Radboud University Nijmegen

Faculty of Social Sciences

Temporal entrainment as a mechanism for conversational turn-taking in dyadic and triadic interactions

Thesis MSc Artificial Intelligence

Author:
Karen Beckers

Supervisor:
Judith Holler

Second reader:
Andrea Ravignani

June 2021

Abstract

Turn-taking is a procedure in human communication that requires an enormous amount of temporal precision. Previous research has proposed that this exact timing is due to the entrainment of neural oscillators that facilitate speech rhythm tracking. Via this rhythmic synchronisation listeners can predict upcoming transition relevance places and anticipate on taking the turn. This research has tested two implications of this proposed model of entrainment. The first implication states that the rhythmic entrainment causes a correlation between the speech rates of the participants involved in a conversation. The second implication entails that, since the entrainment is counterphased, there will be a counterphased alignment of the rhythmic units of the two parties involved in a turn-taking event. Both hypotheses were examined on the stress level, represented by beats, and the syllable level. Analyses were performed on a corpus of dyadic and triadic English conversations. A beat tracking algorithm, originally designed for tracking temporal regularities in music, was applied to extract a sequence of beats from the audio signal. Beat rate is expressed as the average duration between subsequent beats; the mean inter-beat interval (IBI). Syllable rate is expressed in a similar fashion. Repeated measures correlation coefficients showed a negative correlation between the beat rates of two speakers in both experimental setups, though the strength of the effect was larger in dyads than in triads. A systematic relation between beat locations in overlapping turns was found, though this relation did not display counterphased alignment. Some form of entrainment of beat rates between the two parties in a turn-event seems present, though this entrainment does not follow the hypothesised system. No significant correlation between syllable rates in dyads and triads was found. All experimental conditions of the syllables rates showed no systematic pattern of rhythmic alignment. These results show no evidence for syllables as the rhythmic unit of entrainment.

Contents

1 Introduction **3**
1.1 Background 4
1.2 Related work 5

2 Method **7**
2.1 Data 7
2.2 Speech rate models 8
2.3 Analysis 10

3 Results **13**
3.1 Speech rate entrainment 13
3.1.1 Beat rate 13
3.1.2 Syllable rate 14
3.2 Rhythmic alignment 14
3.2.1 Beat phase 14
3.2.2 Syllable phase 17

4 Discussion **19**
4.1 Speech rate entrainment 19
4.2 Rhythmic alignment 22

Appendices **27**

Appendices

A Repeated measures correlation plots **28**

B Rose plots of circular statistics **32**

Chapter 1

Introduction

Human interaction, though it might seem as without order, has a highly organised structure. One of the most dominant factors of this organisation is turn-taking, where participants of a conversation alternate their speaking turns. In casual conversation, it is not explicitly specified how the interaction should be performed. Participants determine the procedure of communication locally without making it explicit and leaving room for evolvement of the procedure. This is in contrast with more structured forms of conversation such as interviews or debates, where there are often several rules of conduct. For instance, the content of the conversation, the length of turns and the order of the speaking turns are often fixed (Sacks et al., 1974). Casual conversation does not have speaking turns of fixed lengths or a predefined speaking order, and yet people are capable of such precise timing that turns seem to follow each other seamlessly. In fact, with an average gap length of around 200 ms, the pause between two turns is even smaller than the time it takes to produce a vocal reaction to a stimulus (600 ms) (Holler & Kendrick, 2015). Moreover, transitions are expected to operate smoothly. Ill-timed turn transitions are immediately noticed and can cause some feeling of awkwardness, when gap durations are relatively long, or rudeness, when there is no gap between the turns (Wilson & Wilson, 2005). This interesting phenomenon leads to the belief of some extensive cognitive processes underlying turn-taking.

To facilitate smooth turn-taking two requirements must be met (Levinson & Torreira, 2015). Firstly, a projection of the end of the current speaker's turn must be made. Secondly, one should be prepared to start speaking. Many principles have been proposed that could be involved in the turn-taking system (Duncan, 1972; Gravano & Hirschberg, 2011; Rochet-Capellan & Fuchs, 2014). A mechanism that might seem the most apparent is the use of cues. These cues can be non-verbal, such as hand gestures or eye movement, or they can be part of the speech signal. In the latter case the indicators will either be acoustic, when there is, for example, a pitch drop at the end of a sentence or an increase in pitch at the end of a question, or textual, such as the punch line in a joke (Gravano & Hirschberg, 2011). Next to cues, it has been shown that specific breathing patterns are involved in the organisation of turns (Rochet-Capellan & Fuchs, 2014). Another mechanism that can accommodate both requirements for smooth turn-taking is speech rhythm entrainment.

Rhythmic entrainment is a phenomenon that can take many forms. During this process body movements or vocalizations are synchronised to an external rhythmic pulse (Wilson & Cook, 2016). This occurs, for example, during dancing and in music when multiple musicians have to play together or when someone taps along to a beat. This requires some means by which the rhythm is transferred to all parties that participate in the synchronisation, such as a metronome. Moreover, in order for two individuals to react in synchrony predictive timing must be employed, similar to the condition for a smooth turn transition (Merker et al., 2009). Rhythmic entrainment is not exclusive to humans. It has been observed in a wide variety of animal species, among which are sea lions, parrots and several types of apes (Cook et al., 2013; Lameira et al., 2019; Large & Gray, 2015; Wilson & Cook, 2016).

Speech is another rhythmic entity that can be sensitive to entrainment. It is proposed that speech rhythm can be described by a hierarchical model consisting of multiple levels (Martinec, 2002). The foot level is the lowest level in the hierarchy, where a foot is the smallest rhythmical unit consisting of only a few syllables. The highest level in the hierarchy are generic stages, such as a chapter in this paper or the method section of a recipe. The rhythms at the different levels are represented as waves and constitute a series of accents. In the lower hierarchical levels, turn transitions are viewed as rhythmic gaps that break the isochronous sequence at this level. However, turns might be connected through one or several of the higher levels. Moreover, cooperation between the parties in a conversation enables the establishment of a

joint rhythm that transcends turns, in contrast to conflictual speech exchange such as political interviews (Martinec, 2002). This research aims to investigate the presence of rhythmic entrainment in speech for dyadic and triadic conversations by considering syllables and beats as rhythmic wave accents.

1.1 Background

Sacks et al. (1974) present a thorough characterisation the organisational nature of turn-taking in spontaneous speech exchange. They observe several aspects that are implicitly present in any conversation. For example, lengths of turns are not organised, the speakers between whom a turn transition takes place are not predefined, the distribution of turns among the participants is not fixed and instances of more speakers at once can occur but are very short. Sacks et al. (1974) note that any model describing the organisation of turn-taking should be flexible to suit this variability of conversation. Moreover, it has to be context free, meaning that it covers all the major features that are involved without imposing a specific context. It should be emphasised that such a system only applies to spontaneous, casual speech. Given that other settings of speech exchange employ different rules, for example in a play the turns are preallocated, these settings call for a different turn-taking scheme.

The proposed model consists of two elements and a set of rules. The first element is the turn-constructional component, which describes the building blocks of a turn. These building blocks are called turn-constructional units and can be of different sizes, such as words, clauses or sentences. Each turn can be formed by a sequence of these units of indefinite length, but it must at least comprise one turn-constructional unit. The completion of an unit indicates a transition relevance place, where a speaker change can take place. The second element of the model is the turn-allocation component, which consists of two techniques for turn allocation. In one of the techniques, the current speaker selects the next speaker, while in the other technique the next speaker is appointed through self-selection. Taking into account these two components, Sacks et al. (1974) define a rule set for speaker appointment.

a) The current speaker explicitly selects the next speaker, who is then obligated to take the following turn.

b) A listener selects him/herself as the next speaker. The turn is appointed to the first person who starts talking.

c) The current speaker selects him/herself by proceeding the turn.

The three situations within the model take place in the presented ordering. If no speaker change occurs the procedure repeats recursively at each upcoming transition relevance place, until a new speaker is appointed. The described model explains the coordination of turn-taking, but it does not specify how the precise timing of a transition is established.

As mentioned before, to ensure a turn-transition runs smoothly, one must be physically ready to speak at the anticipated end of the current speaker's turn (Wilson & Wilson, 2005). This implies that the readiness to take the turn is already developing when the current turn-holder is still speaking, while the upcoming transition relevance place is predicted. Moreover, when arriving at a transition relevance place, each participant of the conversation must be monitoring which of the three rules can be exercised. For example, in dialogues if the current speaker does not select the listener to take the turn, the listener knows he or she can self-select. If the listener then does not take the turn, the current speaker has the opportunity to self-select and continue the turn. Taking into account the periodicity of the turn allocation rule set, it was found that the gap between turns is most often a multiplication of some length S, where S represents one cycle of the rule set and is conversation-specific. These requirements suggest some form of temporal entrainment with a periodic characteristic. Such synchrony has also been established in, for example, breathing patterns at the moment of a turn-transition.

Wilson and Wilson (2005) propose that this temporal prediction is facilitated by endogenous oscillators. These neural populations are known to possess a shared periodicity in their activity that enables them to perform timing-related operations in the brain, which makes them valid candidates for accommodating the precise timing of a turn transition. Next to this, neural oscillators have the capacity to become synchronous if they are allowed to affect each other. This synchronisation of neural populations is in fact so prevalent that it is considered to be an intrinsic principle of the brain (Wilson & Cook, 2016). Taken together, this capacity for timing and synchronisation leads to the belief that in a conversation the neural oscillators of the listener become entrained to those of the speaker, enabling the listener to project upcoming transition relevance places in the utterance of the speaker. As previously mentioned,

for rhythmic entrainment to occur it is required that the rhythm can be perceived by all participants in the conversation. Therefore, in order to allow these oscillators to become phase-locked, a medium is needed to transfer the periodicity. It is argued that this transmission is executed via syllable production (Wilson & Wilson, 2005). The production of one syllable corresponds to the period of the cyclic pattern for the readiness to speak, in which one is maximally ready to speak at the peak of each cycle. The oscillators of the listener synchronise to those of the speaker via speech rate, in terms of syllable production, but the cycles are counterphased, meaning that the listener's readiness is at its highest when the speaker is at a minimum. Finally, if the listener does not take the turn in the first cycle after the speaker finishes talking, the synchronised periodicity remains present for a short period of time, after which the entrainment breaks down due to the absence of the communication of syllables. Considering a speaking tempo of 200 ms per syllable, a commonly used estimate of the average syllable rate in American English (Wilson & Wilson, 2005), and the counterphase synchronisation, the listener will take the turn at roughly 100 ms before or after the end of the speaker's turn, both of which will be perceived as a fluent transition. Additionally, the model implies that the participants should have coordinated syllable rates and that this rate is correlated with S, the length of one cycle in the Sacks et al. (1974) model. Moreover, due to the counterphase entrainment of speaker and listener, the probability for simultaneous starts in dialogues is reduced. However, in multiparty conversations there are multiple listeners who will be in-phase to each other, which in turn increases the possibility of simultaneous attempts to take the turn.

1.2 Related work

Several papers have tested the proposed model for the timing of turn-taking. While Wilson and Wilson (2005) propose the syllable as the rhythmical unit for entrainment, other studies have suggested that the beat might be a better fit (Beňuš, 2009; Schultz et al., 2016). This unit is composed of only the prominent, accented syllables and might thus be considered a higher rhythmic level than syllables in the hierarchy proposed by Martinec (2002). Beňuš (2009) compared pitch accents, as a representation of beats, and syllables as the rhythmic units of the oscillator model. The English corpus was composed of dialogues that emerged while playing collaborative computer games. Even though pitch accents appeared to be better rhythmical structures than syllables, their overall results poorly supported the concept of rhythmical entrainment.

O'Dell et al. (2012) further build on the analysis of Beňuš (2009) and intended to broaden the scope of the neural oscillator model to other languages. By modelling pause duration in a Finnish corpus of one speaker pair they tested the implication of the Wilson and Wilson (2005) model that the duration of a pause will be a multiple of some time unit. Here, the assessed time unit was the syllable rate of the chunk of speech preceding the pause, but no significant results were found.

A comparison of three oscillator models, originally constructed for music perception, was presented by Inden et al. (2012), two of which consisted of a single oscillator and the third model was a network of oscillators adapted to various frequencies. Experiments were executed using a German corpus of spontaneous dyadic conversation. It was shown that, whereas the performance of single oscillators does not exceed chance level, a network of oscillator performed significantly better at capturing the rhythmicity of syllable timing.

Schultz et al. (2016) examined speech synchronisation as the convergence of speech rates over the course of a conversation. Similar to Beňuš (2009) they propose that entrainment is established via speech stress, as opposed to syllable onsets. A beat tracking algorithm, capturing beats in terms of stressed syllables, was applied to an English corpus of scripted dialogues. One person in each speaker pair was associated with the research group and was asked to speak at a certain tempo. The results showed that participants did adapt their speech rate to that of the confederate and beat rates were shown to converge during the conversation, confirming the theory of Wilson and Wilson (2005). However, the reading of a play script can not be considered casual speech, since the content of the conversation, the length of turns and the turn distribution are all predetermined. This setup is therefore not in line with the requirements of the turn-taking scheme (Sacks et al., 1974). In light of the positive results regarding speech rhythm entrainment, it would be a good next step to implement the beat tracking algorithm to a setup that does consist of spontaneous speech.

All previous research has solely focused on two-party conversations. As stated in the second rule of the Sacks et al. (1974) rule set, the first person who self-selects will obtain the turn. In dialogues, there is only one candidate who can carry out this rule. On the other hand, in triadic conversations there

will be two participants who might execute self-selection at the upcoming transition-relevance place. When both are trying to self-select, these two parties will have to compete for the turn. Common tactics for winning this race are breathing in loudly, interjecting signaling words and rush-throughs. This competition might affect the rhythmical alignment of the transition. As multi-party conversations account for a large portion of everyday talk, research on this type of speech exchange is needed to get a clear image of speech synchronisation in all facets of casual conversation.

Understanding the mechanisms underlying the timing of turn-taking could aid the development of spoken dialogue systems and social robots. The precise timing of a turn is one of the most dominant factors in making a conversation feel natural, but it is also one of the most difficult aspects to implement in an artificial agent (Żarkowski, 2019). Recent years have seen vast improvements in natural language understanding, processing and generation, since the arrival of large language models based on the the transformer architecture (Vaswani et al., 2017). Combining these developments with the enhancement of the timing mechanisms in artificial agents and dialogue systems could be the push forward that this field of research needs.

