The Dual of Denial: Two Uses of Disconfirmations in Dialogue and Their Prosodic Correlates

Emiel Krahmer*, Marc Swerts*,†, Mariët Theune*, Mieke Weegels*

* IPO, Center for User-System Interaction,
Eindhoven University of Technology,
P.O.Box 513,
NL-5600 MB, Eindhoven,
The Netherlands
{e.j.krahmer/m.g.j.swerts/m.theune}@tue.nl

† CNTS (Center for Dutch Language and Speech).
UIA, University of Antwerp
Universiteitsplein 1,
B-2610, Wilrijk,
Belgium

Unusual symbols:	F_0, χ^2
Number of pages:	32
Number of figures:	2
Number of tables:	9

Keywords: spoken dialogue systems, prosody, error detection, information grounding, perception

Abstract

In human-human communication, dialogue participants are continuously sending and receiving signals on the status of the information being exchanged. These signals may either be positive ('go on') or negative ('go back'), where it is usually found that the latter are comparatively marked to make sure that the dialogue partner is made aware of a communication problem. This article focuses on the users' signaling of information status in human-machine interactions, and in particular looks at the role prosody may play in this respect. Using a corpus of interactions with two Dutch spoken dialogue systems, prosodic correlates of users' disconfirmations were investigated. In this corpus, disconfirmations can have two uses: they may serve as a positive signal in one context and as a negative signal in another. With the data obtained from the corpus an acoustic and a perception experiment have been carried out. The acoustic analysis shows that the difference in signaling function is reflected in the distribution of the various types of disconfirmations as well as in different prosodic variables (pause, duration, intonation contour and pitch range). The perception experiment revealed that subjects are very good at classifying disconfirmations as positive or negative signals (without context), which strongly suggests that the acoustic features have communicative relevance. The implications of these results for human-machine communication are discussed.

“But does he deserve to be stabbed to death with an ice pick?” I ask.
“No,” says Douglas, and keeps on saying it, “no, no, no, no, no, no, no”[1]

1 Introduction

One of the central shortcomings of current spoken dialogue systems is that they are insufficiently able to spot communication problems (resulting, for instance, from poor recognition or from incorrect default assumptions) and hence have difficulty in responding to them. Things are different in human-human interaction, where communication problems are often easily spotted and solved. This is probably due to the fact that human dialogue participants continuously monitor the dialogue flow, sending negative (‘go back’) signals if there are communication problems, and positive (‘go on’) signals if the dialogue is on track and running smoothly. It seems a reasonable hypothesis that in human-machine interactions, the human dialogue participant similarly sends negative signals in response to problems, and positive signals otherwise. We conjecture that the ability of spoken dialogue systems to distinguish between such signals from the user is linearly correlated with the fluency of the interaction.

The current article investigates to what extent *prosodic* features are beneficial in distinguishing positive and negative cues. We expect that speakers use more prosodically marked features (higher pitch, longer duration, more pauses, marked intonation contours, . . .) in the case of a negative signal than in the case of a positive signal. To test this hypothesis, we concentrated on *one* type of utterance which may serve as a ‘go back’ signal in one context while it serves as a ‘go on’ signal in another context, namely a “no” answer to different types of system prompts. To illustrate this, consider the

[1] Michael Douglas in an interview with Entertainment Weekly on the movie Basic Instinct.

following two questions of a train time table information system, taken from the corpus of Weegels (1999).

(1) a. Do you want to go from Eindhoven to Swalmen?

b. Do you want me to repeat the connection?

Both (1.a) and (1.b) are yes/no questions and to both “no” is a perfectly natural answer. However, the two questions serve a rather different goal. Question (1.a) is an (explicit) attempt of the system to verify whether its current assumptions (about the departure and arrival station) are compatible with the intentions of the user. If this is not the case, the user will signal this using a disconfirmation, thereby indicating that at least one of the system's assumptions is incorrect. Question (1.b), on the other hand, is not an attempt of the system to verify its assumptions, and hence it cannot represent incorrect system assumptions. A subsequent “no” answer from the user thus serves as a ‘go on’ signal. The two uses of disconfirmations, being lexically similar but functionally different, constitute minimal pairs from a dialogue perspective, allowing us to check whether the various occurrences of this kind of utterance vary prosodically as a function of their context. In this way, they form ideal, naturally occurring speech materials for investigating the role of prosody in problem signalling.

In this article the hypothesis that ‘go back’ signals are prosodically marked with respect to ‘go on’ signals is tested both in an acoustic and a perceptual analysis. In the following, we will first make the ‘go on’ and ‘go back’ notions more explicit and present a brief overview of the context of this work (section 2), then describe the specific goals of the current article (section 3) and the speech corpus used (section 4). Sections 5 and 6 report on the acoustic and perceptual analyses respectively. We end with a general discussion (section 7).

2 Positive and negative cues in dialogue

From human-human communication it is known that dialogue participants are continuously sending and receiving signals on the status of the information being exchanged. This process is often referred to as *information grounding* (Clark & Schaeffer 1989, Traum 1994) and typically proceeds in two phases: a *presentation phase* in which the current speaker (the sender) sends a message to his conversation partner, and an *acceptance phase* in which the receiver signals whether the message came across unproblematically or not. In the acceptance phase the receiver may send a positive 'go on' signal or a negative 'go back' signal. We assume that a rule like the following underlies the distinction between 'go on' and 'go back' signals.[2]

> 'GO ON'/'GO BACK' RULE
>
> If the sender's utterance S_i is incompatible with the information state of the receiver, then the receiver's subsequent response U_i is a 'go back' signal, otherwise it is a 'go on' signal.

