# On cue — additive effects of turn-regulating phenomena in dialogue

**Anna Hjalmarsson**
Centre for Speech Technology, KTH
Stockholm, Sweden
annah@speech.kth.se

## Abstract

One line of work on turn-taking in dialogue suggests that speakers react to "cues" or "signals" in the behaviour of the preceding speaker. This paper describes a perception experiment that investigates if such potential turn-taking cues affect the judgments made by non-participating listeners. The experiment was designed as a game where the task was to listen to dialogues and guess the outcome, whether there will be a speaker change or not, whenever the recording was halted. Human-human dialogues as well as dialogues where one of the human voices was replaced by a synthetic voice were used. The results show that simultaneous turn-regulating cues have a reinforcing effect on the listeners' judgements. The more turn-holding cues, the faster the reaction time, suggesting that the subjects were more confident in their judgments. Moreover, the more cues, regardless if turn-holding or turn-yielding, the higher the agreement among subjects on the predicted outcome. For the re-synthesized voice, responses were made significantly slower; however, the judgments show that the turn-taking cues were interpreted as having similar functions as for the original human voice.

## 1 Introduction

This paper describes a perception experiment that investigates the probabilities of who will be the next speaker based on potential "cues" in the behaviour of the previous speaker. The experiment was designed as a game where the subjects were asked to listen to two-party dialogues and, whenever the recording halted, guess who would be the next speaker. The aim is to investigate if combinations of simultaneous cues affect the confidence of listeners' judgments. The results also have implications for spoken dialogue system (SDS) research; If SDS can signal turn completion or non-completion in a way that can be easily discriminated by humans, turn-transitions in such systems could be made more intuitive. Thus, a secondary aim of this study is to test if the cues can be reproduced in a synthetic voice. Both human-human dialogues and dialogues where one of the human voices was replaced with a synthesized voice were tested.

### 1.1 Incremental language processing

Spoken dialogue systems that opt for human-likeness (Edlund et al., 2008) should be flexible and allow their users to hesitate and revise their speech in a way that is similar to interacting with a human dialogue partner. However, turn management in current SDS is in general not very sophisticated. One frequent strategy is to interpret long silences, above a certain threshold (Ferrer et al., 2002), as end of user turn. Thus, the system still risks barging in over its users because of the large variance in silence duration for spontaneous speech (Campione & Veronis, 2002). Faster processing of input only partly solves the problem, since the response delay due to end of turn detection is still not targeted. In fact, perceiving, planning and producing speech is time consuming for humans too, but we have strategies to avoid long ambiguous silences. First, we start to plan new contributions before the other person has stopped speaking. When starting to speak, we typically do not have a complete plan of what to say but yet we mange to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions.

If behavioural cues related to these human strategies can be identified, we can employ similar methods in SDS. The objective is to indicate to the user that the system plans to continue speaking and by doing this avoid user confusion regarding whether the ongoing system utterance is complete or not. In a similar fashion, the system also needs strategies to efficiently signal and detect end of turns.

## 1.2 Turn taking in spoken dialogue

Humans are expected to produce new dialogue contributions within a certain time. Then again, speech is not generated in regular constant pace of vocalized segments, but in streams of fragments in varying sizes (Butterworth, 1975). In addition, spontaneous dialogue involves unexpected interruptions or disfluencies such as pauses, corrections and repetitions that we use to refine, alter, and revise our plans as we speak (Clark & Wasow, 1998). Despite its irregularities, we only talk simultaneously for brief periods of time (Schegloff, 2000). Sacks et al. (1974) suggest that this is viable because humans have a mutual understanding of transition relevance places (TRPs). A frequent assumption is that humans can predict these TRPs almost exactly and that a majority of speaker shifts are directly adjoining without any overlap or silence. If this is true, interlocutors are able to predict approaching end of turns in advance very precisely. These TRPs are claimed to be detected in terms of expected end points of semantic or lexical units (e.g. de Ruiter et al., 2006). Yet, analysis of turn transitions in American English, German and Japanese have shown that pauses and overlaps are normally (Gaussian) distributed (Weilhammer & Rabold, 2003), suggesting that perfectly adjoining transitions are rare.

