

**Manuscript for the thematic issue of *Phonetica* entitled  
Progress in Experimental Phonology: From  
communicative function to phonetic substance and vice  
versa**

**Title of this manuscript: Exploring prosody in interaction  
control**

Authors: Jens Edlund & Mattias Heldner (names in alphabetical order)

Affiliation: KTH Speech, Music and Hearing, Stockholm, Sweden

Short title: Prosody in interaction control

Full address: KTH Speech, Music and Hearing, Lindstedtsvägen 24, SE-100 44 Stockholm,  
Sweden

Phone numbers: +46(0)8-790 7874 (Jens), +46(0)8-790 7563 (Mattias), +46(0)8-790 7854  
(fax)

E-mail: {mattias|[edlund](mailto:jens.edlund@speech.kth.se)}@speech.kth.se

## **Abstract**

This paper investigates prosodic aspects of turn-taking in conversation with a view to improving the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor. It examines the relationship between *interaction control*, the communicative function of which is to regulate the flow of information between interlocutors, and its phonetic manifestation. Specifically, the listener's perception of such interaction control phenomena is modelled. Algorithms for automatic online extraction of prosodic phenomena liable to be relevant for interaction control, such as silent pauses and intonation patterns, are presented and evaluated in experiments using Swedish Map Task data. We show that the automatically extracted prosodic features can be used to avoid many of the places where current dialogue systems run the risk of interrupting their users, and also to identify suitable places to take the turn.

## **1 Introduction**

Conversation is a primary means for human communication. An important function of conversation is to exchange propositional content, and during this exchange the interlocutors must somehow regulate the flow of information to make it proceed smoothly and efficiently. This *interaction control* is a collaborative effort where interlocutors continuously monitor various aspects of each other's behaviour, including semantics, gestures and prosody, in order to for example make decisions about *turn-taking* and *feedback*. The term turn-taking includes how to take the floor without interrupting, how to keep the floor, and how to pass the initiative on to others. Feedback is used to indicate to the speaker that the listener is attentive, understanding, agreeing, etc.

The aim of this work is to improve the interaction control in spoken human-computer dialogue. The primary motivation for this is that if a spoken dialogue system is to be

perceived as a good conversational partner, it has to be able to understand the speaker's organisation of the content, and to know how to time its own contributions to the dialogue appropriately, in addition to recognising and responding to verbal input and generating verbal output.

Current spoken dialogue systems commonly detect where the user ceases speaking in order to find out where they should take their turn. The method is based on the assumption that speakers have finished what they intended to say when they become silent, and that these points in time are also suitable places for the system to speak. Such *endpoint detection* triggers on a certain set amount of silence, or non-speech. The method makes sense; given that a speaker is allowed to complete what she/he intends to say, the end of the utterance is likely to coincide with silence at a place where an interlocutor might take the next turn. The method segments speech into reasonably sized units, in many cases corresponding to sentences or some sentence-like units. However, spontaneous conversational speech frequently contains silent pauses inside what we would intuitively group into turns, complete utterances or sentence-like units, and inside what are indeed semantically coherent units. Typical examples are hesitations such as “*You will get to a eh <long silence> well what shall we call it*”. Silences (or hesitations) may even occur inside prosodic words or compound words: “*There is a eh cycleway and some horse- <long silence> path maybe*”. Thus, dialogue systems using silence-based endpoint detection run into problems with unfinished utterances when encountering spontaneous speech, as the silences following them are not likely to be suitable places for the system to speak, and as unfinished utterances may be difficult to interpret [Bell et al., 2001].

There are a number of common tasks within speech technology and natural language processing where it would be useful to perform automatic segmentation into units that better match what humans perceive as finished utterances. The following two groups summarise our primary motivation for this work, but other applications may apply.

(i) System barge-in. Computers may want to use speech to notify human interlocutors about various events, for example “coffee is ready”. This is perhaps analogous to a meeting secretary, and it is important that the barge-in behaviour is perceived as polite.

(ii) Interaction control, turn-taking, feedback. Segmentation of user input is essential in the *conversational* components of a dialogue system, notably for identifying suitable places to speak [Heldner et al., forthcoming]. A dialogue system is likely to be perceived as a better conversational partner if it has a clearer idea of when human interlocutors have finished talking, and if it is able to respond rapidly.

In advanced spoken dialogue systems, spoken language understanding and interaction control are combined. The AdApt system, for example, uses a semantically based approach in order to deal with the problems that occur as a result of silence-only based utterance segmentation [Bell et al., 2001]. Another semantic approach is used in Skantze & Edlund [2004].

