Repeats, Reformulations, and Emotional Speech: Evidence for the Design of Human−Computer Speech Interfaces

Kerstin Fischer
University of Hamburg

1. INTRODUCTION: THE PROBLEM

At human−computer speech interfaces, irritations caused by system malfunctions cannot be completely avoided. These irritations are not just local problems which can be easily overcome; they constitute severe problems for human−to−computer communication in at least two ways:

Firstly, the acoustic characteristics of the users' utterances have been found to be very different if they constitute repetitions or reformulations of previous utterances. That is, if the system claims not to have understood a contribution by the speaker, the speaker will repeat his utterance, however, usually with a different stress pattern, different phrasal intonation, with a strong emphasis on exact pronunciation or even hyper−articulation, and short pauses between the words. Some of these properties may cause severe problems for current automatic speech processing systems; for instance, Levow (1998) describes that the error rate in speech recognition rises from 16% to 44% for repetitions. That means that the characteristics of an utterance are very different if it constitutes a repetition of a previous contribution, and that these differences cannot be neglected in human−computer interaction (HCI).

The second problem concerns the fact that speakers may become emotionally involved while working with an automatic speech processing system such that the system's malfunctions may provoke emotional responses in the user. Thus, the speakers' attitude towards the system may change during time. This change in attitude may have global consequences on the prosodic, lexical, and conversational properties of the speakers' utterances. For instance, the average pitch may rise, the local properties as the above may occur also when no irritation directly precedes the current utterance, people may start talking to themselves, and words (e.g. four−letter words) may be used the system has not been trained for. Huber et al. (1998) have shown that if a speech recognizer was trained on normal speech and tested on emotional speech or vice versa, the speech recognition rate decreases significantly. Like the local changes observed in direct reaction to system malfunctions, these linguistic properties

(prosodic, lexical, and conversational) thus constitute great problems for current automatic speech processing systems which need to be addressed if HCI speech interfaces are to be successful. This paper will show which irritations can be found in reaction to system malfunction and how these can be addressed.

2. METHOD

In order to get data for the analysis of these speaker reactions, a corpus has been designed especially to provoke reactions to probable system malfunctions. In that scenario, the speakers are confronted with a fixed pattern of (simulated) system output which consists of sequences of acts, such as messages of non−understanding or insufficient perception, rejections of proposals, which are repeated in a fixed order. This allows to compare the speakers' behaviour through time and therefore to analyse their strategies in repetitions, reformulations, and in situations of emotional involvement. For instance, in the dialogues a sequence of a rejection of a date, a misunderstanding and a request to propose a date occurs three times in each dialogue and allows to compare how the speaker's reactions to the system's utterances change through time. After it is clear how speakers react in general, it can be experimented with certain de−escalation strategies. Thus it is possible to initiate clarification dialogues if the system encounters problems with the user's speech, or to generate utterances which may possibly calm down an angry user. Therefore, the fixed dialogue structure not only allows to control for local and global changes in speaker behaviour, but also to experiment regarding the influence of speaker behavior by varying the system output systematically. The dialogues are finally analysed regarding their lexical, conversational, and prosodic properties.

3. LOCAL AND GLOBAL REACTIONS TO SYSTEM MALFUNCTION

One obvious effect of speakers' getting angry is their use of vocabulary which expresses an evaluation of the system. In the corpus described above, this concerns four−letter−words and, for instance, the interjection *hm* which indicates dissatisfaction and divergence. However, more often the speakers become ironical, using vocabulary like *brilliant* or *very interesting*, which do not belong to standard word lists, either. Furthermore, metalinguistic vocabulary is frequently used, such as *this is not a proposal*, or *what I mean is...*. Designing a system on the basis of human−to−human communication (HHC) causes wrong predictions on the probability of these lexical items if they are included in the system's lexicon at all.

Regarding discourse strategies, the dialogues recorded diverge from natural

conversation in a number of ways; a property which may cause problems for current speech processing systems is the use of greeting acts in the middle of the dialogue, possibly to attempt a restart of the system, which is in contradiction with any dialogue model. Furthermore, speakers' metalinguistic statements of what they said and not said may include aspects which in natural dialogues do not occur and which may not have been accounted for in the linguistic models of the domain.

Finally, the acoustic properties of utterances have been found to change considerably if speakers have to repeat their utterances; if they are getting angry, even more changes occur. These changes can be attributed to attempts to make understanding easier for the automatic speech processing system, for example, by hyper−articulation, or to the changing attitude towards the system when the system is unexpectedly unsuccessful. The acoustic properties of repeated, reformulated and emotional speech include the lengthening of syllables, e.g. *mo:::nday*, increasing loudness, hyper−articulation, pausing between syllables and disfluent speech, systematic variation of stress, and the occurrence of laughter, sighing, and audible breathing.

4. ADDRESSING THE PROBLEM

Designing HCI systems always involves finding a way between adapting the system to the users' habitual verbal behaviour and imposing specific requirements on the user (Ogden & Bernick 1996); it is constrained by current technological possibilities on the one hand and by the adaptability (and willingness to adapt) of the speaker on the other. Likewise there are two possibilities to approach the problems described above: On the one hand, the system can be adapted to understand in spite of the properties characteristic of repetitions, corrections and emotional speech. Preparing the system for these peculiarities may mean to train a speech recognizer on Wizard−of−Oz data elicited under conditions similar to those described above; alternatively, it may be investigated what exactely the linguistic features are which differ from the normal training data and how recognition and processing can be adapted to be capable of dealing with these changes. Likewise, lexicon and other linguistic knowledge resources like a dialogue model need to be constituted on the basis of Wizard−of−Oz data.