## 1.3 Turn management signals

An early series of works on turn-taking (cf. Duncan, 1972; Duncan & Fiske, 1977) suggest that interlocutors react to a set of signals employed by the previous speaker to indicate approaching turn endings. According to Duncan (1972 p.283): "The proposed turn-taking mechanism is mediated through signals composed of clear-cut behavioural cues, considered to be perceived as discrete". Analysis of dialogues showed that the number of available turn-yielding cues was linearly correlated with listeners' turn taking attempts. However, if speakers' employed signals to suppress such attempts the number of turn-taking attempts radically decreased, regardless of the number of turn-yielding signals.

#### Cues relevant for turn-taking

Turn-holding cues are those referred to as attempt-suppressing signals by Duncan. This type of cue indicates that the speaker intends to hold the turn. Turn-holding cues reported by Duncan include drawl on the final syllable (phrase-final lengthening), an intermediate pitch level and sociocentric sequences (stereotyped lexical expressions or cue phrases). Turn-yielding cues reported include rising or falling pitch, the termination of a hand gesture, a drop in loudness and completion of grammatical pauses. Recent work by Gravano (2009) presents a number of phenomena found to take place at significantly higher frequencies before speaker switches. These cues include a falling or high-rising intonation, a reduced lengthening, a lower intensity level, a lower pitch level, points of textual completion, a higher frequency of jitter, shimmer and noise-to-harmonics ratio and longer inter-pausal unit duration. Moreover, in line with Duncan's findings, Gravano's show strong support for a linear relationship (positive correlation) between the number of simultaneously available turn-yielding cues and the number of turn-taking attempts. In line with Gravano and Duncan, this work further investigates how discrete cues form a complex signal that guides interlocutors' turn-taking behaviour in dialogue.

Duncan and followers have mainly focused on describing correlates of actual turn-taking behaviour. Nonetheless, there is a range of acceptable behaviours; some may be perceived as impolite, yet effective if speakers get their points across. Consequently, interlocutors have the choice to act "hazardously" and defy the "principles" of turn-taking. For example: speakers can choose to barge in at less suitable places and avoid taking the floor when expected to. Bearing this in mind, in this experiment we explore the probabilities of different outcomes regardless of the outcome of the original dialogue. Schaffer (1983) and Oliveira & Freitas (2008) approached turn-taking issues from a similar perspective, i.e. analyzing the judgments of non-participating listeners in perceptual experiments. However, while their approach was to isolate the signals presented to the subject, our approach is to label the cues separately and subsequently study their combined effect on listeners' judgments in a context that is similar to listening in on someone else's conversation. As pointed out by Oliveira & Freitas (2008), analyzing dialogues outside of their contexts is problematic; yet, by allowing the subjects in our study to follow dialogues incrementally in chronological order rather than listening to disconnected phrases, we hope to overcome some of these problems.

It should also be mentioned that the outcome of turn-yielding signals is difficult to predict. Even if the previous speaker directs questions

with the intention of eliciting a specific response or feedback from a listener, this participant may choose not to take the floor. Then again, often turn-yielding cues merely signal completion of a turn and leave the floor open. This means anyone may take the floor, including the previous speaker, whereas if a speaker signals turn holding intentions, the outcome is more predictable. From a SDS perspective, being able to suppress turn-taking attempts and discriminate users' internal pauses from turn completions is useful knowledge.

