On the use of prosody for on-line evaluation of spoken dialogue systems

Swerts, M.G.J.; Krahmer, E.J.

Published in:
Proceedings of the second International Conference on Language Resources and Evaluation (LREC), Athens, Greece, 31 May - 2 June, 2000

Publication date:
2000

On the Use of Prosody for On-line Evaluation of Spoken Dialogue Systems

Marc Swerts, Emiel Krahmer

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

Abstract

This paper 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 may serve as a signal to 'go on' in one context and as a signal to 'go back' in another. With the data obtained from this 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 interactions are discussed.

1. Introduction

Lewis & Norman (1986) point out that, as soon as human users are involved, it is not possible to design user interfaces that fully eliminate error. Users may make mistakes or provide the system with input that it cannot interpret. Since such failures are unavoidable, it is important that user interfaces are designed in such a way that they include methods to discover errors and to recover from them. This implies that a run time prediction of error is crucial to allow the system to immediately adapt its interaction strategy (Johnson 1999).

Given the state of the art of current speech technology, spoken dialogue systems are especially prone to error, for instance because of user utterances that are misrecognized (Oviatt et al. 1998) or default assumptions which the system makes but which turn out the be incorrect. (For a more exhaustive overview of potential sources of errors, see Dybkjær et al. 1998.) It would be most beneficial if such communication problems could be detected on-line, because then the system has the option of changing its strategy (for instance, it could switch from implicit verification to explicit verification). Unfortunately, there is no known method for accurate on-line signalling of communication problems. It has been argued that ASR confidence scores may serve as the basis of such a method. However, this meets with two problems: (1) there is not a simple one-to-one relation between low confidence scores and (recognition) errors, nor between high confidence scores and correct recognitions (see e.g. Bouwman et al. 1999) and (2) confidence scores are only potentially relevant for speech recognition errors, but do not apply to other sources of miscommunication.

At this point it is expedient to look at how humans cope with apparent communication problems. 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 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 former case (there is no problem), the receiver transmits a positive signal ('go on'), in the latter case (there is a problem), he or she sends a negative signal ('go back'). Various studies of human-human communication (e.g., Swerts et al. 1998) revealed that the negative signals are comparatively marked, as if the speaker wants to devote additional effort to make the other aware of the apparent communication problem. A plausible explanation for this is that missing a negative cue may cause breakdown of the communication.

The hypothesis underlying our work is that in spoken human-machine communication, humans employ essentially the same kinds of positive and negative cues for information grounding as they do in ordinary human-human communication.[1] The problem, however, is that most, if not all, spoken dialogue systems do not systematically pay attention to these cues. We conjecture that the ability of spoken dialogue systems to distinguish between positive and negative cues from the user is linearly correlated with the fluency of the interaction, since these cues provide important information about the status of the information currently under negotiation. We have studied a corpus of human-machine dialogues (Weegels 1999), obtained with two Dutch train time table information systems, in order to find out which cues people actually use in human-machine communication. In Krahmer et al. (1999a) a number of positive and negative cues have been singled out and their (joined) information potential for spotting communication problems was studied. It was indeed found that human speakers who converse with a spoken dialogue system put more effort in 'go back' signals than they do in 'go on' signals.

[1] This in line with the hypothesis put forward in Reeves & Nass (1996) that humans treat computers (and media in general) as social actors. More specifically, Reeves & Nass suggest that users who communicate with a machine in natural language will use their communicative abilities as if they are communicating with another human.

The current paper focuses on the prosodic features of positive and negative cues. We expect that speakers use more prosodic effort (higher pitch, longer duration, more pauses, marked intonation contours, …) in the case of a 'go back' signal than in the case of a 'go on' 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 following two questions 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 some pieces of information that it has recently gathered (the departure and arrival station). If the user would respond to this question with a "no" this would definitely be a 'go back' signal: the user indicates that at least one of the system's beliefs is incorrect. Question (1.b), on the other hand, is not an attempt of the system to verify its beliefs, and hence it cannot represent incorrect system beliefs. A subsequent "no" answer from the user thus serves as a 'go on' signal. The two types of "no" answers, being lexically similar but functionally different, constitute minimal pairs from a dialogue perspective, allowing us to check whether the various occurrences of this 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 information grounding.

The current paper focuses on the hypothesis that 'go back' signals are prosodically marked compared to 'go on' signals, which will be tested both in an acoustic and a perceptual analysis. In the following, we will first present a brief overview of the context of this work (section 2), then describe the speech corpus used (section 3). Section 4 reports on the acoustic analysis that was performed, while section 5 gives an in-depth description of the perceptual analyses. We end with a general discussion (section 6). The results from section 4 are also described in Krahmer et al. 1999b. The results of the perceptual analysis are presented here for the first time.