This study implements the beat tracking algorithm, as applied by Schultz et al. (2016), in a situation of spontaneous speech exchange. Here, rhythmical alignment in speech is viewed from two angles. Firstly, we look at the entrainment of speech rate, or tempo. Wilson and Wilson (2005) indicate that an implication of their model is participants should talk at similar rates during a conversation. The second approach looks at the components of that speech rate. As tempo is a metric based on a sequence of evenly spaced points in time, we inspect the alignment of those points, which are expressed as either syllables or beats. These two angles will be applied to both dyadic and triadic conversations.

The remainder of this paper is structured as follows: chapter 2 describes the architecture and implementation of the applied models. An overview of the data and the statistical methods can also be found in this section. In chapter 3 the statistical results for speech rate correlation and rhythmic alignment are reported. Finally, an interpretation of the results and future aims are presented in chapter 4.

Chapter 2

Method

2.1 Data

The data comprises audio recordings of 12 two-party and 11 three-party conversations of spontaneous speech, each with a duration around 20 minutes. This corpus was originally constructed for a research that examined eye gaze during turn-taking events (Holler & Kendrick, 2015). This study on eye gaze required eye tracking data, video recordings and audio recordings. The research described in this paper only uses the audio files. For each group of participants, the triadic conversation was recorded first, then one participant left the room and the remaining two participants engaged in a dyadic conversation. All participants were native English speakers and the participants within a group were familiar to each other. Their ages ranged from 19 to 68 years, with a mean of 30 years. All conversations were recorded at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. In order to record the individual voice of a speaker, each participant was equipped with a head-mounted microphone. In addition, a ceiling microphone recorded all the participating parties. Thus, for each conversation this leads to either two (for triads) or three (for triads) audio files of individual speakers and one audio file of the entire conversation. For more information on the full recording procedure, see Holler and Kendrick (2015).

Qualified conversation analysts annotated each precise turn beginning and end in Praat, which was then transported into ELAN (Boersma & Weenink, 2021; Wittenburg et al., 2006). The annotations can be divided into two groups. The first group consists of solely question-response pairs (QR) where a question was directed towards a specific individual, who then replied (Holler & Kendrick, 2015). This corresponds to the execution of the first rule in the aforementioned rule set. The other group of annotations, henceforth called TT, are all turn transitions that are not part of a question-response sequence. All analyses are performed on these distinct groups of turn transitions and on a combination of the groups, resulting in the set of all annotated transitions. In total there are 1346 annotated pairs of turns, of which 612 are dyads and 734 are triads (see table 2.1). Within these groups, another distinction has been made between overlapping and non-overlapping turn transitions. Most research on turn-taking has either focused on non-overlapping turns or made no clear division between overlap and non-overlap and it has been argued that more research into overlapping transitions is needed in order to generalise about the turn-taking organisation in both situations (Wilson & Wilson, 2005). Figure 2.1 shows the distribution of latencies for dyads and triads, where latency is the time between the end of the turn of speaker 1 and the beginning of the turn of speaker 2. Negative values thus indicate an overlapping transition. Very short latencies are most common, as is expected for smooth turn-transitioning. In 36% of all transition pairs the time of the transition falls within the range of -200 ms and 200 ms, which will be perceived as a perfect transition without gap or overlap. Moreover, 70% of all latencies are within the range of -500 ms and 500 ms. Beňuš (2009) noted that as an implication of the counterphased entrainment of the neural oscillators, the distribution of the latencies should be bi-modal, with a valley around 0 ms and a peak on either side. However, their data did not show such a distribution and neither does the data in this research, as shown in figure 2.1.

	Dyads			Triads		
	Overlap	*Non-overlap*	*Total*	*Overlap*	*Non-overlap*	*Total*
QR	158	164	322	145	134	279
TT	70	220	290	184	271	455
All	228	384	612	329	405	734

Table 2.1: Distribution of turn transitions

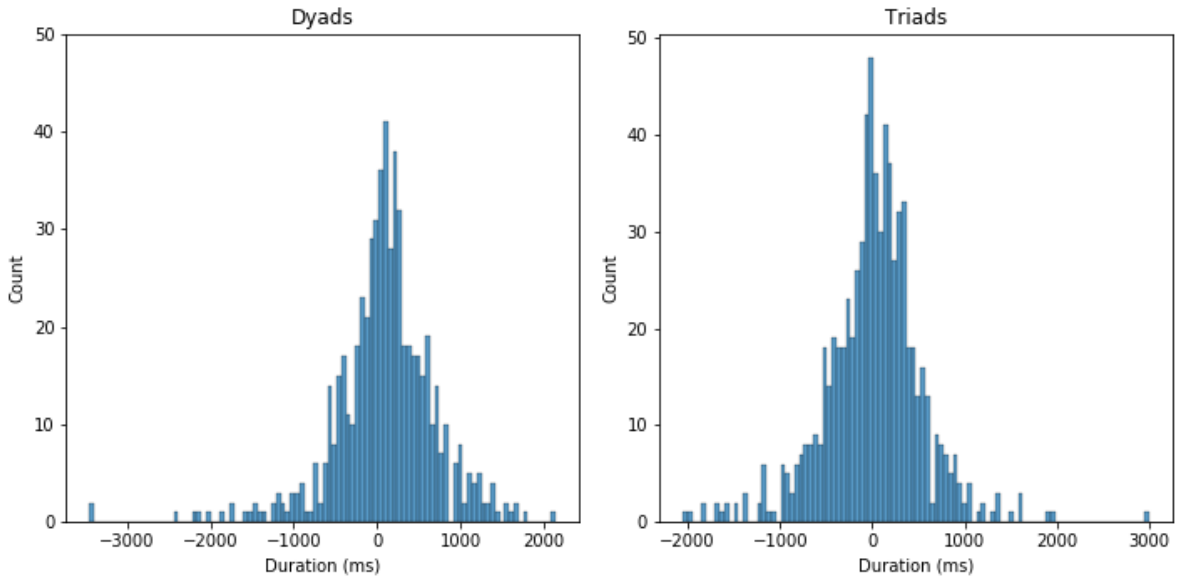


Figure 2.1: Distribution of latencies for dyads (left) and triads (right).

2.2 Speech rate models

The algorithm applied in the Schultz et al. (2016) research was originally developed for beat tracking in music by Ellis (2007) for the 2006 MIREX Audio Beat Tracking task. It was designed to meet two conditions. First, the outputs of the algorithm should be equal to temporal locations in the audio signal where a beat is indicated. Secondly, the detected beats should be somewhat evenly spaced in order to establish a rhythm. The sum of these requirements results is a goal that can be captured by dynamic programming.

In dynamic programming a task is divided into subtasks or stages. At each stage a decision must be made about which action to take. This decision is based on a so-called policy; a rule that determines the next action, given the current state of the system. The optimal policy of a subtask will be the action that either maximizes the profit or minimizes the cost of some constraint applicable to the task at hand. For example, when traveling from A to B, you could choose to minimize the duration of the travel or the travelled distance and they will both result in different routes. The main idea of dynamic programming is that the best solution for the global task will be the accumulation of all the optimal policies (Bellman, 1966).

The beat tracking algorithm has both above-mentioned conditions embedded in its policy. An onset envelope is calculated by first mapping the audio signal onto 40 Mel frequency bands and then summing the intensity values over all bands for every time point (Ellis, 2007). Next, the signal is equalized using a high-pass filter and then smoothed with a Gaussian envelope, resulting in an envelope defining the onset strength. The onset strength is employed as a measure for identifying possible beat locations, where viable beat locations are represented by high onset strength values. The fitness of the identified beats within the rhythmic pattern is determined by comparing the interval of subsequent beats with an ideal inter-beat interval (IBI). Here lies the biggest restriction of this beat tracking algorithm. In order to determine an ideal IBI, the algorithm requires a predefined tempo estimate. For example, in a piece of music with a predetermined tempo of 120 beats per minute, the ideal IBI would be 500 ms. Equation 2.1 shows the function that calculates the fitness of a beat time, where Δt is the current IBI and τ_p is

the ideal IBI.

$$F(\Delta t, \tau) = -\left(\log \frac{\Delta t}{\tau_p}\right)^2 \tag{2.1}$$

The function will return a value of 0 if both intervals are equal and larger differences between the two arguments will return larger negative values. The maximisation of both elements, the onset strength and the beat spacing, comprises the objective function of the algorithm:

$$C(\{t_i\}) = \sum_{i=1}^{N} O(t_i) + \alpha \sum_{i=2}^{N} F(t_i - t_{i-1}, \tau_p) \tag{2.2}$$

Here, $\{t_i\}$ is the returned sequence of beats and C represents the scores at those moments in time. $O(t)$ is the onset strength envelope and the parameter α can be set by the user to determine the importance of isochrony in the beat sequence. Larger values will increase the penalty this term puts on a deviation from the ideal IBI, which will result in a more rigid adherence to the predefined tempo. To apply this algorithm to speech, a low value for α must be implemented, since speech has a far less isochronous nature compared to music (Schultz et al., 2016). The optimal series of beats can be found through recursion, as the optimal policy score at time t is the sum of the local onset amplitude and the score corresponding to a preceding beat location τ where the combination of that score and the transition cost $F(\Delta t, \tau)$ is at a maximum. This can be captured in the following recursive equation:

$$C^*(t) = O(t) + \max_{\tau=0\ldots t} \{\alpha F(t - \tau, \tau_p) + C^*(\tau)\} \tag{2.3}$$

This optimal score C^* is calculated for every point in time on a 4 ms time grid, while simultaneously monitoring the prior beat instance that resulted in the best score. The function that represents this process of monitoring optimal prior beat instances is shown in equation 2.4.

$$P^*(t) = \arg\max_{\tau=0\ldots t} \{\alpha F(t - \tau, \tau_p) + C^*(\tau)\} \tag{2.4}$$

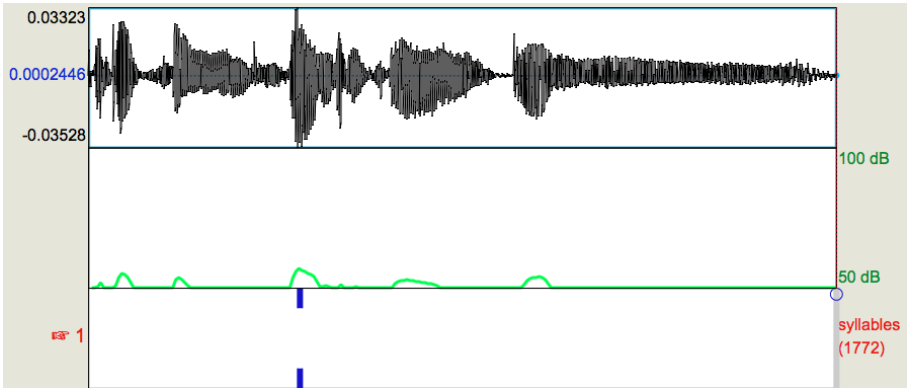
P^* can be viewed as some sort of lookup table that for every beat location t holds the location of the optimal preceding beat. As the resulting scores will get increasingly larger throughout the chunk of audio, the time instance that resulted in the largest value for C^* is set to be the final beat in the sequence. From this point, the algorithm works its way back through all the stored best preceding beats in P^* until it arrives at the beginning of the audio signal and will have found the optimal beat sequence.

Schultz et al. (2016) observe that, due to its compliance with temporal variability in the audio signal, this beat tracking algorithm can be well-suited for analysing speech. Moreover, beats are identified based on acoustic features that are related to speech stress, such as pitch and intensity. In their research, Schultz et al. (2016) take the extracted series of beats to construct a notion of beat rate. They define beat rate as the average inter-beat interval of an audio chunk. A similar approach is applied in the current research.

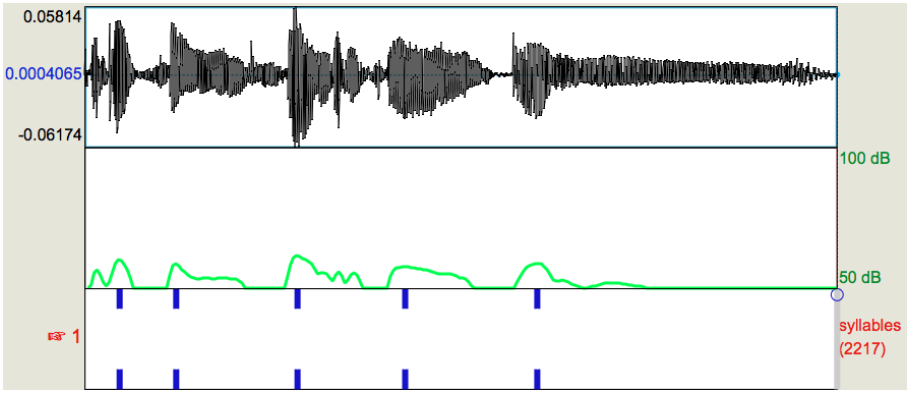
The annotated turns of each speaker in each conversation and each of the three transition types were taken out of the audio signal and were put together to create an audio file that solely consists of speech belonging to turns in that category of the dataset. This results in three audio files per speaker per conversation. The tempo estimate for individual speakers was based on all the audio in the corresponding wavefile, with all silences longer than 500 ms removed (Giannakopoulos, 2021). To allow for tempo variation within the resulting beat sequence, the parameter α was given the low value of 10. The beat tracking algorithm with these two parameter settings was applied to the three series of turns of each speaker. The output sequences were divided in individual turns in retrospect. The beat rate of a turn was calculated by taking the mean difference between adjacent elements in the sequence of beat times.

Since the model proposed by Wilson and Wilson (2005) specifies syllable production as the medium of entrainment, speech tempo is also measured at the syllable level in terms of syllables per second, henceforth called syllable rate, similar to the approach of Schultz et al. (2016). Syllables are distinguished by applying a Praat script that detects syllable nuclei (De Jong & Wempe, 2009). A syllable nucleus is defined as the peak of a syllable, which often corresponds to the vowel. First, the peaks of energy (dB) are detected within the intensity contour. Then, the unvoiced peaks are removed with help of the pitch contour. The script is applied to the complete audio recording of a speaker and afterwards the

resulting syllable locations are assigned to the corresponding turns. It was found that in some cases the peaks of intensity could not be detected, as the intensity level did not exceed a certain threshold and the absolute difference between the peak and the surrounding dips was too small. Therefore, the amplitude of the audio files was multiplied by 1.75, to aid better detection. The result of boosting the amplitude is shown in figure 2.2. The waveform, intensity contour and locations of the detected syllables are shown in the top, middle and bottom panel respectively in both images. Not all peaks in the intensity contour are detected in the original audio, but after boosting the signal all peaks are identified. Similar to the method of obtaining beat rates, the syllable rates are calculated as the average duration between two subsequent syllables.



