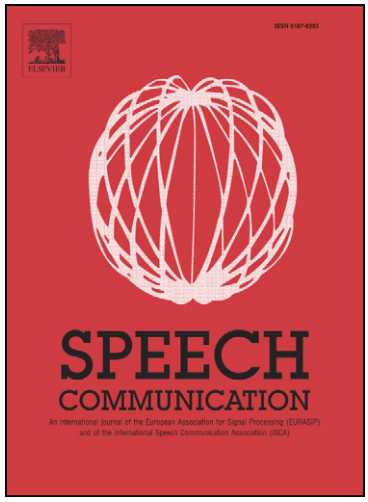
The additive effect of turn-taking cues in human and synthetic voice

Anna Hjalmarsson



The additive effect of turn-taking cues in human and synthetic voice

Anna Hjalmarsson

PII: S0167-6393(10)00137-8
DOI: 10.1016/j.specom.2010.08.003
Reference: SPECOM 1917

To appear in: *Speech Communication*

Received Date: 4 December 2009
Revised Date: 17 June 2010
Accepted Date: 2 August 2010

Please cite this article as: Hjalmarsson, A., The additive effect of turn-taking cues in human and synthetic voice, *Speech Communication* (2010), doi: 10.1016/j.specom.2010.08.003

The additive effect of turn-taking cues in human and synthetic voice

Anna Hjalmarsson[1]
Speech, music and hearing, KTH,
Lindstedsvägen 24, SE-100 44, Stockholm, Sweden
annah@speech.kth.se

Abstract

A previous line of research suggests that interlocutors identify appropriate places to speak by *cues* in the behaviour of the preceding speaker. If used in combination, these cues have an additive effect on listeners' turn-taking attempts. The present study further explores these findings by examining the effect of such turn-taking cues experimentally. The objective is to investigate the possibilities of generating turn-taking cues with a synthetic voice. Thus, in addition to stimuli realized with a human voice, the experiment included dialogues where one of the speakers is replaced with a synthesis. The turn-taking cues investigated include intonation, phrase-final lengthening, semantic completeness, stereotyped lexical expressions and non-lexical speech production phenomena such as lexical repetitions, breathing and lip-smacks. The results show that the turn-taking cues realized with a synthetic voice affect the judgements similar to the corresponding human version and there is no difference in reaction times between these two conditions. Furthermore, the results support Duncan's findings: the more turn-taking cues with the same pragmatic function, turn-yielding or turn-holding, the higher the agreement among subjects on the expected outcome. In addition, the number of turn-taking cues affects the reaction times for these decisions. Thus, the more cues, the faster the reaction time.

Keywords: Turn-taking; Speech synthesis; Human-like interaction; Conversational interfaces

1 Introduction

At the department of Speech music and Hearing, KTH we currently do research in the area of human-like dialogue systems. The motivation is to allow users to interact with a system in a way that is similar to interacting with a human dialogue partner (c.f. Edlund et al., 2008). One crucial aspect of these systems is to control the flow of dialogue contributions between the system and the user. Very few dialogue systems use sophisticated methods to manage turn-taking. These systems are generally poor both at detecting users' end of turns and at generating appropriate turn-taking behaviour to help users discriminate momentary pauses from ends of turns. A frequently used strategy is to interpret long silences as end of turns. Whereas silence is an explicit, unambiguous indication that a speaker is momentarily not vocalizing, it is a crude detector of end of turns as pause length *within* turns varies. For dialogue systems in English, the silence threshold for end of turn detection has been reported to range between 0.5 to 1 second (Ferrer et al., 2002). Yet, analyses of spontaneous dialogue in French show that silences within turns (*pauses*) may be longer than 1 second (c.f. Campione & Veronis, 2002). Moreover, Weilhammer & Rabold (2003) found that the mean duration for silences between turns (*gaps*) in spontaneous face-to-face conversation in American English was 380 milliseconds, which is shorter than 0.5 second. Consequently, if we use a silence threshold (0.5 to 1 second) to detect end of turns, we end up with a system that has a longer mean response time than humans, but which still risks interrupting its users.

Apart from using silence for end of turn detection in spoken dialogue systems, one frequent strategy is to signal turn-taking artificially, as for example in push-to-talk systems, where the user takes and maintains the turn explicitly by pushing a button. However, while push-to-talk has shown to

[1] Corresponding author Tel.: +46 8 790 62 93; fax: +46 8 790 78 54.
E-mail address: annah@speech.kth.se (A. Hjalmarsson).

be an efficient strategy for improving task completion, the extra element of pushing a button appears to affect the way users interact with the system. For example, Fernández et al. (2007) found that, compared to free turn-taking; push-to-talk resulted in longer turns and less positive feedback. Allowing users to interact freely without artificial artefacts such as a button may not be a necessity to build successful spoken dialogue systems, but it is a crucial aspect if we want to build dialogue systems that interact with its users in a human-like manner.

Humans generate speech incrementally and on-line as the dialogue progresses using information from several different sources (Kilger & Finkler, 1995). 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 manage to integrate information from different sources in parallel and simultaneously. Occasionally we need to hesitate and revise our speech as we go along. As a consequence, speech is not generated in regular constant pace of vocalized segments, but in streams of fragments in varying sizes (Butterworth, 1975). These irregularities in pause duration and turn length suggest that interlocutors cannot use silence duration to discriminate momentary pauses from ends of turns. An early theory of turn-taking suggests that speakers identify appropriate places to speak by attending to various behavioural cues or signals in the message of the preceding speaker (c.f. Duncan, 1972, Duncan & Fiske, 1977). 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". Duncan explored such turn-taking cues in a corpus of face-to-face dialogues in American English. Correlation analyses of these data show that the number of available turn-yielding signals is linearly correlated with listeners' turn taking attempts. When several signals are used in combination, there appears to be an additive effect. However, when speakers employed signals to suppress such attempts, the number of turn-taking attempts radically decreased, regardless of the number of turn-yielding signals.

The present study further explores Duncan's hypothesis by examining the effect of turn-taking cues experimentally. The objective is to investigate the possibilities of generating turn-taking cues with a synthetic voice. Whereas the focus is on how to communicate appropriate places for interlocutors to take the turn, the results also have implications for end of turn detection. The experiment is set up as a game, designed to extract judgements based on first intuition rather than afterthought. The stimuli were dyadic dialogues played to the listeners as continues dialogue segments. The motivation behind this setup was to present the dialogue segments in chronological order, which is how they are perceived in their original setting. The experimental design allows us to collect data from naïve users in a controlled experimental setting.

2 Previous work

The aim of the present study is to explore the effect of various behaviours that regulate the flow of interaction in dialogue. The existence of such cues is based on the assumption that listeners attend to *interactional cues* in dialogue. Such cues are verbal and non-verbal behaviours that pragmatically affect the conversation. For example, Clark (2002) claims that dialogue phenomena such as repeats, repairs, fillers and prolonged syllables are *strategies* used by speakers to synchronize their own internal processes with their addressees. In support of this view, prosodic variation, for example intonation and segmental lengthening, appears to mark various segment boundaries in speech (c.f. Wightman et al. 1992). Additionally, it has been shown that hesitation phenomena are likely to occur more frequently before longer utterances (Shriberg, 1994) and that listeners predict hesitations to be followed by words with high complexity (Watanabe et al., 2008).

2.1 Turn-taking in dialogue

Influential work by Sacks et al., 1974) describes human turn management as set of principles motivated by the inclination to avoid gaps or overlaps. These principles provide speakers with a mutual understanding of Transition Relevant Places (TRPs) (Ford & Thompson, 1996). A frequent assumption is that speakers can predict TRPs very precisely and that a majority of speaker changes are directly adjoining without any overlap or silence. The assumption that turn-transitions are completed without gaps or overlaps is not compatible with theories of turn-taking based on behavioural cues near the end of previous turn since speakers need at least 200 milliseconds to verbally react to an auditory

stimulus (Izdebski & Shipp, 1978). Furthermore, as pointed out by de Ruiter et al. (2006), observations of correlations between certain behavioural phenomena and turn-endings do not necessarily imply causality. Instead, it is suggested that humans predict upcoming turn-endings by lexico-syntactic content alone after showing that listeners' accuracy in predicting upcoming turn-endings in Dutch dialogues did not decrease when the intonational contour was removed (de Ruiter et al., 2006).

However, recent analysis of turn transitions in spontaneous face-to-face conversation in American English, German and Japanese have shown that pauses and overlaps are in fact normally (Gaussian) distributed (Weilhammer & Rabold, 2003), suggesting that perfectly adjoining transitions are rare. Moreover, analysis of three different corpora including both face-to-face and telephone conversation in three different languages – Dutch, Swedish and Scottish English – show that 41% to 45% of the speaker transitions are longer than 200 milliseconds (Heldner and Edlund, in press). The large number of speaker turns separated by gaps longer than 200 ms suggests that Duncan's theory of turn-taking based on behavioural cues near the end of previous turn is feasible.

We will now present some of the behaviours that have been suggested as relevant for turn-taking.

