What Makes Speakers Angry in Human-Computer Conversation

Kerstin Fischer and Anton Batliner

Abstract

Often, it cannot be completely avoided that current human-computer conversation systems function in a way that is dissatisfactory for the user. In this paper it is investigated what exactly it is that makes speakers angry and how their linguistic behaviour may change globally, in accordance with their changing speaker attitude, and locally, in reaction to particular system malfunctions. The prosodic peculiarities of the speakers' utterances can serve as indicators for the amount of problems a particular type of system malfunction may create. They can also serve to show which types of interventions by system designers can be useful.[1]

1 Problem

Human-computer conversation systems do not always work as they should. The problem which arises is that if speakers are repeatedly confronted with system malfunctions, the properties of their speech may differ considerably from what normal human-computer conversation systems have been trained with (Levow 1998). The reasons may be that speakers employ local error-resolution strategies (Oviatt et al. 1998) or that their attitude towards the system changes globally which may cause their linguistic behaviour to vary considerably. The current study addresses the question of what exactly makes speakers angry and how such situations can be avoided.

The types of system behaviour that are simulated in order to investigate the speakers' reactions to them are the following:

- rejections of proposals, for instance: "*this date is already occupied.*"
- misunderstandings, for instance: "*an appointment at 4 in the morning is not possible.*"
- failed understanding, for instance: "*I did not understand.*"
- generation errors, for instance: "*bla appointment was soll date*?"
- varying processing time, for instance, pauses of 30 seconds;
- instructions by the system, for instance: "*please concentrate*," or "*please speak more clearly.*"

2 Method

In this investigation, speakers' reactions to the above system malfunctions are

[1]The research for this paper was supported by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil project under grant number 01 IV 701 F7. The responsibility for the contents lies with the author.

first elicited and then analysed. The focus of the analysis is on those prosodic properties that may constitute a problem for automatic speech processing if this kind of deviant language is not considered in the training of automatic speech processing systems. Thus, they may render an already problematic conversation even more problematic. In particular, the speakers' reactions to (simulated) system's output such as failed understanding, misinterpretation or rejection are investigated with respect to features like syllable lengthening, pause inclusion, and hyperarticulation. This method allows to constitute a typology of system malfunctions according to the speakers' reactions to them. In a second step, methods how to avoid an increase in anger, as evidenced in the prosodic peculiarities of the utterances under consideration, for human-computer conversation system designers will be discussed, and an implemented emotion recognizer will be presented which allows to initiate compensating strategies.

3 Data

A corpus has been designed especially to provoke reactions to probable system malfunctions. The speakers are confronted with a fixed pattern of (simulated) system output which consists of sequences of acts, such as messages of failed understanding and rejections of proposals, which are repeated in a fixed order. For instance, in the dialogues a sequence composed of a rejection of a date, a misunderstanding and a request to propose a date occurs three times in each dialogue. The impression the speakers get is that they are talking to a system which does not understand them very well. The uncooperative dialogues according to the fixed schema are preceded by a phase of approximately 20 turns (phase 0) in which the system is cooperative and, by using the same utterances as in the main dialogue, reacts relevantly to the speakers' utterances. The procedure to use a fixed schema of prefabricated system utterances allows to compare how each speaker's reactions to particular types of system malfunctions change over time. It also allows to compare the speakers' use of language interpersonally. As an example[2] consider the speaker's changing reaction to the system's statement that the vacation time is from the 15th of June to the 20th of July while the speakers' task is to schedule appointments in January. This system utterance occurs in three different phases of the dialogues:

(1) s0582202: die Urlaubszeit ist fünfzehnten Juni bis zwanzigsten Juli. [*vacation time is from 15th of June to 20th of July.*]

e0582202: ja, das hat ja auch nicht viel damit zu tun, da wir uns im Januar befinden, ne? [*yes, and this has not much to do with the fact that we are talking about January, has it?*]

(2) s0584102: die Urlaubszeit ist fünfzehnten Juni bis zwanzigsten Juli. [*vacation time is from 15th of June to 20th of July.*]

e0584102: <B> ja, klasse. <P> Dienstag, zwölfter erster, achtzehn *3 bis zweiundzwanzig *2 Uhr *2. [*<B> yes, great. <P> Tuesday the 12th of January, 6 to 10pm.*]

[2]Transcription conventions are <B> = breathing, <P> = pause, *2 = hyperclear speech, *3 = strong emphasis, *4 = pauses between words, *5 = very strong emphasis, *6 = pauses inside words, *7 = syllable lengthening, *8 = hyperarticulation (with phoneme changes), *9 = speech distorted by laughter.

(3) s0587102: die Urlaubszeit ist fünfzehnten Juni bis zwanzigsten Juli. [*vacation time is from 15th of June to 20th of July.*]

e0587102: dich sollte man feuern. <B> sechster *4 Januar *4, <P> zwanzig *2 bis zweiundzwanzig Uhr. [*you should be thrown out. <B> 6th of January, <P> 8 to 10pm*]

Since these changes in linguistic behaviour occur although nothing else in the situation changes, they are interpreted as changes in the speakers' attitude towards the system, i.e. as increasing anger. This is supported by results from the questionnaire speakers fill out after the recording. So far, all participants stated that they have been emotionally engaged. All but five speakers, who have found the interaction with the simulated system amusing, report to have been angry during the recording. The data considered for this study are 36 dialogues of approximately 25-30 minutes length each, which were transcribed and lexically, conversationally, and prosodically annotated (Fischer 1999a). There were 19 female and 17 male speakers whose age ranges from 17 to 61 years.