(a) Without boosting



(b) With boosting

Figure 2.2: Syllable detection in the original audio and with the amplitude boosted

2.3 Analysis

The speech rates are analysed in two different approaches. The first approach inspects speech rhythm entrainment in terms of rate correlation. For each transition pair we have a beat rate of the person who speaks first, speaker 1, and a beat rate of the person to whom the turn is transitioned to, speaker 2. The model proposed by Wilson and Wilson (2005) suggests that the speech rates of the participants in a conversation will be approximately equal. This implies that the mean IBI of speaker 1 will be similar to the mean IBI of speaker 2. This implication is tested using a statistical technique called repeated measures correlation (rmcorr) (Bakdash & Marusich, 2017). Analyses are executed with the rmcorr package for R. More standard approaches for correlation have the assumption of independence as a restriction. However, this assumption does not apply to the current data, since there are multiple paired observations within each conversation. Rmcorr takes into account the non-independence within a conversation and allows one to calculate the association between two variables, examined per conversation on multiple occasions. The two variables at hand are the beat rates of speaker 1 and speaker 2. The rmcorr coefficient r_{rm} is a value between -1 and 1 and represents the linear correlation between the two terms over all conversations. When plotting, each conversation has its own regression line, whose slope is equal to the r_{rm} value. Since the rmcorr coefficient is an overarching value for all conversations, the

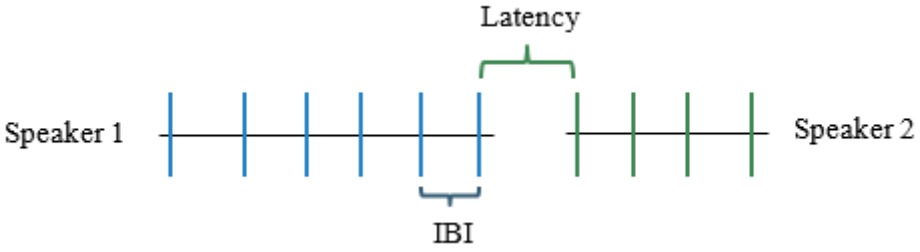


Figure 2.3: Abstract representation of beat alignment in non-overlapping turns

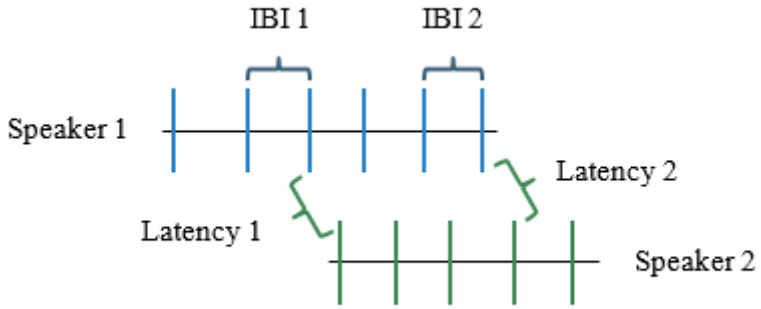


Figure 2.4: Abstract representation of beat alignment in overlapping turns

slope of each regression line is identical. The intercept of each line is conversation-specific. This thus results in a set of parallel regression lines. Since rmcorr can take all observations within a conversations into account, compared to other methods that would need to average over observations, this technique is said to have much more statistical power than these other, simpler methods (Bakdash & Marusich, 2017). This statistic is applied to the syllable rates in an identical fashion.

The second approach looks at entrainment in terms of rhythmical alignment. In the Wilson and Wilson (2005) oscillator model beats of speaker 1 and speaker 2 are expected to be half a period apart, due to the counterphased entrainment. Even when there is a pause between turns, the entrainment continues and the first beat of speaker 2 should be located at roughly $N + 0.5$ times the cycle length S of speaker 1, where $N = \{1, 2, 3, ...\}$. Circular statistics provide a method to examine this beat spacing while taking in account the periodic nature of the data (Cook et al., 2013). In this technique one oscillatory period is viewed as a cycle of 360°and if counterphased, the beats of speaker 1 and 2 will be 180°apart. The relative phase that is used here is calculated as a ratio between latency and IBI. When multiplied with 360, this results in a phase angle vector indicating the angular distance between the beats. The latency and IBI of interest differ for overlapping and non-overlapping turns. Figure 2.3 shows an abstract representation of the non-overlapping situation. Here, the IBI of interest is the final IBI of speaker 1 and latency is the distance between the final beat of speaker 1 and the first beat of speaker 2 (Ravignani, 2018). The group of overlapping turns is divided into two situations, as depicted in figure 2.4. In the first situation, henceforth called overlap (1), we look at the rhythmical alignment of the first beat of speaker 2 with the closest IBI of speaker 1. The second situation, overlap (2), inspects the alignment of the final IBI of speaker 1 and the beat of speaker 2 that is closest. Since the timing of turn transitions in overlapping turns has not been extensively studied, both overlap situations are included in this research. If the two speakers are entrained with regard to both rate and alignment, latency 1 and latency 2 in figure 2.4 should be equal. The calculated ratios can be visualised by circular histograms called rose plots. Each histogram bin accounts for 18°(Cook et al., 2013). Within these plots, a vector can be drawn that has the mean phase angle of all observations and a length between 0 and 1 that represents the spread of the data. If all observations would have the same angle, the length of the vector will be 1.

Three tests from the circular statistics domain are applied. The Rayleigh test of Uniformity analyses the presence of a systematic pattern of beat spacing between speaker 1 and speaker 2. When applied

without specifying a mean phase angle, this statistic tests for the presence of an unimodal distribution. The test can also be implemented with a specific mean angular phase μ to examine a relation between speaker 1 and speaker 2 in that explicit direction. The resulting value z represents the strength of the relation. The Rayleigh test is applied in three different manners; without specified mean, μ is 180°and μ is equal to the mean phase angle. The circular ANOVA is used to compare distributions between the experimental setups (Ravignani, 2018). Since significant results of the circular ANOVA only indicate that there is some distributional difference within the tested group, a Watson U^2 test is applied to gain a better understanding of the difference between each pair in the tested group. Statistical analyses are performed using the circular package for R (Agostinelli & Lund, 2017). The same method is applied to the syllable rate values.

Chapter 3

Results

3.1 Speech rate entrainment

3.1.1 Beat rate

After analysis there were 102 dyadic and 126 triadic transition pairs where either one of the turns or both yielded a mean IBI value of 0. These are all very short turns with an average duration around 500 ms, which is too short to extract a sequence of beats. These transition pairs are therefore excluded from the research.

Table 3.1 shows the rmcorr coefficients. All experimental conditions show a negative correlation between the beat rates of speaker 1 and speaker 2. Only the overlapping set of question-response pairs in the triads shows no significance. All other conditions have a significant relation. The r_{rm} coefficients of the dyads demonstrate a much stronger negative correlation, with values ranging from -0.52 to -0.63. In contrast, the correlation values in the triadic condition lie between -0.15 and -0.39. A manual evaluation shows that in the dyadic transitions, the r_{rm} values of the overlapping and non-overlapping sets do not show any great differences, where the biggest difference of 0.06 is observed in the TT condition. Differences in correlation within the triadic conversations are much larger, with a largest difference of 0.22 between overlapping and non-overlapping turns in the QR set of transition pairs, though the r_{rm} for overlapping QR is not significant. However, the largest difference in significant correlation values is still found in the QR set, where the difference in correlation between the non-overlapping turns and the total set of turns is 0.12. The correlation plots are shown in the appendix. As mentioned before, the slope of each regression line is equal to the rmcorr coefficient of that experimental situation.

Dyads			
	Overlap	*Non-overlap*	*Total*
QR	$r_{rm} = -0.6,\ p < .001$	$r_{rm} = -0.59,\ p < .001$	$r_{rm} = -0.6,\ p < .001$
TT	$r_{rm} = -0.57,\ p < .001$	$r_{rm} = -0.63,\ p < .001$	$r_{rm} = -0.62,\ p < .001$
All	$r_{rm} = -0.55,\ p < .001$	$r_{rm} = -0.52,\ p < .001$	$r_{rm} = -0.53,\ p < .001$

(a) Dyads

Triads			
	Overlap	*Non-overlap*	*Total*
QR	$r_{rm} = -0.15,\ p = .099$	$r_{rm} = -0.37,\ p < .001$	$r_{rm} = -0.25,\ p < .001$
TT	$r_{rm} = -0.39,\ p < .001$	$r_{rm} = -0.28,\ p < .001$	$r_{rm} = -0.32,\ p < .001$
All	$r_{rm} = -0.31,\ p < .001$	$r_{rm} = -0.27,\ p < .001$	$r_{rm} = -0.37,\ p < .001$

(b) Triads

Table 3.1: Results of the rmcorr statistic for beat rates

3.1.2 Syllable rate

Even after boosting the audio signal, the syllable rate script was unable to detect a sequence of syllables in 285 dyads and 258 triads, resulting in a syllable rate of 0. These data points are therefore excluded in the analyses.

The results for the repeated measures correlation of syllable rates can be viewed in table 3.2. Contrasting to the beat rate results, all of the dyadic conditions show a positive correlation between the syllable rates of the two speakers. Most triadic setups obtained a negative r_{rm} value, except for the non-overlapping transitions in the QR and TT sets. Overall, the correlation values are much smaller than the results of the beat rates. The r_{rm} coefficients of the dyads and triads have a range of 0.01 - 0.13 and -0.09 - 0.12 respectively. However, none of the rmcorr results for syllable rates are significant. The plots depicting the described results are in the appendix.

	Dyads		
	Overlap	*Non-overlap*	*Total*
QR	$r_{rm} = 0.09,\ p = .47$	$r_{rm} = 0.11,\ p = .44$	$r_{rm} = 0.09,\ p = .29$
TT	$r_{rm} = 0.13,\ p = .46$	$r_{rm} = 0.01,\ p = .92$	$r_{rm} = 0.04,\ p = .61$
All	$r_{rm} = 0.1,\ p = .29$	$r_{rm} = 0.01,\ p = .87$	$r_{rm} = 0.04,\ p = .43$

(a) Dyads

	Triads		
	Overlap	*Non-overlap*	*Total*
QR	$r_{rm} = -0.09,\ p = .39$	$r_{rm} = 0.12,\ p = .35$	$r_{rm} = -0.01,\ p = .86$
TT	$r_{rm} = -0.08,\ p = .39$	$r_{rm} = 0.12,\ p = .35$	$r_{rm} = -0.1,\ p = .1$
All	$r_{rm} = -0.09,\ p = .18$	$r_{rm} = -0.07,\ p = .35$	$r_{rm} = -0.08,\ p = .09$

(b) Triads

Table 3.2: Results of the rmcorr statistic for syllable rates

3.2 Rhythmic alignment

3.2.1 Beat phase

There were a few rare instances in the overlapping transitions where there was only one beat of speaker 1 before the start of the utterance of speaker 2 or where the turn of speaker 1 continued during and after the turn of speaker 2. These rare situations are removed from the data.

Table 3.3 shows the results of the Rayleigh test for both dyads and triads. All phase angles of the overlapping situations in both dyads and triads lie between 53.59°and 316.15°. In the non-overlapping dyadic transitions, only the QR transition type falls outside of this range. In the triadic conversations, all sets of non-overlapping turns have a mean phase angle outside of this range. In dyads, all three sets of transition types show evidence for an unimodal phase distribution in the overlap (2) situation when applying the Rayleigh test for uniformity ($z = 0.2,\ p = .01$ for QR, $z = 0.29,\ p = .003$ for TT, $z = 0.15,\ p = .01$ for all). In line with these outputs, the resulting values when μ is equal to the mean phase angle of that set show strong significance, where μ is 46.89°($p < .001$), 338.1°($p < .001$) and 53.59°($p = .002$) for QR, TT and all respectively. The same pattern is visible in the TT set ($\mu = 14.39,\ z = 0.14,\ p = .005$) and total set ($\mu = 17.46,\ z = 0.11,\ p = .003$) of triadic conversations. The rose plots of these five experimental conditions are shown in figure 3.1. The plots of all other conditions can be found in the appendix. All mean phase angles of these five situations are within a vicinity of 75.49°.

Several other setups show significance for the Rayleigh test when μ is set to be the mean phase angle. In the dyadic group, these are the sets of QR overlap (1) ($\mu = 38.75,\ z = 0.1,\ p = .05$), TT overlap (1) ($\mu = 316.15,\ z = 0.17,\ p = .02$) and overlap (1) in all transitions combined ($\mu = 11.57,\ z = 0.09,\ p = .05$). The triadic setups are QR overlap (1) ($\mu = 325.57,\ z = 0.11,\ p = .04$), TT overlap (1) ($\mu = 8.5,\ z = 0.12,\ p = .01$), TT non-overlap ($\mu = 135.82,\ z = 0.08,\ p = .04$) and overlap (1) in all transitions combined ($\mu = 335.36,\ z = 0.08,\ p = .02$). However, in all of these situations there is no

significance for the Rayleigh test for uniformity. Not one experimental situation provides support for counterphased alignment ($\mu = 180$).

