Problem Detection in Human-Machine Interactions based on Facial Expressions of Users

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

Published in:
Speech Communication

Publication date:
2005

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 343–359
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Problem detection in human–machine interactions based on facial expressions of users

Pashiera Barkhuysen *, Emiel Krahmer, Marc Swerts

Communication & Cognition, Tilburg University, P.O. Box 90153, Tilburg NL-5000 LE, The Netherlands

Received 12 February 2004; received in revised form 9 July 2004; accepted 6 October 2004

Abstract

This paper describes research into audiovisual cues to communication problems in interactions between users and a spoken dialogue system. The study consists of two parts. First, we describe a series of three perception experiments 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 tests, subjects are capable of performing this task to some extent, but with varying levels of correct classifications. Second, we report results of an observational analysis in which we first attempt to relate the perceptual results to visual features of the stimuli presented to subjects, and second to find out which visual features actually are potential cues for error detection. Our major finding is that more problematic contexts lead to more dynamic facial expressions, in line with earlier claims that communication errors lead to marked speaker behaviour. We conclude that visual information from a user's face is potentially beneficial for problem detection.

1. Introduction

The goal of the investigation presented in this article is to explore to what extent it could be beneficial to use features of a user's facial expression to detect communication problems in his or her interactions with a spoken dialogue system. It is well-known that managing communication problems in spoken human–computer interaction is difficult. One key issue 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 or false default assumptions). The occurrence of problems negatively affects user satisfaction (Walker et al., 1998), but also has an impact on the way users communicate with the

* Corresponding author.
E-mail addresses: p.n.barkhuysen@uvt.nl (P. Barkhuysen), e.j.krahmer@uvt.nl (E. Krahmer), m.g.j.swerts@uvt.nl (M. Swerts).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.10.004

system in subsequent turns, both in terms of their language and speech. For instance, when users notice that a system has difficulties to handle their prior spoken input, they tend to produce utterances with marked linguistic features (e.g., longer sentences, marked word order, more repeated information, etc.) (Krahmer et al., 2002). In addition, human speakers also respond in a different vocal style to problematic system prompts than to unproblematic ones: when speech recognition errors occur, they 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; Levow, 2002; Hirschberg et al., 2004), although more recent studies suggest that systems become less vulnerable to hyperarticulation (Goldberg et al., 2003). 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).

In other words, one could state that dialogue problems lead to a marked interaction style of users, which manifests itself partly in a set of prosodic correlates. 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., Hirschberg et al., 2004; Lendvai et al., 2002). While this has led to some improvements, the extent to which prosody is beneficial differs across studies. Moreover, in all these studies a sizeable number of problems is not detected. In general, it appears that the detection of errors improves if prosodic features are used in combination with other features already available to the system, such as more traditional acoustic or semantic confidence scores, knowledge about the dialogue history, or the grammar being used in a particular dialogue state (Litman et al., 2001; Bouwman et al., 1999; Hirschberg et al., 2001; Danieli, 1996; Ahrenberg et al., 1993). The current paper explores whether it is potentially useful to include yet another set of features, i.e., visual features from the face of the user who is interacting with the computer.

Indeed, it makes sense to assume that a speaker's facial expressions may signal communication problems as well. One obvious reason is that hyperarticulation is likely to be detectable from inspecting more exaggerated movements of the articulators. Erickson et al. (1998) found that speakers' repeated attempts to correct another person are highly correlated with more pronounced jaw movements, which are likely to be clearly visible to their addressees (see also Gagné et al., 2004; or Dohen et al., 2003 about related visual correlates of contrastive stress). In addition, in line with the earlier observation that speakers adapt their language and speech after communication errors to a more marked interaction style, there is evidence that speakers also change their facial expressions in problematic dialogue situations. Swerts et al. (2003) applied the so-called Feeling-of-Knowing paradigm (Hart, 1965; Smith and Clark, 1993; Brennan and Williams, 1995) to investigate how speakers cue that they are certain or rather uncertain about a response they give to a general factual question. It was found that it is indeed often clearly visible when people were insecure about the answer to a response, in that speakers show much more deviations from "normal" facial expressions (e.g., more eyebrow movement and gaze acts). Given such observations, it is worthwhile to investigate whether speakers also exhibit special visual expressions when they are confronted with communication problems in spoken human–machine interactions.

This research fits in a recent interest to try and integrate functional aspects of facial expressions in multimodal systems, with the ultimate goal to make the interaction with such systems more natural and efficient. Some systems already supplement their interface with an embodied conversational agent (ECA), for instance in the form of a synthetic head, to support the communication process with users. Visual cues of such ECA's appear to be functionally relevant in more than one respect. They make the speech more intelligible (e.g., Agelfors et al., 1998; see also

Jordan and Sergeant, 2000), and can give clues about the status of the information a system sends to the user, for instance to signal the difference between negative or positive feedback responses from a system (Granström et al., 2002). An additional advantage of using a synthetic face is that it can give silent cues about the internal state of the system, e.g., to signal that it is paying attention to the user or that it is looking for information, following the general best practice to make a system's behavior and reasoning clear to a user (Sengers, 1999).