The work presented here represents a continuation of work presented elsewhere [Edlund et al., 2005; Heldner et al., forthcoming]. Here, we explore online prosodic analysis with the aim of bringing human-like interaction control capabilities to conversational computers. More specifically, we look at prosodic phenomena liable to be relevant for interaction control, such as silent pauses and intonation patterns. It is worth noting that although our primary goal is to create natural, human-like spoken dialogue systems, human-computer conversation experiments provide good evidence as well as counter-evidence for the relationship between interaction control phenomena and their prosodic manifestations.

## **2 *Prosodic phenomena and interaction control***

Previous work suggests that a number of prosodic or phonetic cues are associated with turn yielding and thus potentially relevant for interaction control. These cues include

phenomena such as silent pauses; various intonation patterns (rises, falls, down-steps, up-steps); decreases in speech rate; final lengthening; intensity patterns; centralized vowel quality; creaky voice quality; and exhalations. Note that both rises and falls have been associated with turn-yielding. These cues are typically located somewhere towards the end of the turn, although not necessarily on the final syllable [e.g. Ford and Thompson, 1996; Local and Kelly, 1986; Local et al., 1985; Local et al., 1986; Ogden, 2001; Wells and MacFarlane, 1998].

Similarly, there are studies suggesting that certain prosodic or phonetic cues are associated with turn keeping, and these cues are of course also potentially relevant for the interaction control. They include phenomena such as glottal or vocal tract stops without audible release; a different quality of silent pauses as a result of these glottal or vocal tract closures; assimilation across the silent pause; and other “held” articulations (e.g. lengthened vowels, laterals, nasals or fricatives) [Local and Kelly, 1986; Ogden, 2001].

Certain intonation patterns are associated with turn-keeping. In particular, level intonation patterns in the middle of the speakers’ fundamental frequency range have been observed to act as turn-keeping cues in several different languages. For example, Duncan [1972] reported that any pattern other than a level tone in the speaker’s mid register (a 2 2 | pattern in the Trager-Smith prosodic transcription scheme [Trager and Smith, 1957]) signals turn-yielding in English. Thus, the mid level pattern acts as a turn-keeping signal, although Duncan did not use that term. Similarly, Selting [1996] reported that level pitch accents before a pause are used to signal a turn-holding (or turn-keeping) in German; Koiso, Horiuchi, Tutiya, Ichikawa, and Den [1998] observed that flat, flat-fall and rise-fall intonation patterns tended to co-occur with speaker holds (i.e. turn-keeping) in Japanese; and in another study on Japanese conversations Noguchi & Den [1998] reported that flat intonation at the end of pause bounded phrases acts as an inhibitory cue for backchannels. Furthermore, in a study of final pitch accents and boundary tones in the turn-taking system of Dutch, Caspers [2003] identified two intonation patterns that seem to be associated with turn-keeping: an accent lending rise followed by level high pitch (H\* %) used for bridging

syntactic breaks between utterances; and a filled pause with a mid level boundary tone (M %) for bridging hesitations within syntactic constituents. However, Caspers could not find any intonation patterns clearly associated with turn-yielding. This observation lead her to conjecture that turn changing is the unmarked case and that only the wish to keep the turn need to be marked with specific intonation patterns.

When delving into this impressive body of work, it is worth noting that the methods employed were impressionistic auditory analyses or manual acoustic analyses for the most part. The observations rely on full access to the knowledge of trained (interactional) linguists. This is perhaps reflected in the fact that several cues in the above mentioned studies were defined with reference to either the metrical structure or to the accentual structure of the utterances [e.g. Local et al., 1986; Wells and MacFarlane, 1998]. In order to capture these cues, accurate classification into metrically heavy and light syllables, as well as into (focally) accented and non-accented words is required. These classifications in turn require access to information that is not present in the acoustic signal alone. Consequently, it may not be possible to operationalise all of the above mentioned cues for use in online, automatic systems. Some cues however, such as intonation patterns immediately before silent pauses are more readily available for automatic analysis.

Silent pauses are frequently used for chunking the speech stream into manageable units for speech technology applications. The end-of-utterance detectors in current automatic speech recognition typically rely exclusively on a silence threshold somewhere between 500 and 2000 ms (sic!) for delimiting the units to be recognised [cf. Ferrer et al., 2002 and references mentioned therein]. That is to say that the output of the recogniser comes in chunks corresponding to speech bounded by 'long enough' silent pauses; that recognisers using these pauses may deliver the users' dialogue contributions to the dialogue system only when there is a 'long enough' pause; and furthermore that the system response times are long.