		Dyads	
		Uniformity	$z = 0.1,\ p = .27$
	Overlap (1)	$\mu = 180$	$z = -0.08,\ p = .9$
		$\mu = 38.75$	$\mathbf{z = 0.1,\ p = .05}$
		Uniformity	$\mathbf{z = 0.2,\ p = .01}$
QR	*Overlap (2)*	$\mu = 180$	$z = -0.13,\ p = .98$
		$\mu = 46.89$	$\mathbf{z = 0.2,\ p < .001}$
		Uniformity	$z = 0.09,\ p = .36$
	Non-overlap	$\mu = 180$	$z = 0.004,\ p = .48$
		$\mu = 267.55$	$z = 0.09,\ p = .08$
		Uniformity	$z = 0.17,\ p = .14$
	Overlap (1)	$\mu = 180$	$z = -0.12,\ p = .92$
		$\mu = 316.15$	$\mathbf{z = 0.17,\ p = .02}$
		Uniformity	$\mathbf{z = 0.29,\ p = .003}$
TT	*Overlap (2)*	$\mu = 180$	$z = -0.27,\ p = .99$
		$\mu = 338.1$	$\mathbf{z = 0.29,\ p < .001}$
		Uniformity	$z = 0.07,\ p = .38$
	Non-overlap	$\mu = 180$	$z = -0.07,\ p = .92$
		$\mu = 355.16$	$z = 0.07,\ p = .08$
		Uniformity	$z = 0.09,\ p = .25$
	Overlap (1)	$\mu = 180$	$z = -0.08,\ p = .95$
		$\mu = 11.57$	$\mathbf{z = 0.09,\ p = .05}$
		Uniformity	$\mathbf{z = 0.15,\ p = .01}$
All	*Overlap (2)*	$\mu = 180$	$z = -0.09,\ p = .96$
		$\mu = 53.59$	$\mathbf{z = 0.15,\ p = .002}$
		Uniformity	$z = 0.04,\ p = .52$
	Non-overlap	$\mu = 180$	$z = -0.04,\ p = .87$
		$\mu = 2.96$	$z = 0.04,\ p = .13$

(a) Dyads

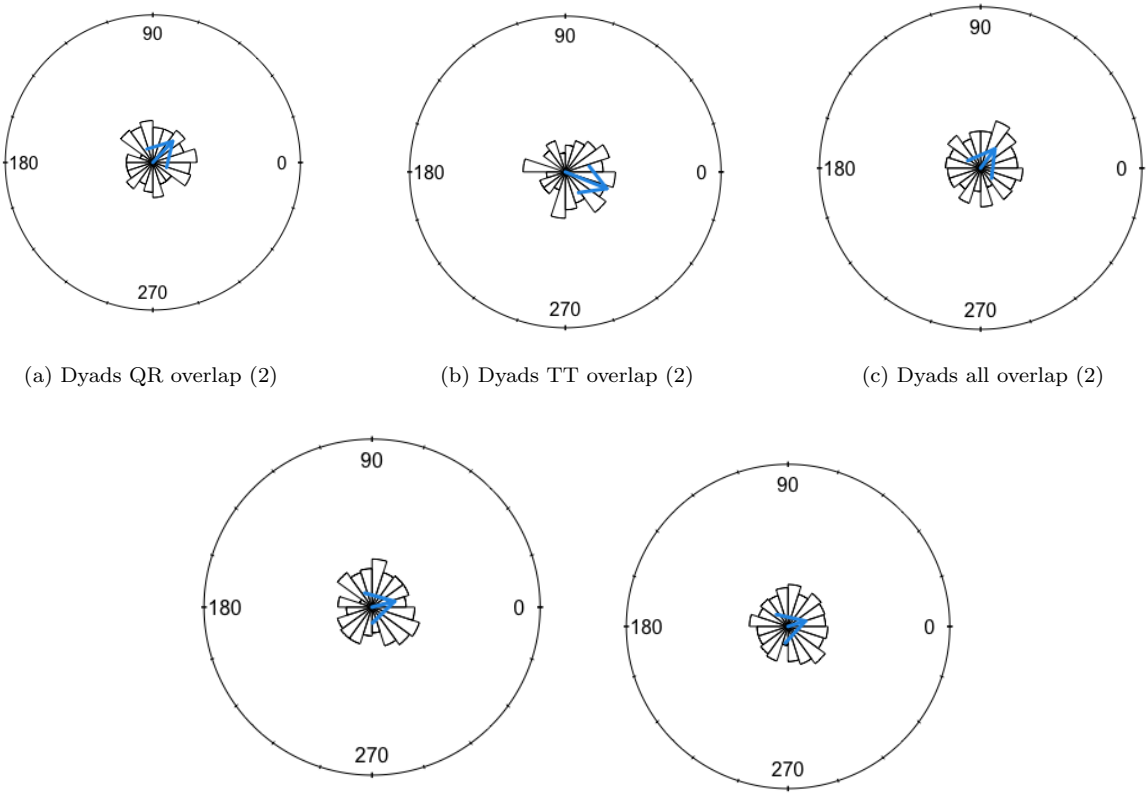
		Triads	
		Uniformity	$z = 0.11,\ p = .21$
	Overlap (1)	$\mu = 180$	$z = -0.09,\ p = .93$
		$\mu = 325.57$	$\mathbf{z = 0.11,\ p = .04}$
		Uniformity	$z = 0.13,\ p = .1$
QR	*Overlap (2)*	$\mu = 180$	$z = -0.13\ p = .98$
		$\mu = 4.48$	$\mathbf{z = 0.13,\ p = .02}$
		Uniformity	$z = 0.07,\ p = .58$
	Non-overlap	$\mu = 180$	$z = 0.04,\ p = .31$
		$\mu = 241.33$	$z = 0.07,\ p = .15$
		Uniformity	$z = 0.12,\ p = .08$
	Overlap (1)	$\mu = 180$	$z = -0.12,\ p = .99$
		$\mu = 8.5$	$\mathbf{z = 0.12,\ p = .01}$
		Uniformity	$\mathbf{z = 0.14,\ p = .04}$
TT	*Overlap (2)*	$\mu = 180$	$z = -0.14,\ p = .99$
		$\mu = 14.39$	$\mathbf{z = 0.14,\ p = .005}$
		Uniformity	$z = 0.08,\ p = .24$
	Non-overlap	$\mu = 180$	$z = 0.06,\ p = .11$
		$\mu = 135.82$	$\mathbf{z = 0.08,\ p = .04}$
		Uniformity	$z = 0.08,\ p = .14$
	Overlap (1)	$\mu = 180$	$z = -0.07,\ p = .96$
		$\mu = 335.36$	$\mathbf{z = 0.08,\ p = .02}$
		Uniformity	$\mathbf{z = 0.11,\ p = .02}$
All	*Overlap (2)*	$\mu = 180$	$z = -0.11,\ p = .99$
		$\mu = 17.46$	$\mathbf{z = 0.11,\ p = .003}$
		Uniformity	$z = 0.06,\ p = .35$
	Non-overlap	$\mu = 180$	$z = 0.05,\ p = .1$
		$\mu = 154.06$	$z = 0.06,\ p = .07$

(b) Triads

Table 3.3: Results of the Rayleigh test for beat phases

Circular ANOVA's tested distributional differences in the different experimental setups. Three conditions were examined; the overlap type (overlap (1), overlap (2), non-overlap), the transition type (QR, TT, all) and the conversation type (dyads, triads). The results are shown in table 3.4. Within dyads, no distributional difference was found between the types of overlap. Conversation type appears to be a predictor of phase within the dyadic subset of overlap (2) ($\chi^2 = 7.02,\ p = .03$). To get more insight in the nature of these distributional differences a Watson U^2 test is applied to all possible pairs of transition types in this subset. This test statistic revealed that, while QR and the combination of all transitions seem to come from the same distribution, the TT type is significantly distinct from both QR ($U^2 = 0.20,\ \textit{critical value} = .19$) and the all type ($U^2 = 0.24,\ \textit{critical value} = .19$). This can also be observed in figure 3.1, where the unit vectors of QR and all have similar directions, but the vector of the TT type has a contrasting phase angle.

The values in table 3.4 also demonstrate a distributional difference between overlap types in the triadic conversations ($\chi^2 = 6.29,\ p = .04$). Three Watson U^2 tests demonstrated that the non-overlapping subset differs from both the overlap (1) subset ($U^2 = 0.19,\ \textit{critical value} = .19$) and the overlap (2) subset ($U^2 = 0.22,\ \textit{critical value} = .19$), but both overlapping subsets seem to draw from the same distribution ($U^2 = 0.06,\ \textit{critical value} = .19$). This is also in line with the mean phase angles of these subsets displayed in table 3.3. Finally, no distributional distinctions between dyads and triads were found.



(a) Dyads QR overlap (2) (b) Dyads TT overlap (2) (c) Dyads all overlap (2)

(d) Triads TT overlap (2) (e) Triads all overlap (2)

Figure 3.1: Rose plots of the beat phases with significant result, where each bin spans 18°and one cycle of 360°corresponds with the duration of the IBI of speaker 1. The direction of each blue vector is the mean phase angle of that specific group of transitions. The length of the vector represents the strength of the relation in that direction.

Groups of interest	Subset	Degrees of freedom	Results
Overlap (1), overlap (2), non-overlap	Dyads all	2	$\chi^2 = 1.47,\ p = .48$
Overlap (1), overlap (2), non-overlap	Triads all	2	$\boldsymbol{\chi^2 = 6.29,\ p = .04}$
QR, TT, all	Dyads overlap (1)	2	$\chi^2 = 2.25,\ p = .32$
QR, TT, all	Dyads overlap (2)	2	$\boldsymbol{\chi^2 = 7.02,\ p = .03}$
QR, TT, all	Dyads non-overlap	2	$\chi^2 = 1.79,\ p = .41$
QR, TT, all	Triads overlap (1)	2	$\chi^2 = 1.07,\ p = .59$
QR, TT, all	Triads overlap (2)	2	$\chi^2 = 0.15,\ p = .93$
QR, TT, all	Triads non-overlap	2	$\chi^2 = 1.74,\ p = .42$
Dyads, triads	All	1	$\chi^2 = 0.21,\ p = .64$

Table 3.4: Circular ANOVA results of beat phases

3.2.2 Syllable phase

Similar to the beat phases, instances in the overlapping transitions where there was only one syllable of speaker 1 before the start of the utterance of speaker 2 or where the turn of speaker 1 continued during and after the turn of speaker 2 are removed from the data.

The results of the Rayleigh tests for syllable phase in dyads and triads are shown in table 3.5. No experimental setups in the dyadic turn transitions display an unimodal distribution when applying the Rayleigh test for uniformity. Only the overlap (2) subset in the TT transition type shows significance when μ is equal to the mean phase angle of that set ($\mu = 85.99,\ z = 0.23,\ p = .01$).

Within the triads, only the phase distribution of overlap (2) in the TT group appears to be unimodal (Rayleigh test with unspecified mean direction; $z = 0.15,\ p = .03$). Rayleigh test with μ set to be the mean angular phase of 44.86°confirmed this relation ($z = 0.15,\ p = .004$). The corresponding rose plot can be viewed in figure 3.2. Two other triadic conditions showed significant results for the Rayleigh test with μ set to the mean phase angle of that set; overlap (1) in the TT transition type ($\mu = 48.56,\ z = 0.14,\ p = .01$) and overlap (2) in the combination of all transitions ($\mu = 44.71,\ z = 0.09,\ p = .02$).

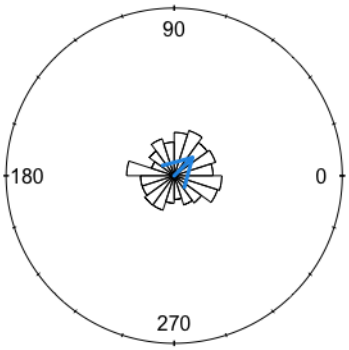


Figure 3.2: Rose plot of the triadic TT overlap (2) condition for syllables

The results of circular ANOVA's for syllable phases are shown in table 3.6. The same conditions as for the beat phases are examined. No distributional differences were found between overlap types in both dyads and triads. Moreover, all transition types seem to draw from similar distributions. Finally, distributions of dyads and triads are not significantly distinct. No further analysis with a Watson U^2 test was needed.

		Dyads	
		Uniformity	$z = 0.08,\ p = .53$
	Overlap (1)	$\mu = 180$	$z = -0.007,\ p = .54$
		$\mu = 274.83$	$z = 0.08,\ p = .13$
		Uniformity	$z = 0.05,\ p = .81$
QR	*Overlap (2)*	$\mu = 180$	$z = -0.01,\ p = .57$
		$\mu = 284.99$	$z = 0.05,\ p = .26$
		Uniformity	$z = 0.1,\ p = .42$
	Non-overlap	$\mu = 180$	$z = 0.1,\ p = .09$
		$\mu = 177.64$	$z = 0.1,\ p = .09$
		Uniformity	$z = 0.06,\ p = .84$
	Overlap (1)	$\mu = 180$	$z = -0.006,\ p = .52$
		$\mu = 84.57$	$z = 0.06,\ p = .28$
		Uniformity	$z = 0.23,\ p = .07$
TT	*Overlap (2)*	$\mu = 180$	$z = -0.02,\ p = .56$
		$\mu = 85.99$	$\boldsymbol{z = 0.23,\ p = .01}$
		Uniformity	$z = 0.04,\ p = .82$
	Non-overlap	$\mu = 180$	$z = -0.03,\ p = .7$
		$\mu = 33.77$	$z = 0.04,\ p = .26$
		Uniformity	$z = 0.03,\ p = .84$
	Overlap (1)	$\mu = 180$	$z = -0.006,\ p = .54$
		$\mu = 280.66$	$z = 0.03,\ p = .28$
		Uniformity	$z = 0.05,\ p = .69$
All	*Overlap (2)*	$\mu = 180$	$z = -0.01,\ p = .59$
		$\mu = 74.51$	$z = 0.05,\ p = .19$
		Uniformity	$z = 0.02,\ p = .87$
	Non-overlap	$\mu = 180$	$z = 0.02,\ p = .34$
		$\mu = 143.48$	$z = 0.02,\ p = .3$

(a) Dyads

		Triads	
		Uniformity	$z = 0.08,\ p = .38$
	Overlap (1)	$\mu = 180$	$z = 0.07,\ p = .14$
		$\mu = 218.06$	$z = 0.08,\ p = .08$
		Uniformity	$z = 0.005,\ p = .99$
QR	*Overlap (2)*	$\mu = 180$	$z = -0.004,\ p = .52$
		$\mu = 38.48$	$z = 0.005,\ p = .47$
		Uniformity	$z = 0.005,\ p = .99$
	Non-overlap	$\mu = 180$	$z = -0.003,\ p = .51$
		$\mu = 61.3$	$z = 0.005,\ p = .47$
		Uniformity	$z = 0.14,\ p = .06$
	Overlap (1)	$\mu = 180$	$z = -0.09,\ p = .94$
		$\mu = 48.56$	$\boldsymbol{z = 0.14,\ p = .01}$
		Uniformity	$\boldsymbol{z = 0.15,\ p = .03}$
TT	*Overlap (2)*	$\mu = 180$	$z = -0.11,\ p = .97$
		$\mu = 44.86$	$\boldsymbol{z = 0.15,\ p = .004}$
		Uniformity	$z = 0.02,\ p = .95$
	Non-overlap	$\mu = 180$	$z = -0.005,\ p = .54$
		$\mu = 73.25$	$z = 0.02,\ p = .37$
		Uniformity	$z = 0.04,\ p = .68$
	Overlap (1)	$\mu = 180$	$z = -0.02,\ p = .67$
		$\mu = 59.61$	$z = 0.04,\ p = .19$
		Uniformity	$z = 0.09,\ p = .13$
All	*Overlap (2)*	$\mu = 180$	$z = -0.06,\ p = .92$
		$\mu = 44.71$	$\boldsymbol{z = 0.09,\ p = .02}$
		Uniformity	$z = 0.01,\ p = .95$
	Non-overlap	$\mu = 180$	$z = -0.004,\ p = .54$
		$\mu = 71.48$	$z = 0.01,\ p = .38$