The perspective in the current paper is different from that of such earlier studies in that it does not concentrate on multimodal features of system utterances, but rather deals with analyses of the users' facial expressions. The exploitation of the users' auditory *and* visual cues 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. Earlier work in 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). In addition, comparable to the silent visual cues from a system, facial expressions of a user may indicate communication problems even when the person is *not* speaking, but for instance when (s)he becomes aware of a communication problem during the system's feedback. Such cues clearly have added value compared to the auditory and linguistic cues to errors used before, because they would enable a very early detection of problems. Obviously, this 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).

Therefore, the general goal of the research described in this paper is to investigate the information value of a speaker's visual cues for problem detection in spoken human–machine interaction. The study consists of two parts. First, we describe three perception experiments in which subjects were shown selected recordings of Dutch speakers engaged in a telephone conversation with a train timetable information system. [1] The recordings constituted minimal pairs as they were very comparable in terms of their words and syntactic structure but differed in that they were excised from a context which was either problematic or not. The recordings were presented without the original context to subjects who had to determine whether the preceding speaker utterance had led to a communication problem or not. The *first* experiment 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* experiment concentrates on speakers uttering "no", either in response to a problematic or an unproblematic yes–no question from the system. The *third* experiment, 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 preceded by an overview of the general experimental procedure. The second part of the paper describes the results of some observational analyses. We attempt to find visual correlates of problematic situations that could have functioned as cues to subjects in the different perception studies described in part 1. Our major finding is that more problematic contexts lead to more dynamic facial expressions, in line with earlier claims that communication errors lead to marked speaker behaviour. We conclude our paper with a general discussion and some perspectives on further research.

2. Perception studies

2.1. General method

2.1.1. Data collection

The stimuli used in the three experiments were all taken from an audiovisual corpus of speakers engaged in telephone conversations with a speaker independent Dutch spoken dialogue system providing train timetable information. The corpus

[1] In the remainder, "speakers" refer to users who were recorded while they interact with a spoken dialogue systems; "subjects" are participants in the different perception tests.

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 speakers 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). Speakers 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, speakers 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 or misunderstood, 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 (e.g., Hirschberg et al., 2004; Carpenter et al., 2001; Nakano and Hazen, 2003; Walker et al., 1998).

2.1.2. 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 speaker was omitted from the experiment). In the perception studies, the stimuli were always presented per speaker and in a random order. 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 (and a reference stimulus), in order to make subjects familiar with the kind of stimuli and the experimental setting. See Fig. 1 for two representative illustrations of speaker ED.

2.1.3. Subjects

A group of 66 subjects (20 male and 46 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.

2.2. Experiment I: System questions

2.2.1. 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?

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

In the first study, subjects have to determine on the basis of the speaker's facial expression during the system's explicit verification questions, 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.

2.2.2. 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 1
Percentage of subjects who classify an instance of a speaker listening to a system utterance as a signal a problem

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	

For 9 speakers, subjects classified two non-problematic stimuli (¬P1 and ¬P2) and two problematic ones (P1 and P2).
[a] $p < .05$.
[b] $p < .01$.
[c] $p < .001$.

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

2.2.3. 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 keep a neutral expression during a system question which verifies misrecognized information. PM differed from the other two speakers in the sense that he also smiled in the film fragment.

2.3. Experiment II: Negations

2.3.1. 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), but instead offers the speaker a choice in the possible course of action taken by the system in the subsequent dialogue, 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 negations. Two speakers were omitted, as it was not possible to obtain a balanced set from their data.

2.3.2. 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 utterances are both significantly classified as signals of a problem, while the two problematic utterances score random (most subjects consider them unproblematic). Closer inspection of the stimuli reveals that LS was frowning in the unproblematic utterances. 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. The significant misclassifications for the problematic cases are due to IB and SB. A first inspection of their recordings shows that IB displayed little or no facial expressions, while SB showed strong head movements and was nodding.

Table 3
Percentage of subjects who classify a "no" utterance as a signal of a problem

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	

For 7 speakers, subjects classified two non-problematic stimuli (¬P1 and ¬P2) and two problematic ones (P1 and P2).
[a] $p < .05$.
[b] $p < .01$.
[c] $p < .001$.

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

2.3.3. Discussion

In general subjects found it difficult to determine on the basis of just the "no" whether this negation 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); albeit that 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). Also the unproblematic negations occurred always at the end of the original conversation, so it may have been possible that the speakers' faces showed irritation after being misunderstood earlier in the conversation.

2.4. Experiment III: Destinations

2.4.1. 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 or misunderstood information (cf. (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.

2.4.2. 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. An inspection of these film fragments show that some of the speakers were frowning, and all were hyperarticulating. Another striking outlier is utterance P1 from EC, which all 66 subjects classified as unproblematic. The fragment shows that this speaker displayed a single head movement, but no further movements.

Table 5
Percentage of subjects who classify an instance of a speaker uttering a destination as a signal of a problem

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	

For 8 speakers, subjects classified two non-problematic stimuli (¬P1 and ¬P2) and two problematic ones (P1 and P2).
[a] $p < .05$.
[b] $p < .01$.
[c] $p < .001$.

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

2.4.3. Discussion

In a majority of cases subjects were capable to correctly classify speaker's utterances of destinations. Inspection of the stimuli suggests the same basic 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. Audiovisual hyperarticulation appears to be a clear cue for this.

