INTERSPEECH
2009 BRIGHTON

Are we ‘in sync’: Turn-taking in collaborative dialogues

Štefan Beňuš

Constantine the Philosopher University, Nitra, Slovakia and Slovak Academy of Sciences,
Bratislava, Slovakia

sbenus@ukf.sk

10.21437/Interspeech.2009-618

Abstract

We used a corpus of collaborative task oriented dialogues in American English to compare two units of rhythmic structure – pitch accents and syllables – within the coupled oscillator model of rhythmical entrainment in turn-taking proposed in [1]. We found that pitch accents are a slightly better fit than syllables as the unit of rhythmical structure for the model, but we also observed weak support for the model in general. Some turn-taking types were rhythmically more salient than others.

Index Terms: turn-taking, rhythm, oscillator, entrainment

1. Introduction

Turn-taking behavior represents a complex, non-deterministic, and dynamically evolving cognitive system that spans across various modalities of human interaction, such as spoken and sign language, gesture, or gaze. Better understanding of this system brings great potential for improving the naturalness of dialogue systems and interactive voice response applications.

Probably the most stable, yet still not sufficiently understood, characteristic of turn-taking in dialogues is the prevalent smoothness of turn-exchanges among the interlocutors. It is common that speakers start planning and often executing their speech before the other speaker finishes speaking. Several studies investigated the prosodic, syntactic, and pragmatic features that speakers use for predicting when the current speaker finishes her turn [e.g. 2, 3, 4, 5].

Some studies suggest that rhythmical entrainment of the interlocutors may also facilitate smooth turn-taking. For example, [6] proposed that conversational partners perceive speech as isochronous, that is, they perceive “constancy of intervals between rhythmic events”. [6: 24]. Interlocutors then synchronize such that their turn-productions fit into this rhythmic isochrony. The fundamental unit of such isochrony in English is thought to be the ‘beat’, which roughly corresponds to the temporal interval between adjacent prominent syllables associated with a pitch accent.

However, these proposals are based on impressionistic data and transcripts with little experimental quantitative support. A corpus study in [7] observed rhythmical entrainment among the interlocutors based on correlation of latencies in turn-exchanges in dialogues for the two interlocutors. However, they compared average latencies over conversations, which is a static measure that could be influenced by various factors such as the topic liveliness. [8] analyzed turn-taking in a subset of HCRC map-task corpus testing if the timing of the current turn initiation corresponds to the rhythmical structure of beats (represented by pitch accents) in the preceding turn. [8] found no support for this dynamically defined rhythmical entrainment among speakers.

[1] argued that the syllable, rather than the beat, is the organizational unit of turn-taking entrainment of rhythm. They proposed that the generation of the sequential structure of turn-taking can be captured with a model utilizing two dynamically defined oscillators that describe the potential for initiating speech at any given moment. In this model, each syllable represents a single phase of oscillation, and the two oscillators are counter-phased: the peaks of one oscillator correspond to the valleys in the other. Hence, speech from a single speaker should display an in-phase pattern while speech at turn-exchange points should display an anti-phase pattern. Although the model is not formally developed, it is intuitively appealing, supported by convincing converging evidence, and importantly, the model makes several testable predictions.

Given the controversial status of the syllable as the rhythmical unit and the absence of a formal model based on beats, we use a corpus of collaborative task oriented dialogues to compare the two units of rhythmic structure – pitch accents and syllables – within the coupled oscillator framework of rhythmical entrainment proposed in [1].