2.1.1 Turn-taking cues

First, we need to define the term *turn-taking cue* and what kind of behaviour this refers to. Whereas Duncan refers to these behaviours as "signals", these phenomena are likely more or less intentional. For instance, there are acoustic phenomena, e.g. drop in energy or inhalations that guide interlocutors in their turn-taking. The 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). Whether conscious or not, these non-verbal phenomena appear to affect the addressees' interpretation of the message. However, since the behaviours are not necessarily deliberate, we choose to use the term *cue* rather than signal. Thus, in the present study, *turn-taking cues* refer to all perceivable phenomena relevant for turn-taking, regardless of whether they are conscious or not.

Duncan (1972) introduces a number of different turn-taking cues. Behaviours that have a turn-yielding effect include a rising or falling pitch contour, the termination of a hand gesture, a drop in loudness, and completion of grammatical pauses. Behaviours that suppress turn-taking attempts include an intermediate pitch level and sociocentric sequences (stereotyped lexical expressions). In a recent corpus analysis of non-face-to-face, spontaneous task-oriented dialogues in American English, a number of phenomena were found to take place at significantly higher frequencies before speaker changes than before speaker holds (Gravano (2009). These turn-yielding cues include 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. A flat or sustained pitch contour has been reported to have turn-holding functions (see for example Selting, 1996, Koiso et al., 1998). Cutler & Pearson (1986) present results that suggest that long segments of speech are more likely to be judged as turn final. In addition, turn-final speech segments have been shown to be significantly longer than turn-medial speech segments (Gravano, 2009). In line with Duncan's findings, Gravano's results show support for a linear relationship (positive correlation) between the number of simultaneously available turn-yielding cues and the number of turn-taking attempts.

There are contradictory findings regarding the effect of some turn-taking cues. For example, according to Duncan, a pitch level terminal-junction combination other than an intermediate pitch level in American English is associated with turn-yielding intentions. A more detailed analysis of a rising intonation suggests that a high-rise (H-H%) has turn-yielding effects and a plateau (H-L%) has turn-holding effects whereas the effects of a low-rising contour (L-H%) is unclear (Gravano, 2009). Local et al. (1986), on the other hand, claim that a rising intonation has both turn-yielding and turn-holding functions in Tyneside English. Swedish has two basic intonation patterns, medial fall (H*L%) and fall-rise (H*LH%) (Bruce, 1977). Thus, in an analysis of the prosodic aspects of turn-taking in Swedish, Edlund & Heldner (2005) makes a distinction between patterns with a final rise and a final fall. This analysis shows that a rising intonation was followed by an equal distribution of speaker

changes and speaker holds (51% and 49% respectively), implying that the turn-taking effects of a rising intonation in Swedish are unclear.

Local et al. (1986) claims that increased phrase-final lengthening have turn-yielding functions in Tyneside English whereas Gravano (2009) presents results that show that increased phrase-final lengthening in American English have turn-holding effects. In line with Gravano, Ferrer et al. (2003) presents results that suggest that the final rhyme of the phrase is lengthened in both cases, but that the lengthening before internal pauses is even longer than before end of turns. Furthermore, the duration of the lengthening is positively correlated with pause length.

2.2 Methods to extract perceptual judgements of turn-taking cues

The contradictory findings discussed in the previous section are possibly due to differences between dialects, languages and dialogue contexts. However, it is difficult to quantify and compare some of the results since theories of turn-taking have been strongly influenced by Conversation Analysis (CA) (c.f. Sacks et al., 1974). Studies following a CA tradition analyze and present in detail descriptions of the relevant phenomena. However, in order to detect and process these behaviours automatically, we also need large-scale quantitative studies that allow us to cover larger sets of data. Duncan (1972) and Gravano (2009) studied correlates of speaker changes in corpora of varying sizes, but relatively few studies have investigated turn-taking experimentally. Some exceptions are experimental studies on the relative contribution of various cues. These studies show that lexical cues have a strong effect whereas non-lexical cues have turned out as less influential. For example, Schaffer (1983) and Oliveira & Freitas (2008) studied the role of prosody in turn-taking by analyzing the judgments of non-participating listeners in perceptual experiments. The stimuli in these experiments were utterances manipulated in order to separate prosodic and lexical cues. The results show a great variability in the listeners' use of intonation and do not support a clear-cut effect of prosody alone.

3 Method

The aim of this work is to investigate experimentally how turn-taking cues form a complex signal and affect listeners' expectations of turn-taking behaviour in dialogue. The cues are investigated in a perception experiment where subjects listen to dyadic dialogues in chronological order and try to anticipate whether a token will be followed by a speaker change or not. In line with Duncan's findings, our hypothesis is that, the more turn-taking cues with a particular pragmatic function – turn-yielding or turn-holding – the faster the reaction time to make the judgement and the higher agreement among subjects on the expected outcome. The aim of this study is to explore the possibilities of using behaviours that affect turn-taking in human-human conversation to generate appropriate turn-taking behaviour in spoken dialogue systems. Thus, in addition to human-human dialogues, the experiment included stimuli where one of the human interlocutors was replaced with a synthetic voice. The motivation to use a synthesis rather than a pre-recorded human voice in a dialogue system is that synthetic voices are easier to update and manipulate on-line (Reiter and Dale, 1997). For example, no new recordings are needed to manipulate prosody or to extend the system's vocabulary.

3.1 The DEAL corpus

The stimuli dialogues were collected in the DEAL domain. DEAL is a spoken dialogue system under development at KTH. The aim of the system is to provide conversation training for second language learners of Swedish. The scene of DEAL is set at a flea market where a talking animated agent is the owner of a shop selling used goods. The objectives are to build a system which is fun, human-like, and engaging to talk to (Hjalmarsson et al., 2007). The recorded dialogues are informal, human-human, face-to-face conversation in Swedish. The recordings were made with close talk microphones with 6 subjects (4 male and 2 female). 8 dialogues were collected. Each customer interacted with the same shop-keeper twice, in two different scenarios. The customers were given a mission: to buy items at a flea market at the best possible price from the shop-keeper. The task was to buy 3 goods for a specific purpose (e.g. to buy tools to repair a house). The shop-keeper sat behind a desk with images of different goods pinned to the wall behind him. Each dialogue was about 15 minutes, making for about 2 hours of speech in total in the corpus. The dialogues were transcribed orthographically and annotated

for entities such as laughter, filled pauses, lip-smacks, breathing and hawks. The annotators could both see the transcriptions and listen to the recordings while labelling.

3.2 Identifying turn-taking cues

First, in order to study the effect of turn-taking cues in a perceptual experiment, we need to identify occurrences of these behaviours. Many of the target behaviours have several dialogue functions and in order to consider them as turn-taking cues, they need to be perceivable to human listeners. Hence, rather than trying to identify the cues automatically, we used human annotators. Duncan (1972) has been criticized for not reporting inter-annotator agreement or formal description of his "signals" (Beattie et al., 1982). An important part of this work is therefore to provide a detailed description of how these were annotated.

The dialogues were first automatically segmented into inter-pausal units (IPUs), a sequence of words surrounded by silence in both channels that was longer than 200 milliseconds (ms). 200 ms was used as a segment criterion since Swedish have long plosives. If we include silences shorter than 200 ms, we risk extracting plosive stops where we aim to extract pauses or gaps between speech segments. The four dialogues contained 2011 silences longer than 200 ms. 85% of these silences were internal pauses and 15% were silences (gaps) between speakers.

The present study explores six different categories of turn-taking cues. The original dialogues were face-to-face interactions but the experiment contained no facial gestures since the focus of the present study is lexical and acoustic cues that can potentially be reproduced in a synthetic voice. Whereas visual cues such as gaze and hand gestures play an important role in turn-taking (c.f. Kendon, 1967), there are, to our knowledge, no findings which suggest that the distribution and characteristics of acoustic turn-taking cues differ between these two conditions. For example, in a comparison of the distribution of prosodic cues between face-to-face and non-face-to-face conversation, it was found that there is large variation in listeners' use of prosodic cues, but the effect of the prosodic cues were similar in both conditions (Schaffer, 1983). Hence, if Duncan's hypothesis is correct, the more cues available, regardless of modality, the more predictable the outcome is.

The turn-taking cues explored in the present study were *intonation*, *semantic completeness*, *phrase-final lengthening*, *disfluencies*, *speech production phenomena* such as perceivable breathing and lip smacks and some frequently occurring *cue phrases* (see Table 1). The cues were chosen to represent a fair distribution of different types of phenomena. That is, we wanted to explore both lexical and non-lexical cues as well as cues that were more or less intentional. For example, the cues include both explicit lexical expressions such as "right?" which were used to elicit responses from listeners as well as speech production phenomena such as lip-smacks and breathing.