(b) Triads

Table 3.5: Results of the Rayleigh test for syllable phases

Groups of interest	Subset	Degrees of freedom	Results
Overlap (1), overlap (2), non-overlap	Dyads all	2	$\chi^2 = 0.54,\ p = .76$
Overlap (1), overlap (2), non-overlap	Triads all	2	$\chi^2 = 0.09,\ p = .96$
QR, TT, all	Dyads overlap (1)	2	$\chi^2 = 0.79,\ p = .67$
QR, TT, all	Dyads overlap (2)	2	$\chi^2 = 1.76,\ p = .42$
QR, TT, all	Dyads non-overlap	2	$\chi^2 = 0.001,\ p = .99$
QR, TT, all	Triads overlap (1)	2	$\chi^2 = 3.18,\ p = .2$
QR, TT, all	Triads overlap (2)	2	$\chi^2 = 0.001,\ p = .99$
QR, TT, all	Triads non-overlap	2	$\chi^2 = 1.74,\ p = .42$
Dyads, triads	All	1	$\chi^2 = 0.27,\ p = .61$

Table 3.6: Circular ANOVA results of syllable rates

Chapter 4

Discussion

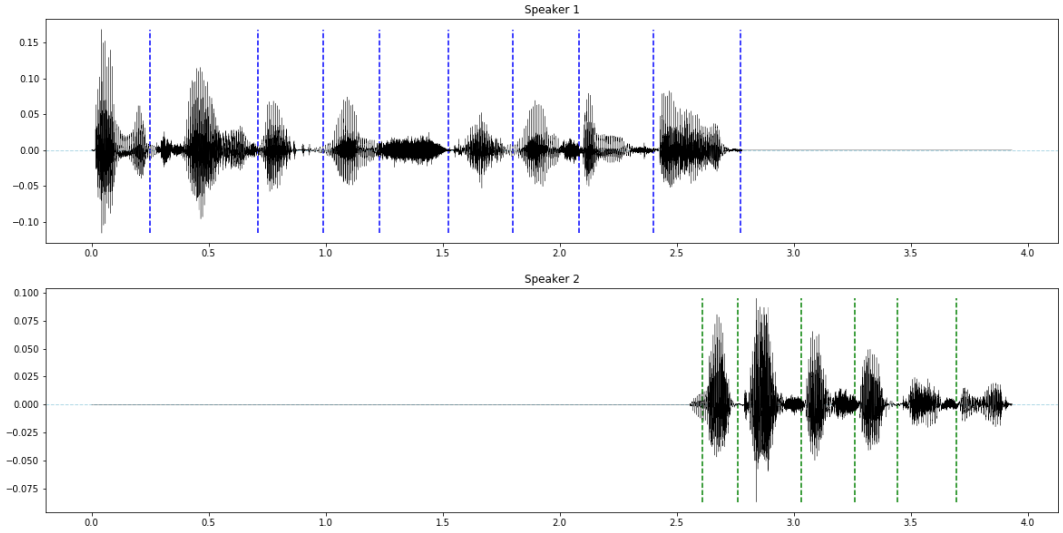
This research explored the presence of temporal entrainment between interlocutors to facilitate turn-taking. Both beats and syllables were considered as the rhythmic unit of entrainment. Two examples of extracted beats and syllables are shown in figures 4.1 (overlapping turn) and 4.2 (non-overlapping turn). In the top panel of figure 4.1, depicting beats in an overlapping turn, the first beat of speaker 2 corresponding to overlap (1) is roughly placed in between the two beats of speaker 1, indicating counterphased alignment, which is almost opposite of the mean phase angle in this subset of transitions (all dyads overlap (1), $\mu = 11.57$). However, the strength of the relation in this direction, although significant, was rather small ($z = 0.09,\ p = .05$). The beats of interest in overlap (2), the final beat of speaker 1 and the second beat of speaker 2, are almost perfectly aligned. This corresponds with the findings for overlap (2) as depicted in the rose plots of figure 3.1, where all vectors are directed towards a location approximately near 0°. For the extracted syllables in the bottom panel of figure 4.1 overlap (1) and overlap (2) are identical. Here, the final syllable of speaker 1 seems to be in time with the first syllable of speaker 2, as opposed to the mean phase angles of the corresponding subsets overlap (1) ($\mu = 280.66$) and overlap (2) ($\mu = 74.51$) in all dyadic transitions.

In the non-overlapping turn of figure 4.2, there is only a small gap between the two beats of interest in the top panel. This slightly deviates from the mean phase angle of the non-overlapping turns in all dyadic transitions ($\mu = 2.96$), although the strength of the relation in this angular direction was small and non-significant ($z = 0.04,\ p = .13$). In the bottom panel, if we would extrapolate the syllables of speaker 1 based on the final inter-syllable interval, the first syllable of speaker 2 falls within two extrapolated syllables of speaker 1, which again would suggest counterphased alignment. Here, this does comply with the mean angular phase of the non-overlapping turns in all dyadic transitions $\mu = 143.48$, though it should be noted that the strength if this relation was small and non-significant ($z = 0.02,\ p = .3$).

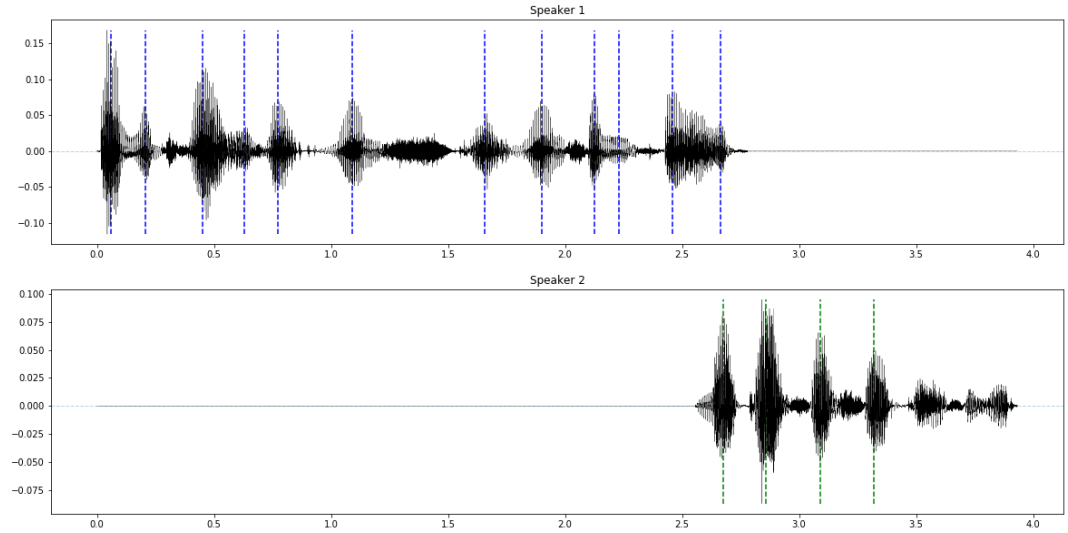
The two figures provide a good visualisation of the different approaches. The extracted beats are more evenly spaced than the syllables, due to the penalty for beat candidates that deviate too much from the ideal IBI. However, the beats are not completely isochronous, since we implemented a low value of α. Since such a penalty is not present in the syllable rate approach, it is possible to have a larger gap between syllables that spans a stretch of unvoiced speech, such as the gap between roughly 1.1 s and 1.6 s in the syllable panel of figure 4.1. Next to this interval spacing, a difference can be observed in the location of the beats and syllables within the waveform. All syllables are located within a chunk of voiced speech, where the amplitude is largest. In contrast, the beats are all located right before a piece of voiced speech where the amplitude is lowest, indicating an onset. These observations are consistent with the implementations of the two approaches, as described in chapter 2.

4.1 Speech rate entrainment

The results presented in chapter 3 are in contrast with the Wilson and Wilson (2005) model for temporal entrainment. One implication of this model is that participants in a conversation will have equal speech rates. Perfect entrainment would be depicted by a repeated measures correlation coefficient of 1. Let us first consider the syllable rates. No indication of speech rate entrainment is present for this rhythmic unit, as there were no significant correlations. Although the theoretical hypotheses regarding syllables as the unit of entrainment that are stated in the research of Wilson and Wilson (2005) seem to be well founded, all experimental results, including those presented in this paper, failed to confirm them (Beňuš, 2009; Inden et al., 2012; O'Dell et al., 2012; Schultz et al., 2016). It should be noted that backchannels



(a) Beats

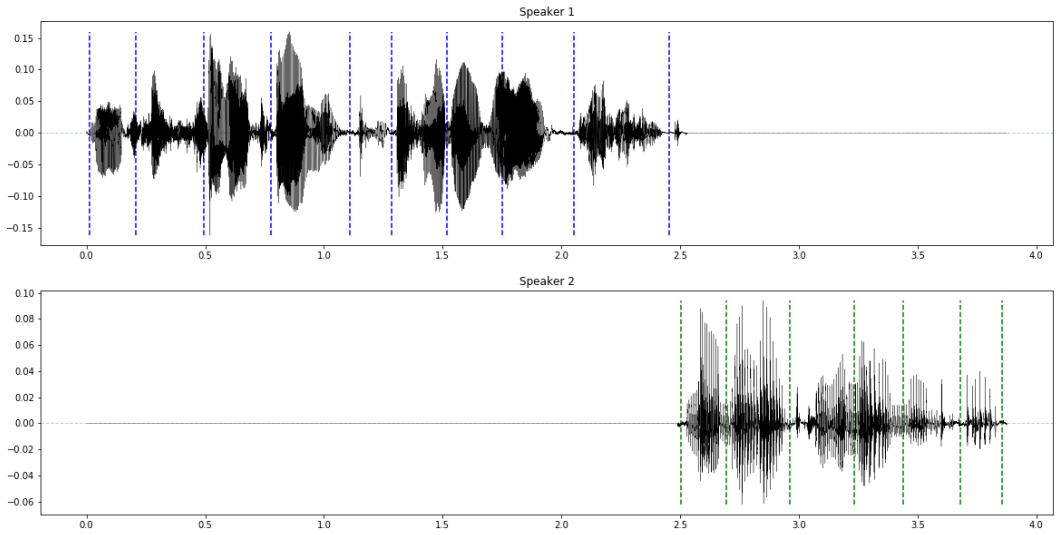


(b) Syllables

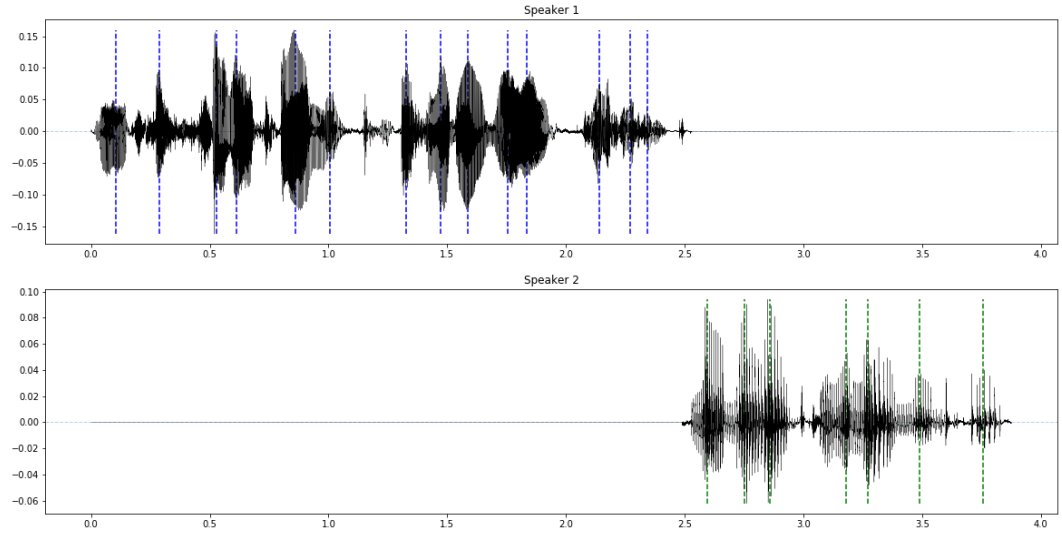
Figure 4.1: Visualization of extracted beats and syllables in an overlapping turn

were not included in the data of the present research. Beňuš (2009) did find patterns of syllable rate correlation in transitions that were associated with backchanneling, which leaves open the possibility that syllables may act as the rhythmic units of entrainment in this niche of turn transitions.

A rather counter-intuitive effect can be observed in the results regarding beat rate entrainment. There is a relation present between beat rates of speaker 1 and speaker 2, only not in the hypothesized direction. All correlation coefficients appear to be negative. A perfect negative correlation of -1 would indicate that as the beat rate of speaker 1 increases, the rate of speaker 2 decreases, and vice versa. For triads the strength of this relation ranges from low to moderate (r_{rm} values are between -0.15 and -0.39) and in dyads all experimental setups depict a high degree of correlation (r_{rm} values are between -0.52 and -0.63). Figure 4.3 shows the rmcorr plot of all dyadic turn transitions. As mentioned before, the rmcorr coefficient, depicted by the slopes of the regression lines, is an overarching value and is equal for all conversations. When disregarding the regression lines and only looking at the scatter plot, the negative relation is still clearly visible in many of the individual conversations, depicted by the different colors. An inspection of the beat rate values per turn revealed that of the two participants in a dyad, one speaks consistently faster than the other. Due to alternating turns, this results in a pattern where, if speaker 1 talks at a lower pace, speaker 2 will talk with a faster tempo, and vice versa. The turn transition corresponding to the highest green dot in figure 4.3 is shown in figure 4.4. Here the interval



(a) Beats



(b) Syllables

Figure 4.2: Visualization of extracted beats and syllables in a non-overlapping turn

between the beats of speaker 2 almost spans two IBI's of speaker 1 and the size of the IBI of speaker 2 is rather large compared to other IBI's in the dataset, but there are clearly no other onsets present in the audio signal. This is one extreme example of different beat rates between the participants. The rmcorr plot of figure 4.3 also shows that most turns have a beat rate value between 0.2 and 0.4. Though there is still a faster and a slower participant, their tempi are generally much closer than in the presented example.

