Adaptation in turn-initiations

Štefan Beňuš
Constantine the Philosopher University, Štefánikova 67, 94074 Nitra, Slovakia
Slovak Academy of Sciences, Institute of Informatics, Dúbravská cesta 9, 84507 Bratislava, Slovakia
sbenus@ukf.sk

Abstract. This study investigates the variability in the temporal alignment of turn initiations and its relationship to the entrainment and power structure between the interlocutors. The data come from spontaneous, task-oriented human-human dialogues in Standard American English, and focus on single-word turn-initial utterances. The descriptive and quantitative analysis of the data show that an emergent asymmetrical dominance relationship is constructed partly through the accommodation (or its absence) to the temporal and rhythmic features of interlocutors' turn-initiations.

Keywords: turn-taking, entrainment, rhythm, affirmative cue words, dominance

1 Introduction

Building adaptive and context sensitive automated communicative systems requires an understanding of the cognitive systems underlying our communicative competence. One of the relevant cognitive systems that are fundamental for human interaction is the organization of turn-taking. Turn-taking is a dynamically evolving, embodied, cross-modal system that is pervasive in both speech and sign language and is strongly linked to paralinguistic domains such as gaze and gestures; e.g. [1], [2], [3]. In general, this floor-management organization underlies the decisions of 'who speaks when' and must include at least three components: 1) ways of signaling and perceiving the cues for transition-relevant places and turn allocation among interlocutors [4], 2) ways of achieving suitable durations of latencies between the turns, avoiding over-long overlaps or silent pauses, and 3) ways of resolving disruptions in the system [5]. In this paper we focus on the second point.

In current state-of-the-art applications of interactive voice-response systems, turn boundary detection is typically based on silence detection with the threshold between 0.5 and 1 second. Multiple problems arise from this implementation such as the occurrence of false positives or hindrance of cohesion. The primary reason for these problems is that the exchange of turns in human-human spoken interactions is not based on detecting silence. Humans have the ability to detect the projected end of the current turn from multiple prosodic, syntactic, pragmatic and gestural cues; e.g. [6]. But even if we could implement all these human abilities and create systems that reliably predict when the interlocutor is about to finish her turn, we still need a model

of when precisely we should start speaking. For this aim, the other feature of human-human interactions is crucial: Interlocutors are assumed to be entrained to each other on a number of linguistic and paralinguistic levels, which greatly facilitates communication [7]. This mutual entrainment then provides a basis for meaningfully modeling the timing of turn-initiations as a dynamic incorporation into the rhythmic patterns of the preceding turns [8]. In support of this approach, entrainment has been also found in the metrical features of utterances [9], in intensity characteristics [10], in phonetic and prosodic characteristics of individual words [11], and in accent and other socio-phonetic variables [12]. At the paralinguistic level, conversational partners entrain their body swaying motions [13], and breathing [14].

The aim of this study is to improve our understanding of the adaptation choices human interlocutors make in the temporal initiation of turns, which is a stepping stone to building more natural human-machine dialogue systems. Our focus is on the timing of turn-initial responses in collaborative task dialogues with special attention to single word responses that pragmatically function as agreements, acknowledgments, backchannels, or filled pauses. We will argue that the adaptive behavior can be observed both sequentially in adjacent turns as well as globally over the entire conversation, and that it contributes to the construction of an inter-speaker power relationship.

2 Corpus description

Data for this study come from a single session of the Columbia Games Corpus [15]. Two female speakers (Spkr1 and Spkr2) played specially designed games without visual contact that involved matching the identity and positions of various objects on their laptop screens. The subjects switched the roles repeatedly. The recordings were then orthographically transcribed, and words were aligned to the source acoustic signal by hand. The analyzed speech in this conversation covers 35.7 minutes and contains 770 turns almost equally distributed between the two speakers (384 for Spkr1 and 386 for Spkr2).

Space restrictions prevent a full description of annotations performed on this corpus but detailed descriptions could be found in [15] or [16]. The prosodic features of the dialogues were labeled using the ToBI annotation scheme [17]. The turn-taking behavior was labeled using a slightly modified scheme from [18] described in [19]. Temporal features such as turn latencies were automatically extracted based on word alignments.