Category	Turn-yielding cues	Turn-holding cues
Intonation	fall	flat
Phrase-final lengthening	No phrase-final lengthening	Long phrase-final lengthening
Speech production phenomena	Audible expirations	Audible inhalations, lip-smacks
Disfluencies	-	Speaker interruptions and repetitions
Cue phrases and filled pauses	Response eliciting	Connectives Filled pauses
Semantic completeness	complete	incomplete

Table 1 : ***Cue categories***

3.3 Annotation of turn-taking cues

Two annotators were used for labelling. The annotators were researchers at the Department of Speech, Music and Hearing at KTH with good knowledge of linguistics and phonetics. In order to avoid

influences from other cues, each annotation task included only the target parameter and no turn-taking issues were considered. That is, the annotators were only instructed to annotate the specific target phenomena (for example intonation as rising, flat or falling) and no reference was made to the task of the actual experiment which was to consider who would be the next speaker. The labelling procedure of the different cues will now be described individually.

3.3.1 Intonation and increased phrase-final lengthening

The turn-taking cues intonation and phrase-final lengthening were annotated in the last 500 milliseconds of IPUs, that is, speech just prior to silences longer than 200 milliseconds. The annotators were provided with these last 500 milliseconds of the IPUs in isolation and in random order in order to reduce influences of the prosodic realization of adjacent speech and the lexical context. For intonation, the target labels were *flat*, *rising* or *falling* pitch contour whereas the target labels for phrase-final lengthening were *long*, *short* and *no* phrase-final lengthening. The inter-annotator agreement for both tasks were 69% overall agreement or kappa coefficient 0.37. Because of the somewhat poor agreement, confusion matrixes were created. For intonation, the confusion matrix revealed that the majority of the confusions were between falling and rising slope. After listening to the stimuli, a possible explanation of these confusions is that a frequently occurring contour in the data was a rising curve with a minor falling slope at the end that annotators may have judged differently. The confusion matrix for phrase-final lengthening suggests that the annotators' boundaries were skewed, since almost all confusions were between neighbouring categories. Still, poor inter-annotator agreement may indicate that the annotations are not reliable. To address this issue, two precautions were taken. First, only stimuli where both annotators agreed were considered to contain cues. Second, the reliability of the manual annotations was further explored in terms of how well these correspond to automatically extracted measures of fundamental frequency (F0) and speaking rate.

As an automatic measure of intonation, the change in F0 during the last 200 ms of the IPU was automatically extracted using Snack (www.speech.kth.se/snack/) and z-score-normalized over speaker and dialogue. As a measure of phrase-final lengthening, speaking rate was calculated over IPUs as the number of syllables per second. Negative durational data is impossible and the distribution of syllable durations will therefore be skewed to the left. This was confirmed by histograms of the distribution of speaker syllable rate per second. Since it has been suggested that the log-normal law is a better fit to duration data (see for example Campione & Veronis, 2002), speaking rate was calculated per second and transformed into a logarithmic scale (base 10). The syllable rate was also z-score-normalized over speaker, dialogue and phoneme.

To explore the relationship between the automatic measures and the manual annotations, ROC (relative or receiver operating characteristic) curves were used (c.f. Metz, 1978). ROC-curves are mainly used to study the accuracy of a diagnostic test in terms of how well it discriminates diseased cases from normal cases. More specifically, ROC-curves illustrate the relationship between true positive rate (TPR) and false positive rate (FPR) as a discrimination threshold is varied. The shape of a ROC-curve illustrates the overall accuracy of a test in terms of the *sensitivity*, the probability that a test result will be positive when the target condition is present, versus the *specificity* (1-FPR) the probability that a test is negative when the target condition is not present. Each point in a ROC-curve represents the sensitivity versus the specificity for a particular cut-off value. A test with perfect discrimination has an area of 1.00. The diagonal lines from the bottom left to the top right in Figure 1 and Figure 2 are so-called *no discrimination lines*. This line illustrates a test with the same discriminative power as random guesses. Points above this line indicate that the test is better than chance at identifying true positives while points below, indicate that the test is useless. The no discrimination line has an area of 0.5.

Here, TPR is the percentage of IPUs with a specific prosodic cue (labelled by both annotators) that was correctly classified as positive based on automatically extracted values of F0 and syllables rate as the threshold for these values are varied. FPR is the percentage of IPUs incorrectly classified as positive as the threshold values are varied. The ROC-curves for intonation and phrase-final lengthening are plotted in Figure 1 and Figure 2 respectively. The aim is to illustrate how well IPUs annotated with a specific prosodic cue can be separated from IPUs that are *not* annotated with that cue using automatically extracted values of F0 and syllable length. The shapes of the curves suggest that threshold values for automatically extracted F0 and syllable rate can be selected to identify the

manually annotated prosodic cues with high accuracy, that is, well above chance. The accuracy for flat intonation (area under the curve 0.84) is higher than for falling intonation (area under the curve 0.72). The area under the curve for long phrase-final lengthening is 0.72 and 0.77 for no lengthening. The discriminative power of these tests, that is, the possibility to identify these manually annotated cues using automatically extracted prosodic features, suggests that the annotators indeed were labelling something dependable, despite the low kappa values. The ROC-curve for rising intonation, however, suggests that the accuracy for this test is poor. For this reason, we choose to exclude this cue from the results analyses.

Note to Publisher: Insert Figure 1 and 2 about here

Figure 1 : ***Receiver Operating Characteristic curve for falling, rising and flat intonation***

Figure 2 : ***Receiver Operating Characteristic Phrase-final Lengthening***

3.3.2 Semantic completeness

Semantic completeness represents the lexical content in the dialogues. This cue corresponds to what is often referred to as lexico-syntactic, lexical or syntactic completion points in a dialogue. However, rather than extracting syntactical completion points we choose to manually annotate semantic completeness. This is since dialogue relies much on context that is not captured by syntax. The labelling procedure was performed as follows:

IPUs were presented incrementally to the annotators and for each segment they were asked to label whether the current IPU "was a complete response to the previous turn". The two annotators were provided with the previous lexical dialogue context, but the label tool only displayed the dialogue up to the target IPU. After each judgment, the dialogue segment up to the next target IPU was provided incrementally. The annotators only had access to the orthographic transcriptions of the dialogues and did not listen to the recordings. Non-lexical elements such as filled pauses and breathing were removed from the transcripts, since they are considered to represent acoustic information – information that is already represented in other cues. Inter-annotator agreement for this task was kappa coefficient 0.73.

3.3.3 Cue phrases

The *cue phrases* considered as turn-taking cues were standardized lexical expressions that appear to be related to turn-taking. That is, lexical expressions closely associated with the termination or continuation of turns.

Cue phrases or so-called *discourse markers* are a class of linguistic devices typically used to signal pragmatic and semantic relations between different segments of speech. Examples in English are: *oh, well, now, then, however, you know, I mean, because, and, but* and *or* (c.f. Schourup, 1999). The definition of what constitute a cue phrases in the present study is broad and includes all types of linguistic entities that speakers use to structure the dialogue at different communicative levels. A rule of thumb is that cue phrases are words or chunks of words that have little lexical impact at the local speech segment level but serve significant pragmatic function.

The DEAL corpus was labelled for cue phrases by two annotators with a high inter-annotator agreement (kappa coefficient 0.82). The labelling of cue phrases included a two-fold task, both to decide if a word was a cue phrase or not – a binary task – but also to classify which functional class it belongs to according to the annotation scheme. The annotators could both see the transcriptions and listen to the recordings while labelling. There were ten different classes of cue phrases (for more details see Hjalmarsson, 2008), but only four of them are explored in the present study. These four categories were lexical expression considered to have explicit turn-taking functions. First, there were three different classes of *connectives*, *additive connectives*, *contrastive connectives* and *alternative connectives* (for example "and", "but", and "or" respectively), which were considered to have turn-holding functions. As the term suggests, the connectives connect segments of speech and are often placed within turns. The fourth cue phase category investigated was *Response eliciting*, that is, lexical expressions (for example "eller hur?" in Swedish or "right?" in English) used to elicit information

from listener(s). This type of cue phrases is typically placed at the end of turns and consequently considered to have turn-yielding functions.

3.3.4 Other verbal cues

There are behaviours that are side-effects of speech production and closely related to either the termination of a turn or an internal pause. Examples of such behaviours are breathing and lip-smacks. Speakers may not be aware of these behaviours, but they may help listeners identify appropriate places to speak and were therefore explored as potential turn-taking cues. Exhalations are associated with the completion of turns and therefore hypothesized to have turn-yielding effects. Inhalations and lip-smacks were considered to indicate an intention to continue speaking and therefore hypothesized as turn-holding. Annotation of these phenomena was already available in the original transcriptions of the DEAL corpus.

Another set of verbal phenomena explored as potential as turn-taking cues in the present study were *lexical repetitions* and *interruptions*. The original DEAL transcriptions include annotations of repeated words or phrases. Such repetitions are often considered as signs of difficulties to plan or produce upcoming utterances (Shriberg, 1994). This makes them potential turn-holding cues.

The transcriptions also included annotation of speaker interruptions; these were annotations of abrupt stops in the middle of the speech flow. According to Levelt's *main interruption rule*, speakers stop the flow of speech immediately when a problem is detected (Levelt, 1989). Hence, speaker interruptions suggest that the speaker has detected a problem in previous segment of speech and that this segment is about to be altered. This makes them potential turn-holding cues,

4 Data preparation

As pointed out by Oliveira & Freitas (2008), manipulating dialogues off-line and analyzing these out of context can be problematic since this may result in stimuli that never would occur in a real dialogue setting. To tackle this problem, the experiment was designed to allow subjects to follow longer dialogue segments chronologically.

