INTERRUPTION IMPOSSIBLE

Mattias Heldner, Jens Edlund and Rolf Carlson

1 Introduction

Most current work on spoken human-computer interaction has so far concentrated on interactions between a single user and a dialogue system. The advent of ideas of the computer or dialogue system as a conversational partner in a group of humans, for example within the CHIL-project[1] and elsewhere (e.g. Kirchhoff & Ostendorf, 2003), introduces new requirements on the capabilities of the dialogue system. Among other things, the computer as a participant in a multi-part conversation has to appreciate the human turn-taking system, in order to time its' own interjections appropriately. As the role of a conversational computer is likely to be to support human collaboration, rather than to guide or control it, it is particularly important that it does not interrupt or disturb the human participants.

The ultimate goal of the work presented here is to predict suitable places for turn-takings, as well as positions where it is impossible for a conversational computer to interrupt without irritating the human interlocutors.

1.1 Turn-taking

In short, the turn-taking system gives a current or active speaker in a conversation at least two possibilities at the end of each utterance: to indicate that he/she wants to continue speaking, that is to keep the turn (turn-keeping); or to indicate that some other participant should take the next turn (turn-yielding). A possible next speaker (which may be the computer) also has several options: he/she/it may await a turn-yielding cue and let the current speaker finish the turn; interrupt the current speaker without wanting to take the turn, for example to backchannel or to make an aside; or interrupt the current speaker to take the turn. A common behavior is apparently that a possible next speaker politely waits until the current speaker has finished.

Turn-taking has furthermore been observed to occur so that the gap, as well as the temporal overlap between speakers is minimized. This observation in turn has

[1] CHIL "Computers in the Human Interaction Loop" is an Integrated Project under the European Commission's Sixth Framework Program. See also http://chil.server.de/.

been taken as evidence that the end of a current speaker's turn is projectable or predictable to the next speaker (e.g. Levinson, 1983). A number of factors have been claimed relevant to the predictability of turn endings, including syntactic, semantic or pragmatic completeness; visual cues such as gaze, head nods, hand gestures, and facial expressions; and prosodic cues including boundary tones and accents, but also silent pauses, speaking rate, voice quality etc.

1.2 Prosodic boundaries

A prosodic function closely related to turn-endings is that of prosodic boundaries. When a speaker is allowed to finish what he/she intends to say, the end of the turn by nature coincides and co-occurs with prosodic boundaries of some kind. Thus, turn-endings and prosodic boundaries share acoustic prosodic cues. However, that is not to say that all prosodic boundaries also constitute possible places for turn-takings. Although there is no general agreement as to the number and types of prosodic boundaries, most researchers agree that there are at least two different boundary strengths apart from no boundary, and hence at least two kinds of prosodic boundary (e.g. Beckman & Ayers Elam, 1997; Buhmann et al., 2002; Wightman & Rose, 1999).

Prosodic boundaries are largely predictable to listeners from the left-hand context only, that is before hearing any silent pause after the word. Perceptual data indicate that the strength of a prosodic boundary, and hence the presence or absence of a boundary, is predictable from a context of only one word (Carlson & Swerts, 2003a, 2003b), and prosodic rather than lexico-grammatical information seems to be the primary cue for making these predictions (Carlson, Hirschberg, & Swerts, 2004).

There have also been several attempts to detect prosodic boundaries automatically using vectors of prosodic features (e.g. Batliner et al., 2001; Campbell, 1993; Heldner & Megyesi, 2003; Wightman & Ostendorf, 1994). These feature vectors are typically intended to capture prosodic phenomena such as accents and boundary tones, final lengthening and silent pauses, whose relevance for turn-taking are also widely recognized (e.g. Caspers, 2003; Wells & MacFarlane, 1998).

In this contribution, we are going to focus on certain prosodic aspects of turn-taking by exploring the relations between turn-taking and prosodic boundaries. Two experiments have been carried out: a listening test where subjects rated the appropriateness of made-up turn-takings, and a production experiment where another group of subjects were asked to indicate suitable places for turn-takings.