Duncan has been criticised for not reporting any inter-annotator agreement or formal description of his "signals" (Beattie et al., 1982). Whether the phenomena that Duncan refers to as "signals" should be considered as conscious or not is problematic. There are for example acoustic cues, e.g. drop in energy or inhalations that guide interlocutors in their turn-taking. However, a likely origin of these "signals" is the anatomy of our speech organs. If we plan to continue speaking, we keep the speech organs prepared and if we plan to finish, we release them (Local & Kelly, 1986). In this paper, all perceivable phenomena relevant for turn-taking are referred to as cues, regardless if they are conscious or not.

## 2 Dialogue data

The dialogues used as stimuli in this experiment were collected in order to obtain data in the DEAL domain. DEAL is a spoken dialogue system for conversation training for second language learners of Swedish under development at KTH. The scene of DEAL is set at a flea market where a talking animated agent is the owner of a shop selling used objects. The objectives are to build a system which is fun, human-like, and engaging to talk to, and which gives language learners conversation training (Hjalmarsson et al., 2007). The recorded dialogues are informal, human-human, face-to-face conversation in Swedish. The task and the recording environment were set up to mimic the DEAL domain and role-play. The corpus includes eight dialogues with six different speakers. All together about two hours of speech were collected. The dialogues were transcribed orthographically including non-lexical entities such as laughter, repetitions, filled pauses, lip-smacks, breathing and hawks. Two annotators labelled the data for cue phrases (CP) with high inter-annotator agreement (kappa 0.82) (Hjalmarsson (2008)). Cue phrases (also frequently referred to as discourse markers) are linguistic devices used to signal relations between different segments of speech. The cue phrases used here were phrases labelled to have either *response eliciting* or *additive* discourse pragmatic functions. Examples of response eliciting CPs are "eller hur" (right) and "då" (then) and examples of additive CPs are "och" (and), "eller" (or) and "men" (but). Though commonly not categorized as such, we also included filled pauses in this category. Response eliciting CPs were expected to have turn-yielding functions, while the additive CPs and the filled pauses were expected to have turn-holding functions.

The transcripts from four dialogues were also time-aligned with the speech signal. This was done using forced alignment with subsequent manual verification of the timings.

### 2.1 Manual labelling of cues

The perception experiment was designed to elicit probabilities of a speaker change versus a hold regardless of the outcome of dialogue, that is, without considering its actual continuation. The four dialogues in the corpus that had been time-aligned were automatically segmented into inter-pausal units (IPUs), a sequence of words surrounded by silence longer than 200 milliseconds (ms). According to Izdebski & Shipp (1978) humans need just under 200 ms to verbally react to some stimulus, which suggests that the speakers in the original dialogues had enough time to react to any potential cues in the end of previous IPU. For shorter silences or in overlapping speech it was impossible to halt the recordings without revealing to the subjects who the next speaker was. The four dialogues contained 2011 such silences, of which 85% were internal pauses and 15% were silences between speakers. Henceforth, silences within speaker turns will be referred to as *pauses* while silence between speakers will be referred to as *gaps*. The terminology is adopted from Heldner, M., & Edlund, J. (submitted).

To distinguish and explore all cues claimed to be relevant for turn-taking is beyond the scope of this paper. Since the focus is on the contributive effect of simultaneously occurring cues, the number of cues was restricted to five categories. The five categories were pitch contour, semantic completeness, phrase-final lengthening, non-lexical elements such as perceivable breathing and lip-smacks and some frequently occurring cue phrases (see Table 1). The dialogues recorded were face-to-face interactions

that most likely contain visual turn-management cues such as hand and facial gestures. However, the visual gestures were not considered here and the labellers and subjects only had access to the audio recordings. The reason for this was to focus on the lexical and acoustic cues that can potentially be reproduced in a synthetic voice. Reported differences between face-to-face and telephone conversation are longer duration of silences in face-to-face interaction (Bosch et al., 2004). However, if Duncan's observations are correct, the more cues available, regardless of modality, the more predictable is the outcome.