1.1. Predictions of [1]

- **Isochrony in turn-internal chunks**. Syllable rates in adjacent inter-pause-units (chunks) from a single speaker should correlate. Latency should positively correlate with the syllable rate of chunk$_1$. Phasing in adjacent chunks (latency/rate$_{ch1}$) should cluster around 0, 1, 2; i.e. it should show an in-phase pattern.
- **Entrainment in turn-exchanges.** In adjacent two chunks from different speakers, the syllable rates should correlate. Latency should correlate with the syllable rate of chunk$_1$. Phasing should cluster around 0.5, 1.5, 2.5; i.e. in an anti-phase pattern.
- **Timing of near-zero gaps and overlaps in turn-exchanges**. Perfect latches with latencies around zero should be dispreferred. The distribution of latencies should be bi-modal with two peaks – one positive, one negative – and both roughly equidistant from zero.
- **Lapses**. Mutual entrainment of interlocutors should last at least 1 second. Hence, simultaneous starts after a silent pause of less than 1s should be in-frequent, and they should rise non-linearly around this value.

2. Corpus

To test these predictions, we use the data from Columbia Games Corpus [2, 9]. The corpus contains speech from 12 dyadic collaborative conversations where participants played a set of computer games designed to elicit conversation. They were seated in a soundproof booth divided by a curtain to control for the modality of interaction ensuring the audio-only mode. Altogether 13 speakers of Standard American English (7 males and 6 females) were recorded.

Subjects were instructed to play two types of collaborative games (CARDS and OBJECTS). In this paper, we only analyze the OBJECTS games, in which one player described the position of a target object with respect to other fixed objects on her screen, while the other tried to move his representation of the target object to the same position on his own screen. Points were given based on the proximity of the target object to its correct location. The subjects switched roles repeatedly.

On average, each OBJECTS game session took 22.4 minutes, totaling 4h 29m of dialogue for this corpus. The recordings were orthographically transcribed, and words were aligned to the source acoustic signal by hand.

2.1. Data annotation and feature extraction

The speech in the OBJECTS corpus has been intonationally transcribed using ToBI [10]. Hence, pitch accented words can be identified. From each pitch-accented word we extracted the time of the energy peak as a rough estimate of pitch-accent alignment. The series of these temporal points was then used for the calculation of pitch-accent rate.

Inter-pausal units (chunks) were automatically identified as a maximal sequence of words surrounded by silence longer than 50ms. Latency was defined as the difference between the end of the chunk and the beginning of the next chunk. The syllable rate in each chunk was automatically computed using the duration of the words from hand-alignment and the number of syllables from dictionaries.

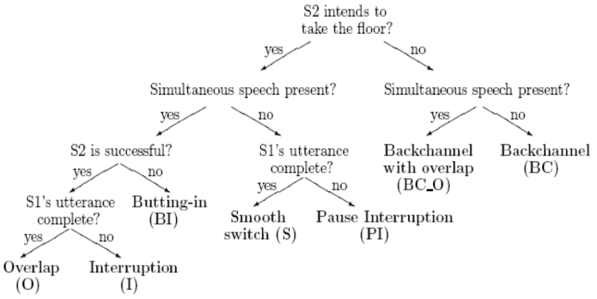

Figure 1: *Schematic diagram for turn-type annotation*

Finally, following a modified annotation scheme based on [11], two labelers annotated each switch between the speakers for a turn-type. First, the presence of simultaneous speech between the speaker turns was determined automatically. Then, the labelers proceeded following the steps illustrated in Figure 1. There were two additional special labels relevant for this study: X2 and X3. X2 is a continuation of previous speech by the same speaker after a backchannel (BC, or BC_O) from the other speaker. X3 marks a simultaneous start. If two turns begin almost simultaneously (formally, within 210 ms of each other [12]) then both speakers are most probably reacting to the preceding turn.

After the two labelers annotated all the files, the remaining disagreements were discussed and if agreement could not be established, these turns were labeled as "?". Finally, a single labeler annotated all X3 exchanges as rhythmically marked or unmarked, and in the marked group, if they were rhythmically non-integrated lapses, or result from other causes such as disfluencies, additions, etc.

3. Results

3.1. Isochrony

The first prediction in 1.1 is that if the speech of a single speaker is isochronous, the rhythms of adjacent chunks should correlate. We call this *chunk-based isochrony*. For syllables, the correlation was significant albeit only moderate, r(4283) = 0.138, p < 0.001. Similar results were obtained for pitch accents, r(3846) = 0.131, p < 0.001.