4.1 Stimuli selection and preparation

Before conducting the actual experiment, we needed to select suitable dialogue segments to use as stimuli from the DEAL corpus. The corpus was first annotated with turn-taking cues as described in the previous section (Section 3.3). Target IPUs were then selected from a list of IPUs with their corresponding cues without listening to the recordings. The selections were made to select IPUs that represent a similar distribution of within-turn and between-turn silences 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. In the end, 125 IPUs were selected as stimuli. The number of cues over IPUs is presented in Table 2. As the table suggests, turn-holding cues occurred more frequently in the stimuli than turn-yielding cues. This reflects the overall distribution of silences within speaker turns and silences between speakers which was 85% pauses within turns and 15% gaps between speakers. This distribution further suggests that after a silence, the likelihood that the current speaker will continue is much higher than the likelihood of a speaker change

Turn-holding cues	**Turn-yielding cues**			
	0	**1**	**2**	**3**
0	6	21	13	3
1	24	11	3	
2	30	7		
3	6			
4	1			

Table 2 : ***Number of turn-yielding and turn-holding cues over stimuli IPUs***

4.2 Re-synthesis of dialogues

The motivation behind this work was to investigate whether cues could be reproduced in a synthetic voice and perceived as having similar functions. In order to create the synthesized stimuli, a corresponding reproduction of the male party in the dialogues was created by replacing his voice with a diphone synthesis. This was done using Expros, a tool for experimentation with prosody in diphone voices (Gustafson & Edlund, 2008). Hence, the dialogues were first transcribed orthographically including non-lexical entities such as laughter, repetitions, filled pauses, lip-smacks, breathing and hawks. Based on these manual transcriptions and the original recordings, Expros automatically extracts fundamental frequency and intensity from the human voice and creates synthetic version using these parameters. In order to extract the timings from the original dialogues, the transcripts were time-aligned with the speech signal. This was done using forced alignment with subsequent manual verification of the timings. Some manual alterations were made to the phonetic transcriptions in order to correct mispronunciations. Since breathing and lip-smacks could not be re-synthesized, the original human realizations were kept and concatenated with the synthetic voice using the manually verified timings. In conclusion, the synthetic version was created to match the original dialogues' timing, intonation, intensity and non-lexical as well as lexical cues.

5 Experiment

The experiment included 4 dialogue segments from 4 different dialogues. The segments were between 116 to 166 seconds long. The dialogues were dyadic dialogues with three different speakers, one male and two female. The male speaker (S_1) participated in all 4 dialogues and the 2 female speakers (S_2 and S_3) in 2 dialogues each. In the experiment, the recording stops playing just subsequent to a target IPU, allowing the subjects to make a judgement. Each subject listened to two human-human dialogues and two dialogues where one party was replaced with the diphone synthesis. The subjects only heard each stimulus once, produced either with a synthetic or human voice. The stimuli presented with a synthetic voice to half of the subjects were presented with human voice to the other half and vice versa. Two movie tickets were awarded to the "best" player.

The experimental setup was designed as a game where the subjects received points if they could figure out who would be the next speaker. The GUI of the test (see Figure 3) included two buttons with "pacmans" and a button for pausing the test. 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. The pacman buttons 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 the dialogues and guess who the next speaker would be by pressing the corresponding button. To make the subjects aware that the playback had halted, the faces changed colour. Each time the recording halted, the mouse pointer was reset to its original position, in the middle of the pause button. Thus, the subjects had to move the mouse pointer from the pause button to one of the pacman buttons in order to make their judgement. This was done to control the conditions before each judgment, enabling comparisons of reaction times.

Note to Publisher: Insert Figure 3 about here

Figure 3 : ***Experiment GUI***

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 the subject was right or wrong was based on which interlocutor vocalized first. Once a judgement was made, the pacman buttons turned red if the answer was wrong and green if the answer was right. This scoring was used to motivate the participants to respond immediately and make the experiment more fun. However, whether they were right or wrong was unimportant for the experimental results of this study.

5.1 Pilot experiment

A pilot experiment was conducted to test the experimental setup and features of the GUI. The pilot experiment included 10 subjects, 5 male and 5 female, between the ages of 31 and 58. Based on the results from this experiment and comments from the subjects, a few changes were made to the experimental design before the final experiment. Training effects were controlled by changing the order of the dialogues. There was also a 210 second long training session to allow the subjects to become familiar with the task.

5.2 Experiment

The 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.

6 Results

This section analyzes the effects of individual as well as combined sets of turn-taking cues. First, we present results on the individual cues. Our motive is to investigate whether these behaviours affect the subjects' judgements as hypothesised.

6.1 The Effect of individual turn-taking cues

To explore the effect of individual turn-taking cues, namely, whether a turn-holding cue increased the expectations of a HOLD and turn-yielding cues increased the expectations of SWITCH, the judgements for all stimuli with a particular cue were compared to the overall distribution of HOLD and SWITCH. The cues investigated were all the cues presented in Table 1. Intonation contour and speaking rate were based on automatic extractions of these features as described in Section 3.3.1. The thresholds were extracted from the ROC-curves (see Figure 1 and Figure 2). To get discrete categories, positive rate (FPR) was prioritized over high true positive rate (TPR).

Figure 4 presents the percentage of judgements for HOLD versus SWITCH hold over the different cue categories. Increased phrase final lengthening is listed separately since we did not have any clear hypothesis of how this cue affect turn-taking.

Note to Publisher: Insert Figure 4 about here

Figure 4 **: *% judgments for HOLD and SWITCH over the different cue categories. Results include both synthetic and natural voice.***

Chi-square tests of independence were employed to investigate whether the judgement distribution between switch and hold for all stimuli containing a particular cue type, differed from judgement distribution of HOLD and SWITCH when this cue was absent. The results from these tests are presented in the top row of Table 3. The distribution of HOLD and SWITCH for all cues except phrase-final lengthening differ significantly (tested individually) from the overall distribution of HOLD and SWITCH for (p<.05, by chi-square test of independence with 2 x 2 contingency tables). These results were also checked for the direction, that is, whether turn-holding cues resulted in a higher number of judgments for HOLD than the overall distribution and vice versa. The results support the conclusion that the turn-taking cues were perceived as hypothesised.

In order to examine the potentials of realizing turn-taking cues with a synthetic voice, Chi-square tests of independence were also calculated comparing the judgement distributions of when a particular cue was present and not for the human and synthetic voice independently (Table 3 rows 2 and 3). These results show (p<.05) that when split over natural and synthetic voice, the same results hold. Chi-square tests of independence were also employed to explore the impact of the different cue types on each individual subject. By doing this, any bias for a particular outcome in the subject's overall judgement distribution is considered. The number of subjects for whom the distribution of HOLD and SWITCH differ when a particular cue is present is presented in Table 3 row 4. It should be

noted that some of the cues were less frequent than others. The total number of stimuli that contain a particular cue is presented in parenthesis after the cue category label (see Table 3).

	Turn-yielding cues			**Turn-holding cues**					**Phrase-final lengthening**
	Falling intonation (27)	Semantically complete (49)	Cue phrase response-eliciting (6)	Flat Intonation (40)	Semantically incomplete (49)	Cue phrase connectives (22)	Disfluencies (8)	Speech production phenomena (18)	Long phrase-final lengthening (23)
All data df=1, N=1993	X^2= 132.95 ***p*=0.00**	X^2= 539.59 ***p*=0.00**	X^2= 137.58 ***p*=0.00**	X^2= 173.14 ***p*=.00**	X^2= 407.11 ***p*=.00**	X^2= 6.38, ***p*=.01**	X^2= 6.58, ***p*=.01**	X^2= 11.04, ***p*=.00**	X^2= 0.19 *p*=0.98
Human voice only df=1, N=1421	X^2= 75.80 ***p*=0.00**	X^2= 306.77 ***p*=0.00**	X^2= 81.24 ***p*=0.00**	X^2= 92.14 ***p*=0.00**	X^2= 238.09 ***p*=0.00**	X^2= 75.86 ***p*=.00**	X^2= 5.80 ***p*=.02**	X^2= 18.57, ***p*=.00**	X^2= 2.03 p=0.16
Synthesis only df=1, N=572	X^2= 58.18 ***p*=0.00**	X^2= 233.07 ***p*=0.00**	X^2= 56.62 ***p*=0.00**	X^2= 81.27 ***p*=0.00**	X^2= 170.02 ***p*=0.00**	X^2= 62.92 ***p*=.00**	-	-	X^2= 0.79 *p*=0.37
n subjects with difference in judgment distribution p<.05	**12/16**	**16/16**	**15/16**	**16/16**	**16/16**	**0/16**	**0/16**	**0/16**	0/16

Table 3 : ***Differences between the judgment distribution (HOLD and SWITCH) per cue compared to the overall judgment distribution (Chi-square test of independence p<.05). Some comparisons could not be made because these configurations did not contain enough data points (cells with a frequency less than 5).***

Previous literature frequently mentions phrase-final lengthening as a turn-taking cue (c.f. Gravano, 2009, Local et al., 1986 and Ferrer, 2003). Since we did not find any turn-taking effects of this cue, this phenomenon was explored in more detail. Phrase-final lengthening was analyzed both over the entire IPU and in more detail at the end of the IPU. First, speaking rate was calculated as both syllables and vowels per second and z-normalized over speaker and dialogue. First, an independent-samples t-test was conducted to explore overall phrase-final lengthening. Regardless of whether the IPU was followed by a speaker change or not, there was a significant difference between the last (M=0.67, SD=0.85) and the penultimate vowel (M=0.12, SD=0.71); t(1440)=13.35, *p*=0.00. The average speaking rate for the two last vowels and syllables before silences longer than 200 ms are displayed in Figure 5. The differences are not significant. Neither was there any significant difference in speaking rate between turn medial and turn final IPUs over the preceding two syllables (compared pair-wise from the end). In order to investigate phrase-final lengthening further, lexical stress was derived from the transcriptions and independent-samples t-tests were conducted between turn-final and turn-medial IPUs for lexically stressed and unstressed syllables separately. Still, no significant differences in phrase-final lengthening between turn-final and turn-medial IPUs were found.