| Category | Turn-yielding cues | Turn-holding cues |
|---|---|---|
| Pitch contour | fall | flat |
| Final lengthening | no | long |
| Non-lexical | Audible expirations | Audible inhalations, lip-smacks |
| Cue phrases | response eliciting CPs | Additive CPs, filled pauses |
| Semantic completeness | complete | incomplete |

Table 1 : Cue categories

Deciding what is a cue is problematic. To consider a parameter as a cue implies that its receiver perceives it or at least that it is perceivable by some other human in the same context. To tackle this problem we used two annotators for all parameters and only parameters that both annotators agreed upon were considered as cues. As follows, the absence of a cue does not necessarily entail its opposite, it simply means the labellers did not perceive the cue or that they did not agree on which category it belonged to. However, the cues are exhaustive and cannot contain yielding and holding functions in the same dimension. As discussed in Ward (2006), knowing where to look and how other prosodic features interact with the relevant cue is problematic. To focus on signals that are perceivable by humans in a dialogue context the labellers did not have any visual representations of the sound. Each labelling task included only the target parameter and no turn-taking issues were considered during labelling. The cues were labelled independently, one by one, in an attempt to avoid influences from other cues. Still, for the prosodic cues, other auditory cues could not be excluded from the recordings used for labelling.

**Pitch slope**

For pitch slope, the task was to label flat, rising or falling pitch contour. This roughly corresponds to ToBi labelling H-L% (plateau), H-H% (high-rise) and L-L% (falling pitch)[1]. The labellers were provided with only the last 500 ms of the IPU to avoid influences of the lexical context. Inter-annotator agreement for pitch slope was rather poor (kappa 0.36). However, a confusion matrix revealed that the majority of the confusions were between falling and rising slope. After listening to the data, a possible explanation is that a frequently occurring contour in the data was a rising curve with a minor slope at the end that labellers may have judged differently. This suggests that a more fine-grained labelling scheme could have been used. Still, as already mentioned, only stimuli where labellers agreed were considered to contain cues. Since the literature provides no clear-cut results of the effects of a rising pitch, which appears to contain both turn-yielding and turn-holding functions (Edlund & Heldner, 2005), this was not considered a cue.

**Phrase-final lengthening**

The labelling procedure for phrase-final lengthening was almost identical to the one of pitch slope except for the target labels, which were long, short and no phrase-final lengthening. Inter-annotator agreement for this task was also poor (kappa 0.37), however, the confusion matrix suggests that the annotators' boundaries were skewed, since almost all confusions were between neighbouring categories. Minor lengthening was not considered a cue.

**Semantic completeness**

Semantic completeness represents the lexical context of the dialogues. To extract syntactically complete phrases using part of speech tagging is not feasible since utterances in dialogue often violate syntactic rules and since dialogue relies much on context that is not captured by syntax. As an alternative, labellers were asked to decide whether the last utterance was pragmatically complete or not considering the previous context. The labelling was done incrementally from the orthographic transcriptions of the dialogues without listening to the recordings. Non-lexical elements such as filled pauses and breathing had been removed from the transcripts, since they are considered to represent acoustic information — information that is already represented in other cues. The label tool only displayed the left context of the dialogue up to the silence just after the target IPU. After each judgment, the dialogue

[1] ToBi is a standard for labelling English prosody (Silverman et al., 1992)

segment up to the next target pause was provided incrementally. Inter-annotator agreement was high for this task (kappa 0.73). The labelling procedure for semantic completeness is very similar to the procedure used by Gravano (2009).

