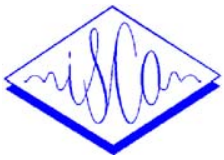

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Audiovisual Perception of Communication Problems

Pashiera Barkhuysen, Emiel Krahmer & Marc Swerts

Communication & Cognition
Tilburg University, The Netherlands
{P.N.Barkhuysen; E.J.Krahmer; M.G.J.Swerts}@uvt.nl

Abstract

We describe three perception studies in which subjects are offered film fragments (without any dialogue context) of speakers interacting with a spoken dialogue system. In half of these fragments, the speaker is or becomes aware of a communication problem. Subjects have to determine by forced choice which are the problematic fragments. In all three studies, subjects are capable of performing this task to some extent, but with varying levels of correct classifications. We conclude that combining auditory with visual information is beneficial for problem detection.

10.21437/SpeechProsody.2004-63

1. Introduction

It is well-known that managing communication problems in spoken human-computer interaction is difficult. One reason for this is that spoken dialogue systems are not good at determining whether the communication is going well or whether communication problems arose (e.g., due to poor speech recognition). Various researchers have shown that human speakers respond in a different vocal style to problematic system prompts than to unproblematic ones. For instance, when speech recognition errors occur, human speakers tend to correct these in a hyperarticulate manner (which may be characterized as longer, louder and higher). This generally leads to worse recognition results ('spiral errors'), since the standard speech recognizers are trained on normal, non-hyperarticulated speech (Oviatt et al., 1998). In a similar vein, when speakers respond to a problematic yes-no question, their denials ("no") share many of the properties typical of hyperarticulate speech, in that they are longer, louder and higher than unproblematic negations (Krahmer et al. 2002).

Based on these observations, it has been suggested that monitoring prosodic aspects of a speaker's utterances may be useful for problem detection in spoken dialogue systems. It has indeed been found that using automatically extracted prosodic features helps for problem detection (e.g., Litman et al. 2001, Lendvai et al. 2002), although the extent to which prosody is beneficial differs across studies. Moreover, in all these studies a sizeable number of problems is not detected. One way to improve the accuracy of problem detection is by including additional features. In this paper, we investigate whether *visual* cues, besides auditory ones, can be used as signals of problematic situations. Earlier work in, for instance, bimodal speech recognition has shown that using automatic lipreading in combination with more standard automatic speech recognition techniques leads to a reduction of the number of recognition errors (see e.g., Petajan 1985). Moreover, the use of both auditory *and* visual cues to problems is becoming a real possibility in advanced multimodal spoken dialogue systems (see e.g., Benoit et al. 2000), which combine speech recognition with facial tracking. An additional potential advantage of using visual information is that visual cues indicating communication problems might also occur when the person is *not* speaking, but for instance when (s)he becomes aware of a communication problem during the system's feedback. Such early detection would be useful from a system's point of view, since the sooner a problem can be detected, the earlier a repair strategy may be started (e.g., a re-ranking of recognition hypotheses or a modification of the dialogue strategy).

In this paper, we describe three perception studies to investigate the informativeness of auditory and visual cues for problem detection in spoken human-machine interaction. In these three studies, subjects were shown selected recordings of Dutch speakers engaged in a telephone conversation with a train timetable information system. The recordings constituted minimal pairs as they were very comparable but differed in that they were excised from a context which was either problematic or not. The recordings were presented without context to subjects who had to determine whether the preceding speaker utterance had led to a communication problem or not. The *first* study (section 3) focuses on subjects' responses during verification questions of the system (i.e., when subjects listen in silence), which either verify correct or misrecognized information. The *second* study (section 4) concentrates on speakers uttering "no", either in response to a problematic or an unproblematic yes-no question from the system. The *third* study (section 5), finally, is devoted to speakers uttering a destination station (filling a slot), either for the first time (no problem) or as a correction (following a recognition error). The descriptions of these three studies are sandwiched between an overview of the general experimental procedure (section 2) and a general discussion (section 7).