2. Effort in Dialogue

Since a spoken dialogue system can never be certain that it understood the user correctly, it is in constant need of verification. If a verification question of the system makes it clear that something is wrong (e.g., because a speech recognition error occurred), users are expected to spend more effort on their signals in order to prevent complete breakdown of the communication. Krahmer et al. (1999a) tried to find support for this claim in a study of responses of Dutch speakers in their interactions with a train time table information system. The following distinction between positive and negative cues was expected (see table 1), based on the idea that speakers want to finish the dialogue successfully as soon as possible and with minimal effort (Zipf 1949). For more details, see Krahmer et al. (1999a). 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).

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

Table 1: Positive vs. negative cues

The central hypothesis of Krahmer et al. (1999a) is that users more often employ the 'go back' signals when the preceding system utterance contains a problem, whereas the 'go on' signals are used in response to unproblematic system utterances. For nearly all of the cues of table 1 this was indeed found. Many of these cues have a high 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. 1999) applied to the annotated data from Weegels (1999) 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. On the one hand, these results are certainly encouraging. They show that taking combinations of cues into account provides a reliable indicator of problems. On the other hand, one has to keep in mind that these experiments were performed with hand-annotated data and that there is a certain 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 from a word graph. 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 to our previous work is to include another set of characteristics of user utterances in our prediction: a number of prosodic features.

To this end, the current paper looks at possible prosodic differences between positive and negative signals, us-

Features	POSITIVE	NEGATIVE
Boundary tone	low	high
Pitch range	low	high
Duration	short	long
Pause	short	long
Delay	short	long

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

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

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

ing different types of disconfirmations as analysis materials. 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. Consequently, we expect that in human-machine interactions the difference in signaling function will also be reflected in a difference in prosodic effort (cf. Swerts & Ostendorf 1997). This expectation is also based on recent work on hyperarticulate speech (e.g., Levow 1998, Oviatt et al. 1998, Soltau & Waibel 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. All this leads to the expectations in table 2. regarding prosodic features and the predicted settings for positive and negative signals. We discuss 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.

3. Data

The stimuli for both the acoustic and the perceptual analyses were taken from a 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', Hockey et al. 1997), (3) responses containing no explicit disconfirmation marker ('stuff').

4. Acoustic analysis

4.1. Method

For the acoustic analysis a random selection of 109 negative answers (by 7 speakers) to yes/no questions from both systems was used, taken from the corpus described above. 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 in 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 1 below.

4.2. Results

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 this distribution is highly significant (χ^2 =

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

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

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)

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

Table 5: Average values for different features of all occurrences of "no" (single no and no+stuff). Standard deviations are given between brackets.

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 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). The results for the continuous prosodic features of interest are given in table 5. 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. 5, Table 5 illustrates 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. 1999).

4.3. 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 1 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). Both

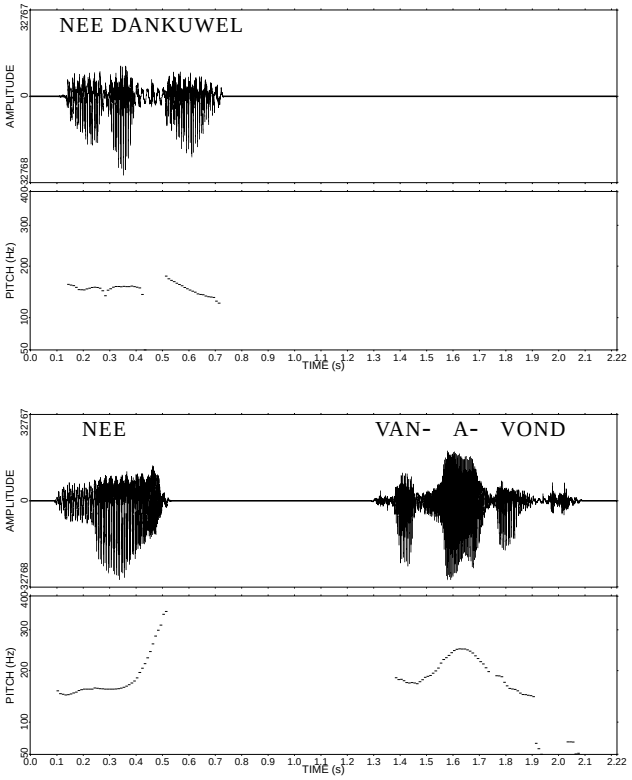

Figure 1: "No plus stuff" responses of one speaker to two different yes/no questions from the system: top is a POSITIVE utterance ("nee dankuwel"(*no thanks*)) and bottom is a NEGATIVE utterance ("nee vanavond" (*no tonight*)).