2.5. Observational analyses

2.5.1. Introduction

The series of perception experiments described above brought to light that subjects are generally capable to detect problematic dialogue events on the basis of observations of recorded film fragments of human–machine interactions. While

subjects also had access to possible speech cues in the video films, there are reasons to believe that visual signals have undoubtedly played a role too in their classification of problematic and unproblematic events. In particular, since the speakers did not talk at all in experiment I, subjects could only have payed attention to facial expressions from the recorded speaker.

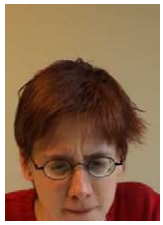
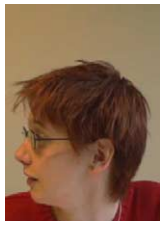
To gain further insight into such visual cues, we annotated all fragments in terms of a number of facial features, that could have functioned as cues to problematic or unproblematic dialogue events. In the next sections, we will first describe the labeling procedure we defined, and then embark on the results of analyses where we correlate the annotated features both with the actual and the perceived problems described in the earlier part of the paper. It will be shown that problematic dialogue sequences are characterized by more dynamically varying facial expressions of users, in line with earlier observations that speakers switch to a marked interaction style in terms of their language and speech in the case of problems.

2.5.2. Labeling

In order to determine which visual cues influenced subjects' judgements we labeled the fragments mentioned above using a set of facial features. The choice of these features was primarily based on the results of pilot observations of a subset of the recorded video fragments (see various discussion sections above). The labels consist of seven different visual features, five of which are defined and visualised in Table 7. The chosen features are roughly comparable with Action Units (AUs) described by Ekman and Friesen (1978), though there is not necessarily a one-to-one mapping to these Action Units. These Action Units constitute the basic ingredients for the influential Facial Action Coding System (FACS) which assumes that every visible facial movement is the result of muscular action. Therefore, a comprehensive coding system can be obtained by discovering how each muscle of the face acts to change a unique visible appearance. With that knowledge it would be possible to analyze any facial movement into anatomically based uniquely discriminable Action Units. Table 7 in particular displays examples of marked settings of smile (AU 12–13), diverted head position (AU 51–58), gaze (AU 61–64), frown (AU 4) and eyebrow raising (AU 1–2). Additional visual features not shown in this table are final mouth opening (AU 25–27) (i.e., whether a speaker silently opened his mouth at the end of the video film to prepare for upcoming speech) and the occurrence of repetitive vertical or horizontal head gestures (basically reflecting a yes or no signal); both are difficult to visualize using a single still image. All of these features were labeled as discrete events, in terms of presence or absence of a marked setting of the feature, except for diverted head position and smiling which were given a number on a small scale between 0 and 2 to reflect different strengths, where 0 stands for a complete absence and 2 represents a very clear presence of of a diverted head position or smiling. The repetitive head gestures, when present, were given a different label according to whether they represented a vertical (''yes'') or horizontal (''no'') gesture. In addition to these purely visual features, we also included one primarily auditory one, i.e., the occurrence of hyperarticulation. The presence of hyperarticulation was largely determined on the labelers' auditory impression of whether the speech was generally spoken with a louder voice, higher pitch, and/or at a slower rate, though it is clear, as already suggested by earlier findings of Erickson et al. (1998), that hyperarticulation was also cued visually. Following procedures outlined by Wade et al. (1992), hyperarticulation was given a number between 0 and 2 to distinguish different degrees of hyperarticulation, where 0 represents complete absence and 2 a very strong form of hyperarticulation.

The labeling was performed by the three authors of this article. The procedure was as follows. The judges watched the film fragments and labelled them using a set of eight features, i.e., the seven visual features plus hyperarticulation. Each judge labelled each feature individually. Comparing the labelers' individual scores showed an agreement in most of the cases (80%), where agreement is computed by counting the number of video fragments which received total consensus (three identical annotations for all eight features) divided by the total number of fragments. If a

Table 7
Selection of a number of annotated features; the description and examples represent the marked settings for each feature

Label	Definition and example
Smile	Speaker produces a clearly visible smile or laughter 
Head movement	Speaker moves head away from its position at onset 
Gaze	Speaker diverts eye gaze from its position at onset, relative to the position of the head 
Frown	Speaker produces a frown, primarily visible in the forehead or between the eyebrows 
Eyebrow raising	Speaker raises one or two eyebrows from neutral position 

feature was labeled on a scale and the individual scores on the scale did not match (e.g., one judge saw minor hyperarticulation ('1') and the two other judges noted very clear hyperarticulation('2')), this was also regarded as disagreement. The film fragments of the destinations invoked the largest amount of disagreement (25%). The features upon which there was most disagreement were: hyperarticulation (48%) and head movements (38%), whereas judges always agreed on the annotation of final mouth opening. One complicating factor in the labelling process was that the different features are not entirely independent and are sometimes difficult to separate, such as the potential co-occurrence of a single head movement (diverted head position) and repetitive head gestures which could result in nodding. Also, it was not always obvious to determine whether the face varied in terms of a head movement alone, or in combination with diverted gaze. For the analyses below, disagreements between labelers were resolved via majority voting for the discrete features, while the scores for the continuous features (diverted head position, smiling and hyperarticulation) were summed resulting in an overall score between 0 and 6 for these respective features.