However, as noted above, silent pauses may be indicative of turn-yielding as well as of turn-keeping, and spontaneous speech frequently contains silent pauses also within segments we would intuitively call utterance units, and within segments that are indeed semantically coherent units. Presumably, the uncomfortably long silence requirement of 2000 ms mentioned above is used in the hope of excluding such utterance internal silences. Part of the goal with this study is simply to assess the extent of this problem. Human listeners can discriminate these utterance-internal pauses from utterance-final ones using other prosodic cues, gestural cues, and knowledge of semantic completeness, but these pauses are often well above the silence thresholds in the end-of-utterance detectors. Moreover, these silences often occur before semantically heavy words without which the unit preceding the pause may be difficult to interpret, as in the examples above.

### **3 Method**

The method used in the present study is in many ways similar to those used by Koiso et al [1998] and Caspers [2003]. That is, Map Task dialogues were segmented into pause bounded units – so called interpausal units (IPUs). The transitions from one IPU to the next were classified in terms of speaker changes and speaker holds. Prosodic features were extracted from the region immediately before the boundary, under the assumption that information relevant to interaction control is localised just before the speaker changes or holds. In addition, we supplemented the speech material with several kinds of markup and offline analyses, and we evaluated the predictive power of the automatically extracted prosodic features for making interaction control decisions.

#### **3.1 Speech material**

The speech data used for the present study consists of Swedish Map Task dialogues recorded and transcribed by Pétur Helgason at Stockholm University [Helgason, 2002].

Map tasks were designed to elicit natural-sounding spontaneous dialogues [Anderson et al., 1991]. There are two participants in a Map Task dialogue: an instruction-giver and an instruction follower. They have one map each, and these maps are similar but not identical. For example, certain landmarks on these maps may be shared, whereas others are only present on one map, some landmarks occur on both maps but are in different positions etc. The task is for the instruction-giver to describe a route indicated on his or her map to the follower.

The Swedish Map Task data used in this study consists of recordings of two pairs of speakers. Within each pair, each speaker acted as giver once and as follower once. They were recorded in an anechoic room at the Phonetics lab, Stockholm University, using close talking microphones, and facing away from each other. They were recorded on separate channels with a good separation of the channels. In other words, the data was obtained under close to ideal conditions. The tasks elicited spontaneous dialogues with a number of disfluencies and hesitations, and where almost every turn was acknowledged by another turn or by means of verbal feedback. There were four dialogues containing about 1100 dialogue contributions (including feedback or backchannels) and the total duration of the four dialogues was about one hour.

These recordings were accompanied by verbatim transcriptions, tentatively segmented into dialogue contributions, but there was no indication of stretches of overlapping speech. For further details on the Swedish Map Task data, see Helgason [2002].

### **3.2 Segmentation into IPUs**

As in Koiso et al [1998] and Caspers [2003], the material was segmented into interpausal units (IPUs). The details of our segmentation, however, differed from the previous ones in several respects. First, each of the giver and the follower channels was automatically segmented into speech and silence using a basic speech activity detector (SAD). The frame-

level decisions produced by the SAD were smoothed and a minimum silence of 300 ms was required. The minimum silence decreases the likelihood of detecting pauses in the occlusion phases in stops (that often exceed the 100 ms threshold used in previous studies). Second, each transition from speech to silence in the giver's channel where there was no overlapping speech in the follower's channel was marked as an IPU boundary. Transitions from follower to giver were left out for simplicity, whereas omitting cases of overlapping speech was motivated by a wish to avoid interruptions and contributions where there is competition for the turn, since these are not what this study aims to model.

### **3.3 Speaker change vs. speaker hold classification**

Each IPU selected was labelled as an instance of either *speaker hold* (HOLD) or *speaker change* (CHANGE). The HOLD label was given to those IPUs that were followed by a contribution from the same speaker, that is the giver (recall that only the giver IPUs were investigated). The IPUs that were followed by a contribution from the other speaker (i.e. the follower) were labelled CHANGE. This markup was also made automatically, based on the acoustic signal in the giver and follower channels. Note, however, that this markup makes no distinction between turns and backchannels. If a giver contribution was succeeded by a follower contribution, there was a CHANGE irrespective of whether the follower contribution was a backchannel. Thus, transitions that Koiso et al [1998] and Caspers [2003] would have classified as HOLD due to backchannels on either side of the silent pause were classified as CHANGE here. The speaker change vs. speaker hold classification was later used as a gold standard for the predictions based on prosody.