Note to Publisher: Insert Figure 5 about here

Figure 5 : ***Vowels and syllables per second z-normalized over speaker and dialogue for the last two vowels and syllables of the IPU for HOLD versus SWITCH. No differences are significant.***

6.2 The additive effect of turn-taking cues

This section presents results from analyzing the effect of turn-taking cues used in combination. All turn-taking cues in Table 1 except phrase-final lengthening were explored. Phrase-final lengthening was excluded since the judgement distribution for this cue did not differ significantly from the overall judgement distribution. For simplicity, all cues were given equal weight (1) and the relative contribution of the different cues was not considered.

Because of properties of durational data as discussed in Section 3.3.1, the reaction times were transformed into a logarithmic scale (base 10). The average reaction times differed considerably between subjects (from 933 ms to 1510 ms) and were therefore z-normalized over each subject. A one-way ANOVA with judgement agreement (75%, 85%, 95% and 100%) as a between-subject factor

was used to test for differences in reaction times over judgement agreement. Stimuli with high agreement, regardless of the number of cues, were judged significantly faster than stimuli with low agreement, F (4, 1028) = 7.29, $p < .00$. The average reaction time for stimuli with 75%, 85%, 95%, and 100% judgement agreement are presented in Figure 6. For completeness, each point is labelled with its average log¹⁰ value (un-normalized) in milliseconds. All differences, except between 75% and 85% agreement, are significant (Tukey's test, p<.05, see Table 4). Analyses were done with four outliers, the two longest and the two shortest reaction times, excluded.

Note to Publisher: Insert Figure 6 about here

Figure 6 : ***Average reaction time log10 z-normalized milliseconds over IPUs with % agreement. Error bars represents the standard error.***

Judgement Agreement		Difference in mean response time log_{10} z-value (log_{10} in ms)	Standard error	p-value
75%	**85%**	0.112 (31 ms)	0.06	0.285
	95%	0.398 (100 ms)	0.07	**0.000**
	100%	0.561(142 ms)	0.06	**0.000**
85%	**95%**	0.286 (70 ms)	0.06	**0.000**
	100%	0.449 (112 ms)	0.06	**0.000**
95%	**100%**	0.163 (42 ms)	0.05	**0.046**

Table 4 *: **Differences in average response time between 75%-85%, 75%-95%, 75%-100%, 85%-95%, 85%-100% and 95%-100% judgement agreement (Tukey's p<.05, df=3). Significant differences in bold.***

To study the additive effect of the turn-taking cues, the distribution of judgements for HOLD and SWITCH was compared over stimuli with different numbers of cues. Thus, stimuli with one turn-holding cue were compared to stimuli with two turn-holding cues and so on. The results of these comparisons are presented using a bubble chart (see Figure 7). Some cue combinations were rare (see 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. The bubble chart is used to enable comparisons of all cue combinations, that is, including stimuli annotated to occupy both turn-holding *and* turn-yielding cues. The number of turn-yielding cues is displayed on the x-axis and turn-holding cues on the y-axis. The diameters in the bubble charts represent the percentage of judgments for HOLD versus SWITCH. Each bubble is labelled with the percentage values for HOLD and SWITCH.

Note to Publisher: Insert Figure 7 about here

Figure 7 : ***The distribution of judgments for HOLD versus SWITCH. Each bubble is labelled with the % HOLD (%SWITCH).***

Chi-square tests of independence were employed to explore the impact of the different number of cues on judgement distribution between HOLD and SWITCH. Thus, the distribution of HOLD and SWITCH was compared between 0 and 1 cue, 1 and 2 cues and so on (see Table 5). Turn-holding cues and turn-yielding cues were compared separately. For the overall data set ("All"), all steps differ significantly except between 2-3 turn-holding cues (Chi-square test of independence p<.05). The impact of different number of turn-taking cues was also compared over the different speakers and for the synthesis separately. There is a significant relationship between the number of turn-taking cues and the judgement distribution over all speakers as well as for the synthetic voice (Chi-square test of independence p<.05).

Chi-square comparison			Speaker				
			S_1	**S_2**	**S_3**	**Synthesis**	**All**
Turn holding cues	**0**	**1**	X^2= (1, N = 511) 90.73, ***p* = .00**	X^2= (1, N = 368) 53.19, ***p* = .00**	X^2= (1, N = 256) 29.82, ***p* = .00**	X^2= (1, N = 311) = 38.62, ***p* = .00**	X^2= (1, N = 1297) = 235.66, ***p* = .00**
	1	**2**	X^2= (1, N = 667) 18.94, ***p* = .03**	X^2= (1, N = 197) 22.87, ***p* = .00**	X^2= (1, N = 95) 5.26, ***p* = .03**	X^2= (1, N = 384) = 14.58, ***p* = .00**	X^2= (1, N = 1196) = 81.87, ***p* = .00**
	2	**3**	-	-	-	-	X^2= (1, N = 687)

							235.66, p = .26
Turn yielding cues	**0**	**1**	X^2= (1, N = 203) 114.72, **p = .00**	X^2= (1, N = 400) 96.68, **p = .00**	X^2= (1, N = 217) 51.59, **p = .00**	X^2= (1, N = 419) 46.19, **p = .00**	X^2= (1, N = 1689) 238.29, **p = .00**
	1	**2**	X^2= (1, N = 350) 55.72, **p = .00**	X^2= (1, N = 272) 44.51, **p = .00**	X^2= (1, N = 193) 41.31, **p = .00**	X^2= (1, N = 196) 20.56, **p = .00**	X^2= (1, N = 881) 57.92, **p = .00**
	2	**3**	-	X^2= (1, N = 64) 14.25, **p = .00**	-	-	X^2= (1, N = 304) 11.77, **p = .00**

Table 5 *: **Differences in the distribution of HOLD and SWITCH judgments between 0-1, 0-2, 0-3, 1-2,1-3 and 2-3 turn management cues. Turn-holding and turn-yielding cues were compared separately. Significant differences in bold (Tukey's p<.05, df=1). Some comparisons could not be made because these configurations did not contain enough data points (cells with a frequency less than 5).***

The results above indicate that the additive effect is similar for turn-holding and turn-yielding cues. Namely, the more turn-taking cues, the higher agreement among subjects on the expected outcome. Since the analyses of judgement agreement suggest that the additive effect of the cues were similar for both types of cues, the reaction times were analyzed over the entire data set. Thus, a one-way ANOVA was conducted with the number of turn-taking cues (regardless if turn-yielding or turn-holding) as a factor (the statistics are calculated on IPUs without contradictory cues). There was a significant effect of number of cues on reaction times; F(4, 1988) = 4.01, p=.00. Post hoc comparisons using the Tukey HSD test was used to explore these differences in more detail. These results are presented in Table 6. Although not all steps differ significantly, there is a strong trend: the more turn-taking cues, the faster the reaction time. An independent-samples t-test was conducted to explore differences in reaction time between IPUs with a majority of turn-holding cues and IPUs with a majority of turn-yielding cues. IPUs with a majority of turn-holding cues (M=-0.23, SD=0.95) were judged significantly faster than IPUs with a majority of turn-yielding cues (M=0.03, SD=1.03); t(1229)=-4.3, p=0.00.

Turn-taking cues		**Difference in mean response time, z-value (log_{10} in ms)**	**Standard error**	**p-value**
0	**1**	0.372 (93 ms)	0.11	**0.00**
	2	0.440 (96ms)	0.11	**0.00**
	3	0.695 (117 ms)	0.13	**0.00**
1	**2**	0.067 (3 ms)	0.07	0.843
	3	0.323(23 ms)	0.10	**0.00**
2	**3**	0.255 (21 ms)	0.10	0.058

*Table 6 : **Differences in average response time between 0-1, 0-2, 0-3, 1-2, 1-3 and 2-3 turn-taking cues (Tukey's p<.05, df=3). Significant differences in bold.***

6.2.1 Differences between synthetic and human voice

An independent-samples t-test was conducted to explore differences in reaction time for human and synthetic voice. For this comparison, only stimuli based on the male speaker's (S_1) voice were included. Stimuli based on the other speakers were excluded since their voices did not have a corresponding synthesized version. No significant differences in reaction times between synthetic and human voice were found.