Here we are particularly interested in the case where the sender is a spoken dialogue system and the receiver is a human user. To avoid a possible confusion, it should be noted that both 'go on' and 'go back' signals are backward-looking dialogue acts in the terminology of Allen and Core (1997), see also Pulman (this volume), in the sense that they are responding acts.

The 'go on'/'go back' rule has some non-trivial consequences. First, the responder sends a 'go on' or 'go back' signal following *any* kind of system utterance. Second, the 'go back' signals are not only triggered by recognition errors, but may also cue other sources of communication problems

[2] Notice that this rule is related to the 2nd rule from Groenendijk et al. (1996) governing hearer-behavior in multi-speaker dialogues:

> **Rule H2** If a sentence is uttered which is incompatible with a participant's information state, then she does not update with it, but signals the incompatibility by uttering a sentence that contradicts the sentence uttered.

(e.g., erroneous presuppositions or incorrect default assumptions). Third, a receiver *immediately* signals any communication problems. This is a simplification in that occasionally receivers accept errors, or change their mind and 'correct' an unproblematic system utterance.[3]

Obviously, it is important for the system to be able to distinguish 'go on' from 'go back' signals, since this information is helpful for determining subsequent system actions. This raises the question whether there is a way to distinguish 'go back' from 'go on' signals. From studies of human-human conversation it is known that 'go back' signals are comparatively marked (see e.g., Swerts et al. 1998) to make sure that the dialogue partner is immediately made aware of a communication problem. By analogy, we may expect that if a system utterance makes it clear that something is wrong, users spend more effort on their signals as well, in order to prevent complete breakdown of the communication. Krahmer et al. (subm.) tried to find support for this claim in a corpus of human-machine dialogues. The characteristics listed in 1 were predicted for positive and negative cues, based on the idea that speakers want to finish the dialogue successfully as soon as possible and with minimal effort.

Table 1 approximately here

In all cases, the positive cues can be seen as the unmarked settings of linguistic features. For instance, the default word order in a sentence is unmarked (thus, no topicalization or extraposition). Similarly, it is a positive signal to present new information (which may speed up the dialogue), but not to repeat or correct information (which will definitely not lead to a more swift conclusion of the conversation).

One central observation of Krahmer et al. (subm.) is that users more often employ the 'go back' signals from table 1 when the preceding system utterance contains a problem, whereas the 'go on' signals are used in response to unproblematic system utterances. Many of these cues have a high

[3] In our corpus (see section 4), the former are analysed as 'go on' signals, the latter as 'go back' signals. Both are highly infrequent.

informativity. For instance, if the user's answer contains a marked word-order, then it is highly likely that the preceding system utterance contained a problem. The downside is that some of the highly informative cues occur rather infrequently. However, *combinations* of features can compensate for this and thus serve as good indicators of information status. Experiments using memory based learning techniques (with the IB1-GR algorithm, see Aha et al. 1991 and Daelemans et al. 2000) showed that it is possible to predict in 97% of the cases whether or not the preceding system utterance was problematic on the basis of the user's utterance, by looking at all features. This shows that taking combinations of cues into account provides a reliable indicator of problems. But one has to keep in mind that these experiments were performed with hand-annotated data and that there is a considerable gap between such data and the raw output of a speech recognition engine (a word graph). It remains an empirical question to what extent the positive and negative signals from table 1 can be recovered automatically. In any case, it is to be expected that shifting the analysis from hand-annotated data to word graphs will worsen the percentage of correctly predicted communication problems. This implies that there is definitely room for improvement. Therefore, one possible extension is to include another set of characteristics of user utterances in our prediction: a number of prosodic features.

3 Goal

The current paper looks at possible prosodic differences between positive and negative signals, using disconfirmations as analysis materials. For a variety of reasons, we expect that negative signals are prosodically marked compared to positive signals. A previous study of repetitive utterances in Japanese human-human dialogues (Swerts et al. 1998) showed that speakers more often provide negative signals with marked or prominent prosodic features than they do with positive signals. In particular, reactions to problematic utterances tend to be higher and slower than reactions to unproblematic utterances, and they are more likely to be preceded by a relatively long delay and to have high H% boundary tones. Consequently, we expect that in human-machine interactions the difference in signaling function will also be reflected in a difference in prosodic features (cf. Swerts & Ostendorf 1997). This expectation is confirmed by a bulk of recent work on hyperarticulate speech (e.g., Levow 1998, Oviatt et al. 1998a, 1998b, Soltau & Waibel 1998, Erickson et al. 1998), a speaking style which can be seen both as the result of speech recognition errors and as an important source of such errors. Typically, hyperarticulate speech has an increased pitch and longer duration. Such marked prosodic features also show up in human-human conversations with a relatively higher risk of communication failures, such as conversations in a stressed and/or noisy environment (Lombard 1911, Bou-Ghazele and Hansen 1998), over a relatively long distance (Traunmüller & Eriksson 2000) or in child-directed speech (Lindblom et al. 1992). This leads to the expectations in table 2 regarding prosodic features and the predicted settings for positive and negative signals.