This markup shows the actual turn of events in the dialogue: it is a direct reflection of on the interlocutors' behaviour ensuring that the speaker changes and holds were perceived as such by the participants as well. It does not, however, show how things must be by necessity. A speaker change may, for example, be an unsuitable place for a polite contribution if the follower interrupted the giver in the actual dialogue. Similarly, a speaker

hold may be a suitable place for the follower to give a contribution, but one where the follower simply refrained from saying something. In addition, it is far from obvious that a place where the follower contributed a backchannel is a suitable place for anything but a backchannel. The opposite, a backchannel instead of some other turn, is probably more acceptable.

A speaker CHANGE then, in our approach, is speech in the giver channel followed by at least a 300 ms silence in the same channel, and non-overlapping speech in the follower channel (see Figure 1). Thus, we exclude a small number of potentially interesting cases where there is a speaker change with a slight non-competitive overlap. For reasons of temporal resolution, the minimum amount of silence between giver and follower in a speaker change transition is 10 ms. Correspondingly a speaker HOLD is speech in the giver channel followed by at least a 300 ms silence and then more speech in the same channel. The minimum amount of silence between contributions in HOLD transitions is thus 300 ms.

In addition to the segmentation of IPUs, we extracted the *inter-contribution* intervals (ICI) – the actual durations of silent pauses between the IPUs (i.e. inter-speaker intervals in the case of speaker change, and intra-speaker intervals in the case of speaker holds). Again, this was done automatically and with a temporal resolution of 10 ms.

INSERT FIGURE 1 ABOUT HERE!
-----------------------------

### 3.4 Perceptual judgments

As the classification of speaker change vs. speaker hold only illustrates the actual events, not how things have to be, judgments of how each IPU was perceived by human listeners were added as a reality check. The task for the judges was to decide, based on a presentation of two seconds of speech prior to the silence (but not what followed after the silence), whether the speaker was finished or not at the end of the IPU. Three judges (two

of which are the present authors) independently judged every IPU in the speech material (824 cases) on a five-point scale where 1 represented definitely unfinished; 2 probably finished; 3 could be unfinished or could be finished; 4 probably finished; and 5 definitely finished. 18 judgments were lost for technical reasons yielding a total of 2454 judgments.

Assuming that it is suitable to take the turn after IPUs that are perceived as finished, these judgments may be seen as a markup of potential places for speaker changes. Correspondingly, assuming that it is unsuitable to take the turn after IPUs judged to be unfinished, such judgments indicate places perceived as unsuitable for speaker changes.

### **3.5 Prosodic feature extraction with /nailon/**

Primarily, this research aims at finding and describing the turn-taking information that is encoded in the speech signal, and how this information can be extracted. One way of accomplishing that is to test extracted results against some gold standard – a flawless record often assembled by asking human judges or by somehow describing human behaviour. In our case, the matter is made more complicated by our wish to extract only such information that humans have access to in real dialogue situations – we are for example limited to left context only. Limiting ourselves to such information as is accessible to humans is not just a method to safe-guard against some possibly confounding variables – it is also necessary if the resulting analysis is to be useful in practical applications, such as spoken dialogue systems. The prosodic analysis tool described here is on-line in the sense that it uses no acoustic right context, or look-ahead. On the acoustic level, this goes well with human circumstances. Humans rarely need acoustic right context to make decisions about speech segmentation. On the contrary, they often seem to be able to predict turn endings and suchlike. Naturally, semantic expectations provide quite considerable “look-ahead” to humans, and in an ideal system these should be used in conjunction with acoustic analysis. Furthermore, albeit not a theoretical requirement, any implementation must run in real-time in order to be useful as far as live user studies are concerned. The present implementation is

real-time in the sense that it performs in real time, with a small and constant latency, on a standard PC.

The on-line real-time prosodic analysis is implemented within /nailon/ (a phonetic anagram for online), a Tcl/Tk package based on the Snack Sound Toolkit [<http://www.speech.kth.se/snack/>]. /nailon/ uses Snack to manage sounds and to extract intensity, voicing and pitch information. In its present state, the package also captures speech duration, voiced speech duration, silence duration, and the relative position of intonation patterns in an online estimation of the speakers F0 range. The analysis is in some ways similar to that used by Ward & Tsukahara [2000], and is performed in several consecutive steps, including speech activity detection (SAD), voice, pitch and intensity extraction, pseudo-syllabification, and intonation pattern classification.