### 2.2 Stimuli selection

The task in the experiment was to guess who the next speaker was whenever the dialogue play-back halted. To allow the subjects to get familiar with the dialogue context, i.e. getting a fair understanding of the left context, the dialogue segments could not be too short. At the same time, the test should include segments from more than one dialogue, with different speakers and still not be exhaustingly long. In the final test, segments from four different dialogues ranging from 116 to 166 seconds were used based on their richness in variety of cue types and variety in cue quantity. The four dialogues included three different speakers, one male and two female. The male speaker participated in all four dialogues. In a first pilot experiment, target IPUs, i.e. stimuli in the experiment, were randomly selected, which resulted in a stimuli set that were weighted neither for the number of cues nor for the distribution of gaps and pauses in the overall dialogue. For the final experiment, all IPUs were labelled with cues in advance. Target IPUs were then selected from a list with cue labels without listening to the recordings. The selections were made to get IPUs that represent a weighted distribution of gaps and pauses over speakers and a variety of cues. However, it was difficult to find segments in the data that fulfilled all requirements and a perfect weighted range was impossible to obtain because some combinations did not occur in the data and it is questionable whether these are very frequent in any type of dialogues. In the end, 128 IPUs were used as stimuli (see Table 2).

| | Turn-yielding cues | | | | |
|---|---|---|---|---|---|
| Turn-holding cues | 0 | 1 | 2 | 3 | 4 |
| 0 | 8 | 18 | 17 | 4 | 1 |
| 1 | 15 | 10 | 2 | 1 | |
| 2 | 22 | 8 | | | |
| 3 | 14 | 3 | | | |
| 4 | 4 | | | | |
| 5 | 1 | | | | |

Table 2 : Cue distribution over stimuli IPUs

### 2.3 Re-synthesis of dialogues

One motivation for this work was to investigate whether the cues could be reproduced in a synthetic voice and perceived as having similar functions. In order to test this, one party in the dialogues was replaced with a diphone synthesis. The synthetic voice was reproduced with timings from the manually verified forced alignments and fundamental frequency automatically extracted from the human voice using Expros, a tool for experimentation with prosody in diphone voices (Gustafson & Edlund, 2008). Only the male party in the dialogues was re-synthesized, since we only had access to a male diphone synthesis. Since breathing and lip-smacks could not be re-synthesized, we kept the original human realizations from the recordings.

## 3 Method

The GUI of the test (see Figure 1) included two buttons with "pacmans" and a button where the subjects could pause the test. The pacmans represented the speakers in the dialogues and, when the corresponding interlocutor spoke, the pacman opened and closed its mouth repeatedly. The subjects' task was to listen to recordings and, at each time when the recording halted, guess who the next speaker was by pressing the corresponding pacman button. The speakers in the dialogues were recorded on different channels and the movements of the face with the left position on the screen corresponded to the sound in the subject's left ear, and vice versa. To make the subjects aware that the play-back had halted, both faces turned yellow. The subjects had 3 seconds to make the response or else the dialogue would continue. Each time the recording halted, the mouse pointer was reset to its original position, in the middle of the pause button. This was done to control the conditions before each judgment to enable comparisons between the trajectories of the subjects' movements and their reaction times. The motivation was to track users' mouse events and use these as a confidence measure similar to Zevin & Farmer (2008).

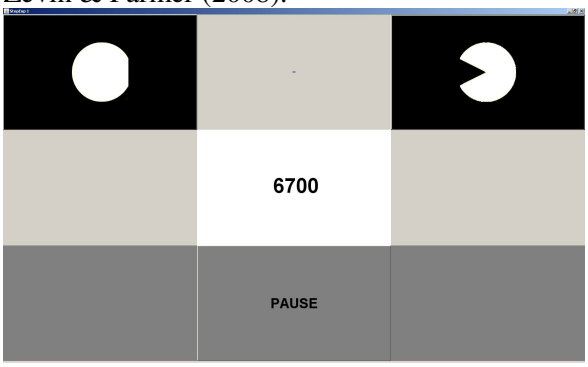


Figure 1 : Experiment GUI

The experimental setup was designed as a game where the subject received points based on whether they could guess the actual continuation

of the dialogue. To elicit judgements based on first intuition rather than afterthought, speed was rewarded. The faster subjects responded, the fewer minus points they incurred when they were wrong and the more bonus points they received if they were right. Whether they made the right choice or not was not important, but it was used as an objective rewarding system to motivate the users. Who was considered the next speaker was based on which interlocutor vocalized first, regardless of whether this was a turn-yielding attempt or only short feedback responses (back-channels). Two movie tickets were awarded to the "best" player.