Table 2 approximately here

This article discusses two experiments that have been carried out to find empirical evidence for these expectations. The first one consists of a set of acoustic analyses of prosodic features in disconfirmations. The second one is a perception experiment which aims at verifying whether human hearers can use some of the prosodic features to distinguish positive from negative cues, without having access to context information. First, the speech materials used in these analyses are further described.

4 Data

The stimuli for both the acoustic and the perceptual analyses were taken from the aforementioned corpus of 120 dialogues with two speaker-independent Dutch spoken dialogue systems which provide train time table information (see Weegels 1999). The systems prompt the user for unknown slots, such as departure station, arrival station, date, etc., in a series of questions. The two systems differ mainly in verification strategy (one primarily uses implicit verification, the other only uses explicit verification), length of system utterances and speech output (concatenated vs. synthetic speech). Twenty subjects were asked to query both systems via telephone on a number of train journeys. They were asked to perform three simple travel queries on each system (in total six tasks). Two similar sets of three queries were constructed, to prevent literal copying of subjects' utterances from the first to the second system. The order of presenting systems and sets was counterbalanced.

The stimuli used in the two analyses consisted of negative answers to yes/no questions from both systems. If the preceding yes/no question was a verification of the system's assumptions (e.g., (1.a) above), then the user's disconfirmation indicates that the yes/no question contained a problem (due to a speech recognition error or an incorrect assumption on the system's part). If the yes/no question was not a verification (such as example (1.b), but also questions like *Do you want other information?* or *Do you want information about another connection?*), then the user's disconfirmation just serves as an answer to that question and does not indicate problems.

Regarding their structure, the users' disconfirmations were divided into three categories: (1) responses consisting of an explicit disconfirmation marker "no" ("nee") only (we shall refer to these cases as 'single no'), (2) responses consisting of an explicit disconfirmation marker followed by other words ('no+stuff' in the terminology of Hockey et al. 1997), (3) responses containing no explicit disconfirmation marker ('stuff').

5 Acoustic analysis

5.1 Method

For the acoustic analysis a random selection of 109 negative answers to yes/no questions from both systems was used, taken from the corpus described above. The proportion of 'go back' signals is .62. The utterances were produced by 5 female and 2 male speakers. The proportion of 'go back' signals is .52 for male and .63 for female speakers, both similar to the overall proportion. The speech data were digitized with a 16 kHz sampling frequency. Fundamental frequency (F_0) was determined using a method of subharmonic summation (Hermes, 1988). Durations of speech segments and of pauses were measured directly in the digitized waveform. The users' responses to the yes/no questions were analysed in terms of the following features: (1) type of boundary tone in "no" (high or not high); (2) duration (in ms) of "no"; (3) duration (in ms) of pause after "no" before stuff; (4) duration (in ms) of pause between system's prompt and user response; (5) F_0 max (in Hz) at energy peak of major pitch accent in stuff; (6) number of words in stuff. It was our original intention to also investigate pitch range in the "no" part of the different responses, but this turned out to be too difficult given that many of the cases were realized with a low-anchored pitch accent followed by a high boundary tone (L*H-H%). For these utterances, it was not possible to adequately measure pitch range, given that the F_0 maximum in the energy peak of the pitch accent basically undershoots the perceived pitch range, whereas the real F_0 maximum at the end of the high boundary tone overshoots it. See the discussion of figure 2 below.

5.2 Results

Table 3 approximately here

Table 3 gives the distribution of different types of disconfirmations following either an unproblematic system utterance or one which contains one or more problems. A χ^2 test reveals that these numbers significantly differ from chance level ($\chi^2 = 22.146$, df = 2, $p < 0.001$). First, this table shows that the minimal response, a single no, is in the majority of the cases used as a positive signal. Second, single stuff responses are exclusively reserved for responses following a system utterance with one or more problems. The majority of the responses to yes/no questions in our data, however, is of the no+stuff type, which may serve either as a positive or as a negative cue. The lexical material in the stuff is quite different for the two signals: for the positive cases, the subsequent words are mostly some polite phrases ("thank you", "that's right"); for the negative cases, the stuff usually is an attempt to correct the information which is misrecognized or which is wrongly assumed by the system.

Table 4 approximately here

Table 4 displays the presence or absence of high boundary tones on the word "no" (for the single no and no+stuff cases) for positive and negative signals. A χ^2 test reveals that this distribution is again well above chance level (χ^2 = 33.004, df = 1, $p < 0.001$). In responses following a problematic system question, "no" is generally provided with a question-like H% boundary tone, which is absent when "no" follows an unproblematic system question. These results are in agreement with observations in Japanese human-human conversations (Swerts et al. 1998).

Table 5 approximately here

One might hypothesize that in the 'go on' case the "no" and the stuff (which consists primarily of politeness phrases) are prosodically integrated, thus explaining the absence of high boundary tones in this case. However, this is not true. For both positive and negative cues the "no" forms a separate intonational phrase in the majority of the cases (see table 5).

Table 6 approximately here

The results for the continuous prosodic features of interest are given in table 6. Taking the utterances of all subjects together, a t-test reveals a significant difference for each of these features. Intra-individual differences could not be tested because the numbers of unproblematic and problematic utterances are insufficient and/or too unequally distributed. However, when looking at the mean within-subject differences, the findings mostly point in the expected direction, thus warranting an overall t-test. For all speakers, the mean duration of “no” and of pauses, F_0 max in stuff, and the number of words in stuff are usually higher in problematic than in unproblematic cases.