3 Descriptive and quantitative observations

In this paper we concentrate on two symptomatic uses of speaker adaptation in the timing of turn-initial single-word utterances. The first type is the sequential local adaptation in adjacent turns. The second is a global adaptation of the timing pattern developed during the entire conversation. We will discuss each of these observations

in the following subsections presenting first the descriptive analysis of representative example followed with quantitative tests for the validity and robustness of patterns.

3.1 Local adaptation: Affirmative cue words in adjacent turns

Consider the following example in which Spkr1 describes the position of the *iron* on her screen and Spkr2's role is to match the position of this object on her screen with the position of Spkr1's screen. Bold numbers show the turn latencies (i.e. duration of silences across turns). Utterance-final rising, falling and level intonational contours are shown with arrows ↑, ↓, and → respectively; square brackets show overlapped speech.

1. Spkr2: okay, lines up↑ **(0.36)**
2. Spkr1: yeah it's it's almost it's just barely (0.27) like over↓ **(0.45)**
3. Spkr2: o[kay]↑
4. Spkr1: [but] it's basically that same line um so the black part at the bottom of the iron↑ **(0.08)**
5. Spkr2: mmhm↑ **(0.13)**
6. Spkr1: not necessarily like on the same line as the white foot it's just a little bit over↓

The timing of *okay* and *mmhm* in lines 3 and 5 from Spkr2 is different. Two preceding instances of *mmhm* from Spkr2 (not shown in the excerpt) came with the latencies of around 0.2s. Perhaps realizing that her acknowledgment in line 3 was 'too late', Spkr2 avoids another overlap by perfectly aligning her backchannel in line 5 with only a 0.08s latency. Fig. 1 gives a visual representation of this adjustment.

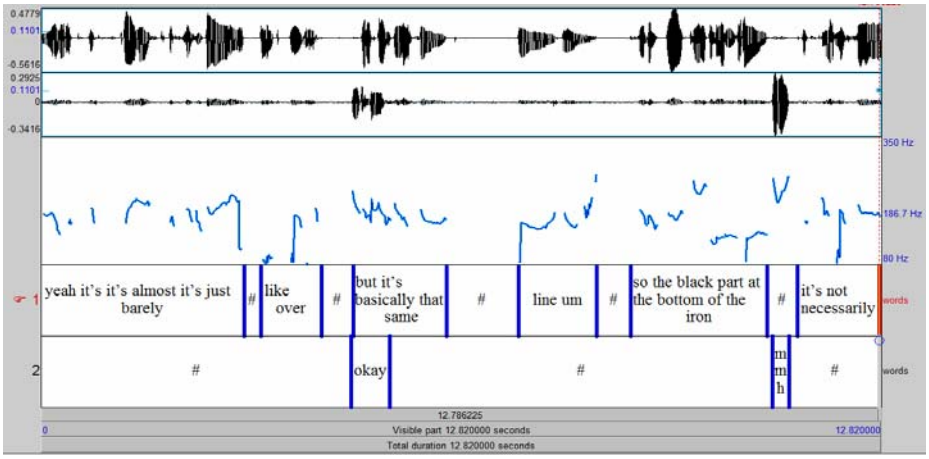

Fig. 1. Adjustment in the temporal alignment of two consecutive turn-initial acknowledgment/backchannels. The top two panels show the sound waves for both speakers, the middle panel fundamental frequency between 80 and 350 Hz, and the bottom two panels show the transcript.

Another type of local adaptation can be observed in the timing of *Adjacency triplets*. In this unit of interaction, the first speaker provides some information or poses a question; the second speaker acknowledges, backchannels, or provides a short answer; and then the first speaker acknowledges this response. In the short excerpt below we see that the length of the silent intervals preceding and following *yeah* from Spkr2 in line 2 are greater than similar intervals surrounding subsequent *no* from the same speaker.