2. General procedure

2.1 Data collection The stimuli used in the three experiments were all taken from an audio-visual corpus of subjects engaged in telephone conversations with a speaker independent Dutch spoken dialogue system providing train timetable information. The corpus consists of 9 speakers (5 male and 4 female) who query the system on 7 train journeys (63 dialogues in total). Each dialogue took approximately 5 minutes. In 76% of the dialogues subjects finish the task successfully (i.e., they obtain the correct advice). The original recordings were made with a digital video camera (25 frames per second). Subjects were led to believe they were involved in the data collection required for a new kind of "video-phone", hence they were instructed to face the camera at all times. Also, to ensure an optimal view of the face without a phone device blocking important visual features, subjects had to interact via a mobile phone positioned in front of them on a table. Afterwards the recordings were read into a computer and transcribed. On the basis of the transcriptions it could be decided which speaker utterances were misrecognized,

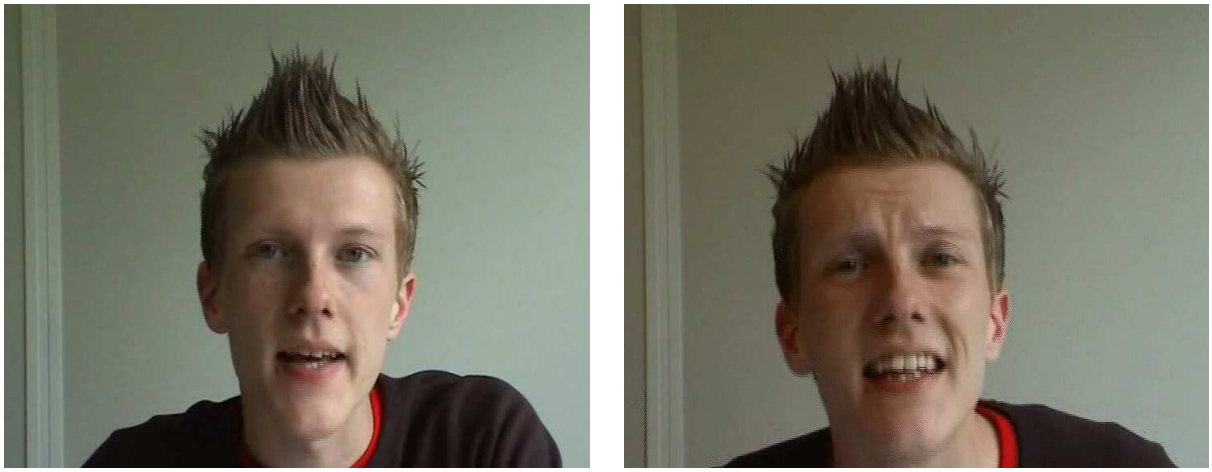

Figure 1: Two stills from speaker ED uttering the phrase "nee" (*no*) in an unproblematic (left) and a problematic situation (right).

and thus led to communication problems. It turned out that 374 out of 1183 speaker turns were misunderstood by the system (32%). These figures are representative of speaker independent spoken dialogue systems in real life settings.

2.2 Procedure For all three perception studies, the stimuli (verification questions, negations and slot-fillers respectively) were randomly selected on the basis of the transcribed dialogues. Per speaker, two problematic and two unproblematic instances were selected (if this turned out to be impossible for a speaker, that one was omitted from the experiment). In the perception studies, the stimuli were always presented per speaker and in a random order. The order in which speakers were presented was randomized in each study as well. Each block of four stimuli per speaker (two problems, two non-problems) was preceded by a reference stimulus showing that speaker in an unproblematic situation. Each study started with a short exercise session containing two unproblematic and two problematic stimuli, in order to make subjects familiar with the kind of stimuli and the experimental setting. See Figure 1 for two representative illustrations of speaker ED.

2.3 Subjects A group of 66 subjects (20 male and 47 female, all students from Tilburg University) participated in the three experiments, all but one native speakers of Dutch. The subjects were between 19 and 47 years old.

3. Study I: System questions

3.1 Task In the first study, subjects saw speakers listening to verification questions. These verification questions can be unproblematic, such as the system question in example (1).