4 Results

There are two types of results regarding the speakers' reactions to different classes of system malfunctions: global, dependent on changes in speaker attitude, and local changes of linguistic behaviour, dependent on the type of system malfunction.

4.1 Global Changes in Speaker Behaviour

There is a global development throughout the dialogues such that the prosodic deviations of the utterances increase in the course of time. This is true of the prosodic properties of utterances in general which change during the unfolding dialogues, and of conversational strategies such as reformulations and repetitions, which are distributed differently among the dialogue phases. This global development in prosodic properties is exemplified in figure 1 (χ^2-test: $p<0.001$). It may be argued that speakers employ these strategies locally as procedures to increase understandability, yet there are also global changes with respect to these strategies if an increase in understandability is not locally relevant. For instance, while after being rejected there is no need for particular procedures which support the understandability of one's utterances, the number of turns containing prosodic peculiarities increases after some interaction with the system and decreases slightly again; this trend, which can also be found in reaction to other malfunctions, can be attributed to a change in speaker attitude such that speakers become angry after some time and give up later.

4.2 Speaker Behaviour Dependent on Types of Malfunctions

Besides global changes in the speakers' linguistic behaviour, there are also differences with respect to individual system malfunctions. For instance, speakers' reactions to misunderstandings, claims of complete failure to understand, and rejections differ regarding their prosodic realization. A local error resolution strat-

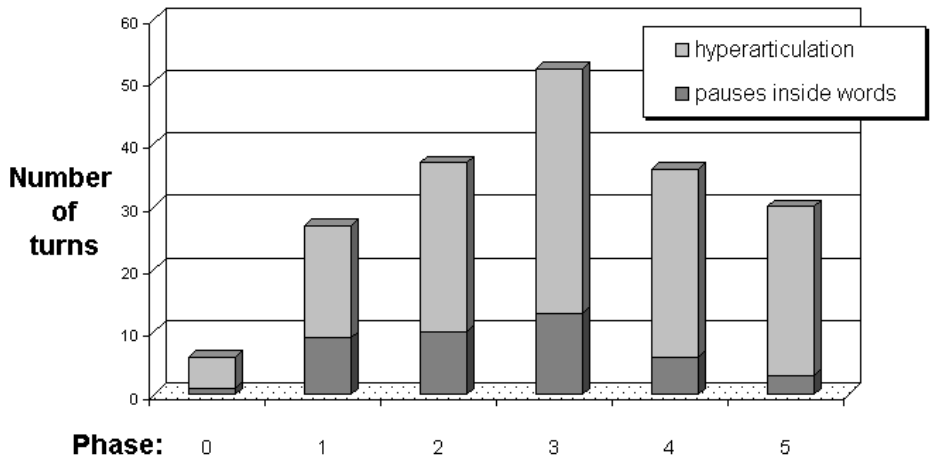

Figure 1: Hyperarticulation and Pauses inside Words in Different Phases of the Dialogues (Approx. 20 Turns per Phase)

egy such as a repetition (cf. Oviatt et al. 1998) in reaction to the system's utterance of complete failure to understand produces an increase of prosodic peculiarities irrespective of the speakers' attitude. Thus, speakers react by means of particular error resolution strategies when confronted with rejections, misunderstandings, or failures to understand already for the first time (.65 vs. .94 vs. 1.18 average number of different prosodic peculiarities per turn for the first occurrence of a rejection, misunderstanding, and failure to understand respectively). In contrast, in later phases, when speakers have given up using conversational strategies such as reformulations and metalanguage as reactions to misunderstandings, strategies, which normally contain only few prosodic peculiarities, the prosodic peculiarities observable outrange those found in reaction to complete failures to understand. The different system malfunctions can thus be distinguished according to their effects on the speaker in time and consequently also according to the degree to which they may be problematic for human-computer conversation systems. Reactions to instructions by the system will be discussed in section 5.

5 Avoidance Strategies

The different types of malfunctions can be classified according to the problems they cause for human-computer conversation; for instance, while rejections have been found to be principally unproblematic, misinterpretations and complete recognition failures by the system should be particularly avoided because of the difficult to process prosodic peculiarities by means of which speakers react to them. In any case, however, long term effects such as the speakers' emotionality have to be taken into account; therefore, unless a human-computer conversation system can avoid misunderstandings and failures to understand completely, there is a need to identify when the speaker is angry. This can be done by means of an automatic classifier which has been trained on the above data. This classifier does not only rely on prosodic prop-

erties which are accounted for by means of 27 prosodic features (cf. also Huber et al. 1998) that model logarithmic F0, energy and durational aspects, it also includes conversational information such as the detection of repetitions and prospectively also other forms of trouble in communication. Using the combined automatic classifier MoUSE, 90% recall and 70% precision could be achieved, while prosody alone yielded 56% precision and 84% recall on the current data (cf. Batliner et al. forthcoming).