Figure 1 approximately here

Figure 1 visualizes the results given in table 6, illustrating that the trend is the same in all cases: negative signals are comparatively marked. First, negative signals differ from positive ones, in that the word “no” —when it occurs— in these utterances is longer. Second, compared to positive signals, there is a longer delay after a problematic system prompt before users respond. Both results are in line with the data for Japanese (Swerts et al 1998). Third, in the no+stuff utterances, the interval between “no” and the remainder of the utterance is longer following a problematic system utterance than following an unproblematic one. Fourth, after a problematic yes/no question, the stuff part of the answer usually contains a high-pitched narrow focus accent to mark corrected information, whereas in the unproblematic case the stuff is usually prosodically unmarked. Finally, in reaction to a problem, the stuff part tends to be longer in number of words, which is in agreement with our previous, more general finding (Krahmer et al. subm.).

5.3 Discussion

The acoustic results given above clearly indicate that there is a marked prosodic difference between positive and negative signals. To illustrate some of these effects more clearly, consider figure 2 which visualizes the waveforms and corresponding F_0 contours of two typical disconfirmations produced by one of our speakers, one being a ‘go on’ signal (top), the other a ‘go back’ signal (bottom).

Figure 2 approximately here

Both utterances consist of a disconfirmation marker (“no”) followed by stuff, but it is clear that they are realized with quite different prosody. In line with our hypothesis, the word “no” in the ‘go on’ case is comparatively short (185 ms), it is not provided with a prominent high boundary tone, and it is immediately followed by the stuff without a clear silence interval. In addition, the stuff part of this response does not contain a prominent pitch accent. On the other hand, the utterance at the bottom of the figure is a ‘go back’ signal and accordingly contains a relatively long “no” (441 ms), which is produced with a clear high boundary tone, and is followed by a fairly long pause of 762 ms. Note that the contour on the word “no” is of the type referred to above, L*H-H%, which does not permit a straightforward specification of pitch range. Also, the stuff contains a clear narrow focus pitch accent which serves to highlight corrected information. What cannot be derived from this figure is that in the ‘go back’ mode speakers generally tend to produce their responses after a longer delay than in ‘go on’ mode, and also that the stuff part is generally longer in words in the former case.

6 Perceptual analysis

6.1 Method

In a second experiment we investigated whether the acoustic findings have perceptual relevance. For this experiment we used 40 “no”s, all taken from no+stuff disconfirmations. We opted for no+stuff disconfirmations since these are the most frequent and are equally likely to occur after either a problematic or an unproblematic utterance from a distributional perspective (see table 3), and are thus least biased in terms of their function as positive or negative cues. The 40 “no”s were taken from the utterances of 4 speakers. The speakers were selected on the basis of the fact that they produced no+stuff in both conditions (positive and negative). For the perception study, we only used the “no”-part of these utterances, given that the stuff-part would be too informative about their function as positive or negative cues (see the two no+stuff answers analysed in section 5.3). Of the 40 “no”s, 20 functioned as a positive and 20 as a negative signal. Unfortunately the corpus did not allow us to get equal numbers of positive and negative signals for all speakers. Subjects of the perception experiment were 25 native speakers of Dutch. They were presented with 40 stimuli, each time in a different random order to compensate for any potential learning effects. They heard each stimulus only once. The experiment was self-paced and no feedback was given on previous choices. In an individual, forced choice task, the subjects were instructed to judge for each "no" they heard whether the speaker signaled a problem or not. They were not given any hints as to what cues they should focus on. The subjects were first presented with four "exercise" stimuli to make them aware of the experimental platform and the type of stimuli. It is worth stressing that the choice to use only "no"s extracted from no+stuff answers implies that not all the acoustic features studied in the previous section survive in the current perceptual analysis. In particular, we lose the features delay (time between end of prompt and start of user's answer), pause (time between end of "no" and beginning of stuff) as well as any possible cue in the stuff part (e.g., number of words, narrow-focused pitch accents).

6.2 Results

Tables 7, 8 and 9 approximately here

The results are presented in tables 7 and 8, and summarized in table 9. A χ^2 test was used to determine whether a distribution is above chance level. Table 7 focuses on the perception of positive signals. It turned out that 17 out of the 20 positive signals were correctly classified as cases in which the speaker did not signal a problem. The remaining three cases were in the expected direction, though not significant. Table 8 zooms in on negative signals. Here 15 out of 20 negative signals were classified correctly as instances of "no" signaling problems. Interestingly one negative signal was significantly misclassified as a positive signal. A post-hoc acoustic analysis of this "no" revealed that it shared its primary characteristics with positive signals. In particular: the "no" was relatively short, and lacked a high boundary tone.

6.3 Discussion

It seems a reasonable hypothesis that when speakers systematically dress up their utterances with certain features, hearers will be able to attach communicative relevance to the presence or absence of these features. To test if this is indeed the case for the acoustic properties of utterances of “no” found in section 5, the perception experiment was carried out. Of course, from a system perspective it is not really important whether or not people are able to use acoustic features as cues, as long as the acoustic features are easily measurable and consistent. However, we do believe that a perception test provides additional evidence for the relevance of prosodic features for signalling communication problems.