(1) User: Amsterdam.
System: So you want to travel to Amsterdam?

But they can also verify misrecognized information as in (2):

(2) User: Rotterdam.
System: So you want to travel to Amsterdam?

In the first study, subjects have to determine on the basis of the speaker's facial expression during the system's verification, whether the verified information is correct (as in (1)) or not (as in (2)). They were shown 4 verification questions for all 9 speakers (36 stimuli in sum). For each speaker, two verification questions followed a recognition error and two did not.

Table 1: *Percentage of subjects who classify an instance of a speaker listening to a system utterance as signaling a problem. For 9 speakers, subjects classified two non-problematic stimuli (¬P1 and ¬P2) and 2 problematic ones (P1 and P2). (* $^a = p < .05$; $^b = p < .01$; $^c = p < .001$*)*

Speaker	¬P1	¬P2	P1	P2
AA	$.00^c$	$.01^c$	$.73^c$	$.94^c$
CH	$.80^c$	$.20^c$	$.99^c$	$.99^c$
DB	$.24^c$	$.30^b$	$.94^c$	.50
EC	$.20^c$	$.00^c$	$.62^a$	.59
ED	.61	.58	$.97^c$	1.0^c
IB	$.03^c$	$.23^c$	$.36^a$	.56
LS	$.28^c$	.53	$.94^c$	$.29^c$
PM	$.20^c$	.46	$.99^c$	$.38^a$
SB	$.06^c$	$.03^c$	$.88^c$	$.99^c$
Mean	.26		.75	

3.2 Results The results are presented in Table 1. All tests for significance were performed using a χ^2 test. Inspection of the table reveals that most speakers' reactions to unproblematic verification questions are indeed classified by the majority of the subjects as unproblematic. The overall mean of subjects who perceive unproblematic stimuli as problematic is only 26%. On the other hand, most subjects indeed classify speakers' reactions to problematic verification questions as signals of a problem (overall mean 75%). Table 2 summarizes the classifications from Table 1: for 12 of the 18 problematic verification questions and for 13 of the 18 unproblematic ones a statistically significant number of subjects made the correct classification. Note that some of the stimuli were systematically misclassified (in particular, utterance ¬P1 of speaker CH, utterance P1 of speaker IB, utterance P2 of speaker LS and utterance P2 of speaker PM).

Table 2: *Contingency table summarizing the number of significant classifications from Table 1, non-significant classifications are counted as random.*

Condition	Classification			Total
	Problem	¬Problem	Random	
Problem	12	3	3	18
¬Problem	1	13	4	18
Total	13	16	7	36

Table 3: *Percentage of subjects who classify a "no" utterance as signaling a problem. For 7 speakers, subjects classified two non-problematic stimuli (¬P1 and ¬P2) and 2 problematic ones (P1 and P2). (* $^a = p < .05$; $^b = p < .01$; $^c = p < .001$*)*

Speaker	¬P1	¬P2	P1	P2
AA	.49	$.27^c$	.59	.50
CH	$.08^c$	$.26^c$	$.76^c$	.53
EC	.59	.58	.41	.39
ED	.39	.46	$.88^c$	$.68^b$
IB	$.18^c$	.52	$.18^c$	$.65^a$
LS	$.71^c$	$.68^b$	.45	.42
SB	$.38^a$	$.27^c$	$.24^c$	$.70^c$
Mean	.41		.52	

3.3 Discussion The results of the first study show that subjects are generally capable of correctly determining whether a verification question contained a problem or not, solely on the basis of a speaker's facial expression during the verification. This shows that keeping track of facial expressions during spoken human-machine interactions can be helpful, even when speakers are silent. Closer inspection of the stimuli suggests that during unproblematic verification questions, subjects maintain a neutral facial expression throughout, while they become more expressive (e.g., moving, laughing or frowning) during problematic verification questions. Interestingly, the aforementioned systematic misclassifications support this informal observation, in that speaker CH frowns during an unproblematic system question, while speakers IB, LS and PM remain a neutral expression during a system question which verifies misrecognized information.