1. Spkr1: basically on the right side of his hat↑ **(0.35)**
2. Spkr2: yeah↓**(0.17)**
3. Spkr1: okay (0.42) and but it's not touching the hat↑**(0.09)**
4. Spkr2: no↑ **(0.12)**
5. Spkr1: okay (0.07) and then the distance →

Adjacency triplets also present suitable material for comparing relative and absolute time in describing the temporal alignment patterns. While the absolute latency duration used in the preceding two examples might be suggestive of temporal entrainment, a much stronger case could be made if these patterns are supported with relative timing as well. This is because relative timing measures are less influenced by the variability in speech rate that is orthogonal to the rhythmical structure of turn-taking. Hence, if the same generalizations can be drawn from both absolute and relative timing measures, the observed rhythmical alignment patterns can be considered robust and general in nature.

One way of relativizing time in turn-taking research is to study the timing of peaks of prosodic prominence in relation to such peaks in the preceding utterance [9], [16], [20]. These prominent syllables, or *P-centers* [8], are assumed to be linked most tightly to the amplitude (loudness) of the syllables [21]. Therefore, we define such prominence peaks as the amplitude peaks of the stressed syllables in all words that received a pitch accent mark in the labeling of the prosodic structure using the ToBI scheme [17]. Fig. 2 illustrates the adaptation of Spkr2 to the rhythmical pattern established by her interlocutor in two consecutive adjacency triplets from the excerpt above. We see that the spacing of the pitch accents in the first question is greater than in the second question, to which Spkr2 adjusts by aligning the prominence peak of the second response more tightly than the first response.

Although the above examples are representative of the analyzed conversation, such context-sensitive descriptive analysis should be complemented with quantitative analyses to assess the validity and robustness of the patterns identified with the examples. Frequent impressionistically-based claims about wide-spread rhythmical entrainment of interlocutors [9] might result from perceptual 'mirage' in which listeners interpret speech as necessarily rhythmical [22], and impressionistic transcriptions of speech may be mis-perceived and unreliable [23].

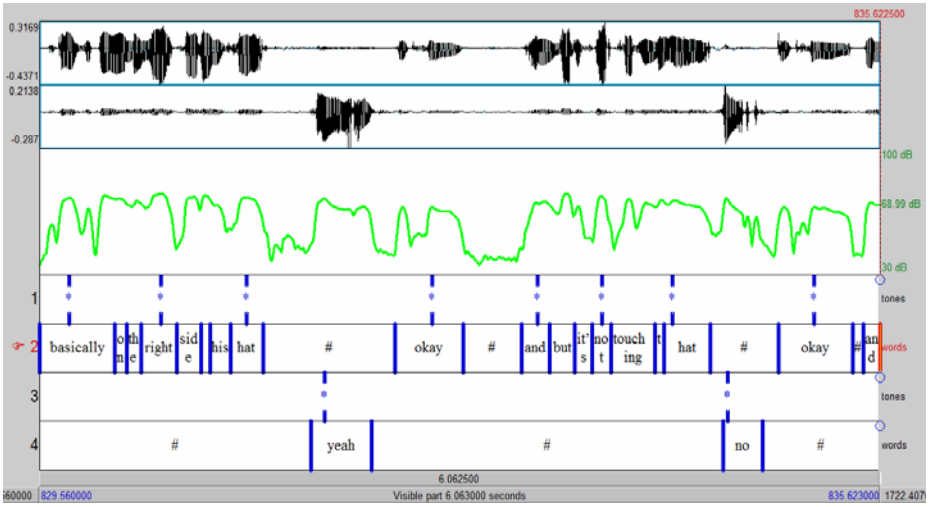

Fig. 2. Adjustment in the rhythmical alignment of two consecutive adjacency triplets. The top two panels shows the sound waves, the middle panel shows intensity between 30 and 100 dB, and the bottom 4 panels show the transcript with pitch accent labels as "*"