2 Listening test

2.1 Method

Both the listening and the production experiments used speech material from a seminar on speech technology. A lecturer of German origin gave this seminar in English, that is in something we could call 'European English'. The lecturer spoke most of the time, but he was interrupted by questions from the audience on a few occasions. An excerpt of the first five minutes of the seminar (667 words) was used in the experiments.

To enable an exploration of the relations between turn-taking and perceived prosodic boundaries, the speech material was manually annotated using a three-level convention developed within the GROG-project (Heldner & Megyesi, 2003). Each orthographic word was classified as being followed by either a weak or a strong boundary, or as not followed by any boundary. The first author annotated the entire material in three independent sessions, timed a few days apart. The majority votes of the three sessions were taken as the final classification of the prosodic boundaries. The agreement and Kappa figures for this task were 92%, and 79%, respectively. This annotation procedure resulted in 503 words being classified as not followed by any boundary, 118 words as followed by a weak boundary, and 46 words as followed by a strong boundary.

The listening test was carried out to evaluate made-up turn-takings in different prosodic boundary conditions. Each stimulus in the listening test consisted of a turn-taking made-up of a fragment from the seminar spoken by the lecturer followed by a fragment of a question from somebody in the audience.

The lecturer parts were determined using the annotations of perceived boundaries. Ten words followed by strong boundaries, ten words followed by weak boundaries, and ten words not followed by any prosodic boundary were randomly selected among those that received unanimous classifications in the annotation sessions. The right-hand word boundaries of these words formed the endpoints of the thirty lecturer parts, that is the turn-taking positions to be evaluated. The starting points were simply the endpoints of the previous part, or in the case of the first one, the beginning of the file. Thus, any silent pauses after the words were excluded from the end of the lecturer parts and included in the beginning of the next part.

The question that followed was always the same rendering of *"what about <um> could you give us some <hrm> rough idea what"*. Thus, the stimuli contained made-up turn-takings where the lecturer parts were always different and the question part was always the same.

These stimuli were presented in a listening test where the task was to rate whether the questioner enters the conversation at an appropriate place on a five-point scale, where 1 is an inappropriate and 5 an appropriate occasion for asking a

question. The subjects were asked to try to ignore the fact that the same speaker repeatedly uttered the same question, as well as the relevance of the question in the context. The stimuli were presented over headphones, and the order was randomized individually for each subject. The subjects could repeat each stimulus as many times they wanted before making their judgment.

Twenty subjects, eleven men and nine women, were recruited from the staff at the Department of Speech, Music and Hearing at KTH to serve as subjects in the listening experiment. The experiment took about 10 minutes for each subject. They were not rewarded for their participation.

2.2 Results

The listening test was carried out to obtain scores of the appropriateness of the turn-takings occurring in no boundary, weak boundary and strong boundary positions, as well as for the individual stimuli. Figure 1 shows the distribution of judgments for the three boundary conditions.

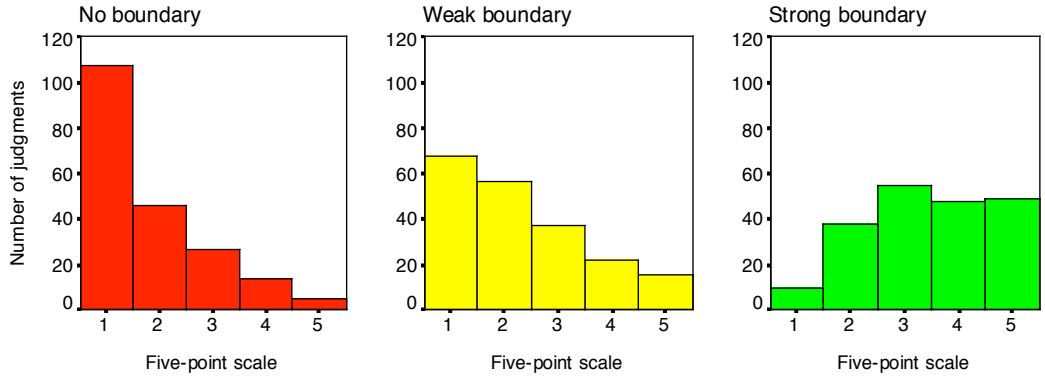

Figure 1. The distribution of judgments on a five-point scale (where 1 represents an inappropriate and 5 an appropriate place for asking a question) for turn-takings at no boundaries, weak boundaries and strong boundaries.