In order to be on-line as well as efficient, each step is performed on a small, continuously moving window consisting of a small number of 10 ms frames. In the experiments presented here, the window size is 300 ms, or thirty 10 ms frames. The latency of the system is a function of the frame size plus whatever processing time is needed for each step and each frame. The algorithms used here are implemented incrementally, keeping both processing and memory footprint to a minimum. Each processing step is described in detail below.

### **3.5.1 SAD**

A basic speech activity detector (SAD) is used to discriminate speech from non-speech (or silence). This decision is based on a noise threshold determined from the intensity distribution (simply the local minimum following the first local maximum in the distribution). A measure of the intensity (in dB) is computed for every 10 ms sound frame and the intensity distribution is updated continuously. Any frame with more energy than the threshold is marked as speech. The sequence of frame-level decisions is converted into

durations of speech and silence segments by requiring that a minimum number of consecutive frames (ten frames or 100 ms in the present experiments) be given the same classification in order for a change to be reported. This padding removes some of the effect of various low-energy components of speech such as fricatives, short silences such as the occlusion part in stops, and various short high energy segments embedded in silences. Although this SAD is simplistic, it is so far sufficient for our needs. SAD is a vivid research topic, but here, we are interested in the extraction of prosodic features, and from that point of view SAD is a prerequisite, not a research topic in itself. It will not be discussed further here.

### **3.5.2 Voice, pitch, and intensity extraction**

A pitch extractor (the ESPS `get_f0` program included in Snack) acquires information about voiced and unvoiced speech frames, and the F0 values of the voiced frames. This sequence of frame-level voicing decisions is used to compute durations of voiced and unvoiced speech, again under the requirement that a change is stable over a number of frames for it to be reported, to allow for artefacts introduced by the pitch tracker. The F0 values in Hertz are transformed into semitones relative to a fixed value and smoothed using a median filter (currently over 9 frames). The median filter is applied separately to correct for the pitch extractors inability to smooth the pitch curve over such short speech segments. The semitone transformed F0 data are then used to estimate speaker F0 range based on the cumulative distribution of F0 data. The F0 range is bounded by a topline and a baseline defined as the cumulative mean  $\pm 2$  standard deviations (also calculated cumulatively). The semitone scale is used to ensure that +1 standard deviations interval is the same musical and perceptual interval as  $-1$  standard deviations. The F0 range is divided into three equal parts: high, mid and low. Intensity is treated in a similar manner: the median filtered intensity in dB in each frame is used to incrementally build up a cumulative mean  $\pm 2$  standard deviations.

### **3.5.3 Pseudo-syllabification**

The pseudo-syllabification algorithm is loosely based on Mermelstein's [Mermelstein, 1975] technique to find intensity minima in the speech signal by using convex hulls. Convex hulls are continuously tracked over the median filtered intensity values. Note, however, that only voiced segments are included. The convex hulls are taken to correspond to pseudo-syllables, than can be retrieved and analysed as the need arises. In the present experiments, the last convex hull that has been seen over the intensity curve is extracted each time a sufficiently long (300 ms) silence is detected.

### **3.5.4 Intonation pattern classification**

Whenever a sufficient silence is found, the information from the pseudo-syllabification is used to point out a number of frames that roughly corresponds to the last syllable nucleus (i.e. minimally the vowel) before the silent sequence. The information from the pitch extractor pertaining to the pseudo-syllable is then used to classify the intonation patterns. These intonation patterns are classified in terms of their position in the F0 range, currently as high, mid or low tones, and in terms of their shapes: rises, falls, and level tones etc.

## **3.6 Turn-keeping and turn-yielding decisions**

Finally, the speech and silent pause durations in combination with the intonation patterns classification are used to make decisions about interaction control. Any silent pause in the giver channel exceeding the pause threshold (i.e. 300 ms), that is preceded by a mid level intonation pattern, belongs to the TURN-KEEPING category. Silent pauses followed by low intonation patterns are categorised as TURN-YIELDING. All other places belong to the DON'T-KNOW category.

The thresholds and parameter values were manually set in the initial implementation used in the tests reported here, but they were set before the tests were performed and have not been subsequently altered to improve results on the current speech material. In future versions, they should be optimised based on corpus studies.

### **3.7 Evaluating the predictive power of the automatically extracted prosodic features**

To evaluate the predictive power of the prosodic features extracted with /nailon/, the interaction control decisions made by /nailon/ were compared with what actually happened in the dialogues, that is the automatic classification into speaker change vs. speaker hold described above.