The perceptual study clearly shows that subjects are good at correctly classifying instances of “no”, extracted from no+stuff utterances, as positive or negative signals. There was only one instance of a “no” which was consistently misclassified: this concerned a “no” which followed a problematic system utterance but was perceived by most subjects as a positive signal. Interestingly, this “no” shared its primary characteristics (relatively short and no high boundary) with the positive signals.

It is no surprise that no “no” was correctly classified as a positive or a negative signal by *all* subjects. After all, only some of the acoustic features found in the acoustic analysis of section 5 were part of the stimuli presented to the subjects. In particular, subjects could not use for their classification: (i) the delay between the end of the preceding system question and the start of the user’s disconfirmative answer, (ii) the pause between the “no” and the stuff nor (iii) any features present in the stuff (such as length and presence or absence of narrow focused pitch accents, besides, of course, the lexical content). Looking specifically at tables 7 and 8 suggests that classifying ‘go on’ signals as ‘go on’ signals is somewhat easier than classifying ‘go back’ signals as ‘go back’ signals. This might be due to the fact that the word “no” has the most communicative import in the 'go on' case, while for the 'go back' case it is rather the stuff part (which usually contains a correction) which is most informative (see below). In other words, it might be the case that for the 'go back' signals the prosodic features to be found in the stuff part of the disconfirmation, which was not presented to the listeners, are relatively more important. Yet, even given a subset of the potentially relevant acoustic features, subjects perform very well for both positive and negative signals.

7 General discussion

The main finding of this article can be summarized as follows: in the case of communication problems, speakers more often employ prosodically marked features in their reaction. If the preceding system utterance contained a problem (either a speech recognition error or an incorrect default assumption), then (1) the user's utterance of the word "no" has a longer duration, (2) there is a longer pause between the system's utterance and the user's reaction, (3) in the case of a no+stuff answer, the delay between the "no" and the stuff is longer, (4) the stuff part contains a narrow focus, high-pitched (corrective) accent and (5) the stuff contains more words. Various distributional differences between 'go on' and 'go back' signals were found: for instance, single stuff answers are solely reserved as responses to problematic system utterances and, in addition, users who respond to problematic utterances primarily use H% boundary tones. The perception study revealed that subjects are very good at correctly classifying instances of "no" (taken from no+stuff utterances) as positive or negative signals, without having access to the utterance context.

These findings can easily be related to the respective functions of the two uses of disconfirmation. A 'go on' disconfirmation is simply an answer to the question and does not address any underlying assumptions of the system. In principle, a single "no" is a sufficient answer. The stuff is exclusively reserved for politeness phrases, which follow more or less automatically, tend to be short and provide no further information. This explains the short pauses between the "no" and the stuff as well as the lack of accents in the stuff. If a yes/no question from the system contains a problem, just answering "no" might be sufficient but is not very cooperative. Assuming that the user wants the dialogue to be over as soon as possible it is more efficient to immediately *correct* the system. To do that, single stuff adequately serves the purpose, whilst an explicit "no" may be added to strengthen the problem signaling. Since the stuff is not meant to be polite, but really aims at furthering the dialogue in an efficient way by correcting information, it typically contains more words. In the case of communication problems, it may be assumed that cognitive load is relatively high, since the user has to reconstruct where the system's assumptions do not match her own intentions, and has to formulate an adequate reaction for that particular context. As Levelt (1989) argues, there is a close correspondence between cognitive load and length of pauses, which might explain that both the delay and the pause between the "no" and the stuff are longer in the case of problems.

The findings related to prosodic markedness are in line with our earlier findings, in which it was shown that subjects use the negative ('go back') variants of the features described in table 1 more often when the preceding system utterance contains a problem, whereas the positive cues ('go on') are more often used in response to unproblematic system utterances. Taking these two results in combination provides evidence for the claim that people devote more effort to negative cues on various levels of communication.

An interesting question is how generalizable the results are. We contend that our findings are not specific for "no" nor for Dutch nor for the domain of train travelling. Support for this claim is found, for instance, in the work by Swerts et al. (subm.). One of the findings from their study of American-English human-machine dialogues is that utterances following speech recognition errors can be reliably distinguished from 'normal' utterances using a set of automatically obtained acoustic/prosodic characteristics (pitch range, amplitude, timing, *inter alia*). For instance, 'corrections' appear to be more prosodically marked than other utterances (higher, longer, louder, slower, ...), which is in agreement with our current results. For some of the features discussed in this article there is a clear correspondence between the kind of system question and the precise marked setting of that feature. For instance, it was found that the stuff following problematic utterances is generally longer, which is in line with the earlier finding of Krahmer et al. (subm.). However, there it was also found that reactions to an implicit verification question (i.e., an open question) are on average twice as long as reactions to an explicit verification (i.e., yes/no) question. Yet following both system questions the answers were longer in the case of communication problems. This implies that long problem signalling answers to implicit verification questions will generally be longer than long problem signalling answers to explicit verifications. We also found that high boundary tones are more likely to arise following problems. However, it should be noted that the presence of a high H% boundary tone *by itself* is not necessarily a signal of problems, for high boundary tones have been claimed to have multiple functions (see Pierrehumbert & Hirschberg 1990).