As can be seen, while the most frequent judgment for the no boundary as well as for the weak boundary conditions was ‘an inappropriate place for asking a question’ (i.e. a 1-vote), the most frequent one for the strong boundary condition was intermediate between inappropriate and appropriate (i.e. a 3-vote). The strong boundary stimuli also received a fair number of 5-votes, that is ‘an appropriate place for asking a question’.

Consequently, the mean scores for the three boundary conditions (shown in Table 1) show that the listeners generally judged the turn-takings in strong boundary positions to be more appropriate than those in weak or no boundary conditions, and that weak boundaries were slightly better than no boundaries. The

mean of the whole experiment was below the mid-point of the scale, thus there was a negative bias. Furthermore, as the distributions for no boundary and weak boundary were positively skewed (1.23 and 0.70), while the distribution for strong boundary was slightly negatively skewed (-0.21), the mean scores underestimate the differences between boundary conditions.

Table 1. Mean scores and standard deviations for the three boundary types, as well as for the whole experiment (All positions).

	Mean score	Std dev
No boundary	1.81	1.07
Weak boundary	2.31	1.26
Strong boundary	3.44	1.19
All positions	2.52	1.36

Yet Figure 1, as well as the standard deviations in Table 1 reveal that there was considerable variation within the boundary conditions. An analysis of the mean scores for the individual stimuli shows that all the strong boundary stimuli received higher scores than the mean of the whole experiment. Furthermore, nine out of ten no boundary stimuli got lower mean scores than the mean for the experiment. The weak boundary stimuli displayed more variation with one stimulus receiving the second highest score and another the second lowest score. Still, seven out of ten weak boundary stimuli got lower mean scores than the mean of the experiment.

2.3 Discussion

The listening test showed that turn-takings at strong prosodic boundaries were generally judged to be somewhere between acceptable to appropriate, while those at no and weak boundaries were generally judged inappropriate to acceptable. There were a few exceptions from the general trend. Two of the no and weak boundary stimuli that received higher scores than the mean of the whole experiment gave the impression of the questioner starting simultaneously with the lecturer after a long silence. Many of the listeners commented on these stimuli, and those who interpreted them as unintentional simultaneous starts judged them to be fairly appropriate. Furthermore, the turn-takings in the two other weak boundary stimuli that received higher scores than the mean of the whole experiment both occurred in enumerations where semantic and pragmatic factors might have made it plausible to interrupt, for example with additions to the list. However, for a conversational computer to be able to predict that these weak

boundary positions are acceptable places for turn-takings it is likely that more than just prosodic information will be needed.

3 Production experiment

3.1 Method

The production experiment was carried out to let the subjects indicate possible places for turn-takings in the same speech material that was used in the listening experiment. Here, each trial consisted of the subjects listening to a lecturer part of the seminar and as soon as they thought it was appropriate to take the turn they pressed a key, the sound of the lecturer stopped, and the question *"what about <um> could you give us some <hrm> rough idea what"* was played. Subsequently, the sequence of lecturer part and question was repeated and the subjects had to rate whether the trial was successful (discard, keep), the timing of the turn-taking on a micro level (early, ok, late), as well as the politeness of the turn-taking (rude, neutral, polite). The subjects could repeat the lecturer-question sequence as many times they wanted before making their judgment but they could not change the placing of the turn-taking.