The next prediction was that slower speech should result in longer latencies, which we call *initiation isochrony*. For the syllable as the rhythmical unit, the correlation was significant albeit only moderate, r(4283) = 0.176, p < 0.001. Worse but still significant results were obtained for pitch accents, r(4036) = 0.097, p < 0.001.

The final prediction involved the phasing of chunk initiation with respect to the rhythm of the preceding chunk. To test this, we calculated the phasing measure as latency/rate$_{ch1}$ for both syllables and pitch accents and analyzed its distribution in histograms. Figure 2 shows the histograms for the phasing values between 0 and 5 based on syllables (top) and pitch accents (bottom). Both variables show non-normal distribution, supported by Kolmogorov-Smirnov tests, but this is due to skewness rather than to the presence of multiple peaks, supported by S-shaped PP-plots. Here we analyze the histograms descriptively and leave the computational characteristics of the distributions for future work. Neither of the histograms peaks at predicted whole number integers. While the syllable-based distribution has a clear single peak around 0.5, and a minor discontinuity around 1.5, the bottom histogram shows two peaks very close to one another around 1.3 and 1.7 respectively, and another minor one around 2.2.

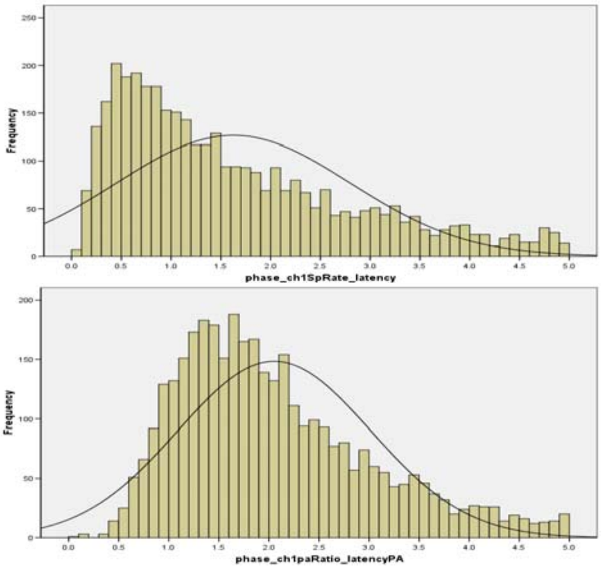

Figure 2: *Phasing for adjacent chunks from a single speaker based on syllables (top) and pitch accents (bottom).*

3.2. Entrainment

If speakers entrain to one another's rhythm, adjacent chunks representing a turn-exchange should correlate. We call this *chunk-based entrainment* since the rates over whole chunks are compared. For syllables, the correlation was not significant; r(3314) = -0.027, p = .119. However, looking separately at turn-types, both backchannels (BC) and continuations after them (X2), and only these two, had significant positive correlations, r(393) = 0.18, p < .001 and r(340) = 0.189, p < .001 respectively. Hence, only turn-taking associated with backchanneling displays chunk-based entrainment at the level of syllable rate. The chunk-based entrainment at the level of pitch accents was also not significant. Separately for turn-types, significant positive correlations were found for Overlaps (O) and Pause Interruptions (PI), r(438) = 0.136, p = .004 and r(173) = 0.2, p = .008 respectively.

[6] suggest that at least 3 metrical units are required for the perception of isochrony, and consequently, for entrainment to occur. Hence, we tested the above correlations for chunks with at least 3 words and pitch accents respectively. While the syllable-based correlation remained non-significant, the positive pitch-accent based correlation became significant, r(1087) = 0.089, p = .003.