The current analysis suggests that the presence of cues such as a prolonged delay before answering or a high-pitched narrow focus accent are good indicators of problems. In combination with the findings of Krahmer et al. (subm.), the present results provide potentially useful information for spoken dialogue systems which monitor whether or not the communication is in trouble: if a question is followed by a user's utterance which has various marked properties (such as relatively many words, disconfirmations, corrections, long delays, words with a narrow focus, high-pitched accent), the system can be fairly certain that there are communication problems. If, on the other hand, the user's utterance does not contain such features, then it is highly likely that the dialogue is running smoothly. Using a systematic and reliable strategy to decide whether or not there are communication problems may be very useful in a number of situations. It can be used as a basis for choosing the verification strategy employed by the system, but it may also be a cue to switch to a different recognition engine. Levow (1998) found that the probability of experiencing a recognition error after a correct recognition is .16, but immediately after an incorrect recognition it is .44. This increase is probably caused by the fact that the speakers used hyperarticulate speech when they noticed that the system had a problem recognizing their previous utterance. Similar findings are reported in a number of studies, such as Shriberg et al. (1992) and Litman et al. (2000). This implies that it might be beneficial to switch to a speech recognizer trained on hyperarticulate speech if there are communication problems (cf. Hirschberg et al. 1999).

Acknowledgments

Thanks are due to Antal van den Bosch, Olga van Herwijnen, Stephen Isard, Esther Klabbers, Elizabeth Shriberg, Jacques Terken, and an anonymous referee, as well as to the audience of the ESCA Workshop on Dialogue and Prosody (Veldhoven, 1999). The authors are mentioned in alphabetical order. Weegels and Theune were supported by the Priority Programme Language and Speech Technology (TST), sponsored by NWO (The Netherlands Organization for Scientific Research).

References

Aha, D., Kibler, D., Albert, M., 1991. Instance-based learning algorithms. Machine Learning 6, 37-66.

Allen, J., Core, M. 1997. Damsl: Dialogue markup in several layers. Draft contribution for the Discourse Resource Initiative.

Bou-Ghazele, S., Hansen, J., 1998. HMM-based stressed speech modeling with applications to improved synthesis and recognition of isolated speech under stress. IEEE transactions on speech and audio processing 6(3):201-216.

Clark, H.H., Schaeffer, E.F., 1989. Contributing to discourse. Cognitive Science 13:259-294.

Daelemans, W., Zavrel, J., van der Sloot, K., van den Bosch, A., 2000. TiMBL: Tilburg Memory Based Learner, version 3.0, reference guide. ILK Technical Report 99-01,
`http://ilk.kub.nl/∼ilk/papers/ilk0001.ps.gz`.

Erickson, D., Fujimura, O., Pardo, B., 1998. Articulatory correlates of prosodic control: emotion and emphasis. Language and Speech 41(3-4):399-418.

Groenendijk, J., Stokhof, M., Veltman, F., 1996. Corefence and modality in the context of multi-speaker discourse. In: Context dependence in the analysis of linguistic meaning. H. Kamp and B. Partee (eds.), Stuttgart, IMS, 195-216.

Hermes D.J. 1988. Measurement of pitch by subharmonic summation. Journal of the Acoustical Society of America 83, 257-264.

Hirschberg, J., Litman, D., Swerts, M., 1999. Prosodic cues to recognition errors. In: Proceedings of the International Workshop on Speech Recognition and Understanding (ASRU-99), Keystone, CO, USA.

Hockey, B., Rossen-Knill, D., Spejewski, B., Stone, M., Isard, S., 1997. Can you predict answers to y/n questions? Yes, no and stuff. In: Proceedings Eurospeech'97, Rhodos, Greece, pp. 2267-2270.

Krahmer, E., Swerts, M., Theune, M., Weegels, M., submitted. Error-detection in spoken human-machine interaction.

Levelt, W.J.M., 1989. Speaking. From Intention to Articulation. MIT Press, Cambridge, Massachusetts.

Levow, G.-A., 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In: Proceedings COLING-ACL, Montreal, Canada, pp. 736-742.

Lombard, E., 1911. Le signe de l'elevation de la voix. Ann. Maladies Oreille Larynx, Nez. Pharynx 37, 101-119.

Lindblom, B., Brownlee, S., Davis, B., Moon, S., 1992. Speech transforms. Speech Communication 11 (4-5), 357-368.

Litman, D., Hirschberg, J., Swerts, M., 2000. Predicting automatic speech recognition performance using prosodic cues. In: Proceedings of the First Meeting of the North-American Chapter for Computional Linguistics (NAACL'00), Seattle, Washington.

Oviatt, S., Bernard, J., Levow, G.-A., 1998a. Linguistic adaptations during spoken and multimodal error resolution. Language and Speech 41(3-4): 419-442.

Oviatt, S., MacEachern, M., Levow, G.-A., 1998b. Predicting hyperarticulate speech during human-computer error resolution. Speech Communication 24, 87-110.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P., Morgan, J., Pollack, M. (Eds.), Intentions in Communication, Cambridge MA: MIT Press, pp. 342-365.

Pulman, S., 2000. Relating dialogue games to information state. Speech Communication. *This volume.*

Shriberg, E., Wade, E., Price, P., 1992. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In; Proceedings of the DARPA Speech and Natural Language Workshop. San Mateo CA: Morgan Kaufmann Publishers, pp. 49-54.

Soltau, H., Waibel, A., 1998. On the influence of hyperarticulated speech on recognition performance. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia.

