide enough detail. Figure 2, only providing the peak delay of the recurrence, is not informative regarding the delay at which the recurrence starts and ends. Knowledge about the exact start and end delays can be translated into specific rules, that take into account how tightly or loosely, the two channels are coupled and thus the time frame during which two events should be generated together. Therefore, we report those start and end delays (which can be interpreted as a measure for spread) for every significant combination in Table 1. The spread is specified by the start and end intervals of the cross-recurrence effect. The start interval indicates the interval at which the cross-recurrence rate starts to rise above the random recurrence rate, and the end interval indicates the latest interval before the cross-recurrence rate crosses the random recurrence rate again. In case a combination signifies a mutually exclusive relation, it is exactly the opposite: the start interval indicates the first interval below the baseline and the end interval indicates the latest interval before the cross-recurrence rate intersects with the baseline again. For example, *beat single* gesture events have been found above baseline 5.50 seconds before an *explain* dialog act starts, *explain* dialog act events have been found 11 seconds before a *beat single* gesture starts.

5 CONCLUSION

In the current study, we investigated the temporal dependencies between verbal and nonverbal behaviors. We specifically looked at which dialog acts and gestures do (or do not) occur in the same time frame.

The results provide multiple insights. First of all, significant effects for relationships within the gesture modality exist, such as deictic and beat gestures, and iconic and beat gestures being synchronized while relationships within dialog acts are mutually exclusive. Secondly, significant effects between dialog acts and gestures exist. These results function partly as a proof for the theoretical conjectures that the pragmatic function of speech is linked to the accompanying gestures [24], but also show that agent developers should not ignore those dependencies as they can help build more accurate multimodal behavior generation systems.

Interestingly, and contrary to the existing literature [32], we found that some dialog acts start before the gesture starts (e.g., people start with their explanation first, and then start to use deictic gestures). Furthermore, we found some instances of dialog acts and gestures that do not go together. Especially deictic gestures do not go with specific dialog acts (*ready*, *acknowledge*, *reply-no*, *reply-y*), which probably relates to the low information density of a typical instance of one of those dialog acts. In other words, deictic gestures are probably more appropriate with dialog acts that are more information-rich. However, more research is needed to further analyze this relation. Acknowledgments and beat gestures do not go together either.

We encourage ECA developers to compare their behavior generation rules with the measurements as presented in Figure 2 and Table 1 in order to get ECAs behavior generation closer to human behavior. Finally, we hope that not only this article provides insights into the relationships between intentions and gestures, but that it also shows the agent community the possibilities of using cross-recurrence analysis to translate human behavioral data into patterns to be used for agent dialog systems.

6 ACKNOWLEDGEMENTS

This research has been funded by grants NSF-IIS-0416128, NSF-BCS-0826825, and PROJ-007246 from the European Union, OP Zuid, the Ministry of Economic affairs, the Province of Noord-

Brabant, and the municipality of Tilburg awarded to Max M. Louwerse. We thank the raters of the recurrence plots for their time and effort. The usual exculpations apply.

REFERENCES

[1] John Langshaw Austin. 1962. *How to do things with words.* Oxford University Press, Oxford.

[2] Janet Beavin Bavelas, Nicole Chovil, Douglas A Lawrie, and Allan Wade. 1992. Interactive gestures. *Discourse processes* 15, 4 (1992), 469–489.

[3] Geoffrey Beattie and Heather Shovelton. 1999. Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica* 123, 1-2 (1999), 1–30.

[4] Kirsten Bergmann and Stefan Kopp. 2009. GNetIc–Using bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 76–89.

[5] Pieter A Blomsma, Julija Vaitonyte, Maryam Alimardani, and Max M Louwerse. 2020. Spontaneous Facial Behavior Revolves Around Neutral Facial Displays. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*.

[6] Angelo Cafaro, Brian Ravenet, and Catherine Pelachaud. 2019. Exploiting evolutionary algorithms to model nonverbal reactions to conversational interruptions in user-agent interactions. *IEEE Transactions on Affective Computing* (2019), 1–12.

[7] Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics* 23, 1 (1997), 13–31.

[8] Erica A Cartmill, Sian Beilock, and Susan Goldin-Meadow. 2012. A word in the hand: action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1585 (2012), 129–143.

[9] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (2000), 70–78.

[10] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[11] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of SIGGRAPH '01: The 28th International Conference on Computer Graphics and Interactive Techniques*. Association for Computing Machinery, New York, NY, United States, 477—-486.