The starting points of the lecturer parts were the same as in the listening experiment. However, as the sound continued playing until the subjects pressed a key, the lecturer parts could both be longer and shorter than those in the listening test. Consequently, some parts of the seminar may have been played more than once while others might never have been played. The experiment was presented over headphones, there were thirty trials (or starting points), and the order of the lecturer parts was randomized individually for each subject.

Another group of subjects from our department, this time five men and three women, were recruited to participate in the production experiment. This experiment took about 15 minutes for each subject. They were not rewarded for their participation.

3.2 Results

There were eight subjects doing thirty trials each, and a total of 240 trials. However, sixteen out of these were discarded because the subjects had indicated that their productions were not successful, leaving 224 trials for analysis. The possible places for turn-takings indicated by our subjects were analyzed in terms of whether they occurred at a weak or a strong boundary, or not at any boundary. In addition, the gap at the turn-taking was measured.

It turned out that there was a clear relation between turn-taking positions and prosodic boundaries: strong boundary positions were selected in 173 or 77% of

the trials; weak boundaries were selected in 43 or 19% and no boundaries in 8 or 4% of the cases. Furthermore, most of the strong boundaries occurring in the experiment were also marked as possible places for turn-taking. Due to the experimental setup, only 37 out of the 46 strong boundaries in the annotation were ever presented to any subject, and 31 or 84% of these constituted possible places for a turn-taking according to at least one of the subjects. A comparison with the listening test furthermore showed that eight out of the ten strong boundary positions evaluated there were also picked as possible turn-taking positions in the production experiment. The remaining two strong boundary positions were preceded by several other strong boundaries and these were picked instead.

Table 2. Mean gap at the turn-taking (in seconds) and standard deviations for the three boundary types.

	Mean	Std dev
No boundary	0.02	0.07
Weak boundary	0.64	0.50
Strong boundary	0.99	0.53

The analysis of the timing of the turn-takings revealed that the gap at the turn-taking (i.e. the distance from the end of the lecturer fragment to the onset of the question) was sometimes substantial, while there were no cases of temporal overlap. The analysis moreover indicated that the duration of the gap was dependent on the boundary type. As can be seen in Table 2, the average gap was considerably longer at strong boundaries than at weak ones, while it (as might have been expected) was negligible at no boundaries. Although there were substantial differences between the subjects (accounting for some of the variance in Table 2), the relations remained the same for all of them. There were no indications of training effects. The observed gap durations, furthermore, resemble those for the silent pauses occurring in the same positions. The silent pauses at strong boundaries were on average 1.2 seconds, those at weak boundaries 0.5 seconds, while there were not any silent pauses at the no boundaries.

3.3 Discussion

The production experiment support the findings from the listening test by showing a strong preference for turn-takings at strong boundaries, although turn-takings may also occur at certain weak and no boundaries according to our subjects. Furthermore, as nearly all strong boundaries occurring in the experiment were marked as possible turn-taking positions, it is reasonable to assume that strong

boundaries generally make up appropriate turn-taking positions, at least in this communicative situation.

Furthermore, the tendency to minimize the gap at turn-takings (c.f. Levinson, 1983) does not seem to be the only principle at work here, and clearly, it is not a matter of minimization at the fraction of a second level. Rather, it seems that our subjects varied the duration of the gap depending on the boundary strength at the turn-taking position. We can only speculate about the reasons for this, but the gap durations may be governed by rhythmical principles. First, the shorter gaps at weak boundaries and longer gaps at strong boundaries resemble the realization of (within speaker) boundaries in spontaneous dialogue, where weak boundaries have relatively shorter silent intervals and more final lengthening than strong boundaries (e.g. Heldner & Megyesi, 2003). Second, Fant & Kruckenberg (1989) suggest that the sum of final lengthening and silent interval duration is planned to conform with one or two or more rhythm units. Following this line of thought, it might be that the speaker taking the turn somehow tries to get into time with the previous speaker by adjusting the gap duration.

Furthermore, it is possible that the experimental situation introduced a delay from the moment our subjects decided on a turn-taking position until they actually pressed the key. If that is the case, the gap durations reported here will not reflect those in a genuine conversation. However, we see no reason why such a delay in the motor control should be affected by boundary strengths in somebody else's speech. So at least the relations between gaps at weak and strong boundaries ought to remain constant.