2.5.3. Results

In the results section, we explore to what extent there is a relation between the *perceived* problems in the three experiments and the annotated audiovisual features described above. In addition, we also investigate the relation between the audiovisual cues and the *actual* presence or absence of problems in the stimuli.

Audiovisual features and the perception of problems. First, we will look at various correlations of these features with the proportion of subjects who classify a film fragment as problematic. To this end, we will take a purely perception-oriented approach, in the sense that we do not take into account whether or not the fragment was originally extracted from a problematic or unproblematic dialogue context. In other words, what matters is how that fragment is classified by a subject, irrespective of whether that classification was correct or not. The results are shown in Table 8, which gives the overall results for the stimuli used in

Table 8
Distribution of utterances from experiments I–III perceived as problematic or not problematic as a function of the presence or absence of a marked feature setting

Feature	Present	Perceived as		Statistics	
		¬Problem	Problem	χ^2	Cramér's V
1. Hyperarticulation	No	854	466	221.7[a]	.237
	Yes	1046	1594		
2. Smiling	No	2455	2231	119.0[a]	.137
	Yes	607	1043		
3. Diverted head pos.	No	1015	1097	.1	.004
	Yes	2047	2177		
4. Frowning	No	2539	1883	484.4[a]	.277
	Yes	523	1391		
5. Eyebrow raising	No	2665	2615	58.4[a]	.096
	Yes	397	659		
6. Eye movements	No	1671	1563	29.6[a]	.068
	Yes	1391	1711		
7. Mouth opening	No	2256	2628	38.9[a]	.078
	Yes	806	646		
8. Repeated head gest.	No	2452	2762	99.4[a]	.125
	Horiz.	144	252		
	Vert.	466	260		

The significance and the strength of the associations are expressed in terms of χ^2 (df = 1, except for 8 where df = 2) and Cramér's V tests, respectively.

[a] $p < .001$.

experiments I–III, respectively. Hyperarticulation does not play a role in experiment I (the speaker silently listens to the system), and is treated as a missing value in that experiment. For the purpose of simplicity we recoded the scalar features to binary ones in this table (but see below). The results are presented in the form of different 2-by-2 matrices, which give the distributions of utterances perceived as problematic or not problematic as a function of the presence or absence of a marked feature setting. The significance and the strength of the associations are expressed in terms of χ^2 and Cramér's V tests, respectively. The overall results show that almost all features had a significant impact on the way an utterance is perceived as problematic or not: the presence of a marked setting leads to a higher proportion of problem perceptions, with the exceptions of (1) final mouth opening, which, when present, has a higher relative number of non-problem classifications and (2) diverted head positions, which did not have an overall influence on problem perception.

If we look at the stimuli used in experiment I (System questions), we see that all audiovisual features have a significant influence on the perception judgements (with $p < .001$). In order of strength: frowning ($\chi^2 = 453.2, V = .437$), repeated head gestures ($\chi^2 = 305.2, V = .358$), eyebrow raising ($\chi^2 = 154.3, V = .255$), smiling ($\chi^2 = 130.9, V = .235$), eye movements ($\chi^2 = 129.2, V = .233$), mouth opening ($\chi^2 = 26.8, V = .106$) and, finally, diverted head position ($\chi^2 = 16.7, V = .084$) (recall that hyperarticulation plays no role in this experiment). It is worth noting that even though diverted head position had no overall significant effect (see Table 5), there is a small but significant effect of this feature in the first experiment. In general, the presence of a marked audiovisual feature implies that more subjects perceive problems, only for mouth opening this trend is reversed.

For the stimuli from experiment II (Negations), the results are less clear. Only three features had a significant influence on problem perception, and in general, the scores on the Cramér's V test showed much weaker associations than reported for experiment I. Ordered by strength the significant cues were: frowning ($\chi^2 = 43.0, V = .153$), hyperarticulation ($\chi^2 = 31.3, V = .130$), and smiling ($\chi^2 = 17.0, V = .096$). This outcome is consistent with the results of the perception study in experiment II; apparently the stimuli in this part contained few cues which subjects could use to determine whether a speaker's "no" came from a problematic or an unproblematic turn.

The situation for experiment III (Destinations) is subtly different again. All features have a significant effect, apart from repeated head gestures. And again, if a marked audiovisual feature setting is present, this leads to an increased proportion of perceived problems, unless the feature is mouth opening which, as above, seems to have an effect in the opposite direction. Interestingly, the relative importance of the features (in terms of strength of association) is somewhat different here: hyperarticulation ($\chi^2 = 224.6, V = .326$), mouth opening ($\chi^2 = 87.3, V = .203$), frowning ($\chi^2 = 65.2, V = .176$), diverted head position ($\chi^2 = 62.8, V = .172$), eye movement ($\chi^2 = 7.9, V = .061$), eyebrow raising ($\chi^2 = 6.9, V = .057$), smiling ($\chi^2 = 6.5, V = .055$). For destinations, hyperarticulation is clearly the single most important cue that subjects based their perceptual judgements on.