6.2.2 Differences between speakers

As shown in Table 5, the judgement distribution between HOLD and SWITCH for different speakers shows that there is an additive effect of the turn-taking cues regardless of speaker. To explore if there is any differences in reaction times for judgements of stimuli produced by different speakers, a one-way ANOVA was conducted with speaker as a factor. The results show that there is a significant effect of speaker, F(2, 1452)=16.06, p=.00. Post hoc comparisons using the Tukey's HSD test show that the mean reaction time for speaker S_2 (M=0.14, SD=1.04) and S_3 (M=0.03, SD=1.02) differed significantly from speaker S_1 (M=-0.19, SD=0.99), p=.00 for S_1*S_2 and p=.00 for S_1*S_3. However, no differences were found between speakers S_2 and S_3. To explore if this difference was an effect of differences in cue frequency over IPU and speaker, a Kruskal-Wallis test was conducted for average number of cues per IPU over speakers, but no significant differences were found.

7 Conclusion

Duncan (1972) has previously shown that a number of verbal and non-verbal behaviours affect turn taking in dialogue. If used in combination, the number of turn-taking cues is linearly correlated with listeners' turn-taking attempts. The present study further explores these findings by examining the effect of such turn-taking cues experimentally. The objective of the present study was to investigate the possibilities of generating turn-taking cues with a synthetic voice. In order to explore this issue, the experiment included dialogues realized with a human voice as well as dialogue where one of the speakers was replaced with a synthesis. Analyses of the reaction times show that stimuli with high annotator agreement, regardless of the number of turn-taking cues, were judged significantly faster than stimuli with low agreement. The judgment agreement and reaction times were further used as measures to analyze the perceptual effect of the turn-taking cues.

First, the effect of individual turn-taking cues was explored. For each cue, the judgement distribution between HOLD and SWITCH was analyzed. The results show that all except one of the turn-taking cues explored in the present study affected the judgements as hypothesized. The exception was **phrase-final lengthening** which did not have a significant effect on the listener's judgements. The judgement distribution for different cues further suggests that some cues had a major impact, affecting a large majority of the judgements, whereas some cues were less influential. These differences between cues suggest that some cues are more central than others are. If so, the additive effect of turn-taking cues is not necessarily linear.

The aim of the present study is to explore the potentials of using turn-taking cues in spoken dialogue systems. Primarily, dialogue system designers should consider cues that affect a majority of judgements and users accordingly. Such cues include **semantic completeness**, **turn-yielding cue phrases** and a **falling** and **flat intonation**. It should be mentioned that some cues might be easier to employ in dialogue systems than others are. For example, **cue phrases** and **falling** and **flat intonation** are all discrete behaviours that can be produced locally just prior to a pause or turn-ending without the need for syntactic or semantic representation. **Semantic completeness** is an influential cue, but in order to employ semantic completeness as a cue, the system needs keep track of whether a dialogue segment is complete or not. If the state differs from the system's continued plan of generation, it needs to add words in order to change semantic completeness in way that is consistent with the current dialogue context.

The present study further explores the additive effect of turn-taking cues. The results show that the more cues with the same pragmatic function, the faster the reaction time and the higher the agreement on the expected outcome. Thus, as hypothesized, the higher number of turn-yielding cues, the higher the expectations of a turn-change and the higher number of turn-holding cues, the higher the expectations of a speaker continuation. This is in line with Duncan's findings.

The objective of the present study was to identify human-like turn-taking strategies that can be produced with a synthetic voice in order to communicate appropriate places for dialogue system users to take the turn. The results show that turn-taking cues presented with a synthesis have a similar effect as cues presented with a human voice. As for cues presented with a human voice, an increased number of simultaneous turn-holding cues increased the expectations of a HOLD and an increased number of turn-yielding cues increased the expectations of a SWITCH. No differences in reaction times were found between the two conditions. Furthermore, analyses of the judgement distribution indicate that the effects of the individual cues as well as the additive effect of the cues are very similar for the synthetic and the human voice (see Table 3 and Table 5 respectively).

The experiment was designed to allow the subjects to follow the dialogues in chronological order and get familiar with the speakers and the dialogues in a way that is similar to how dialogue is perceived in a real conversation. However, this restricted the number of dialogues and speakers used as stimuli. The analyses of the reaction times suggest that one speaker was judged more easily than the other speakers were. A possible explanation is that speaker S1 occurred more frequently and the subjects got familiar with this speaker's particular turn-taking strategies. Still, the additive effect of the cues was similar for all speakers. Differences between speakers and how speakers adjust their turn-taking strategies to their dialogue partners are interesting areas for future research.

Finally, turn-holding cues were judged significantly faster than turn-yielding cues and the judgement distribution show that 87% of the listeners expected stimuli annotated with one turn-

holding cue to be followed by a HOLD, whereas only 62% of the listeners expected stimuli annotated with one turn-yielding to be followed by a SWITCH. These results suggest that the outcome of turn-holding cues is more predictable than turn-yielding cues. For stimuli with contradictory cues, that is stimuli with both turn-yielding and turn-holding cues, the judgements were almost equally distributed between HOLD and SWITCH.

In conclusion, we have shown experimentally that there are a number of behaviours near the end of the previous speaker turn that affect listeners' expectations of a speaker change. Furthermore, the synthesis affects listeners' expectations of a turn change in a way that is similar to a human voice.

8 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 valuable comments and annotation of data. Many thanks also to the reviewers for valuable comments that helped to improve the paper.

References

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

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. (2002). Speaking in time. *Speech Communication, 36*(1-2).

Cutler, A.., & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In Johns-Lewis, C. (Ed.), *Intonation and discourse* (pp. 139–155). London: Croom Helm.

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.

Fernández, R., Schlangen, D., & Lucht, T. (2007). Push-to-talk ain't always bad! Comparing Different Interactivity Settings in Task-oriented Dialogue. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 25-31). Trento, Italy.

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

Ferrer, L., Shriberg, E., & Stolcke, A. (2003). A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*. Hong Kong.

Ford, C., & Thompson, S. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, E., Schegloff, E., & Thompson, A. (Eds.), *Interaction and grammar* (pp. 134-184). Cambridge: Cambridge University Press.

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 Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008) (pp. 293-296). Berlin/Heidelberg: Springer.

Heldner, M., & Edlund, J. (In press). Pauses, gaps and overlaps in conversations, Journal of Phonetics.

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

Hjalmarsson, A.. (2008). Speaking without knowing what to say... or when to end. In Proceedings 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.

Kendon, A.. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

Kilger, A.., & Finkler, W. (1995). Incremental Generation for Real-Time Applications. Technical Report RR-95-11, German Research Center for Artificial Intelligence.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech, 41*, 295-321.

Levelt, W. (1989). *Speaking-From Intention to Articulation.* The MIT Press.

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.

Schourup, L. (1999). Discourse markers. *Lingua, 107*(3-4), 227-265.

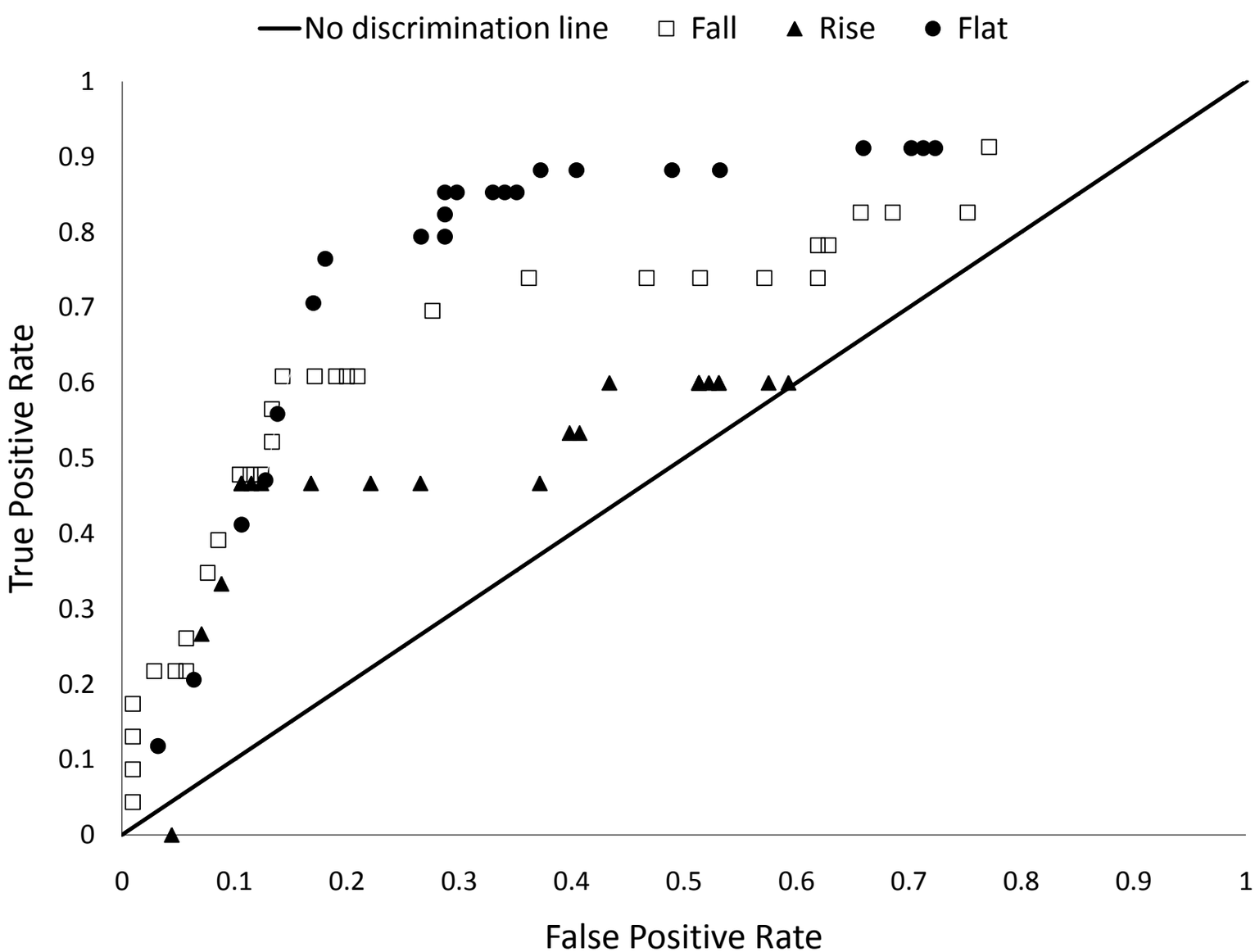
Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turnconstructional units and turns in conversation. *Pragmatics, 6*, 357-388.