### 3.1 Pilot experiment

A pilot experiment with ten subjects was conducted to test the experimental setup and features of the GUI. The reset of mouse pointer before each response did not seem to affect the subjects noticeably. In fact, some of them even claimed that they had not noticed that the pointer moved. There were, however, obvious training effects; i.e. the response times were significantly faster at the end of the test. In the final experiment, training effects were controlled for by changing the order of the dialogues. There was also a 210 seconds long training session to allow the subjects to become familiar with the task.

### 3.2 Experiment

The final experiment included 16 subjects, 9 male and 7 female, between the ages of 27 and 49. All were native Swedish speakers except for two who had been in Sweden for more than 20 years. Five of the subjects were working at the department of Speech Music and Hearing, but the majority had no experience in speech processing or speech technology. Each subject listened to two human-human dialogues and two dialogues where one party was replaced with the diphone synthesis. The re-synthesized dialogues differed between subjects.

## 4 Results

It was difficult to find dialogue segments with an equal distribution of cue types and cue type combinations. All cues were considered as having equal weight and the relative contribution of the different cues was not considered. Some cue combinations were rare (Table 2) and since small variances in the data will affect the results for these cues, cue combinations represented in less than five IPUs were excluded. Moreover, since the results from the human–human condition and the human-synthesis condition appeared to be very similar, both conditions are included in the overall results presentation.

First, IPUs with a majority of turn-holding cues were judged significantly faster than IPUs with a majority of turn-yielding cues (t-test $p<.05$). However, as already discussed in Section 1.3, the outcome of turn-holding cues is more predictable. This is also confirmed by the overall distribution of pauses versus gaps (85% respectively 15%) and the extent to which the subjects agreed on the expected outcome for the different cue categories.

### 4.1 Reaction times

Reaction times can never be negative and the maximum value (3 seconds) was set generously, well above the time needed for most judgments (the geometric mean was 1166 ms). The distribution of reaction times is therefore skewed to the left. As suggested by Campione & Veronis (2002) the log-normal law is a better fit to duration data. Reaction times were therefore transformed into a logarithmic scale (base 10). Moreover, the average reaction times differed considerably between subjects (from 933 ms to 1510 ms). The reaction times were therefore also z-normalized for each subject. The reaction times for the judgments are a likely indication of how confident the subjects were in their decision. This was supported by the fact that stimuli with high agreement, regardless of cues, were judged significantly faster by subjects (Tukey's test $p<.05$) (see Figure 2).

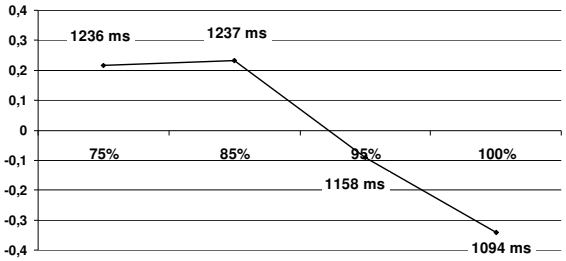


Figure 2 : Average reaction time $\log_{10}$ z-normalized over IPUs with % agreement.

For completeness, each point is labelled with its average $\log_{10}$ value (un-normalized) in milliseconds. All differences are significant, except for between 75% and 85% agreement.