In the presentation of the results we have treated hyperarticulation as a binary cue, whereas in

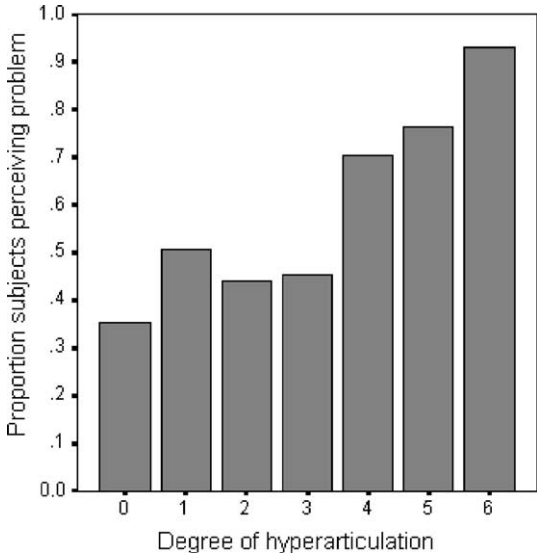


Fig. 2. Barcharts with the average proportion of subjects perceiving a stimulus as problematic as a function of different degrees of hyperarticulation.

fact it was coded on a 7 point scale (the summed score of the 3 judges). Fig. 2 shows the average proportion of subjects perceiving a fragment as problematic as a function of different degrees of hyperarticulation (ranging from 0 to 6), for the stimuli from experiment II and III. This figure shows a clear trend, where stimuli that get more extreme values in terms of hyperarticulation, also are perceived as more problematic. Correlational analysis reveals that the proportion of perceived problems increases as a function of the degree of hyperarticulation ($r = .679$, $p < .001$).

In general, it appears that the presence of a marked audiovisual feature setting gives rise to more subjects perceiving a problem. While the results show that there are significant effects of various features, the 'sizes' of these effects are often rather minimal as can be seen from the Cramér's V scores. This suggests that the perception of problem status does not seem to be the result of a *single* factor in isolation. Indeed, when we checked all 2-way interactions between the various factors on the whole dataset using a multinomial logistic regression analysis, we found that all these interactions were above chance level, which suggests that perceived problem status results from a combination of cues. More detailed interaction analyses are unfortunately not feasible given the unbalanced nature of the data set and the resulting data sparseness.

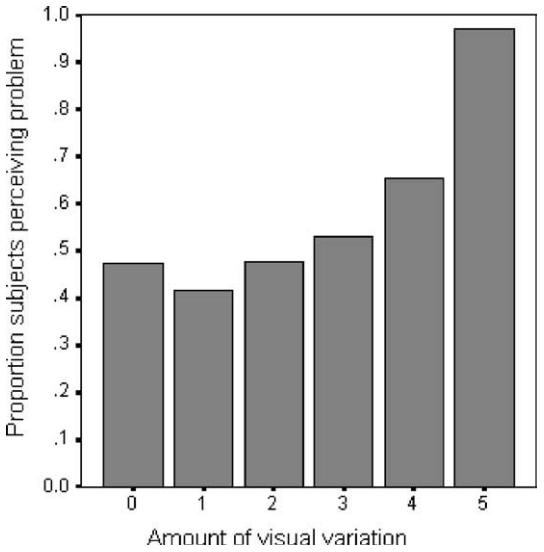


Fig. 3. Barcharts with the average proportion of subjects perceiving a stimulus as problematic as a function of amount of visual variation.

As an alternative way to get a view on the effect of combinations of features, we determined if and how the perceived problem status of a stimulus depended on the number of marked features in an utterance. By focussing solely on visual variation, we get a better insight in the contribution of the visual factors to problem perception. To this end, we calculated the average proportion of subjects perceiving a fragment as problematic as a function of the degrees of visual variation, where visual variation was computed by summing over the presence of marked settings of each visual feature, where smiling, head movements and repetitive head gestures were recoded in terms of presence or absence. [2] This gave a range that varied between the theoretical extremes of 0 and 7 (though we actually did not get any case where all visual features were present at the same time). The results are visualised in Fig. 3. Interestingly, the resulting picture is very similar to that in Fig. 2; more problematic fragments get more extreme values both in terms of visual variation and in terms of hyperarticulation. Correlational analyses bring to light that the proportion of perceived problems increases as a function of the amount of visual information ($r = .294$, $p < .01$). Summarizing: Film fragments that are perceived as more problematic are also more dynamic in terms of speech and facial features.

Audiovisual features and the presence of problems. So far we have taken a purely perceptive perspective, yet it is also interesting to take a more system-oriented perspective and investigate the relation between the audiovisual cues and the actual presence or absence of communication problems.

[2] Note that some repetitive head gestures do not appear to cue problems (e.g., nodding). In a similar vein, we saw that mouth opening is not perceived as a cue for problems either. A more sophisticated analysis to visual variation might leave out these cues, but here we simply summed over *all* visual variation.

To find out, we redid the analysis with problem instead of perceived problem as our class of interest. The results of this analysis can be found in Table 9, which gives the distribution of utterances from experiments I–III that are either problematic or not as a function of the presence or absence of a marked feature setting. The first thing to note is that we have much less datapoints here than in the perceptual analysis. Still, there are some significant features, namely hyperarticulation ($\chi^2 = 4.8, V = .283$) and smiling ($\chi^2 = 6.5, V = .261$). Thus, when a speakers hyperarticulates or smiles, chances that a communication problem had occurred increase. Frowning, eyebrow raising and gaze show a similar pattern, although not statistically significant. Repeated head gestures and mouth opening do not seem to correlate with problem status. It is interesting to note that even though frowning occurs relatively often in unproblematic stimuli (12 times), subjects in the perception test have a strong tendency to interpret frowning as a cue for problems. A somewhat similar observation can be made with respect to nodding, which occurs almost as often in unproblematic as in problematic stimuli (6 and 5 times respectively), while subjects have relatively strong tendency to interpret this behavior as a cue for the absence of communication problems.