## **4 Results and discussion**

### **4.1 The extent of the problem with silence based segmentation**

By combining the results of the perceptual judgments and the automatic speaker change vs. speaker hold classification we can get an estimate of the extent of the problem with silence-based segmentation in spontaneous dialogues. Table 1 shows the distribution of perceptual judgments of the IPU's tabulated against the speaker change vs. speaker hold classification, that is, whether the IPU's were perceived as finished or not (on a five-point scale) tabulated against whether there actually was a speaker change or not.

INSERT TABLE 1 ABOUT HERE!
----------------------------

Several observations can be made. First, to state the obvious, all the speaker hold cases represent pause-bounded units where no speaker change occurred, and these cases correspond to 52.3% of all silent pauses in the material. Some of these speaker hold cases were judged as probably or definitely finished utterances, and may be seen as potential places for speaker changes only that no speaker change occurred. These cases are not part of the problem with silence based segmentation. However, by collapsing the votes for probably and definitely unfinished as well as those for probably and definitely finished and taking the majority vote across the three judges, we find that the speaker hold cases judged to be unfinished correspond to 36.7% of all pauses (cf. Table 2). These cases pose a real problem for silence-based segmentation since they are the cases where a dialogue system using silence-based segmentation runs the risk of interrupting its users. These are the cases we want to avoid.

INSERT TABLE 2 ABOUT HERE!
----------------------------

There was also a substantial number of pause bounded units where speaker changes actually occurred (47.7% of all pauses). By collapsing the votes, this time for probably and definitely finished, and then taking the majority vote across the three judges, we find that the majority of the speaker change cases (83.5%) were indeed perceived as finished utterances (cf. Table 2). So these are the places where we want the system to speak. Finally, there were also a few cases where speaker changes occurred although the judges felt that they were probably or definitely unfinished (15%). Some of these were probably deliberate interruptions, and therefore not places where we want to detect an opportunity for the system to speak.

In addition to illustrating the problem with silence based segmentation, Tables 1 and 2 give an indication of the amount of noise in the data. In the majority votes by the three human listeners 27.8% of the HOLD cases were perceived as finished, and 15% of the CHANGE cases were perceived as unfinished (cf. Table 2). In our view, this points to the minimum

error we can expect from an automatic categorisation, unless it has access to more information than the human judges were given.

## **4.2 Would increasing the silence threshold solve the problem with silence based segmentation?**

The minimum amount of silence required for IPU segmentation in this study is 300 ms. Humans often respond even faster than that, yet longer minimum pauses are often used in spoken dialogue systems in the hope of ensuring correct turn-taking behaviour. The question, then, is: does an increased silence threshold solve any problems involved in silence based segmentation? Table 3 presents the proportion of IPUs that resulted in a speaker hold in five subsets of the material, each consisting of the IPUs with an ICI above a certain threshold.

INSERT TABLE 3 ABOUT HERE!
----------------------------

In order to assess the amount of errors created by using silence as the sole basis for turn-taking decisions, and to find out how successful the strategy of lengthening the required silence is, we analysed the Map Task dialogues to see how HOLD and CHANGE depended on ICI. As noted before, most current systems use silence thresholds of 500 - 2000 ms, and our system currently requires 300 ms of silence. First of all, it is worth noting that less than 50% of the turns in the Map Task dialogues had an ICI of more than 500 ms, implying that continuously waiting for 500 ms or more before responding is not natural. Considerably more – over 70% of the turns, are captured if the minimum ICI is lowered to 300 ms. Even more interestingly, the percentage of HOLDS is virtually unchanged whether one looks at all IPUs with a minimum ICI of 300, 500, 1000, or 1500 ms. At ICIs of 2000 ms or more, the percentage of HOLDS falls somewhat, but the material is too small to be reliable – we only have 19 instances to go on. The numbers suggest that the only thing a spoken dialogue system would achieve by using a silence threshold of 2000 ms is a sluggish behaviour.

### 4.3 Can prosody help in solving the problem with silence based segmentation?

Requiring longer silences obviously does nothing to solve the problems involved in silence based segmentation, but can prosody and intonation patterns help? Recall that level intonation patterns in the middle of the speakers' fundamental frequency range have been observed to act as turn-keeping cues in several different languages.

Based on the prosodic features extracted with /nailon/, the automatic classifier classified each IPU into one out of three categories: TURN-KEEPING, DON'T-KNOW, and TURN-YIELDING. Table 4 presents the results tabulated against the speaker CHANGE/HOLD classification. The material contained 824 IPUs, 28% of which were classified as TURN-YIELDING, 16% AS TURN-KEEPING, and 56% ended up in the garbage category DON'T-KNOW.