The reaction times for stimuli with more turn-holding cues were significantly shorter (ANOVA $p<.05$, $df=3$). The differences are displayed in Table 3 (Tukey's test $p<.05$). IPUs with contradictory cues, i.e. both turn-yielding and turn-holding cues, are not included here. Al-

though not all steps differ significantly, there is a strong trend; the more turn-holding cues, the faster the reaction time.

| Turn-holding cues *i* | *j* | Difference in mean response time, *i* − *j* $\log_{10}$ z-value ($\log_{10}$ in ms) | Standard error | p-value |
|---|---|---|---|---|
| 0 | 1 | 0.141 (32.0 ms) | 0.07 | .382 |
| | 2 | **0.363 (89.3 ms)** | **0.06** | **.000** |
| | 3 | **0.562 (138.5 ms)** | **0.08** | **.000** |
| 1 | 2 | 0.222 (57.3 ms) | 0.08 | .079 |
| | 3 | **0.420 (106.5 ms)** | **0.09** | **.000** |
| 2 | 3 | 0.198 (49.2 ms) | 0.09 | .183 |

Table 3 : Differences in average response time between 0-1, 0-2, 0-3, 1-2, 1-3, 2-3 turn-holding cues (Tukey's p<.05, df=3). Significant differences in bold.

### 4.2 Synthesis versus natural

To present all cue combinations, including IPUs with both turn-yielding and turn-holding cues visually, three dimensional bubble charts will be used from now on. The charts display the number of turn-yielding cues on the x-axis and turn-holding cues on the y-axis.

Overall, reaction times for the synthetic voice are significantly longer (t-test p<.05). However, the reaction times decrease with an increased number of turn-holding cues in a very similar fashion as for the natural voice. This is illustrated in Figure 3. The width of the bubbles represents the z-normalized reaction times on a logarithmic scale. Unfilled bubbles represent the synthetic voice and black bubbles the human voice (the bubbles lay on top of each other). As in the overall data set (see Table 3), the reaction times for IPUs with more turn-holding cues were also significantly shorter for the synthetic voice (Tukey's p<.05).

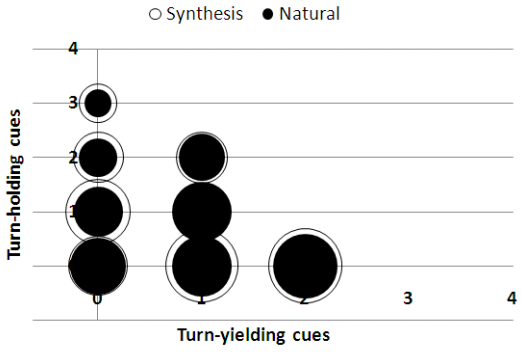


Figure 3 : Average reaction time $\log_{10}$ z-normalized for natural and synthetic voice

### 4.3 Agreement

The experiment can be viewed as a series of Bernoulli trials with dichotomous response, SWICH or HOLD. To study the effects of simultaneous cues on the actual judgments, binary stepwise logistic regression was used. The results show that there are significant relationships between turn-yielding cues and SWITCH and turn-holding cues and HOLD (p<.05). The diameters in the bubble charts in Figure 4 and Figure 5 represent % judgments for SWITCH versus HOLD for human and synthetic voice. The results show that cues are perceived as hypothesized.

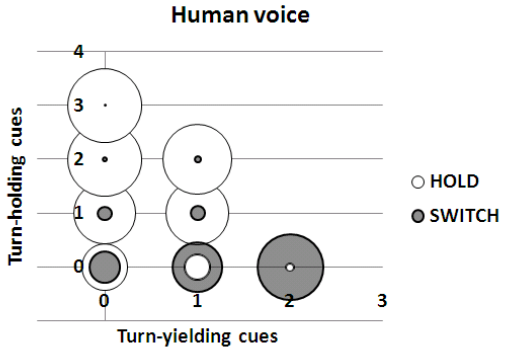


Figure 4 **:** The distribution of judgments for SWITCH versus HOLD for Human voice