As above, it is interesting to look at both the amount of hyperarticulation and at the amount of visual variation as cues for communication problems. Figs. 4 and 5 show the average proportion of problematic stimuli as a function of the amount of hyperarticulation and the degrees of visual variation, respectively. Correlational analyses reveals that the proportion of problems increases as a function of both degree of hyperarticulation and of the amount of visual variation, though the latter is not significant, probably due to sparse data (hyperarticulation: $r = .914$, $p < .01$; visual variation: $r = .601$, $p = .207$). As one would expect, hyperarticulation is a clear cue for problems. But the data show a similar trend for visual variation: it appears to be a cue for problems as well, in the sense that if two or more visual cues are present in a stimuli, the chances that the utterance was

Table 9
Distribution of utterances from experiments I–III that are either problematic or not as a function of the presence or absence of a marked feature setting

Feature	Present	Fragment was		Statistics	
		¬Problem	Problem	χ^2	Cramér's V
1. Hyperarticulation	No	14	6	4.8[a]	.283
	Yes	16	24		
2. Smiling	No	41	30	6.5[a]	.261
	Yes	7	18		
3. Diverted head pos.	No	13	19	1.7	.133
	Yes	35	29		
4. Frowning	No	36	31	1.2	.113
	Yes	12	17		
5. Eyebrow raising	No	43	37	2.7	.168
	Yes	5	11		
6. Eye movements	No	26	23	.4	.063
	Yes	22	25		
7. Mouth opening	No	37	37	0	0
	Yes	11	11		
8. Repeated head gest.	No	39	40	.1	.033
	Horiz.	3	3		
	Vert.	6	5		

The significance and the strength of the associations are expressed in terms of χ^2 (df = 1, except for 8 where df = 2) and Cramér's V tests, respectively.

[a] $p < .05$.

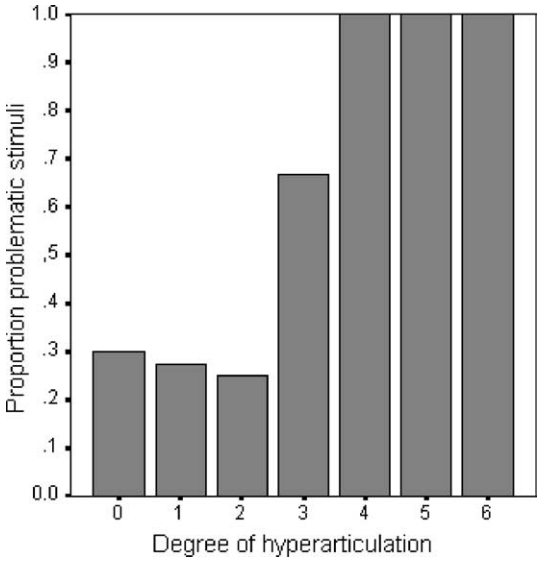


Fig. 4. Barcharts indicating the percentage of problematic stimuli as a function of different degrees of hyperarticulation.

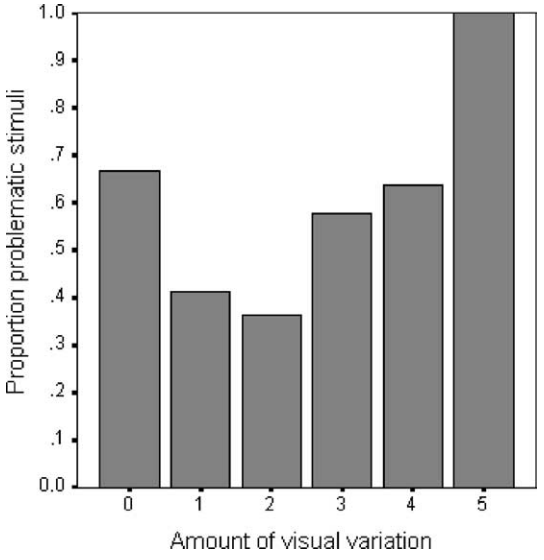


Fig. 5. Barcharts indicating the percentage of problematic stimuli as a function of the amount of visual variation.

problematic increase as well. This latter bargraph also illustrates that it is not feasible to detect errors on the basis of visual cues alone, since a sizeable number of stimuli contained no visual cues but were problematic nevertheless.

2.5.4. Discussion

The main finding of the correlational analyses presented here is that the perceived problem status of a user utterance is not only reflected in a particular speech feature, i.e., in different degrees of hyperarticulation, but also in the visual domain, i.e., in changes in overall facial movement. In particular, the more problematic a fragment is perceived, the more likely it has more dynamically changing auditory and visual correlates. As one would expect, there are also clear correspondences between audiovisual features and actual problem status. In particular, the *combination* of visual features is a good cue for errors. The current experiment does not allow us to determine which combinations of audiovisual features are particularly relevant for error detection, since we did not have sufficient datapoints to get full insight into possible interaction effects.