To test our hypotheses, we employed several quantitative analyses. For example, the distribution of turn types based on the labeling scheme described in section 2 showed that Spkr1 initiated her turns as overlapped with the end of the preceding turn more often than Spkr2; a Pearson chi-square test $r(1, 511) = 6.45$, $p = 0.011$. Spkr2 also produced her turn-initial backchannels, agreements, acknowledgments, and filled pauses with longer latencies than Spkr1; an Anova test $F(1, 397) = 4.6$, $p = 0.032$ with mean latencies 0.33 for Spkr2 and 0.22 for Spkr1. Finally, we tested the correlation between the rate of pitch accents in the last pause-defined unit before the turn-exchange and the latency between the last pitch accent in the turn before the exchange and the first pitch accent after the exchange. The rate of pitch accents was calculated as the number of pitch accents divided by the length of the pause-defined unit and represents a rough measure of speech rhythm. If entrainment takes place, we expect shorter latencies following units with faster rate and longer latencies following units with slower rate, hence a significant positive correlation. Despite moderate correlation coefficients, showing a rather low degree of rhythmical entrainment, the values for Spkr2 were consistently significant $r(293) = 0.22$, $p < 0.001$ in all data, and $r(127) = 0.3$, $p = 0.001$ for turn-initial backchannels, agreements and acknowledgments, while significance was never reached for Spkr1.

These findings provide quantitative support for our descriptive analysis that characterized Spkr2 as more accommodating and willing to adapt her turn-taking behavior to her interlocutor.

3.2 Global adaptation: Timing of turn-initial filled pauses

Turn-initial filled pauses facilitate both production and perception of linguistic material because they allow speakers to plan their intended message and listeners to prepare to perceive important content [4], they mark discourse and prosodic boundaries [24], and signal planning difficulties associated with cognitive load and the presence of choice [25]. In all those functions, turn-initial filled pauses tend to be produced with significant latencies after the end of the preceding turn. Multiple examples of this default temporal alignment are also present in our corpus. However, we also observed instances of a tight temporal alignment of the filled pauses with the end of the preceding turn. Consider the excerpt below.

1. Spkr1: okay, how about the little black part, um, where the beak starts, do you see [that]→
2. Spkr2: [**um**] it's like blinking in and out let me see, um yeah there's like black above the beak righ[t]↑
3. Spkr1: [o]kay [just a little bit of that]↑
4. Spkr2: [yeah you can see that]↑
5. Spkr1: okay and um anything el[se]↑
6. Spkr2: [**um**] let me think, mm, see
7. Spkr1: is the tail sticking out from th- b- where the branch is like it's not aligned↑ **(1.18)**
8. Spkr2: [**um** yeah it's] not [aligned with the] branch→
9. Spkr1: [the tail of the lion]↑ [okay]↑
10. Spkr1: and either is the foot like it's ?-[sticking] out a little bit more↑ **(0.25)**
11. Spkr2: [the feet]
12. Spkr2: **um** (2.11) oh the branch on the left side↑

The turn-initial *ums* in lines 2 and 6 overlap the end of the preceding turn. They follow a question, and there is thus no need to hasten to grab the floor, because Spkr1 has explicitly yielded the floor and selected Spkr2 to continue. There is also no need in this context to signal that the interlocutor needs to attend to the speaker, which is another common function of turn-initial filled pauses. This is because Spkr1 is presumably fully attending to Spkr2 as she is expecting an answer to her question. Finally, if these filled pauses signaled planning difficulties, they would be probably preceded by a relatively long silent pause representing cognitive processing, and not temporarily aligned almost perfectly with the end of the preceding turn.

This temporal pattern contrasts with the alignment of *um* in line 8. Here, Spkr2 seems to signal hesitation and aligns her filled pause with the latency longer than one second. Spkr1 seems to detect difficulty in processing her question and adds more information that overlaps with the filled pause from Spkr2. Multiple examples of the contrast in our corpus between this default 'loose' alignment of filled pauses and the 'tight' alignment exemplified in lines 2, 6, and 12 support the analysis that the 'tight' alignment of turn-initiations evolved as a global adaptation of Spkr2 over the entire conversation to avoid the overlaps from Spkr1.