The measure of *initiation entrainment* at the time of turn-exchange describes the timing of a chunk start with respect to the rate of the preceding chunk from another speaker. For syllables, the correlation between raw latency and syllable rate of chunk$_1$ in exchanges with reliable latency (S, O, X2, BC, PI) was not significant; r(3060) = 0.026, p = 0.143. Looking separately at turn-types, only PIs have significant (and comparatively high) correlation, r(208) = 0.318, p < .001. For separate sessions, only 1 out of 12 sessions showed positive significant correlation.

The initiation entrainment based on pitch accents showed better results on syllables. The correlation of pitch accent rates and latency of pitch accents was significant; r(2863) = 0.131, *p* < 0.001. In this measure of entrainment, limiting the data to cases where chunk$_1$ had at least 3 pitch accents did not improve the results. Table 1 shows that pause interruptions (PIs) have again the strongest correlation, followed by continuations after backchannels (X2) and overlaps (O). Backchannels with smooth switches have the weakest correlations. For separate sessions, 9 out of 12 show significant positive correlations suggesting initiation entrainment based on pitch accents in these sessions.

Table 1. *Initiation entrainment for a subset of turn-types.'*': p < 0.05; '**': p < 0.001.*

	N	R
BC	389	.14*
X2	331	.32**
S	1532	.19**
O	438	.32**
PI	173	.47**

Finally, the model in [1] predicted that phasing of adjacent chunks for two different speakers should show the anti-phase coupling, i.e. phasing should cluster around 0.5, 1.5, 2.5, etc. Figure 3 shows the histograms for the phasing values between -2 and 5 based on syllables (top) and pitch accents (bottom). Compared to the histograms in Figure 2, they show a slightly less skewed and less peaked non-normal distributions. The syllable based distribution peaks around 0.5 and pitch-accent based one peaks at 1.5, which is in line with the prediction of the model. In addition to the peak at 0.5, there are discontinuities around 0, 1, and 1.5 in the top histogram. The bottom one has a wider spread of most frequent values with the peak around 1.6, and minor discontinuities can be observed around 2.2, and -0.1. Hence, the predicted peaks roughly separated by one unit of phasing are not observed. Crucially, we also don't see a qualitative difference between the histograms in Figures 2 and 3 predicted by the model. Both syllable based and pitch-accent based measures of phasing seem to peak at similar values for adjacent chunks in turn-internal positions as well as at turn-exchange points.

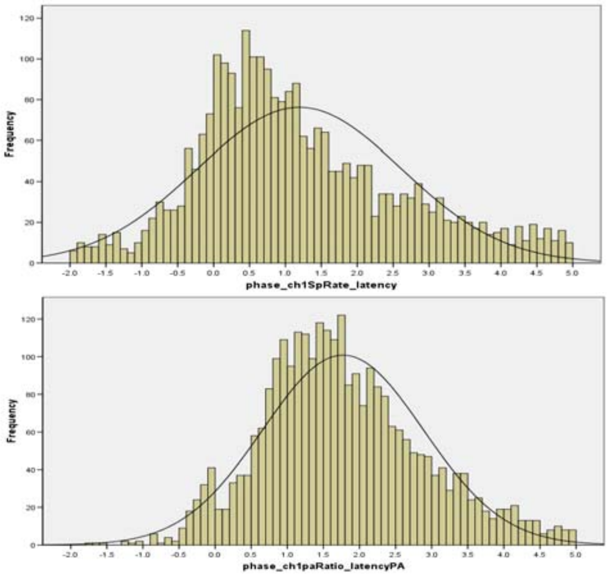

Figure 3: *Phasing for exchanges in syllables (top) and pitch accents (bottom).*

3.3. Near-zero latencies

Due to the assumed anti-phase relationship of the interlocutors at turn-exchanges, the model in [1] predicts that the latches with latencies around zero should be dispreferred. Hence, the distribution of latencies should be bi-modal with two peaks – one positive, one negative – and both roughly equidistant from zero. Fine-grained histograms in Figure 4 with bin sizes of 0.05 show the relevant distributions of raw latencies for syllables (top), and latencies between the final pitch accent in chunk$_1$ and initial accent in chunk$_2$ (bottom). None of the histograms show clear bi-modal distribution, but when compared to each other, the one based on pitch accents on the bottom displays a minor 'valley' around 0.1, resembling thus bi-modal distribution slightly more than the top histogram.