INSERT TABLE 4 ABOUT HERE!
----------------------------

Among the IPUs classified as suitable for turn-taking (i.e. TURN-YIELDING), the speaker CHANGES were in the majority (69%). Correspondingly, the speaker HOLDS were in the majority in the IPUs classified as TURN-KEEPING (82%). From a different point of view, the classifier identified at least 41% of the possible places for turn-taking (recall that some of the speaker HOLDS were perceived as finished by the human listeners – these may well be suitable places although no speaker change actually occurred, cf. Table 1). Furthermore, if only the TURN-YIELDING IPUs are seen as suitable places for turn-taking (and that TURN-KEEPINGS and DON'T-KNOWS are pooled), the classifier identified 84% of the places where interruptions are impossible.

In a spoken dialogue system, the fairly large garbage category produced by the parameter settings in this experiment can be pooled in two ways. If used conservatively, DON'T KNOW

and TURN-KEEPING would be pooled and the system only takes turn at IPUs categorised as TURN-YIELDING. Such a system would avoid 84% of the IPUs unsuitable for system utterances, and detect roughly every second opportunity for a system utterance. The approach is suitable for a system that barges in into human conversations, perhaps in order to give notifications et cetera. Conversely, a more aggressive system would pool DON'T KNOW with TURN-YIELDING and only avoid taking turns when the categorisation gave TURN-KEEPING. This would avoid 24% of the unsuitable places whilst detecting 94% of the suitable places. The latter strategy would be used in a system that engages in a *dialogue* directly with a user.

It is possible to improve these results if the thresholds used in the classifier are tuned with machine learning techniques, and further subdivision of the DON'T KNOW category, for example to add categories such as backchannels, listings, feedback on various grounding levels [Allwood et al., 1993; Clark, 1996]. We intend to explore these possibilities in future work.

## **5 Conclusions**

In this paper, we have explored online prosodic analysis as a means to improve the interaction control in spoken human-computer dialogue. The experiments have shown that dialogue systems relying on silence-based segmentation run the risk of interrupting its users in as much as 35% of all silent pauses, at least if they encounter speech of the kind investigated here.

Furthermore, we have shown that the number of incorrect turn-taking decisions can be reduced substantially by combining standard silence based endpoint detection with an automatic classification of intonation patterns. In the process, it is also possible to decrease the length of the required silence without any loss in performance. This can be used to make a conversational computer more responsive by allowing it to reply faster without simultaneously making it more obtrusive.

Level intonation patterns in the middle of the speakers' fundamental frequency range were found to act as turn-keeping cues, and may thus be used to avoid interrupting human interlocutors with high precision. Although there are several observations of the function of these mid level intonation patterns, to our knowledge they have never been used for avoiding interrupting users of spoken dialogue systems before.

Rising intonation before a silent pause can be associated with turn-yielding as well as with turn-keeping [e.g. Local et al., 1986]. As 51% of the rising intonation patterns co-occurred with actual speaker changes and 49% with speaker holds, we opted to have /naillon/ classify these as DON'T KNOW. However, it is evident that the classification would benefit from a more thorough analysis of rising intonation preceding silent pauses. We suspect that a fourth *conditional turn-yielding* category is needed (in addition to TURN-YIELDING, TURN-KEEPING, and DON'T KNOW) to capture places where only certain types of contributions, such as positive feedback, objections, and clarification requests, are permitted. We intend to explore these possibilities presently.

On a final note, it is clear that prosodic analysis is not enough to create conversational computers with interaction control skills at near-human levels. Humans use higher levels of understanding and a variety of information. We feel, however, that online access to prosodic information provides a valuable source of information that should be combined with other sources to guide the interaction control in conversational dialogue systems.

## **6 Acknowledgments**

We are grateful to the two anonymous reviewers for their comments on an earlier version of this paper. This work was done within the Project CHIL "Computers in the Human Interaction Loop" (IP 506909). CHIL is an Integrated Project under the European Commission's Sixth Framework Program.

## **7 References**

Allwood, J.; Nivre, J.; Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. *J Semant.* 9 (1993).

Anderson, A.H.; Bader, M.; Bard, E.G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H.; Weinert, R.: The HCRH Map Task Corpus. *Lang. Speech* 34: 83-97 (1991).

Bell, L.; Boye, J.; Gustafson, J.: Real-time handling of fragmented utterances; in *Proceedings of NAACL 2001* 2001).