Interestingly, the same pattern of faster and slower participants are visible in the corresponding triadic conversations. As stated in chapter 2, the triadic conversations between speakers A, B and C were recorded first, after which one participants was asked to leave and the two remaining speakers participated in the dyadic conversation. The data show that speaker B in the dyad corresponding to the example in figure 4.4 has a slower beat rate than speaker A. The same is true for the triadic conversation that these two interlocutors are involved in; speaker B overall has a larger mean IBI than speakers A and C and therefore a slower beat rate. When comparing speakers A and C, a distinction can again be made between a faster participant (speaker C) and a slower participant (speaker A), even though the beat rates are very close. This could also be a possible explanation for the less strong correlations in the triadic experimental setups. If the participants each talk at a different rate, either fast, medium or slow, then the overall difference between IBI's in triads will be smaller. In the situations where either

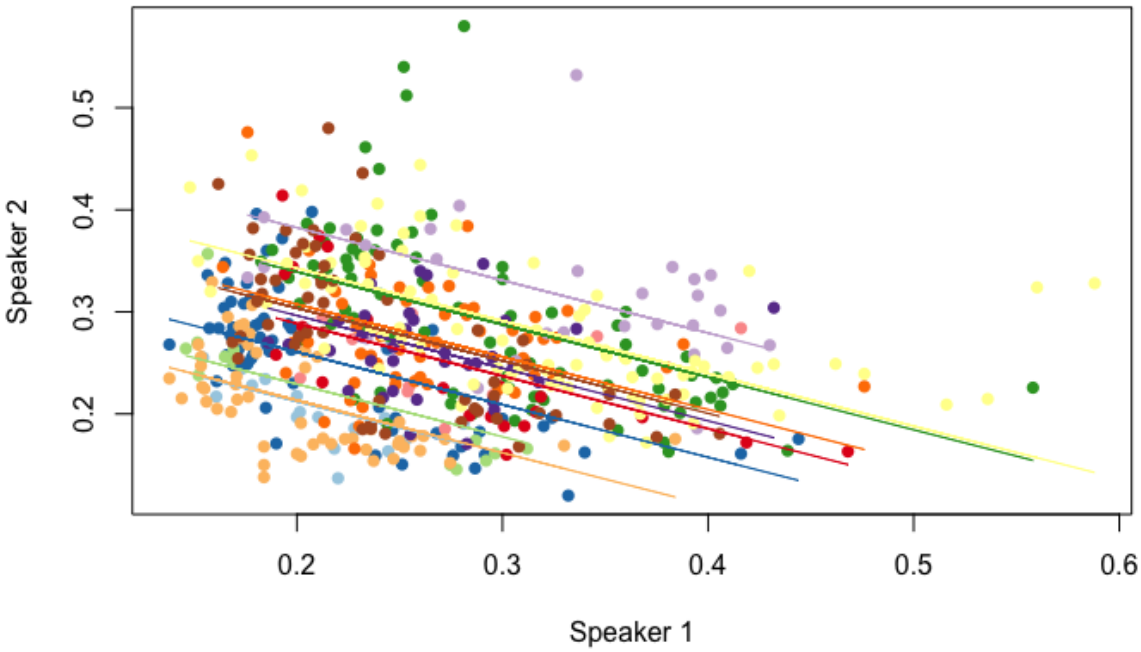


Figure 4.3: Rmcorr plot of all turn transitions of dyads

the slow participant or the fast participant was asked to leave the room after the triadic conversation, the dyadic counterpart is likely to show smaller differences in the beat rates of turn pairs. This is true, for example, for the conversation that matches to the light blue dots in figure 4.3. The participant who had the slowest beat rate in the corresponding triad was not part of this dyadic conversation.

4.2 Rhythmic alignment

The second implication of the model proposed by Wilson and Wilson (2005) that was tested in this paper, is that of counterphased entrainment. Again, no confirmation of syllables as the rhythmical units that facilitate entrainment could be provided. Only a few specific situations in the overlapping transitions showed a slight relation between the syllable locations of the two participants in a turn transition. However, none of these relations pointed towards 180°, showing no evidence of counterphased alignment. This is in agreement with the findings of Beňuš (2009), who inspected a relation between latency and speech rate and found no significant correlation.

The results of Beňuš (2009) did present a correlation between these two arguments when placed in the context of beats, where the overlapping subset of transitions exhibited one of the strongest correlations. However, no specific direction of the relation was stated. Similarly, a relation in the overlapping transitions for beats is also visible in the results presented in chapter 3, especially in the situation of overlap (2). The different experimental setups all show a significant relation for overlap (2) with mean angular phases ranging between 338.1°(or -21.9°) and 53.59°. With a mean IBI of 0.3 s this would entail that the beat of interest of speaker 2 is placed somewhere between 18.25 ms before or 45 ms after the final beat, or extrapolated beat, of speaker 1. In most conditions within the overlap (2) situation the beat of speaker 2 is placed slightly after the beat of speaker 1, as depicted in the rose plots of figure 3.1. Only in the dyadic TT transition type (figure 3.1 (b)) is a clear pattern visible of placing the beat in the second turn right before that of the first turn ($\mu = 338.1$). This pattern is shown to be significantly distinct from the other two sets of transitions in the overlap (2) situation for dyads, as analysed by the circular ANOVA ($\chi^2 = 7.02,\ p = .03$). As the speaker shift in this transition type is executed through self-selection, this could possibly be a way of signaling that speaker 2 intends to take the floor, before speaker 1 can continue, conform rules b) and c) of the rule set for speaker appointment. It would then be expected that this system is also present in triads, maybe even to a larger extent since there are two parties who can invoke this rule and try to take the turn. However, there is no indication of such a

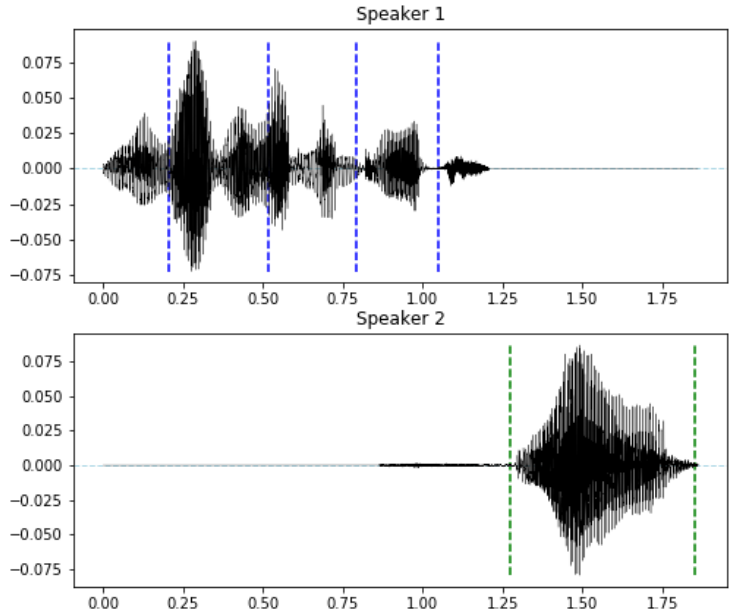


Figure 4.4: Example of transition with a large difference in beat rate between participants

pattern in triads. In fact, the triadic TT transitions in the overlap (2) situations have a mean phase angle of 14.39°, placing the beat of speaker 2 right after the beat of speaker 1.

One could argue that there might be some form of mutual entrainment during the moment of simultaneous speech, where both parties affect each other. This would then result in a stronger relation between the beat placement of the two participants. Possibly, equal speech rhythms are not important as long as there is a point where the beats align so that the turn can be taken over in a very smooth fashion, somewhat like passing the baton in a relay race. This would provide an explanation of the found relation in all types of overlap (2). These results provide interesting insights in to the mechanisms of turn-taking in overlapping turns. Still, not many conclusions about non-overlapping turns can be inferred.

The circular ANOVA revealed a significant distributional difference between overlap types in the total set of triadic transitions ($\chi^2 = 6.29,\ p = .04$). It was shown that this result was due to the distribution of the non-overlapping turns being substantially different from those of the overlap (1) and overlap (2) situations. This is also visible in the mean phase angles of these groups, where the non-overlapping transitions have a μ of 154.06°and overlap (1) and overlap (2) have mean angular phases of 335.36°and 17.46°respectively. The latter two phase angles would roughly correspond to the beats of the two speakers being "in time", while the phase angle of the overlapping turns is closer to counterphased alignment. This could be seen as a subtle indication of the hypothesized counterphased entrainment within this specific set of transitions, be it not that the results for the non-overlapping turns were non-significant, while those of the other two overlap types did show significance. This distributional distinction between overlap and non-overlap was not present in the dyadic conversations ($\chi^2 = 1.47,\ p = .48$). Furthermore, the presented results showed no distributional difference between dyads and triads ($\chi^2 = 0.21,\ p = .64$). This shows that, even though there are more participants that can compete for the floor, this increase in competition does not have a substantial effect on the timing of turn-taking.

Overall, no evidence was found that could confirm the claims of correlated speech rates or counterphased entrainment. No support for the theoretical model of Wilson and Wilson (2005) could be provided. Still it remains true that turn-taking requires an incredible amount of precision. Moreover, oscillators seem a plausible vessel for facilitating this timing, since it is known to be involved in both the productional and perceptual systems of speech (Poeppel & Assaneo, 2020). It might be possible that oscillators are

involved in this mechanism, only not in the hypothesised fashion. So far, speech rhythm entrainment has only been explored on the syllable level and the stress level. It has been proposed that speech rhythm adheres to a system of hierarchical levels (Martinec, 2002). In the lower levels, turn transitions are considered breaks in the isochronous sequence and a speech rhythm is not able to extend over these gaps. However, rhythmicity might be transmitted across turns in a higher hierarchical level. Given the results of this paper and that of previous research, is it likely that the syllable level is too low in the hierarchy and a turn transition breaks the established rhythm. Beats can be considered one hierarchical level higher than syllables and this level already shows more signs of a rhythmic relation between turns. Possibly, there is an even stronger practice of entrainment more levels up in the hierarchy.

The presented research aimed to take a first step at broadening the scope of the temporal foundations of turn-taking from dialogues to multiparty conversations. Most of the literature that has researched temporal entrainment to facilitate turn-taking has focused on two-party conversations (Beňuš, 2009; Inden et al., 2012; O'Dell et al., 2012; Schultz et al., 2016). Moreover, recent research on turn-taking in multiparty conversations is mostly directed towards the domain of artificial dialogue systems (Hara et al., 2019; Żarkowski, 2019). Within this domain, the aim is not to understand the foundations of this mechanism, but to replicate its functioning. To fully grasp this concept in the context of multiparty communication, future work will need to address conversations between more than three parties.

Possibly the largest drawback of using a beat tracking algorithm based on dynamic programming, is the need for regular intervals between beats. Even though employing a small value for the tightness parameter α allows the algorithm to deviate from a strict isochronous sequence, it will never be possible to have large gaps in the sequence, even though there might be a long stretch without onsets in the audio signal. Developing an algorithm based on a transformer architecture would be interesting to see in future work. The transformer is a rather new development in the field of Artificial Intelligence and is rapidly changing the field of Natural Language Processing (Vaswani et al., 2017). Given its great results, this technique is now starting to be adopted in other domains, including music generation (C.-Z. A. Huang et al., 2018; Y.-S. Huang & Yang, 2020). If such a model is able to develop an understanding of musical features to create music, it is an interesting thought to turn this process around and use these models to extract the features from the music. There has even been a development of a transformer dedicated to the prediction of turn-taking (Ekstedt & Skantze, 2020). Considering the rapid evolution of transformer models, the field of psycholinguistics is sure to benefit from these new developments in the following years.

Bibliography

Agostinelli, C., & Lund, U. (2017). *R package `circular`: Circular statistics (version 0.4-93)*. Retrieved June 2, 2021, from https://r-forge.r-project.org/projects/circular/

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in psychology*, *8*, 456.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Beňuš, Š. (2009). Are we in sync': Turn-taking in collaborative dialogues. *Tenth Annual Conference of the International Speech Communication Association*.

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]. Retrieved May 28, 2021, from http://www.praat.org/

Cook, P., Rouse, A., Wilson, M., & Reichmuth, C. (2013). A california sea lion (zalophus californianus) can keep the beat: Motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *Journal of Comparative Psychology*, *127*(4), 412.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2), 385–390.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, *23*(2), 283.

Ekstedt, E., & Skantze, G. (2020). Turngpt: A transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874*.

Ellis, D. P. (2007). Beat tracking by dynamic programming. *Journal of New Music Research*, *36*(1), 51–60.

Giannakopoulos, T. (2021). Silence removal in speech signals. *MATLAB Central File Exchange*. Retrieved June 1, 2021, from https://nl.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601–634.

Hara, K., Inoue, K., Takanashi, K., & Kawahara, T. (2019). Turn-taking prediction based on detection of transition relevance place. *INTERSPEECH*, 4170–4174.

Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency. *Frontiers in psychology*, *6*(98), 1–14.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer. *arXiv preprint arXiv:1809.04281*.

Huang, Y.-S., & Yang, Y.-H. (2020). Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. *Proceedings of the 28th ACM International Conference on Multimedia*, 1180–1188.

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34).

Lameira, A. R., Eerola, T., & Ravignani, A. (2019). Coupled whole-body rhythmic entrainment between two chimpanzees. *Scientific reports*, *9*(1), 1–8.

Large, E. W., & Gray, P. M. (2015). Spontaneous tempo and rhythmic entrainment in a bonobo (pan paniscus). *Journal of comparative psychology*, *129*(4), 317.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, *6*, 731.

Martinec, R. (2002). Rhythmic hierarchy in monologue and dialogue. *Functions of language*, *9*(1), 39–59.

Merker, B. H., Madison, G. S., & Eckerdal, P. (2009). On the role and origin of isochrony in human rhythmic entrainment. *Cortex*, *45*(1), 4–17.

O'Dell, M. L., Nieminen, T., & Lennes, M. (2012). Modeling turn-taking rhythms with oscillators. *Linguistica Uralica*, *48*(3), 218–227.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature reviews neuroscience*, *21*(6), 322–334.

Ravignani, A. (2018). Timing of antisynchronous calling: A case study in a harbor seal pup (phoca vitulina). *Journal of Comparative Psychology*, *133*(2), 272.

Rochet-Capellan, A., & Fuchs, S. (2014). Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130399.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Studies in the organization of conversational interaction* (pp. 7–55). Elsevier.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wilson, & Cook, P. F. (2016). Rhythmic entrainment: Why humans want to, fireflies can't help it, pet birds try, and sea lions have to be bribed. *Psychonomic bulletin & review*, *23*(6), 1647–1659.