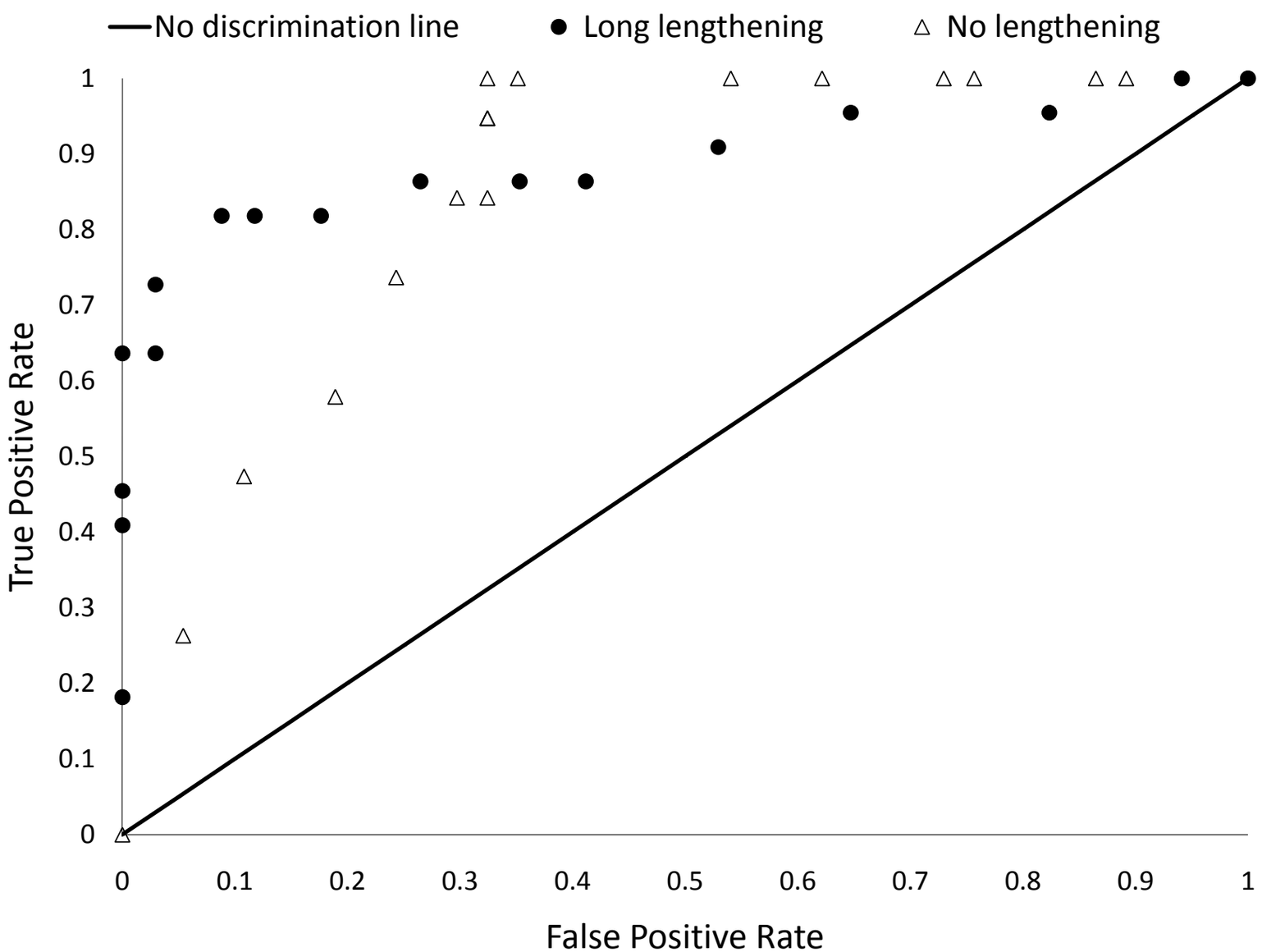
Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. Doctoral dissertation, University of California.

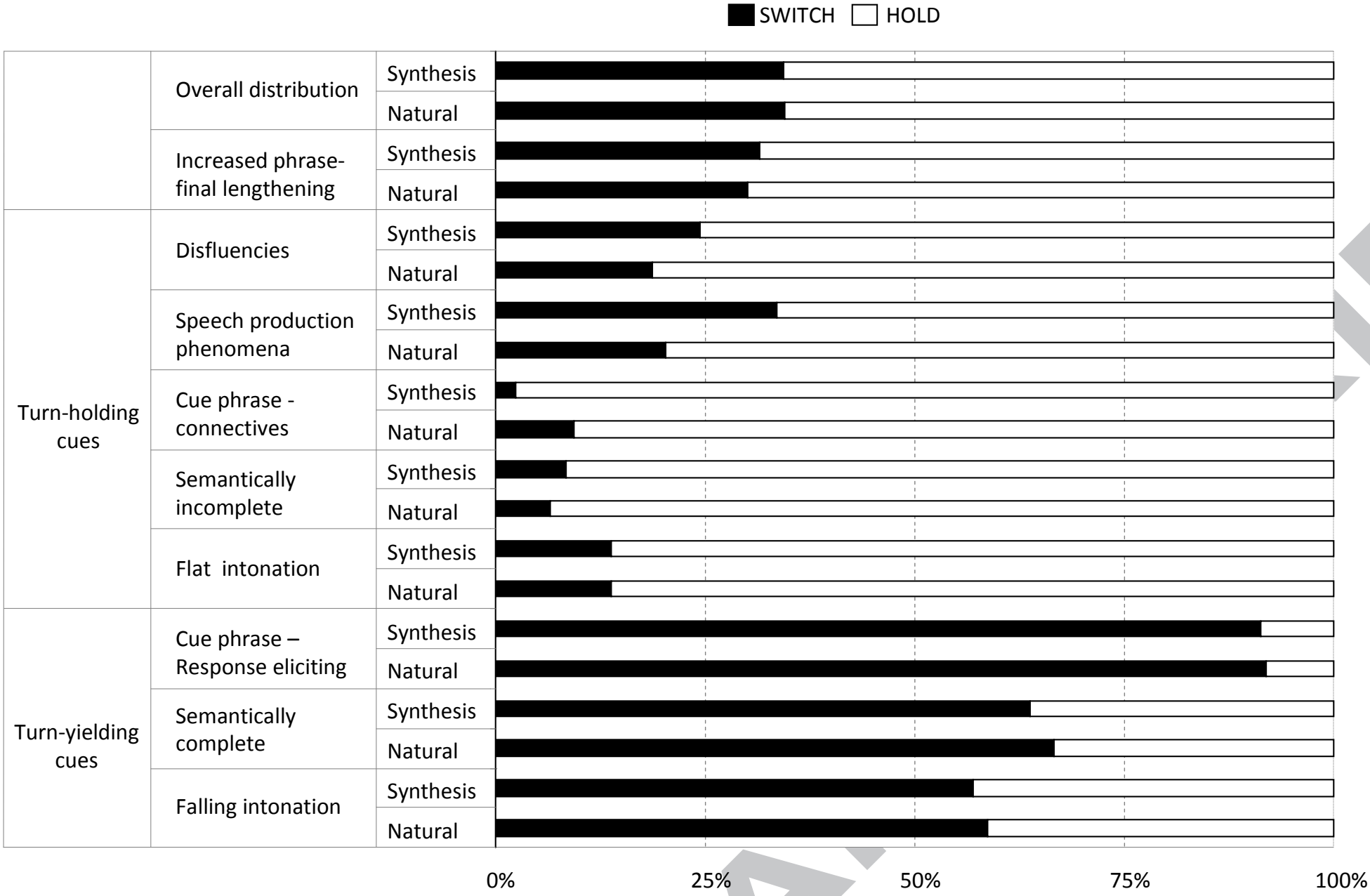
Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech communication, 50*(2), 81-94.