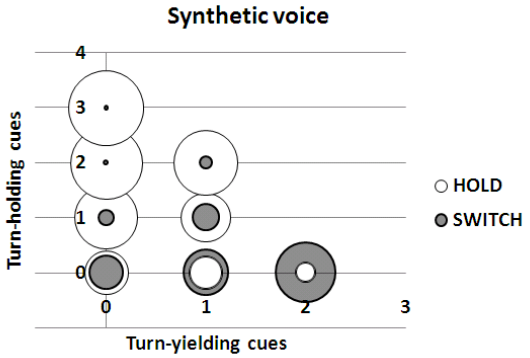


Figure 5 : The distribution of judgments for SWITCH versus HOLD for Synthetic voice

## 5 Final remarks

The results show that the turn-regulating cues are perceived as expected and in line with previous work. The novel contributions in this work include the reported reinforced effect of simultaneous lexical and non-lexical turn-regulating cues on non-participating listeners. Moreover, whereas previous research has focused on turn-yielding cues, we have also been able to present results that support a combined effect of turn-holding cues. Another important contribution is the results from re-synthesizing the human voice which suggests that these behavioural cues can be reproduced in a synthetic voice and perceived accordingly.

## 6 Acknowledgements

This research was carried out at Centre for Speech Technology, KTH. The research is also supported by the Swedish research council project #2007-6431, GENDIAL. Many thanks to Rolf Carlson, Jens Edlund, Joakim Gustafson, Mattias Heldner, Julia Hirschberg and Gabriel Skantze for help with labelling and valuable comments.

## References

Beattie, G. W., Cutler, A., & Pearson, M. (1982). Why is Mrs. Thatcher interrupted so often?. *Nature, 300*(23), 744-747.

Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Proc. of the 7th International Conference TSD 2004* (pp. 563-570). Heidelberg: Springer-Verlag.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research, Volume 4* (Number 1).

Campione, E., & Veronis, J. (2002). A large-scale multilingual study of silent pause duration. In *ESCA-workshop on speech prosody* (pp. 199-202). Aix-en-Provence.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37*(3), 201-242.

Duncan, S., & Fiske, D. (1977). *Face-to-face interaction: Research, methods and theory.* Hillsdale, New Jersey, US: Lawrence Erlbaum Associates.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica, 62*(2-4), 215-226.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

Ferrer, L., Shriberg, E., & Stolcke, A. (2002). Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody. In *Proc. of ICSLP* (pp. 2061-2064).

Gravano, A. (2009). *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. Doctoral dissertation, Columbia University.

Gustafson, J., & Edlund, J. (2008). expros: a toolkit for exploratory experimentation with prosody in customized diphone voices. In *Proc. of PIT 2008, Kloster Irsee, Germany,* (pp. 293-296). Berlin/Heidelberg: Springer.

Heldner, M., & Edlund, J. (2008). *Pauses, gaps and overlaps in conversations.* Manuscript submitted for publication.

Hjalmarsson, A., Wik, P., & Brusk, J. (2007). Dealing with DEAL: a dialogue system for conversation training. In *Proc. of SigDial* (pp. 132-135). Antwerp, Belgium.

Hjalmarsson, A. (2008). Speaking without knowing what to say... or when to end. In *Proc. of SIGDial 2008*. Columbus, Ohio, USA.

Izdebski, K., & Shipp, T. (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research, 21*, 638-651.

Local, J., & Kelly, J. (1986). Projection and "silences": Notes on phonetic and conversational structure. *Human studies, 9*(2-3), 185-204.

Oliveira, M., & Freitas, T. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. In *Speech Prosody 2008* (pp. 485). Campinas, Brazil.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics, 11*, 243-257.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29*(1), 1-63.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. In *ICSLP'92*. Banff, Canada.

Ward, N. (2006). A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic, In *Proc. of Interspeech*, Pittsburgh, Pennsylvania, USA.

Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. In *ICPhS 2003*. Barcelona, Spain.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-578). Chicago.

Zevin, J., & Farmer, T. (2008). Similarity Between Vowels Influences Response Execution in Word Identification. In *Proc. of Interspeech*. Brisbane, Australia.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language, 82*(3), 515-535.