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

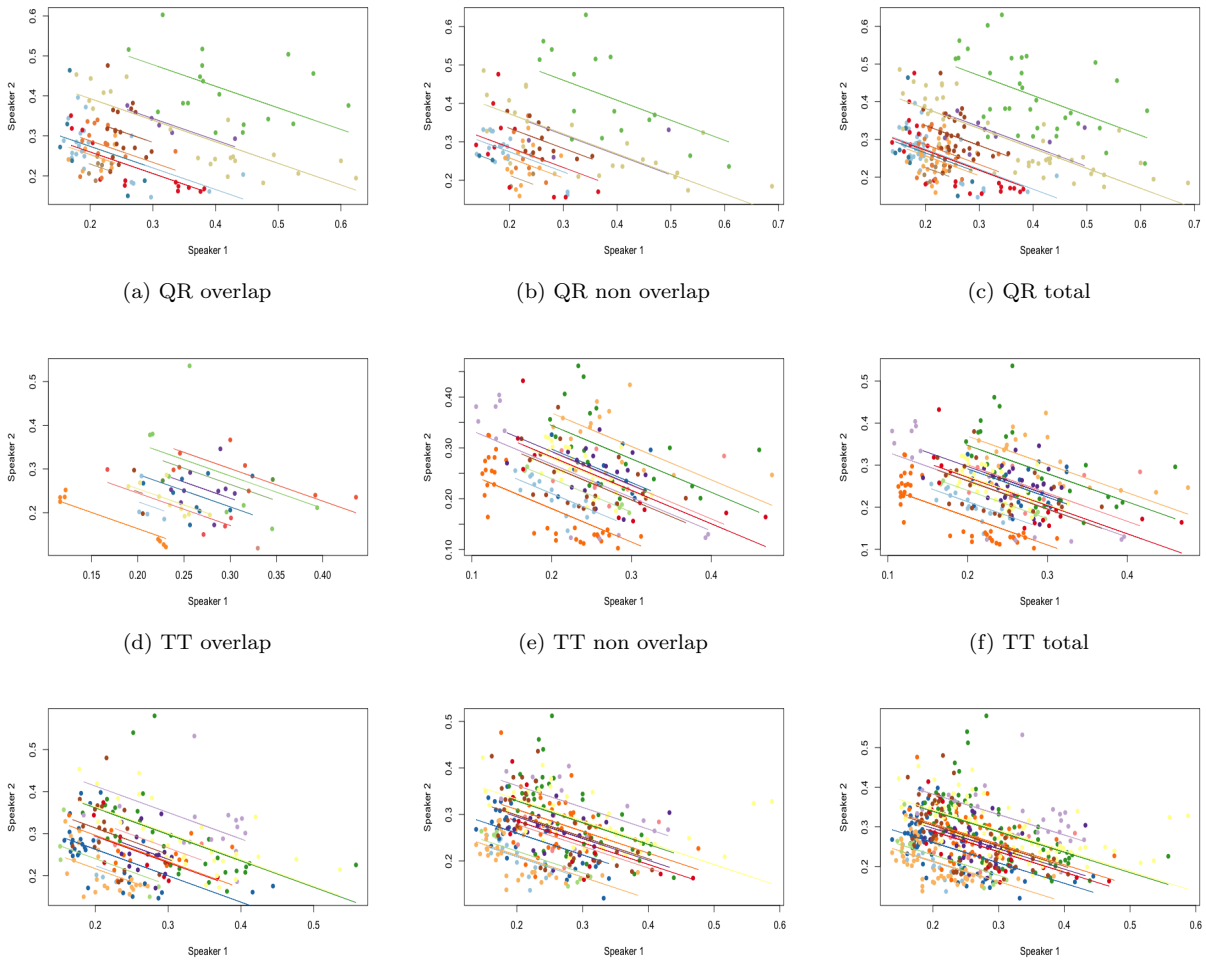
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Żarkowski, M. (2019). Multi-party turn-taking in repeated human–robot interactions: An interdisciplinary evaluation. *International Journal of Social Robotics*, *11*(5), 693–707.

Appendices

Appendix A

Repeated measures correlation plots

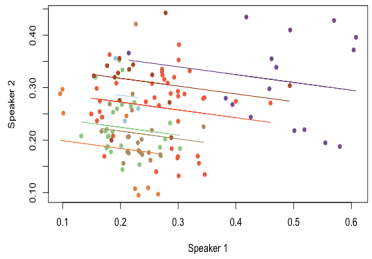


(a) QR overlap (b) QR non overlap (c) QR total

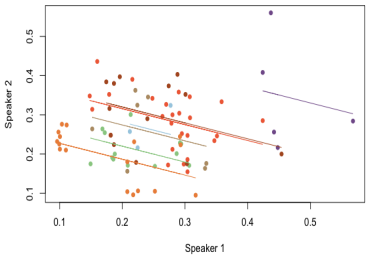
(d) TT overlap (e) TT non overlap (f) TT total

(g) All overlap (h) All non overlap (i) All total

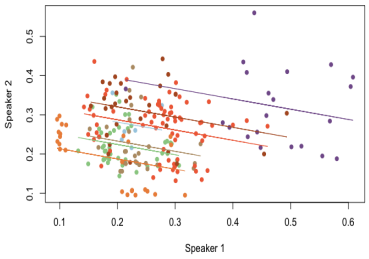
Figure A.1: Rmcorr plots of the beat rates of dyads



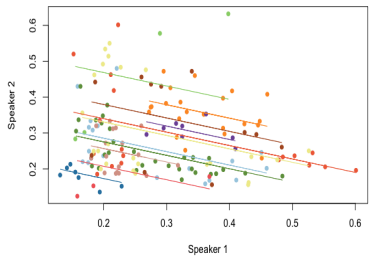
(a) QR overlap



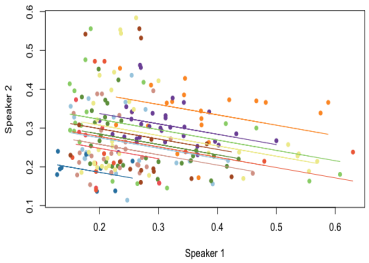
(b) QR non overlap



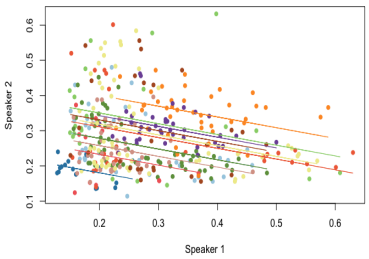
(c) QR total



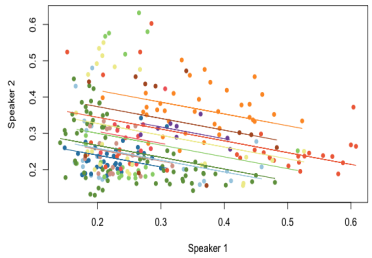
(d) TT overlap



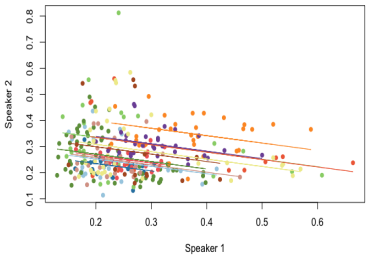
(e) TT non overlap



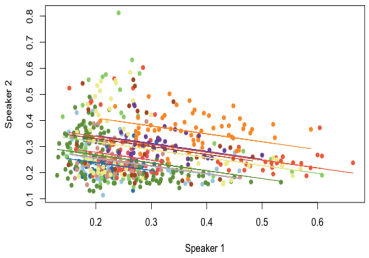
(f) TT total



(g) All overlap

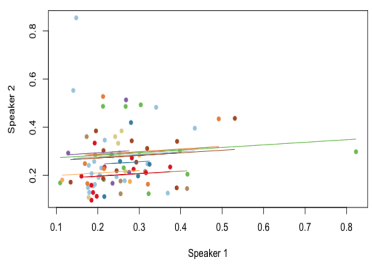


(h) All non overlap

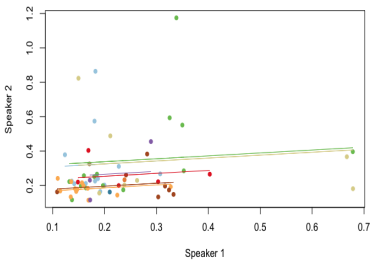


(i) All total

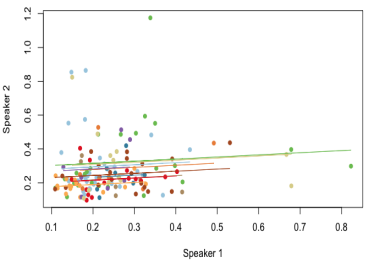
Figure A.2: Rmcorr plots of the beat rates of triads



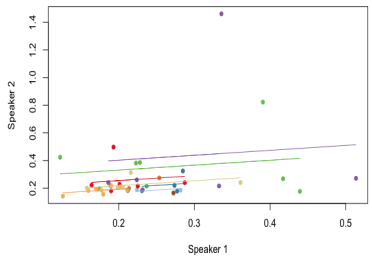
(a) QR overlap



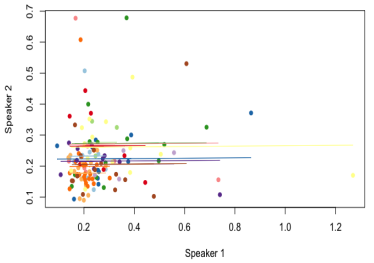
(b) QR non overlap



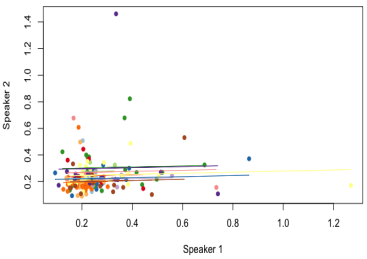
(c) QR total



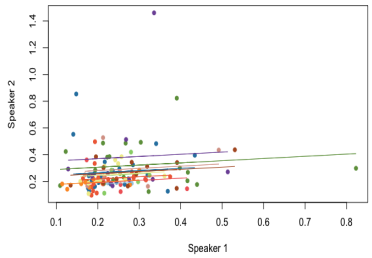
(d) TT overlap



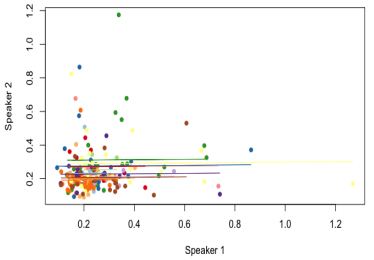
(e) TT non overlap



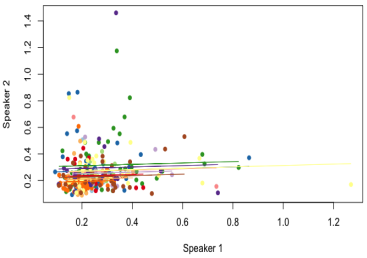
(f) TT total



(g) All overlap

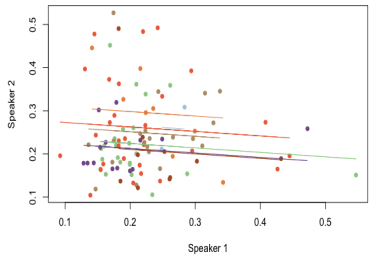


(h) All non overlap

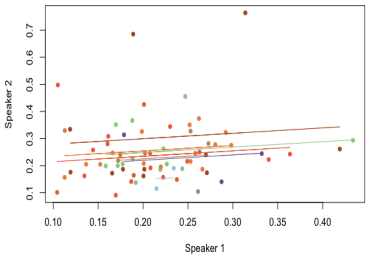


(i) All total

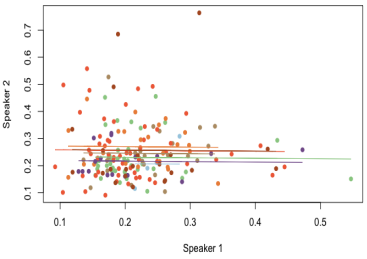
Figure A.3: Rmcorr plots of the syllable rates of dyads