Swerts M., Koiso, H., Shimojima, A., Katagiri, Y., 1998. On different functions of repetitive utterances. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia.

Swerts, M., Ostendorf, M., 1997. Prosodic and lexical indications of discourse structure in human-machine interactions. Speech Communication 22, 25-41.

Swerts, M., Hirschberg, J., Litman, D., submitted. Corrections in spoken dialogue systems.

Traum, D.R. 1994. A computational theory of grounding in natural language conversation. Ph.D thesis, Rochester.

Traunmüller, H., Eriksson, A., 2000. Acoustic effects of variation in vocal effort by men, women, and children. Journal of the Acoustic Society of America, to appear.

Weegels, M., 1999. Users' (mis)conceptions of a voice-operated train travel information service. IPO Annual Progress Report, Eindhoven, The Netherlands, pp. 45-52.

List of figure captions

Figure 1: No+stuff responses of one speaker to two different yes/no questions from the system. **Above**: following the system question "Wilt u dat ik de verbinding nog eens herhaal?" (*Do you want me to repeat the connection?*), the speaker responds with a positive ('go on') utterance: "Nee dankuwel"(*No thank-you*). **Below**: following the system verification question "U wilt dus vamiddag reizen? (*So you want to travel this afternoon?*) , the speaker responds with a negative ('go back') utterance "Nee vanavond" (*No tonight*).

Figure 2: Average values for different features (cf. Table 6).

Table 1: Positive vs. negative cues

POSITIVE ('go on')	NEGATIVE ('go back')
short turns	long turns
unmarked word order	marked word order
answer	no answer
no corrections	corrections
no repetitions	repetitions
new info	no new info

Table 2: List of prosodic features and their expected settings for positive and negative cues

Features	POSITIVE ('go on')	NEGATIVE ('go back')
Boundary tone	low	high
Duration	short	long
Pause	short	long
Delay	short	long
Pitch range	low	high

Table 3: Numbers of negative answers following an unproblematic system utterance (¬ PROBLEMS) and following those containing one or more problems (PROBLEMS)

Type	¬ PROBLEMS	PROBLEMS	TOTAL
no	18	11	29
stuff	0	24	24
no+stuff	23	33	56
TOTAL	41	68	109

Table 4: Presence or absence of high boundary tones following occurrences of "no" (single no and no+stuff) for positive and negative cues.

High boundary tone	¬ PROBLEMS	PROBLEMS	TOTAL
Absent	32	7	39
Present	9	37	46
TOTAL	41	44	85

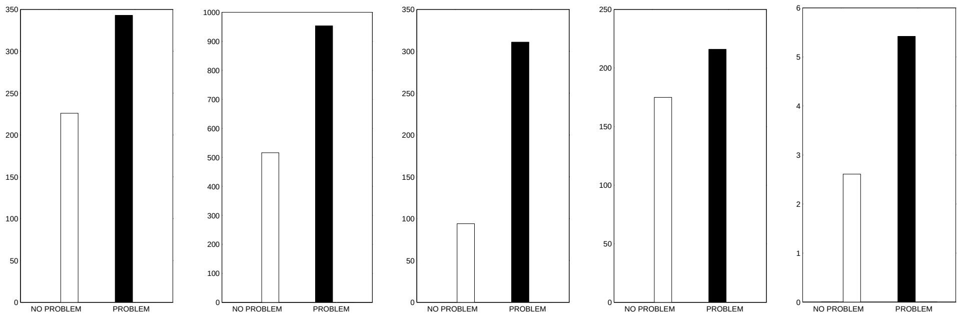
Table 5: Number of no+stuff utterances in which "no" is realized as a separate intonational phrase, for both positive and negative cues.

	¬ PROBLEMS	PROBLEMS	TOTAL
"no" separate phrase	17	30	47
"no" no separate phrase	6	3	9
TOTAL	23	33	56

Table 6: Average values for various features. Duration of "no" (for all occurences of "no": single no and no+stuff), delay between end of system utterance and beginning of user's disconfirmation (all cases), pause between "no" and stuff (for no+stuff cases), F_0 max in stuff and number of words in stuff (both for no+stuff and stuff). Standard deviations are given between brackets.

Feature	¬ PROBLEMS	PROBLEMS
Duration of "no" (ms)**	226 (83)	343 (81)
Preceding delay (ms)**	516 (497)	953 (678)
Following pause (ms)*	94 (93)	311 (426)
F_0 max in stuff (Hz)*	175 (37)	216 (46)
Words in stuff**	2.61 (3.65)	5.42 (8.14)

$^{**}p < 0.001$, $^{*}p < 0.05$



Duration (in ms.) | Delay (in ms.) | Pause (in ms.) | F_0 max (in Hz.) | # Words (in integers)

(Figure 1)

Table 7: Number of *positive* signals which are perceived as positive signals (¬ PROBLEMS) or as negative ones (PROBLEMS).