4. Study II: Negations

4.1 Task In the second study, subjects saw speakers only uttering a negation ("nee", *no*). This could be a response to a yes-no question which does not verify recognized information (so speakers by definition do not become aware of a communication problem), as in example (3):

(3) System: Do you want me to repeat the connection?
User: No.

On the other hand, if the question verifies a misrecognition (cf. example (2) above), subjects' "no" signals a communication problem:

(4) System: So you want to travel to Amsterdam?
User: No.

Subjects of the perception study saw only the "no" utterances, presented without any further context, and had to determine whether the speaker signalled a communication problem (as in 4) or not (as in 3). Stimuli from seven speakers were used in the second study, with a total of 28 disconfirmation answers. Two speakers were omitted, as it was not possible to obtain a balanced set from their data.

Table 4: *Contingency table summarizing the number of significant classifications from Table 3, non-significant classifications are counted as random.*

Condition	Classification			Total
	Problem	¬Problem	Random	
Problem	5	2	7	14
¬Problem	2	6	6	14
Total	7	8	13	28

4.2 Results The results of the second study can be found in Table 3. All tests for significance were performed using a χ^2 test. The results show that subjects found this test much harder than the first one. Overall, the unproblematic negations are perceived as problem indicators by 41% of the subjects, while the problematic ones are perceived as signalling a problem by 52% as the subjects. Clear differences between speakers exist. Speaker LS is often misclassified: the two unproblematic utterance are both significantly classified as signaling a problem, while the two problematic utterances score random (most subjects consider them unproblematic). Overall, in about half of the cases no significant preference in either direction exists (see Table 4). Of the 15 stimuli for which the classification showed a significant pattern, the majority is in the expected direction. The significant misclassifications for the unproblematic cases are both due to LS.

4.3 Discussion In general subjects found it difficult to determine on the basis of just the "no" whether this negation marker signalled a communication problem or not. In roughly half of the cases, there was no significant tendency in either direction. Of the remaining cases most of the classifications were correct. This outcome weakly confirms earlier work on the perception of negations (Krahmer et al. 2002); subjects had more difficulty in classifying the negations in the current experiment. This could be due to the fact that the negation phrases in Krahmer et al. (2002) were always cut from longer utterances (e.g., "no, thanks" or "no, to Rotterdam!"). Alternatively, it could also be that the visual modality distracts listeners from the prosodic cues (compare Doherty-Sneddon et al. 2001).

5. Study III: Destinations

5.1 Task In the third study, subjects saw speakers uttering a destination. This could be in a no-problem context like (5):

(5) System: To which station do you want to travel?
User: Rotterdam.

Or, it could be a correction in response to a verification question of misrecognized information (compare (2) above):

(6) System: So you want to travel to Amsterdam?
User: Rotterdam.

For the third study 8 speakers were selected, with a total of 32 stimuli. One speaker was omitted, as it was not possible to obtain two problematic and two unproblematic stimuli from his dialogues.

Table 5: *Percentage of subjects who classify an instance of a speaker uttering a destination as signaling a problem. For 8 speakers, subjects classified two non-problematic stimuli (¬P1 and ¬P2) and 2 problematic ones (P1 and P2). ($^a = p < .05$; $^b = p < .01$; $^c = p < .001$)*

Speaker	¬P1	¬P2	P1	P2
AA	$.68^b$	.53	$.73^c$	$.65^a$
CH	$.14^c$	$.67^b$	.61	$.94^c$
DB	$.11^c$	.47	$.99^c$	$.97^c$
EC	.53	$.70^b$	$.00^c$	.39
ED	.61	$.70^b$	.61	1.0^c
IB	$.05^c$	$.26^c$	$.99^c$	$.80^c$
LS	$.06^c$	$.26^c$	.56	$.70^b$
PM	$.20^c$	$.32^b$	$.79^c$	1.0^c
Mean	.39		.73	

Table 6: *Contingency table summarizing the number of significant classifications from Table 5, non-significant classifications are counted as random.*

Condition	Classification			Total
	Problem	¬Problem	Random	
Problem	11	1	4	16
¬Problem	4	8	4	16
Total	15	9	8	32

5.2 Results Table 5 displays the results per speaker, and table 6 summarizes these results. Significance was tested with the χ^2 method. The overall results are closely related to those of the first study: most subjects classify most non-problematic destinations as unproblematic, and they classify most problematic destinations as problematic. Again differences between speakers are found, most notable here is that 4 unproblematic slot-fillers are significantly classified as problematic. Another striking outlier is utterance P1 from EC, which all 66 subjects classified as unproblematic.