Once the speakers' anger is recognized by the classifier, a number of different forms of intervention are possible. For instance, the system may ask the speaker to speak more clearly:

(4) e0605104: Montag achtzehnter <P> Januar von acht bis zehn Uhr? [*Monday 18th <P> of January from 8 to 10 am?*]

s0605201: bitte sprechen Sie deutlicher. [*please speak more clearly*]

e0605201: Montag *7 achtzehnter *4 <P> Januar von acht *3 bis zehn *7 Uhr. [*Monday 18th <P> of January from 8 to 10 am.*]

However, as the example shows, the prosodic peculiarities in the speaker's utterance even increase after the request to speak more clearly. That is, what the speaker believes to be helpful for the system's understanding, to pause between words, to emphasise and to lengthen particular syllables, may even make her utterance more difficult to process. An alternative may be to ask the speaker to speak clearly but not hyper-clearly:

(5) e0605206: ich hätte gerne einen zweistündigen *3 Termin <P> am Freitag dem zweiundzwanzigsten *3 Januar <P> von acht *3 bis zehn Uhr [*I'd like to have an appointment <P> on Friday January 22nd <P> from 8 to 10 am.*]

s0605207: bitte sprechen Sie deutlicher, aber nicht überdeutlich. [*please speak more clearly but not hyperclear*]

e0605207: <;<ich spreche bereits deutlich> laugh>. Ich hätte gern einen Termin *7 <P> am Freitag dem zweiundzwanzigsten Januar von acht bis zehn Uhr [*<;<I'm speaking clearly already> laugh>. I'd like to have an appointment <P> on Friday January 22nd from 8 to 10 am.*]

The effect is not much different from the request to speak more clearly, since the speaker employs syllable lengthening and pausing, besides the fact that parts of her turn are distorted by laughter. Furthermore, as the example shows, speakers may react unwillingly when they are asked to change the linguistic properties of their utterances. In contrast, it has turned out that an extremely simple and very successful strategy is to have the system apologize (Fischer 1999b); the number of prosodic peculiarities of utterances decreases immediately and remains low over a number of turns:

(6) s0375102: bitte machen Sie einen Vorschlag. [*please make a proposal*]

e0375102: ja, mach' ich doch die ganze Zeit schon. also, <Laugh> Dienstag *9, der zwölfte erste, ab *4 <B> achtzehn *2 Uhr *2, zwei *3 Stunden *2. [*yes, that what I've been doing the whole time. well <Laugh> Tuesday the 12th, from <B> 6 pm, two hours*]

s0375103: es tut mir leid, daß ich Sie so schlecht verstehe. Wären Sie so freundlich, Ihren Beitrag noch einmal zu wiederholen? [*I'm sorry that I understand you so badly. would you be so kind to repeat your utterance?*]

e0375103: so. wie wäre es am Dienstag, dem zwölften ersten neunzehnhundertneunundneunzig, ab achtzehn Uhr, nachmittags, mitteleuropäischer Ortszeit? [well. how about Tuesday the 12th, 1999, from 6 pm, in the afternoon, middeleuropean time?]

Comparing the speakers' reactions to the last recognition failure of the system (in turn 4302) with their linguistic behaviour after the system's apology in turn 5103, it turns out that on the average the previous turn contains more than the double amount of prosodic peculiarities than turn 5103 with a comparable function, namely to repeat the previous proposal. The apology thus influences the speakers' linguistic behaviour systematically.

6 Conclusions

In this paper, the speakers' reactions to particular types of system malfunctions were analysed. While there are differences with respect to the prosodic peculiarities that occur in reaction to the different kiinds of system behaviour, there are also global changes in speaker behaviour that cannot be attributed to attempts to increase the understandability of one's utterances. For instance, even after rejections of proposed dates the speakers' linguistic behaviour changes over time, due to changes in speaker attitude. An automatic classifier is used to identify when the speaker is angry. Different types of compensating strategies were discussed; the most successful way seems to be to influence the speakers' attitude towards the system directly.

References

[1] Batliner, A., Huber, R., Nöth, E., Spilker, J., Fischer, K. (forthcoming): Desperately Seeking Emotions or: Actors, Wizards, and Human Beings. Proceedings of ISCA-Workshop on Prosody and Emotion, September 2000.

[2] Fischer, K. (1999a): Annotating Emotional Language Data. Verbmobil Report 236.

[3] Fischer, K. (1999b): Repeats, Reformulations, and Emotional Speech: Evidence for the Design of Human-Computer Speech Interfaces. In: Bullinger, H.-J. & Ziegler, J. (eds.): Human-Computer Interaction: Ergonomics and User Interfaces, Proceedings of HCI '99 International. Lawrence Erlbaum Ass., London, pp. 560-565.

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

[5] Oviatt, S., MacEachern, M., Levow, G.-A. (1998). Predicting Hyperarticulate Speech During Human-Computer Error Resolutions. Speech Communication 24: 87-110.