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 on 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.

5. Perceptual analysis

5.1. 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 4.3). 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 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. Each subject was 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).

		Perceived as		
Sp.	Utt.	¬ PROBLEM	PROBLEM	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 6: Number of *positive* signals which are perceived as positive signals (¬ PROBLEM) or as negative ones (PROBLEMS).

5.2. Results

The results are presented in tables 6 and 7, and summarized in table 8. A χ^2 test was used to determine whether a distribution is above chance level. Table 6 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. Table 7 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 consistently misclassified as a positive signal. A post-hoc acoustic analysis of this "no" revealed that it shared the primary characteristics of positive signals, in particular: the "no" was relatively short, and lacked a high boundary tone.

5.3. 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 4, 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 to signal positive and negative cues.

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.

Sp.	Utt.	*Perceived as* ¬ PROBLEM	PROBLEM	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 7: Number of *negative* signals which are perceived as positive signals (¬ PROBLEM) or as negative ones (PROBLEMS).

	Perceived ¬PROBL.	no pref.	PROBL.	Total
¬PROBL.	17	3	0	20
PROBL.	1	4	15	20
Total	18	7	15	40

Table 8: Summary of the perceived classification of positive and 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 important to keep in mind that only some of the acoustic features found in the acoustic analysis of section 4 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). Thus, even given a subset of the potentially relevant acoustic features, subjects perform very well.

6. General discussion

6.1. Summary and discussion of the results

The main finding of this article can be summarized as follows: in the case of communication problems, speakers put much more prosodic effort 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 kinds 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 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.

The findings related to prosodic effort are in line with the findings of Krahmer et al. (1999a), 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, we have found evidence for the claim that people devote more effort to negative cues on various levels of communication.

An interesting question is how generalizable the prosodic 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 is found, for instance, in the recent collaboration of the first author with Hirschberg and Litman. 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.

6.2. On on-line evaluation

In many evaluation schemes the frequency of errors is one of the ingredients (e.g., Nielsen 1993; Walker et al. 1997). Arguably, the most useful kind of evaluation is *on-line* evaluation, since this gives the option of automatically adapting to the current situation. The analyses of this paper suggest 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. (1999a), these 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 the information it tried to verify is not in agreement with the user's intentions. If, on the other hand, the user's utterance does not contain such features, then it is highly likely that the verified information is correct. 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 speakers use hyperarticulate speech when they notice that the system had a problem recognizing their previous utterance, thus 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, Esther Klabbers, Mariet Theune and Mieke Weegels. We would like to thank Elizabeth Shriberg for urging us to do the perceptual experiment. Swerts is also affiliated with UIA and with the FWO - Flanders.

References

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

Bouwman, A.G.G., Sturm, J. & Boves, L., 1999. Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the Arise Project. In Proceedings International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, Vol. 1, pp. 493-496.

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., 1999. TiMBL: Tilburg Memory Based Learner, version 2.0, reference guide. ILK Technical Report 99-01, http://ilk.kub.nl/∼ilk/papers/ilk9901.ps.gz.

Dybkjær, L., Bernsen, N., Dybkjær, H., 1998. A methodology for diagnostic evaluation of spoken human-machine dialogue. International Journal Human-Computer Studies, 48:605-625.

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.

Johnson, C., 1999. Why human error modeling has failed to help system development. Interacting with computers 11 (5): 517-524.

Krahmer, E., Swerts, M., Theune, M., Weegels, M., 1999a. Problem spotting in human-machine interaction. In: Proceedings Eurospeech'99, Budapest, Hungary, pp. 1423-1426.

Krahmer, E., Swerts, M., Theune, M., Weegels, M., 1999b. Prosodic Correlates of Disconfirmations. In: Proceedings ETRW on Dialogue and Prosody, Eindhoven, The Netherlands, pp. 169-174.

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

Lewis C. & Norman, D.A., 1986. Designing for error. In: User-centered system design, ed. by D.A. Norman & S.W. Draper, Hillsdale, London: Lawrence Erlbaum, pp. 411-432.

Nielsen, J., 1993. Usability Engineering. Academic Press.

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

Reeves, B. & Nass, C., The Media Equation: How people Treat Computers, Television, and New Media like Real People and Places, CSLI Publications/Cambridge University Press.

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. 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.

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

Walker, M., Litman, J., Kamm, C., Abella, A. 1997. PARADISE: A framework for evaluating spoken dialogue agents. Proceedings ACL-EACL, Madrid, Spain, pp. 271-280.

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

Zipf, G.K., 1949. Human behavior and the principle of least effort. Addison-Wesley, Cambridge, MA.