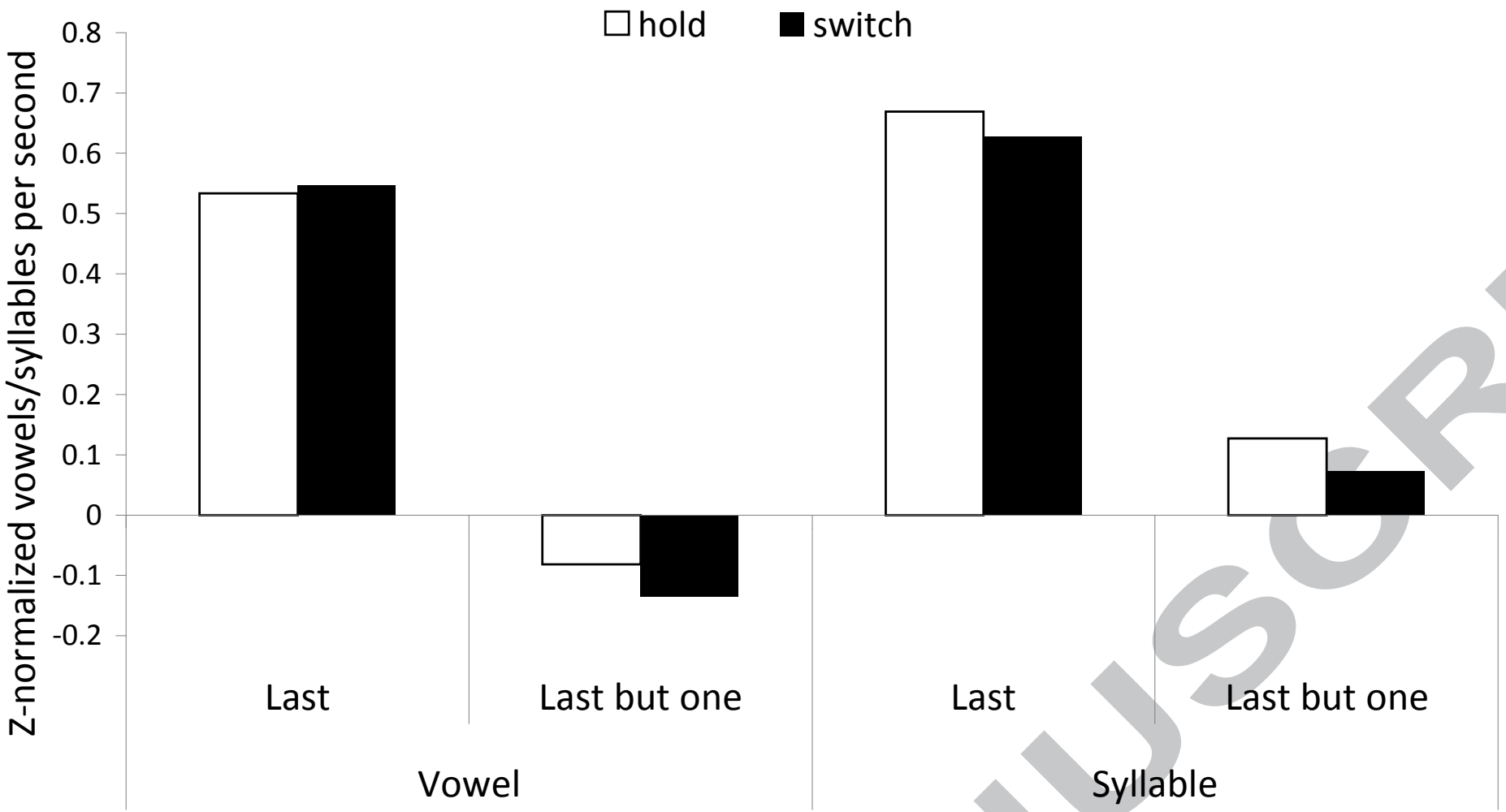
Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. In *ICPhS 2003*. Barcelona, Spain.

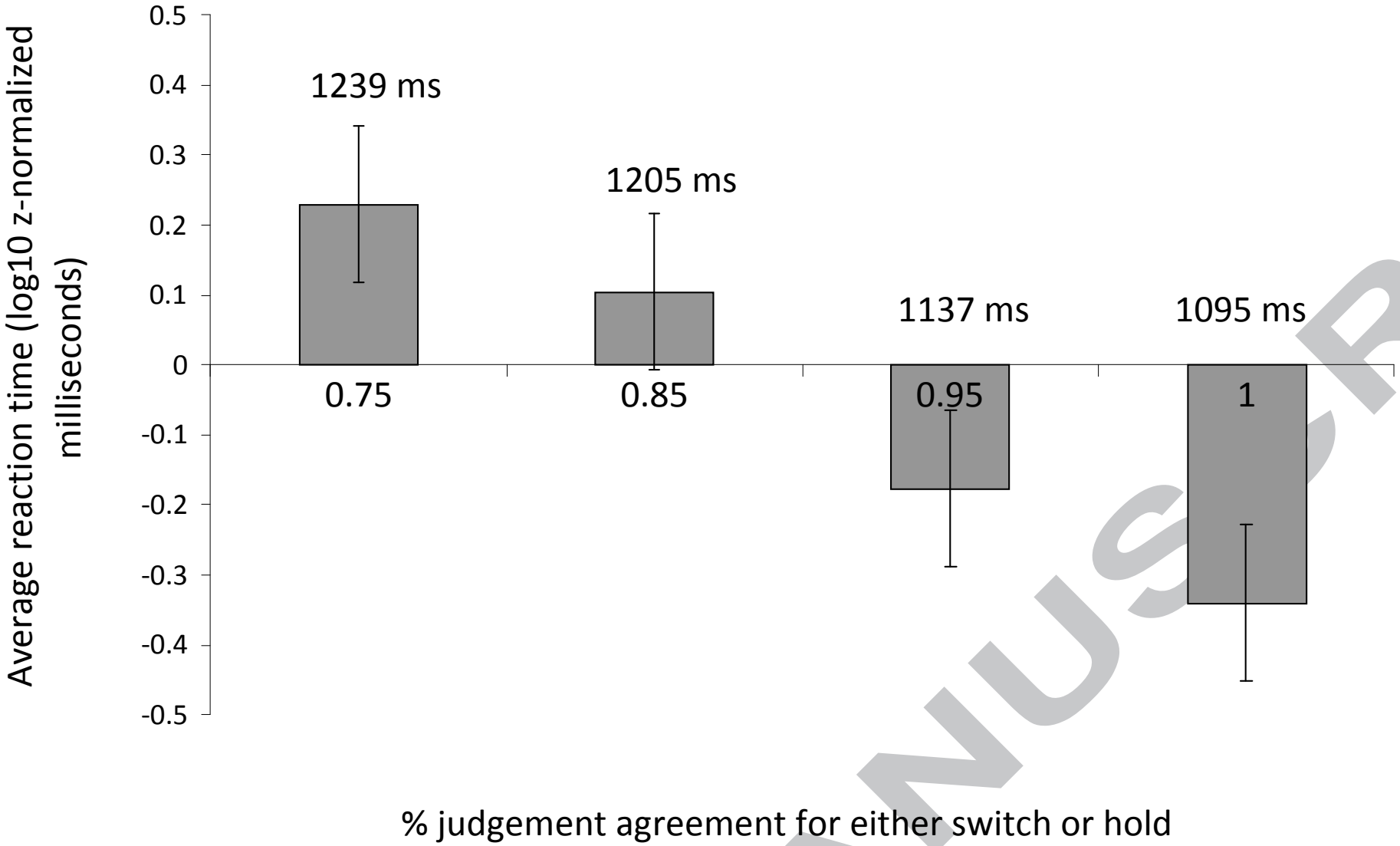
de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language, 82*(3), 515-535.






SWITCH
HOLD
Turn-holding cues
Turn-yielding cues
Overall distribution
Increased phrase-final lengthening
Disfluencies
Speech production phenomena
Cue phrase - connectives
Semantically incomplete
Flat intonation
Cue phrase – Response eliciting
Semantically complete
Falling intonation
Synthesis
Natural
0%
25%
50%
75%
100%


hold
switch
Z-normalized vowels/syllables per second
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
Last
Last but one
Last
Last but one
Vowel
Syllable


Average reaction time (log10 z-normalized milliseconds)
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
1239 ms
1205 ms
1137 ms
1095 ms
0.75
0.85
0.95
1
% judgement agreement for either switch or hold