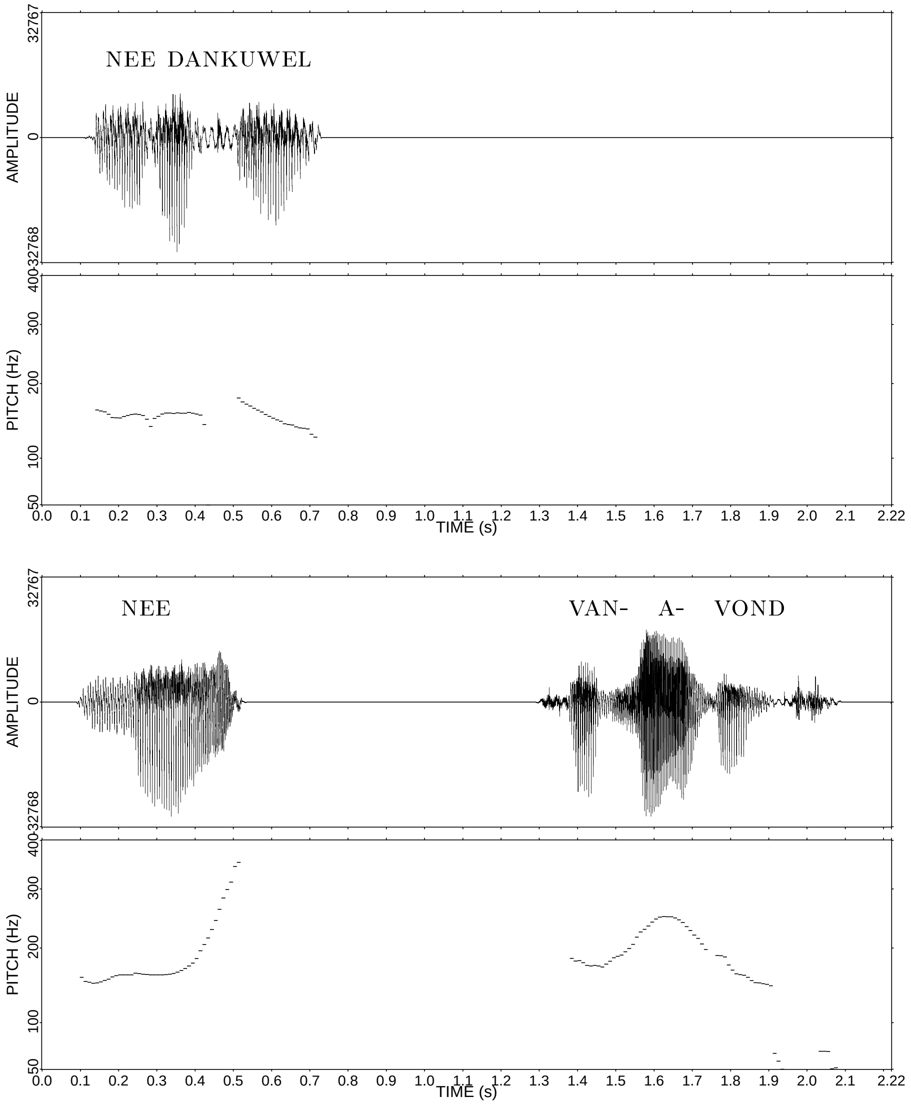
		Perceived as		
Sp.	Utt.	¬ PROBLEMS	PROBLEMS	Sign.
A	1	20	5	$p < 0.01$
	2	22	3	$p < 0.01$
	3	22	3	$p < 0.01$
	4	22	3	$p < 0.01$
	5	23	2	$p < 0.01$
	6	19	6	$p < 0.01$
	7	20	5	$p < 0.01$
	8	20	5	$p < 0.01$
	9	21	4	$p < 0.01$
	10	20	5	$p < 0.01$
	11	19	6	$p < 0.01$
B	1	20	5	$p < 0.01$
	2	20	5	$p < 0.01$
	3	14	11	n.s.
	4	21	4	$p < 0.01$
	5	20	5	$p < 0.01$
C	1	13	12	n.s.
	2	20	5	$p < 0.01$
	3	20	5	$p < 0.01$
D	1	17	8	n.s.

Table 8: Number of *negative* signals which are perceived as positive signals (¬ PROBLEMS) or as negative ones (PROBLEMS).

		Perceived as		
Sp.	Utt.	¬ PROBLEMS	PROBLEMS	Sign.
A	1	6	19	$p < 0.01$
	2	22	3	$p < 0.01$
	3	15	10	n.s.
	4	7	18	$p < 0.05$
	5	2	23	$p < 0.01$
B	1	4	21	$p < 0.01$
	2	3	22	$p < 0.01$
	3	12	13	n.s.
	4	3	22	$p < 0.01$
	5	5	20	$p < 0.01$
	6	11	14	n.s.
C	1	5	20	$p < 0.01$
	2	6	19	$p < 0.01$
	3	6	19	$p < 0.01$
	4	10	15	n.s.
	5	2	23	$p < 0.01$
	6	3	22	$p < 0.01$
	7	7	18	$p < 0.05$
D	1	4	21	$p < 0.01$
	2	1	24	$p < 0.01$

Table 9: Summary of tables 7 and 8 of the perceived classification of positive and negative signals.

	Perceived as ¬PROBLEMS	No significant difference	Perceived as PROBLEMS	Total
¬PROBLEMS	17	3	0	20
PROBLEMS	1	4	15	20
Total	18	7	15	40


NEE DANKUWEL
32767
AMPLITUDE
0
-32768
400
300
200
100
50
PITCH (Hz)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.22
TIME (s)
NEE
VAN- A- VOND
32767
AMPLITUDE
0
-32768
400
300
200
100
50
PITCH (Hz)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.22
TIME (s)

(Figure 2)