4 General discussion

This study gives experimental support for a close relation between turn-taking and prosodic boundaries. In the communicative situation represented by the speech material, that is a seminar with mostly monologue speech and a few questions from the audience, the most appropriate place to take the turn (i.e. to pose a question) is at a strong prosodic boundary. Furthermore, it seems that interruptions are generally if not impossible, then at least less appropriate at no and weak boundaries, and that semantic and pragmatic factors need to be taken into account in order to decide whether no or weak boundaries are possible places for a turn-taking.

Thus, if we can predict (or detect) strong prosodic boundaries based on acoustic prosodic features – something which previous research makes plausible (e.g. Carlson et al., 2004; Carlson & Swerts, 2003a, 2003b) – we can also predict suitable positions for turn-takings. For the purpose of a conversational computer, a rule restricting the interjections by the computer to strong boundaries will surely miss a few possible places, but the ones it will select will be appropriate.

In future work we will explore what acoustic prosodic features to use for prediction of strong prosodic boundaries, as well as the relative importance of phenomena such as boundary tones, final lengthening, creaky voice, and silent pauses.

Another line of future work will be to look into whether listeners prefer that the timing of turn-takings be governed by the boundary strength or perhaps by some underlying rhythm in speech. To this end, other listening tests with systematic variation of the gap duration will be needed, as well as a working operationalization of speech rhythm.

Acknowledgments

This work was done within the Project CHIL "Computers in the Human Interaction Loop" (IP 506909). CHIL is an Integrated Project under the European Commission's Sixth Framework Program. The research was carried out at the Centre for Speech Technology supported by Vinnova (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

5 References

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V. and H. Niemann (2001). Duration features in prosodic classification: Why normalization comes second, and what they really encode. In M. Bacchiani, J. Hirschberg, D. Litman & M. Ostendorf (Eds.), *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding: Prosody 2001* (pp. 23-28). Red Bank, NJ: ISCA.

Beckman, M. E. and G. Ayers Elam (1997). *Guidelines for ToBI labelling*. Retrieved 2002-01-11, from http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/

Buhmann, J., Caspers, J., van Heuven, V. J., Hoekstra, H., Martens, J.-P. and M. Swerts (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In *Proceedings LREC*. Las Palmas.

Campbell, N. (1993). Automatic detection of prosodic boundaries in speech. *Speech Communication, 13*, 343-354.

Carlson, R., Hirschberg, J. and M. Swerts (2004). Prediction of upcoming Swedish prosodic boundaries by Swedish and American listeners. In *Proceedings Speech Prosody 2004*. Nara, Japan.

Carlson, R. and M. Swerts (2003a). Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials. In *Proceedings ICPhS 2003* (pp. 79-82). Barcelona, Spain.

Carlson, R. and M. Swerts (2003b). Relating perceptual judgments of upcoming prosodic breaks to F0 features. In *Proceedings from Fonetik 2003* (pp. 181-184). Umeå: Dept. Philosophy and Linguistics, Umeå University.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics, 31*, 251-276.

Fant, G. and A. Kruckenberg (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*(2), 1-83.

Heldner, M. and B. Megyesi (2003). Exploring the prosody-syntax interface in conversations. In *Proceedings ICPhS 2003* (pp. 2501-2504). Barcelona.

Kirchhoff, K. and M. Ostendorf (2003). Directions for multi-party human-computer interaction research. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing* (pp. 7-9). Edmonton.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Wells, B. and S. MacFarlane (1998). Prosody as an interactional resource: turn projection and overlap. *Language and Speech, 41*(3-4), 265-294.

Wightman, C. W. and M. Ostendorf (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing, 2*(4), 469-481.

Wightman, C. W. and R. C. Rose (1999). Evaluation of an efficient prosody labeling system for spontaneous speech utterances. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Keystone, Colorado, USA.