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

Clark, H.H.: *Using language.* (Cambridge University Press, Cambridge 1996).

Duncan, S., Jr.: Some signals and rules for taking speaking turns in conversations. *J Personal. Soc. Psychol.* 23: 283-292 (1972).

Edlund, J.; Heldner, M.; Gustafson, J.: Utterance segmentation and turn-taking in spoken dialogue systems; in Fisseni, Schmitz, Schröder, Wagner, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, pp. 576-587 (Peter Lang, Frankfurt am Main, Germany 2005).

Ferrer, L.; Shriberg, E.; Stolcke, A.: Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog; in *Proceedings ICSLP'02*, pp. 2061-2064, Denver 2002).

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

Heldner, M.; Edlund, J.; Carlson, R.: Interruption impossible; in Horne, Bruce, *Proceedings of Nordic Prosody IX* (Peter Lang, Frankfurt am Main forthcoming).

Helgason, P.: Preaspiration in the Nordic languages: Synchronic and diachronic aspects. (Department of Linguistics, Stockholm University, Stockholm 2002).

Koiso, H.; Horiuchi, Y.; Tutiya, S.; Ichikawa, A.; Den, Y.: An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Lang. Speech 41*: 295-321 (1998).

Local, J.; Kelly, J.: Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies 9*: 185-204 (1986).

Local, J.K.; Wells, W.H.G.; Sebba, M.: Phonology for conversation: Phonetic aspects of turn delimitation in London Jamaican. *J Pragmat. 9*: 309-330 (1985).

Local, J.K.; Kelly, J.; Wells, W.H.G.: Towards a phonology of conversation: turn-taking in Tyneside English. *J Linguist. 22*: 411-437 (1986).

Mermelstein, P.: Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America 58*: 880-883 (1975).

Noguchi, H.; Den, Y.: Prosody-based detection of the context of backchannel responses; in *Proceedings of the 5th International Conference on Spoken Language Processing*, pp. 487-490, Sydney, Australia 1998).

Ogden, R.: Turn transition, creak and glottal stop in Finnish talk-in-interaction. *J IPA* 31: 139-152 (2001).

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

Skantze, G.; Edlund, J.: Robust interpretation in the Higgins spoken dialogue system; in *Proceedings of Robust 2004, Norwich 2004*.

Trager, G.L.; Smith, H.L.: *An outline of English structure*. (American Council of Learned Societies, Washinton, D.C. 1957).

Ward, N.; Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32: 1177-1207 (2000).

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

## 8 Tables

Table 1. The distribution of perceptual judgments by the three judges as to whether the IPUs were finished or not tabulated against the automatic speaker change vs. s speaker hold classification.

	1 (definitely unfinished)	2 (probably unfinished)	3 (could be either finished or unfinished)	4 (probably finished)	5 (definitely finished)	Total
Speaker change	169	66	40	200	705	1180
Speaker hold	806	91	25	112	240	1274
Total	975	157	65	312	945	2454

Table 2. The distribution of the majority votes for the perceptual judgments. Before determining the majority votes across the three judges, the probably and definitely unfinished cases were collapsed into the category unfinished; and the probably and definitely finished votes were collapsed into the category finished. The column labeled disagree represent the cases were no majority vote could be established.

	Unfinished	Finished	Disagree	Total
Speaker change	59	328	6	393
Speaker hold	303	120	8	431
Total	362	448	14	824

Table 3. Number of inter-pausal units (IPUs) and the proportion of speaker HOLDS amongst these as a function of increased thresholds in silence based segmentation.

	> 300 ms	> 500 ms	> 1000 ms	> 1500 ms	> 2000 ms
# IPUs	634	441	168	69	22
% HOLD	68	71	68	67	55

Table 4. Automatic classification of IPUs into turn-keeping, turn-yielding and don't-know tabulated against the classification of speaker changes and speaker holds.

	TURN-KEEPING	DON'T-KNOW	TURN-YIELDING	TOTALS
CHANGE	23	212	158	393
HOLD	105	255	71	431
TOTALS	128	467	229	824

## 9 *Figure 1 (formatted using a MS Word table)*

Figure 1. Schematic illustration of a speaker change (from giver to follower). Long enough silent pauses are pauses that exceed 300 ms, and the minimum amount of silence between giver and follower contributions (ICI) is 10 ms.

Giver channel:	[...] Speech	Long enough silent pause [...]
		ICI
Follower channel:	[...] Long enough silent pause	Speech [...]