Similarly to section 3.1, we wanted to test the robustness of these observations quantitatively for the whole conversation. First, we observed that 12% of all turns

started with a filled pause. Two thirds of these turns (66%) were produced by Spkr2, and this difference between the speakers was significant; $r(1, 770) = 10.51$, $p < 0.001$. Additionally, Spkr2 produced these turn-initial filled pauses with greater normalized mean pitch and tended to produce them also with greater normalized mean intensity than Spkr1; $F(1, 66) = 4.59$, $p = 0.036$ and $F(1, 66) = 3.14$, $p = 0.081$ respectively. This difference signals greater pragmatic importance of turn-initial filled pauses for speaker B than for speaker A.

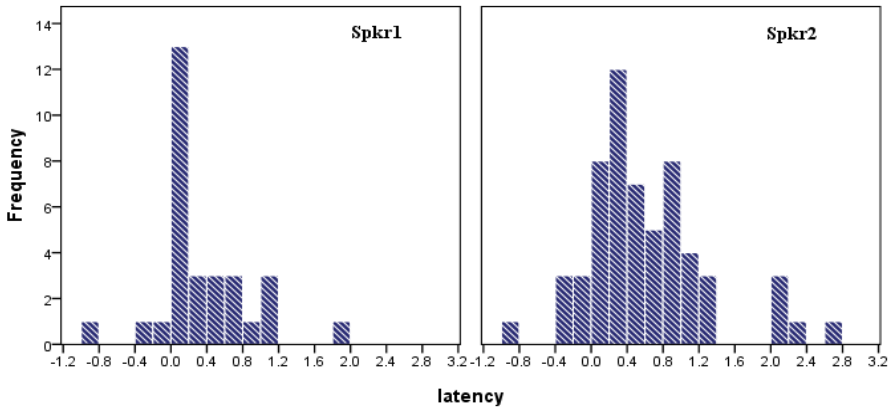

Fig. 3. Latency distributions (between the end of turn-final word and the start of the turn-initial word) in seconds for turn-initial filled pauses separately for two speakers.

Finally, Fig. 3 shows the temporal alignment of turn-initial filled pauses for the two speakers. Both speakers have a very clear peak for latencies between 0 and 0.2s, which corresponds to extremely tight temporal alignment. However, Spkr2's histogram also shows wider distribution of latencies and a second discontinuity around the values of 1 second. We interpret Spkr2's behavior as adapting her default 'loose' temporal alignment of turn-initial filled pauses to a more 'tight' alignment in an effort to secure the floor for her as a reaction to Spkr1's tactics to add more information or take the floor in the absence of such signal. The histograms together with other descriptive and quantitative observations thus support the conclusion that Spkr2 adapted her timing pattern globally during the whole conversation.

An interesting question is how this global adaptation developed over time through the dialogue. Some studies suggest that entrainment to the dialectal features of the speech of the interlocutor happens rapidly at the beginning of the conversation [26]. Yet other results point to continuous entrainment and detrainment of prosodic feature during the course of the dialogue [27]. The temporal evolution of the rhythmical entrainment patterns remains an challenging question for further research.

4 Discussion and conclusion

We presented two strategies for the adaptation in timing of turn initiations in collaborative tasks. First, speakers used the sequential local adaptations in their

adjacent turns to entrain better with the rhythm and turn-taking style of the interlocutor. Second, speakers used a global adaptation of the timing pattern developed during the entire conversation. Both of these adaptations aimed at decreasing the overlap in turn-taking. We also note that the patterns of adaptations discussed above were not automatic since they were prevalent in Spkr2's speech and almost non-existent in Spkr1's speech. Although space limitations prevent a detailed analysis of the two conversations in which the target speakers played the game with different interlocutors, we observed that the patterns we described tend to carry into the other dialogues [28].

Our results lead to the proposal that the temporal patterns in turn-initiations are cognitively meaningful and play a role in constructing an asymmetrical dominance relationship between the speakers. This research shows that future autonomous, adaptive, and context-sensitive dialogue systems should be flexible enough to incorporate both local and global adaptations to the rhythmical and temporal patterns of the interlocutor.

