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