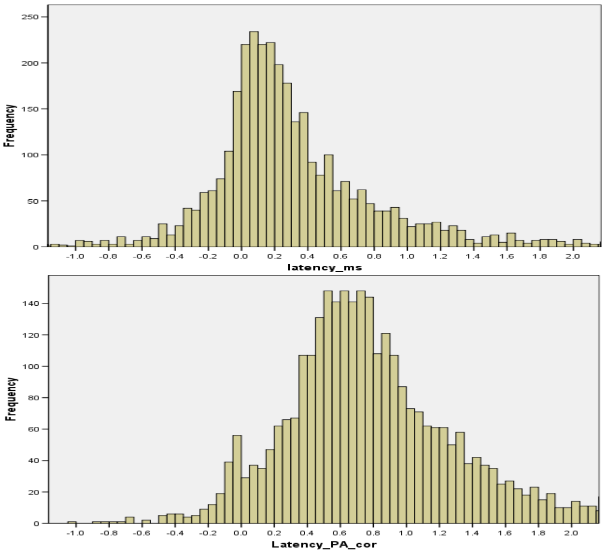

Figure 4: *Distribution of latencies for syllables (top) and pitch accents (bottom).*

3.4. Existence of lapses

The model in [1], together with the original turn-taking model [13], predict that simultaneous starts signal the breakdown of the speakers' entrainment, and that the function describing the length of silent pauses preceding simultaneous starts should be discontinuous with a sharp increase somewhere after 1 second. In our corpus (near) simultaneous turns are labeled as X3. There are 371 (9.3%) such turn-exchanges. Compared to other exchange types, the X3 frequency is similar to those of continuations after backchannels (X2) and backchannels (BC + BC_O), and greater than the rates of all interruption types (BI, I, PI). The X3 exchanges were distributed relatively evenly among the 13 speakers, and they made at least 5% of all the turn-exchanges for each speaker.

X3 exchanges result in overlapped synchronous speech but exclude interruptions and cooperative overlaps. Hence, they signal a problem in the timing of one's speech and might serve as a window into the temporal organization of turn-taking behavior. X3s in our corpus tended to result in brief overlaps: on average the overlap lasted 344ms and contained 2 words from the reference speaker. The average silent pause preceding the first speech in X3 was 430ms.

Following the scheme described in Section 2.1, 20% of the X3 exchanges we judged as not being rhythmically problematic. Of the remaining simultaneous starts, 34% (N = 101) were judged as true lapses and rhythmically non-integrated sequences. As expected, they have significantly longer preceding silent pauses than other X3s. The distribution of silent pauses in these exchanges shows that most simultaneous starts occur within 1 second, and the distribution peaks around 0.5s.

4. Discussion & conclusions

Our results show that pitch accents are a slightly a better fit as the unit of rhythmical structure for the oscillator model in [1] than the originally proposed syllables. This is because the measures of isochrony and entrainment based on pitch accents were comparable and significant, which was not the case for syllables. However, none of these measures were robust, and experimental support for the model in our data was rather weak. This is mostly due to the moderate correlation values for isochrony and entrainment, predicted but not observed patterns of in-phase and anti-phase coordination for turn-internal and turn-exchange chunks respectively, and to the absence of bi-modal distribution of raw latencies.

Despite weak support for rhythmical isochrony and entrainment in general, our results also identified several turn-taking strategies with more salient rhythmical component such as pause interruptions (PI) and backchanneling. For PIs, unfinished speech of the first speaker leaves the possibility of simultaneous start open, increasing thus the importance of suitable timing. For backchanneling, the nature of the task naturally increased the frequency of adjacency pairs *Information-Backchannel*, and they many times occurred in clusters of 3-4 such pairs. This may have prompted a more stylized, and thus rhythmically entrained, nature of these pairs. Several descriptive examples in our data support this analysis.