[12] Chung-Cheng Chiu and Stacy Marsella. 2011. How to Train Your Avatar: A Data Driven Approach to Gesture Generation. In *Intelligent Virtual Agents*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 127–140.

[13] Chung-Cheng Chiu, Louis-Philippe Morency, editor="Brinkman Willem-Paul Marsella, Stacy", Joost Broekens, and Dirk Heylen. 2015. Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach. In *Intelligent Virtual Agents*. Springer International Publishing, 152–166.

[14] Herbert H Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

[15] Moreno I Coco and Rick Dale. 2014. Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in psychology* 5 (2014), 510.

[16] Diana Dimitrova, Mingyuan Chu, Lin Wang, Asli Özyürek, and Peter Hagoort. 2016. Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience* 28, 9 (2016), 1255–1269.

[17] Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is believing: body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–9.

[18] Maria Graziano and Marianne Gullberg. 2018. When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology* 9 (2018), 879.

[19] Boukje Habets, Sotaro Kita, Zeshu Shao, Asli Özyurek, and Peter Hagoort. 2011. The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience* 23, 8 (2011), 1845–1854.

[20] Sarah L Haywood, Martin J Pickering, and Holly P Branigan. 2005. Do speakers avoid ambiguities during dialogue? *Psychological Science* 16, 5 (2005), 362–366.

[21] Judith Holler and Geoffrey Beattie. 2003. Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture* 3, 2 (2003), 127–154.

[22] Carlos Ishi, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. Analysis of relations between hand gestures and dialogue act categories. In *Proceedings of the 9th International Conference on Speech Prosody 2018*. 473–477.

[23] Spencer D Kelly, Dale J Barr, R Breckinridge Church, and Katheryn Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language* 40, 4 (1999), 577–592.

[24] Adam Kendon. 2017. Pragmatic functions of gestures: Some observations on the history of their study and their nature. *Gesture* 16, 2 (2017), 157–175.

[25] Sotaro Kita. 2000. *How representational gestures help speaking.* Cambridge University Press, 162–185.

[26] Sotaro Kita and Thomas Stephen Davies. 2009. Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes* 24, 5 (2009), 761–775.

[27] Dawn Knight. 2011. The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada* 11, 2 (2011), 391–415.

[28] Stefan Kopp, Paul Tepper, and Justine Cassell. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th international conference on Multimodal interfaces*. 97–104.

[29] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience* 12, 5 (2009), 535.

[30] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 243–255.

[31] Margaux Lhommet and Stacy C Marsella. 2013. Gesture with meaning. In *International Workshop on Intelligent Virtual Agents*. Springer, 303–312.

[32] Max M Louwerse and Adrian Bangerter. 2010. Effects of ambiguous gestures and language on the time course of reference resolution. *Cognitive Science* 34, 8 (2010), 1517–1529.

[33] Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive science* 36, 8 (2012), 1404–1426.

[34] Gary Lupyan and Sharon L Thompson-Schill. 2012. The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General* 141, 1 (2012), 170.

[35] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. 2007. Recurrence plots for the analysis of complex systems. *Physics reports* 438, 5-6 (2007), 237–329.

[36] David McNeill. 1985. So you think gestures are nonverbal? *Psychological review* 92, 3 (1985), 350.

[37] David McNeill. 1992. *Hand and mind: What gestures reveal about thought.* University of Chicago press.

[38] David McNeill. 2006. Gesture: a psycholinguistic approach. *The encyclopedia of language and linguistics* (2006), 58–66.

[39] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.

[40] Regina Pally. 2001. A primary role for nonverbal communication in psychoanalysis. *Psychoanalytic Inquiry* 21, 1 (2001), 71–93.

[41] Matthias Rehm and Elisabeth André. 2008. From annotated multimodal corpora to simulated human-like behaviors. In *Modeling Communication with Robots and Virtual Humans*. Springer, 1–17.

[42] Hannes Rieser. 2011. Gestures indicating dialogue structure. In *Proceedings of SEMdial 2011, 15th Workshop on the Semantics and Pragmatics of Dialogue*. 9–18.

[43] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100.

[44] John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language.* Vol. 626. Cambridge University Press, Cambridge.

[45] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592.

[46] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview.

[47] Roel M Willems and Peter Hagoort. 2007. Neural evidence for the interplay between action, gesture and language: a review. *Brain Language* 101 (2007), 278–289.

[48] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.