Acknowledgments. This work was supported in part by the Slovak Ministry of Education grant KEGA 3/6399/08 and the Slovak Research and Development Agency project APVV-0369-07, and was done in collaboration with Augustín Gravano and Julia Hirschberg.

References

1. Schegloff, E.: Sequence Organization in Interaction. CUP, Cambridge (2007).
2. Coates, J., Sutton-Spence, R.: Turn-taking patterns in deaf conversation. Journal of Sociolinguistics 5, 507-529 (2001).
3. Goodwin, Ch.: Transparent vision. In Ochs, E., Schegloff, E., Thompson, S. A. (Eds) Interaction and Grammar, pp. 370-404, CUP, Cambridge, (1996).
4. Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematic for the organization of turn-taking for conversation. Language 50, 696-735 (1974).
5. Schegloff, E.: Overlapping talk and the organization of turn-taking for conversation. Language and Society, 19, 1-63 (2000).
6. Ford, C., Thompson, S.: Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, E., Schegloff, E., Thompson, S. A. (Eds) Interaction and Grammar, pp. 134–184, CUP, Cambridge, (1996).
7. Pickering, M., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169-226 (2004).
8. Couper-Kuhlen, E.: English Speech Rhythm. John Benjamins, Amsterdam (1993).
9. Auer, P., Couper-Kuhlen, E., Müller, F.: Language in Time. OUP, Oxford (1999).
10. Ward, A., Litman, D.: Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. Proceedings of SLaTE Workshop, Farmington, PA (2007).
11. Pardo, J.: On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119(4), 2382–2393 (2006).
12. Aubanel, V., Nguyen, N.: Automatic recognition of regional phonological variation in conversational interaction. Speech Communication (in press).
13. Shockley, K., Santana, M., Fowler, C.: Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception & Performance, 29, 326-332 (2003).

14. McFarland, D.: Respiratory markers of conversational interaction. Journal of Speech, Language, & Hearing Research, 44, 128-143 (2001).
15. Gravano, A.: Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue. Unpublished Ph.D. thesis, Columbia University, NY (2009).
16. Beňuš, Š.: Are we 'in sync': Turn-taking in collaborative dialogues. Proceedings of 10th INTERSPEECH. ISCA, Brighton, UK, pp. 2167-2170. (2009)
17. Beckman, M., Hirschberg, J., 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, pp. 9–54, OUP, Oxford, (2004).
18. Beattie, G.: Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. Semiotica, 39(1/2), 93-114 (1982).
19. Gravano, A., Hirschberg, J.: Turn-Yielding Cues in Task-Oriented Dialogue. In Proceedings of SIGDIAL, Association for Computational Linguistics, pp. 253–261 (2009).
20. Bull, M.: An analysis of between-speaker intervals. In Cleary, J., Moll&aacute-Aliod, D. (Eds.) Proceedings of the Edinburgh Linguistic Conference. pp. 18-27 (1996).
21 Cummins, F., Port, R.: Rhythmic constraints on stress-timing in English. Journal of Phonetics 26(2), 145-171 (1998).
22. Laver, J.: Principles of phonetics. CUP, Cambridge (1994).
23. Carpenter, S., O'Connel, D.: More than meets the ear: Some variables affecting pauses. Language & Communication, 8(1), 11-27 (1998).
24. Swerts, M.: Filled pauses as markers of discourse structure. J. of Prag., 30, 485-496 (1998).
25. Steward, O., Corley, M.: Hesitation disfluencies in spontaneous speech: The meaning of um. Language and Linguistics Compass, 4, 589-602 (2008).
26. Delvaux, V., Soquet, A.: The influence of ambient speech on adult speech productions through unintentional imitation, Phonetica, 64, 145–173 (2007).
27. Edlund, J., Heldner, M., Hirschberg, J.: Pause and gap length in face-to-face interaction. In Proc. of Interspeech 2009. Brighton, UK (2009).
28. Beňuš, Š., Gravano, A., Hirschberg, J.: Pragmatic aspects of temporal alignment in turn-taking, Journal of Pragmatics (submitted).