On the other hand, methods may be developed to prevent these irregularities. This may be done in two ways: firstly speakers could be taught to behave in a particular way; this may include instructions regarding syntax, vocabulary and certain conversational strategies (Ogden & Bernick 1996). However, there may still be aspects which are hardly controllable, and the usability of such a system may be affected since speakers need to be trained before they can use the system, which is not suitable, for instance, for telephone use. Furthermore, it is not unlikely that instructing the user about the system's restrictions may

even trigger behaviour such as hyper−articulation and syllable−lengthening.

Besides explicitely instructing the speakers, preventing the occurrence of the above peculiarities can also be attempted by subtly guiding the speakers' behaviour and influencing their attitude towards the system, preventing them from getting angry. Here it may be useful to see which strategies speakers in HHC employ to ensure a harmonious flow of information and how these can be applied to HCI design.

5. USING DE−ESCALATION STRATEGIES FROM HUMAN−TO−HUMAN COMMUNICATION

Speakers devote almost every tenth word in natural conversation to anchor their utterances in the communicative situation (Fischer 1998). Thus speakers constantly provide feedback for their partners on the one side and make sure that their partners have understood on the other by means of discourse particles, tag questions, and speech routines. Furthermore, speakers display their understanding of the other's turn to each other (Sacks et al. 1974). Normally, they also make sure themselves that their utterances are understandable, for instance by self−initiating repair (Schegloff et al. 1977). If problems occur, such that the speaker has to reject a proposal, these possibly face−threatening acts are presented very carefully and are usually accompanied with accounts of this behaviour (Brown and Levinson 1987). Consequently, misunderstandings, caused by, for instance, recognition errors, which are very frequent in HCI, are very rare in HHC. The transfer of some of these practices to HCI design, however, is not trivial; for instance, the employment of discourse particles in system output demands not only an explicit description of their use in spontaneous spoken language dialogues and a system which can process the relevant higher level dialogue information, but also a speech synthesizer which is capable of generating the appropriate intonation contours for these lexical items. Furthermore, it is questionable whether signalling perception and understanding by the system is useful if it actually has not understood. In contrast, direct explicit accounts, such as explanations of the system's malfunctions, are more straightforwardly employed and only presuppose the recognition of critical situations. For instance, comparable to speakers' accounts of rejections in conversation, if the system detects changes in speaker behaviour which may be caused by a changing attitude towards the system, it may calm down the user by explaining its shortcomings or apologize for them. It may also be very effective to sum up the current state of the discussion, but this again requires elaborate capabilities including dialogue memory and the ability to compare the current state to a projected goal. Consequently, considering the difference between HHC and HCI, the least costly and most efficient way to influence the speakers's attitude towards the system, which also has been proven to be effective in the

Wizard−of−Oz experiments, may be to use explicit accounts which can be previously generated and which only require that the system recognizes situations which make such accounts necessary.

6. CONCLUSION

Three ways of addressing the problems caused by speakers' reactions to system malfunction which are manifest in repetitions, reformulations, and emotional reactions have been presented: Automatic speech processing systems can be adapted to the peculiarities of utterances of this type, speakers can be explicitely instructed, and they can be subtly guided by means of features of natural conversation. For the latter alternative it was examined in how far strategies from HHC can be useful for HCI design, how costly an implementation would be and what could be gained. It can be concluded that while it is often impractiable to instruct speakers before they begin their interaction with an automatic speech processing system, both alternatives, the adaptation of the system to real conversational conditions such as reformulations and emotional involvement and the employment of some less costly strategies speakers use in HHC should both be followed.

REFERENCES

Brown, P. & Levinson, S. (1987). Politeness. Some Universals in Language Usage. 2nd edition, Cambridge: Cambridge University Press.

Fischer, K. (1998). A Cognitive Lexical Pragmatic Approach to the Functional Polysemy of Discourse Particles. PhD Thesis, University of Bielefeld.

Levow, G.−A. (1998). Characterizing and Recognizing Spoken Corrections. *Proceedings of Coling/ACL '98*, Montreal, Canada.

Huber, R., Nöth, E., Batliner, A., Buckow, J., Warncke, V., and Niemann, H. (1998). You BEEP Machine − Emotion in Automatic Speech Understanding Systems. Proceedings of TDS '99, Brno, Czech Republik, pp. 223−228. Masaryk University Press.

Ogden, W.C. & Bernick, P. (1996). Using Natural Language Interfaces. In: Helander, M. (ed.). *Handbook of Human−Computer Interaction*. Amsterdam: Elsevier.

Sacks, H., Schegloff, E. & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn−Taking for Conversation. Language, 50 (4), 696−735.

Schegloff, E., Jefferson, G. & Sacks, H. (1977). The Preference for Self−Correction in the Organization of Repair in Conversation. Language, 53, 361−382.