5.3 Discussion In a majority of cases subjects were capable to correctly classify speaker's utterances of destinations. Inspection of the stimuli suggests the same basis picture as for the first study: when there are no problems, subjects have a neutral facial expression, when they need to correct misrecognized information they become more expressive. The clearest cue appears to be audiovisual hyperarticulation.

6. General Discussion and Conclusion

We have described three perception studies in which subjects were offered film fragments (without any dialogue context) of speakers interacting with a spoken dialogue system. In half of these fragments, the speaker is or becomes aware of a communication problem. Subjects had to determine by forced choice which are the problematic fragments. It was found that in all three studies, subjects were capable of performing this task to a certain degree, but that the number of correct classification varies across the three studies. As it turned, subjects had most difficulty with the second study, in which the stimuli consisted only of negation phrases ("no"). Surprisingly, the results were best in the first study, in which subjects silently listen to a verification question of the system. Speculating on why the different tests have led to different results, we hypothesize that this is partly due to the fact that the stimuli in experiments 1 and 3 were longer than in experiment 2, which consisted of only a very short fragment (the word "no"). Accordingly, the longer clips may have contained more cues than the shorter ones. Next, in order to gain more insight into the audiovisual features that may have served as possible signals to problematic and unproblematic utterances and to support our preliminary informal observations, we intend to label our stimuli in terms of a detailed coding scheme, such as the FACS system (Ekman & Friesen, 1975). We also plan to experiment with (semi-)automatic procedures to detect audiovisual cues in our recordings, such as automatic measurements of the amount of variation in a clip which is potentially useful to distinguish neutral from more dynamic faces. Finally, results of this type of research could be beneficial for improving human-machine interactions in that audiovisual correlates of problematic utterances allow systems to monitor the level of frustration of a user (Picard & Klein, 2002) or to use them as a resource for error detection.

Acknowledgments This research was conducted as part of the VIDI-project "Functions Of Audiovisual Prosody" (FOAP), sponsored by the Netherlands Organization for Scientific Research (NWO). Marc Swerts is also affiliated with Antwerp University and with the FWO-Flanders.

7. References

[1] Benoit, C., Martin, J.-C., Pelachaud, C., Schomaker, L. & Suhm, B., 2000, Audio-Visual and Multimodal Speech Systems, in: *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, I. Mertens, R. Moore (eds.), Kluwer Academic Publishers.

[2] Doherty-Sneddon, G., Bonner, L., Bruce, V., 2001, Cognitive demands of face monitoring: Evidence for visuospatial overload. *Memory & Cognition* 29 (7):909-919.

[3] Ekman, P., & Friesen, W. v., 1975, *Unmasking the face: a guide to recognizing emotions from facial expressions*. Englewood Cliffs, N.J.: Prentice Hall.

[4] Krahmer, E., Swerts, M., Theune, M., & Weegels, M., 2002, The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication*, 36, 133-145.

[5] Lendvai, P., van den Bosch, A., Krahmer, E., Swerts, M., 2002, Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting. In: Kuebler, S. & Hinrichs, E. (Eds.), *Machine Learning Approaches in Computational Linguistics*, 1-15, Trento, Italy.

[6] Litman, D, Hirschberg, J., & Swerts, M., 2001, Predicting user reactions to system error, ACL 2001, 362-369.

[7] Oviatt, S., MacEachern, M., & Levow, G.-A., 1998, Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24(2), 1-23.

[8] Petajan, E., 1985, Automatic Lipreading to Enhance Speech Recognition. *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 40-47.

[9] Picard, R., & Klein, J., 2002, Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications. *Interacting with Computers*, 14 (2), 141-169.