On the level of individual features, one interesting finding is that different features are relevant for the different experiments. For example, frowning and repeated head gestures played a significant part in the first experiment, but had little or no effects in the third experiment. One possible explanation for this might be that in the first experiment the user listens or responds to a verification question, and thus might become *aware* of a communication problem. The stimuli in the first experiment consist of user's feedback reactions to these system verifications, and users may show surprise (frown) or may (dis-)confirm the recognized information using head nodding or shaking. In the third experiment, by contrast, the users *respond* to a question from the system to provide a station name. This could be a correction, in which case hyperarticulation is an important cue. This implies that a system that uses audiovisual cues for the detection of errors should look for different (combinations of) cues depending on contextual information, such as the most recent system question.

Another thing worth observing is that for nearly all individual features, the marked feature setting is associated with problems. This is perhaps surprising since many of these features are multi-interpretable. Smiling is a good example. In the current experiment, smiling, perhaps counterintu-

itively, showed a positive correlation with the perception of problems. Fridlund (1993, pp. 152–155) describes an experiment of Kraut and Johnston (1979), where bowlers' facial displays were analyzed after the play. The bowlers smiled more while facing friends then when looking at the pins, but also smiled less when they had a bad play. This suggests that a negative emotional stimulus influences smiling. In the current experiment, the speaker smiled regularly (in 25 of the 96 film fragments, 26%). However, their smiling suggested problematic interactions (17 out of the 25 fragments). A possible explanation is that there seemed to be a lot of user frustration. The smiling could have been an expression of disbelief (about the capacities of the speech recognition system). The smiling functions thus as a meta-gesture, making comments about the discourse (Kendon, 2001). In that case, the smiling might have been accompanied by other expressions as raising one's brows or frowning, resulting in a so-called blend emotion (Ekman and Friesen, 1975). As mentioned above, the feature frowning also had a significant correlation with the perception of problems. However, it is not clear what kind of problems the frown indicates. It is possible that it reflects the state of the discourse (the speech recognition system may just have misunderstood the speaker), but it could also reflect memory problems. It would be interesting to investigate in future studies whether the frown is the reflection of the inner state (memory overflow), or serves as a discourse signal (misunderstanding problems).

While seven of the eight features were purely labeled on a visual basis, hyperarticulation was not. It would be interesting to see whether hyperarticulation can also be detected visually. It seems likely that it is indeed visible in the articulatory region. But perhaps other visual cues correlate with hyperarticulation as well. It has been pointed out, for instance, that eyebrow movements are associated with accentuation (and thus perhaps with hyperarticulation as well). The current (limited amount of) data do not support this hypothesis. There are raised brows in 8 of the 40 fragments in which hyperarticulation occurs (on a total of 16 raised brows), while raised brows occur in 4 of the 20 non-hyperarticulated fragments (with exclusion of 4 raised brows in study 1, as hyperarticulation was there not possible). [3]

3. 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 classifications varies across the three studies. As it turned out, 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 (the mean number of marked visual cues was three for the system questions, as opposed to two in the other two studies).

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 labelled the stimuli in terms of a detailed coding scheme, comparable with

[3] In the mean time, we have replicated the three experiments in a vision-only setting (i.e., subjects could not hear the speech from system or user) which confirmed our global finding that visual information has cue value to observers about problems in a spoken dialogue system, though the results are somewhat less clear than in the vision + sound experiments reported here. A paper with details about these additional results is currently in preparation.

(part of) the FACS system (Ekman and Friesen, 1975). It was found that, in general, each of the features had a significant effect on whether an utterance is perceived as problematic or not. The presence of a marked setting leads to a higher proportion of problem perceptions, with the exceptions of (1) final mouth opening, which, when present, has a higher relative number of non-problem classifications and (2) diverted head position, which did not have an overall influence on problem perception. In addition, *combinations* of marked feature settings are better indicators of problems than single features in isolation; more problematic fragments get more extreme values both in terms of visual variation and in terms of hyperarticulation. Similarly, the marked feature settings also occur to a larger degree in *actual* problems, though some of the findings, due to fewer datapoints, represent trends rather than real significant effects.

On the basis of these results, we believe that visual information may provide a useful source for error detection, next to existing sources such as linguistic and prosodic cues. In future research, we would like to experiment with (semi-)automatic procedures to detect audiovisual cues in recordings, for instance on the basis of automatic measurements of the amount of movement and visual variation in a clip, which is potentially useful to distinguish neutral from more dynamic faces. We conjecture that such automatic facial tracking 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 and Klein, 2002) or to use them as a resource for error detection.

Acknowledgements

This research was conducted as part of the VIDI-project "Functions Of Audiovisual Prosody" (FOAP) (see also foap.uvt.nl), sponsored by the Netherlands Organization for Scientific Research (NWO). Marc Swerts is also affiliated with Antwerp University and with the Fund for Scientific Research-Flanders (FWO-Flanders). We thank Lennard van de Laar (Tilburg University) for invaluable technical assistance.

References

Ahrenberg, L., Jönsson, A., Thure, A., 1993. Customizing interaction for natural language interfaces. Workshop on Pragmatics in Dialogue, XIVth Scandinavian Conf. of Linguistics and the VIIIth Conf. of Nordic and General Linguistics, Göteborg, 1993.