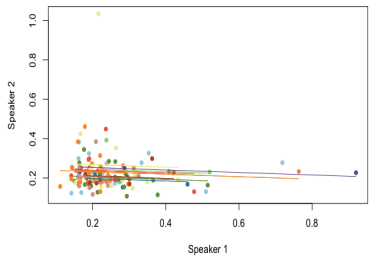
(a) QR overlap



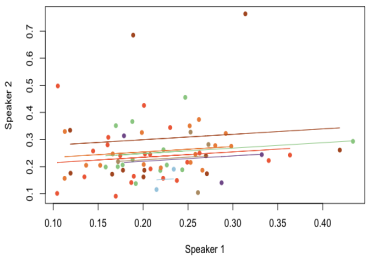
(b) QR non overlap



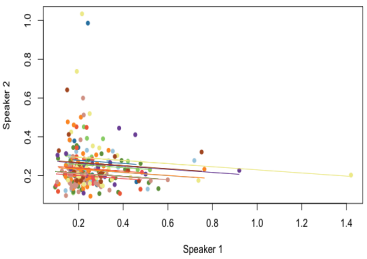
(c) QR total



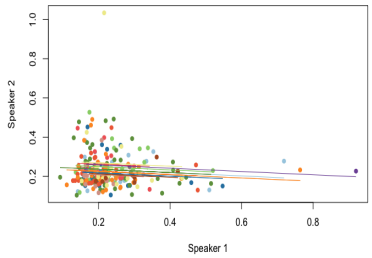
(d) TT overlap



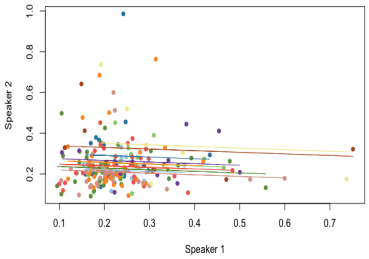
(e) TT non overlap



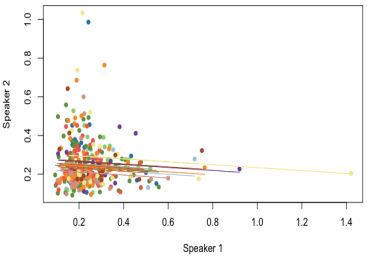
(f) TT total



(g) All overlap



(h) All non overlap

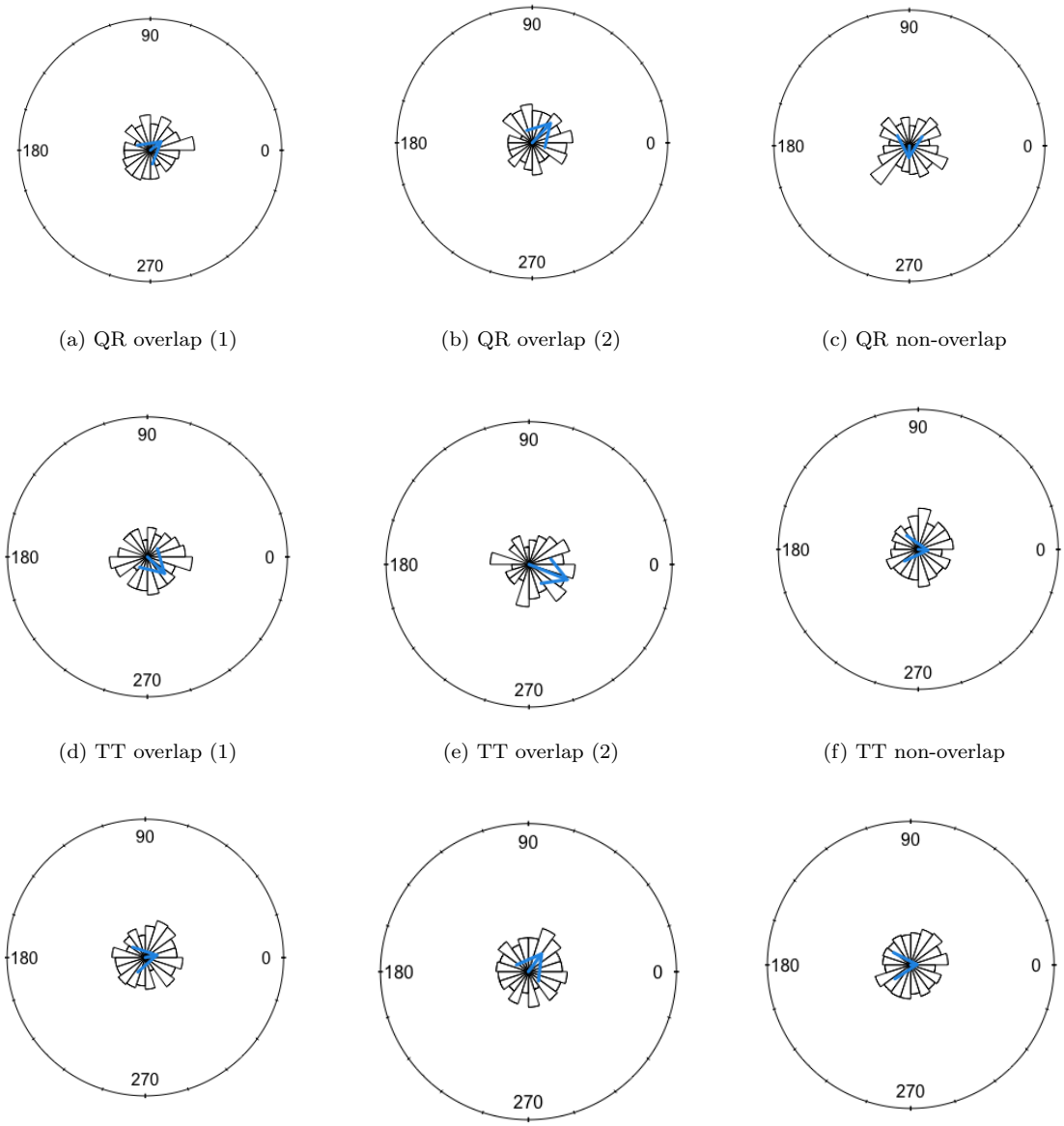


(i) All total

Figure A.4: Rmcorr plots of the syllable rates of triads

Appendix B

Rose plots of circular statistics

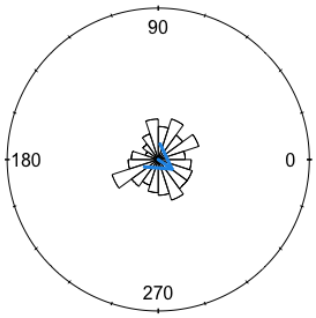


(a) QR overlap (1) (b) QR overlap (2) (c) QR non-overlap

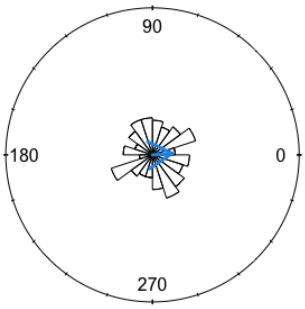
(d) TT overlap (1) (e) TT overlap (2) (f) TT non-overlap

(g) All overlap (1) (h) All overlap (2) (i) All non-overlap

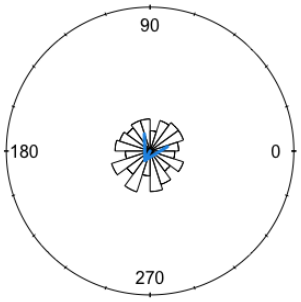
Figure B.1: Rose plots of the beat rates of dyads



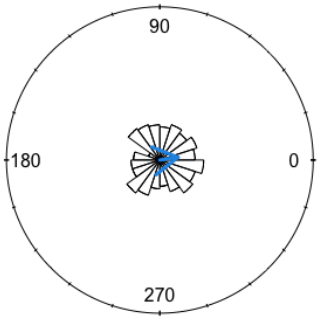
(a) QR overlap (1)



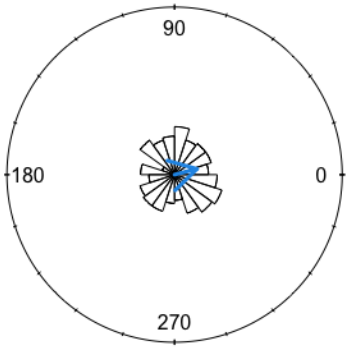
(b) QR overlap (2)



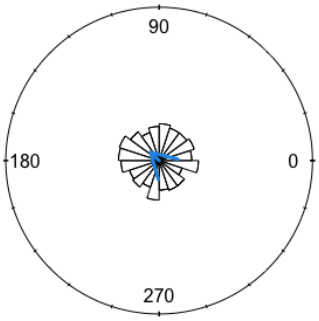
(c) QR non-overlap



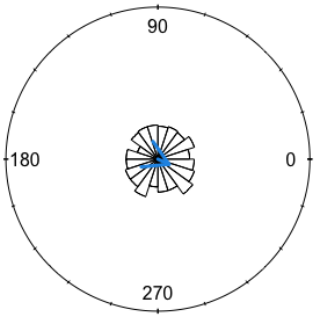
(d) TT overlap (1)



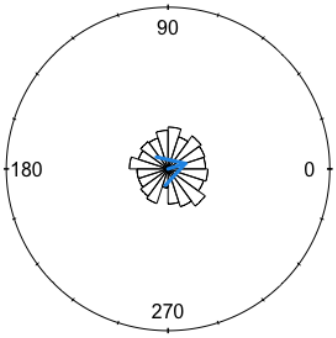
(e) TT overlap (2)



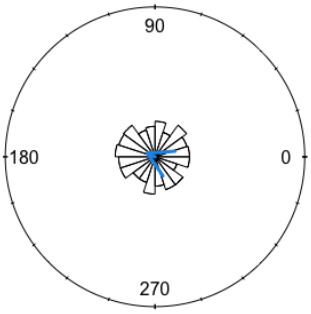
(f) TT non-overlap



(g) All overlap (1)

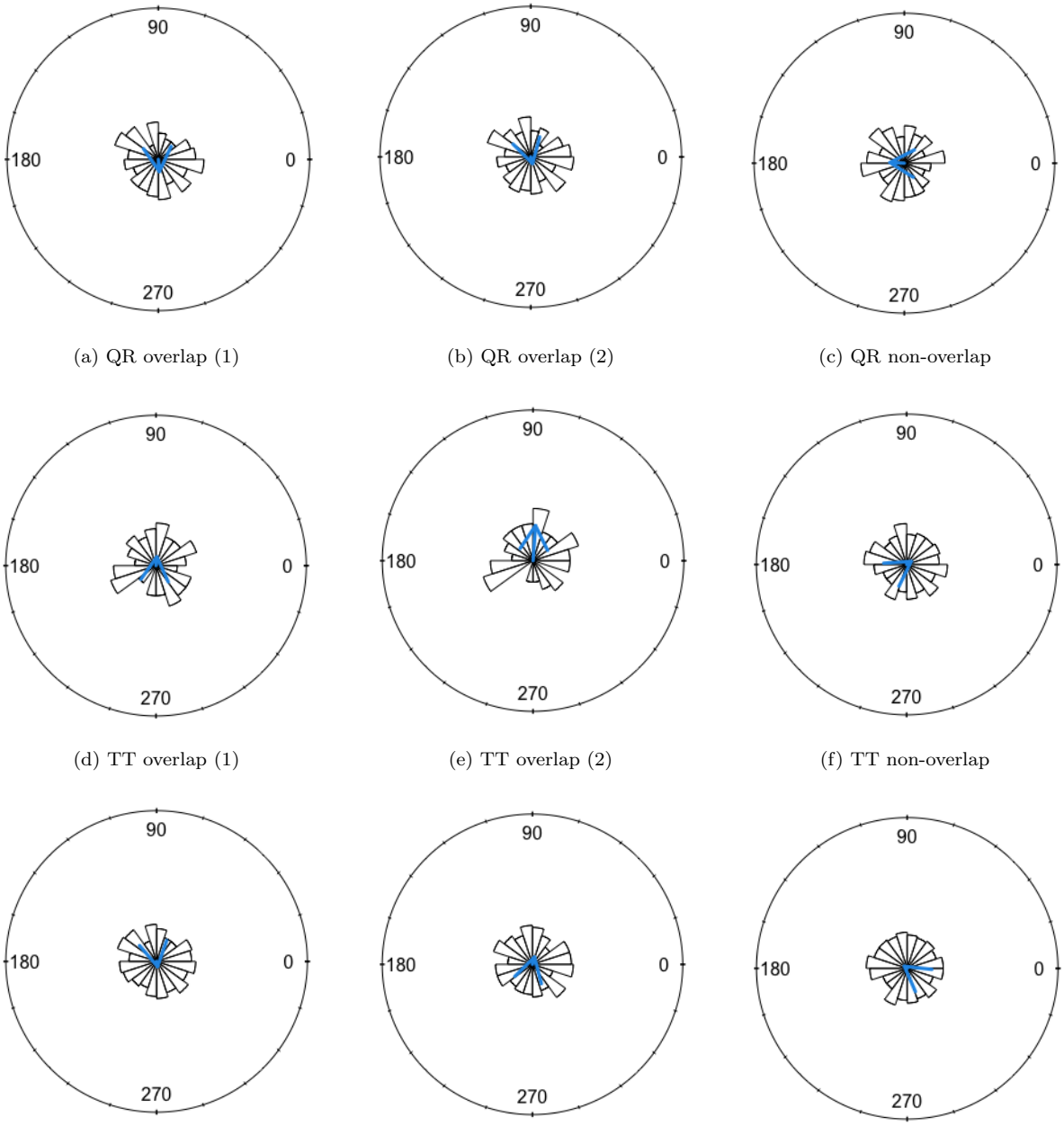


(h) All overlap (2)



(i) All non-overlap

Figure B.2: Rose plots of the beat rates of triads

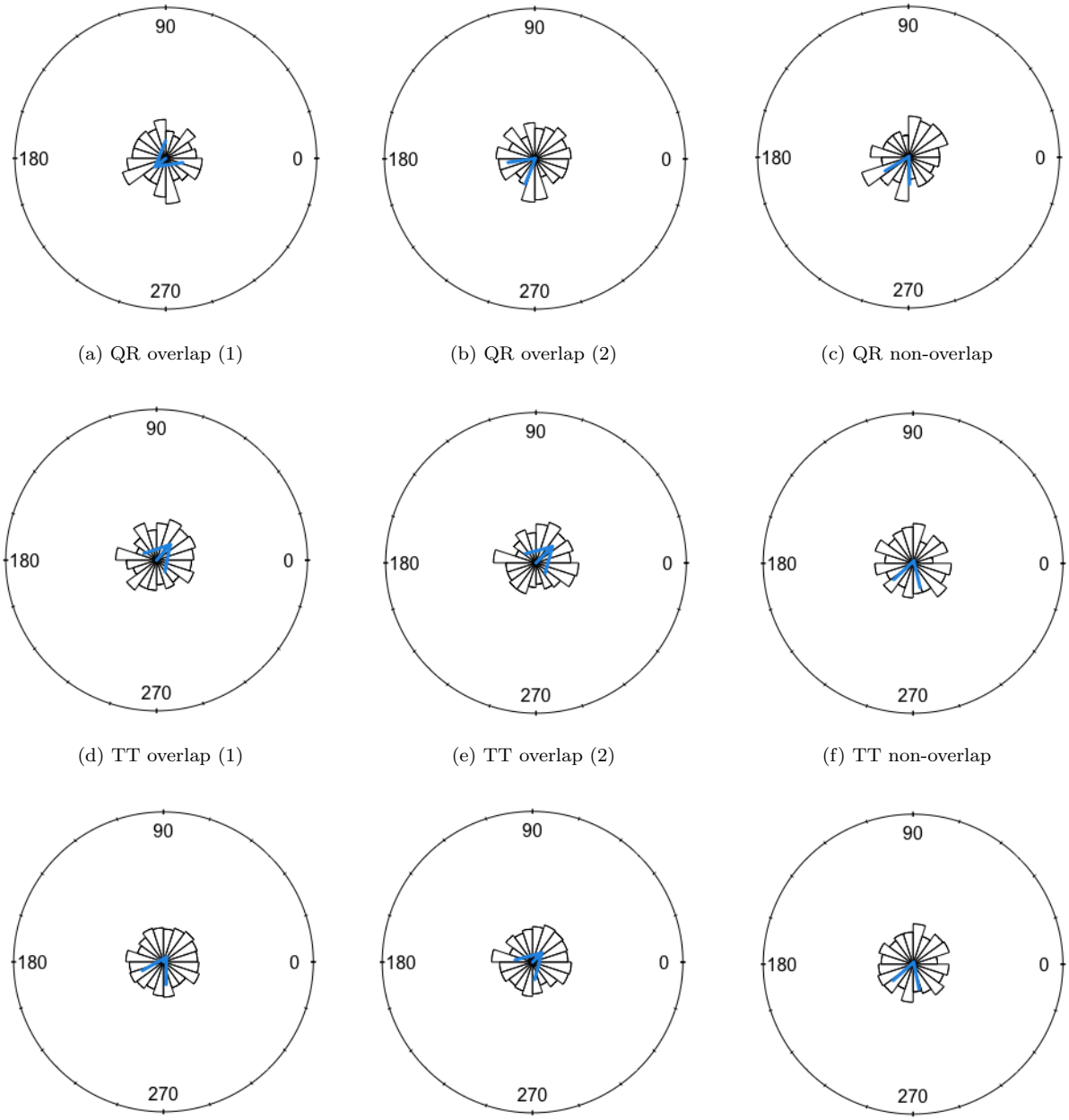


(a) QR overlap (1) (b) QR overlap (2) (c) QR non-overlap

(d) TT overlap (1) (e) TT overlap (2) (f) TT non-overlap

(g) All overlap (1) (h) All overlap (2) (i) All non-overlap

Figure B.3: Rose plots of the syllable rates of dyads



(a) QR overlap (1) (b) QR overlap (2) (c) QR non-overlap

(d) TT overlap (1) (e) TT overlap (2) (f) TT non-overlap

(g) All overlap (1) (h) All overlap (2) (i) All non-overlap

Figure B.4: Rose plots of the syllable rates of triads