Finally, there could be multiple reasons for the scarcity of robust positive support for rhythmical isochrony and entrainment in turn-taking. These include the type of data, the selection of features describing rhythm, or the assumptions and details of the oscillator model. For example, it may be that speech-only conversations with significant cognitive load associated with the collaborative tasks are less suitable than spontaneous face-to-face conversations. Or, the stress foot may prove a better unit of analysis in English than the syllable or the pitch accent. Alternatively, as suggested in [14], rhythmical entrainment might be based on the continuous flow of information between interlocutors rather than the perception of isochrony. Finally, machine learning experiments clustering the features of isochrony and entrainment might provide a better handle for the noisy data. An anonymous reviewer makes a useful suggestion to correlate the acoustic/metric measurements with some measure of perceived smoothness of the recorded dialogues, which could also provide a baseline for the machine learning experiments. All of these considerations provide avenues for future research. But, it could also be that the oscillator model needs to be adjusted or re-thought. Either rhythm entrainment does not require constant overt reinforcement for the oscillators' entrainment, or, dynamically defined entrainment between speakers may be based on stable relationships between more complex landmarks involving breathing and other articulatory gestures.

5. Acknowledgements

The work presented in this paper was supported in part by the Slovak ministry of education grant KEGA 3/6399/08, the Slovak Research and Development Agency project APVV-0369-07, and A. W. Mellon scholarship by the IASH at the University of Edinburgh. This research was done in part in collaboration with J. Hirschberg and A. Gravano at Columbia University, USA.

6. References

[1] Wilson, M. and Wilson, T., "An oscillator model of the timing of turn-taking", Psychonomic Bulletin and Review 12 (6): 957-968, 2005.

[2] Gravano, A., Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue. PhD thesis, Columbia University, 2009.

[3] Ford, C.E., Thompson, S.A., "Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns", in: E. Ochs, E. Schegloff, S.A. Thompson [Ed] Interaction and Grammar, 134–184, CUP, 1996.

[4] Selting, M., 1996, "On the interplay of syntax and prosody in the constitution of turn. Constructional units and turns in conversation", Pragmatics 6: 357–388.

[5] Ward, N. and Tsukahara, W., "Prosodic features which cue back-channel responses in English and Japanese", J. Pragmatics 23: 1177–1207, 2000.

[6] Auer, P., Couper-Kuhlen, E., and Müller, F. Language in Time, OUP, 1999.

[7] Bosch, L. ten, Oostdijk, N. and Boves, L. "On temporal aspects of turn taking in conversational dialogues", Speech Communication, 47:80-86, 2005.

[8] Bull, M., "An analysis of between-speaker intervals", Proceedings of the Edinburgh Linguistic Converence, 18-27, 1996.

[9] Gravano, A., Benus, S., Hirshberg, J., Mitchell, S., Vovsha, I., "Classification of Discourse Functions of Affirmative Words in Spoken Dialogue", *Proceedings of Interspeech*, 1613-1616, 2007.

[10] Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", in S.-A. Jun, [ed] Prosodic Typology: The Phonology of Intonation and Phrasing, 9-54. OUP, 2005.

[11] Beattie, G., "Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted", Semiotica, 39(1/2): 93-114, 1982.

[12] Fry, D.B., "Simple reaction-times to speech and non-speech stimuli", Cortex, 11(4): 355-60, 1975.

[13] Sacks, H., Schegloff, E., and Jefferson, G. "A simplest systematics for the organization of turn-taking for conversation", Language, 50:696–735, 1974.

[14] Cummins, F., "Rhythms as entrainment: The case of synchronous speech", Journal of Phonetics, 37: 16-28, 2009.