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T., 1998. Synthetic faces as a lipreading support. In: Proc. Internat. Conf. on Spoken Language Processing, Sydney, Australia.

Benoit, C., Martin, J.-C., Pelachaud, C., Schomaker, L., Suhm, B., 2000. Audio-visual and multimodal speech systems. In: Gibbon, D., Mertens, I., Moore, R. (Eds.), Handbook of Standards and Resources for Spoken Language Systems. Kluwer Academic Publishers.

Brennan, S.E., Williams, M., 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. J. Memory Lang. 34, 383–398.

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: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing, Phoenix, USA, vol. 1, pp. 493–496.

Carpenter, P., Jin, C., Wilson, D., Zhang, R., Bohus, D., Rudnicky, A., 2001. Is this conversation on track? In: Proc. Eurospeech 2001, Aalborg, Denmark, September 2001, pp. 2121–2124.

Danieli, M., 1996. On the use of expectation for detecting and repairing human–machine miscommunication. Paper presented at the AAAI Workshop on Detecting, Repairing and Preventing Human–Machine Miscommunication, Portland, OR.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A., Schwartz, J.-L., 2003. Audiovisual perception of contrastive focus in French. In: Proc. ISCA Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP), 2003, St-Jorioz, France, September 47, 2003, pp. 245–250.

Doherty-Sneddon, G., Bonner, L., Bruce, V., 2001. Cognitive demands of face monitoring: Evidence for visuospatial overload. Memory Cognition 29, 909–919.

Ekman, P., Friesen, W.V., 1975. Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions. Prentice Hall, Englewood Cliffs, NJ.

Ekman, P., Friesen, W.V., 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Psychologists Press, Palo Alto, CA.

Erickson, D., Fujimura, O., Pardo, B., 1998. Articulatory correlates of prosodic control: Emotion versus emphasis. Language and Speech; Special Issue on Prosody and Conversation 41, 399–417.

Fridlund, A.J., 1993. Human Facial Expression. An Evolutionary View. Academic Press, San Diego, CA.

Gagné, J.-P., Rochette, A.-J., Charest, M., 2004. Auditory, visual and audiovisual clear speech. Speech Comm. 37, 213–230.

Goldberg, J., Ostendorf, M., Kirchhoff, K. 2003. The impact of response wording in error correction subdialogs. In: ISCA Workshop on Error Handling in Spoken Dialogue Systems, Château-d'Oex, Switzerland, August 2003, pp. 101–106.

Granström, B., House, D., Swerts, M., 2002. Multimodal feedback cues in human–machine interactions. In: Proc. Speech Prosody 2002 Conf., Aix-en-Provence, pp. 347–350.

Hart, J.T., 1965. Memory and the feeling-of-knowing experience. J. Educat. Psychol. 56, 208–216.

Hirschberg, J., Litman, D., Swerts, M., 2001. Identifying user corrections automatically in spoken dialogue systems. In: Proc. NAACL-01.

Hirschberg, J., Litman, D., Swerts, M., 2004. Prosodic and other cues to speech recognition failures. Speech Comm. 43, 155–175.

Jordan, T.R., Sergeant, P., 2000. Effects of distance on visual and audiovisual speech recognition. Lang. Speech 43, 107–124.

Kendon, A., 2001. Gesture as a communication strategy. Semiotica 35, 191–210.

Krahmer, E., Swerts, M., Theune, M., Weegels, M., 2002. The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. Speech Comm. 36, 133–145.

Kraut, R.E., Johnston, R.E., 1979. Social and emotional messages of smiling: An ethological approach. J. Personality Social Psychol. 37, 1539–1553.

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. Trento, Italy, pp. 1–15.

Levow, G.A., 2002. Adaptations in spoken corrections: Implications for models of conversational speech. Speech Comm. 36, 147–163.

Litman, D.J., Hirschberg, J.B., Swerts, M., 2001. Predicting automatic speech recognition performance using prosodic cues. Paper presented at the NAACL-01.

Nakano, M., Hazen, T.J., 2003. Using untranscribed user utterances for improving language models based on confidence scoring. In: Proc. Eurospeech 2003, Geneva, Switzerland, September 2003, pp. 417–420.

Oviatt, S., MacEachern, M., Levow, G.-A., 1998. Predicting hyperarticulate speech during human–computer error resolution. Speech Comm. 24, 1–23.

Petajan, E., 1985. Automatic lipreading to enhance speech recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 40–47.

Picard, R., Klein, J., 2002. Computers that recognise and respond to user emotion: Theoretical and practical implications. Interact. Comput. 14 (2), 141–169.

Sengers, P., 1999. Designing comprehensible agents. In: Sixteenth Internat. Joint Conf. of Artificial Intelligence, Stockhom, Sweden.

Smith, V.L., Clark, H.H., 1993. On the course of answering questions. J. Memory Lang. 32, 25–38.

Swerts, M., Krahmer, E., Barkhuysen, P., van de Laar, L., 2003. Audiovisual cues to uncertainty. In: Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems, Château-d'Oex, Switzerland, 2003.

Wade, E., Shriberg, E.E., Price, P.J., 1992. User behaviors affecting speech recognition. In: Proc. 2nd Internat. Conf. on Spoken Language Processing, Banff, Alberta, Canada, pp. 995–998.

Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A., 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. Comput. Speech Lang. 12, 317–347.