"I SAID TWO TI-CKETS": HOW TO TALK TO A DEAF WIZARD

Hannes Pirker *Georg Loderer*[1]

Austrian Research Institute for Artificial Intelligence (OFAI), Schottengasse 3A, A-1010 Vienna
[1]Department of Medical Cybernetics and Artificial Intelligence (IMKAI), Freyung 6, A-1010 Vienna

ETRW on Dialogue and Prosody
Veldhoven, The Netherlands
September 1-3, 1999

ABSTRACT

So-called Wizard-of-Oz (WOZ) simulations are a popular framework for investigating the nature of human machine interaction in general and for the development and evaluation of designs for spoken dialog systems in particular. In this paper a WOZ simulation of a speech based ticket reservation system is presented. In contrast to most of the studies performed in this framework we are not concerned with the evaluation of different dialogue designs. In our experiment the WOZ frequently rejected or misrecognised user utterances. The findings on the prosodic properties of repeats and corrections triggered by these recognition errors are presented. In addition some data on the lexical content of the user utterances is discussed.

1. INTRODUCTION

Wizard-of-Oz (WOZ) simulations are a popular framework in the field human computer interaction. As dialogues in human computer interaction may vary significantly from natural conversations and fully fledged spoken dialogue systems are still rare WOZ simulations can be a valuable tool for gaining insights on the peculiarities of this interaction in general and for the development and evaluation of designs for spoken dialog systems in particular [1].

Typically, WOZ experiments tend to concentrate on the latter subject, namely the evaluation of different system designs (e.g. [4], [9]). Also the WOZ experiment that produced the empirical basis for the present study was designed and performed for just this very same purpose of evaluating design options ([5],[7]).

In this paper however we will deal with the issue of analysing some aspects of the speakers' utterances. Throughout the experiment no perfect recognition was simulated by the WOZ. This resulted in a high number of correction turns. It is analyzed if and how speakers modify their utterances if they were not correctly understood by the WOZ. In [3] a decision tree was presented in order to automatically classify original utterances and two different types of error corrections on the basis of prosodic features. It will be discussed how our data conforms to the these results. Variations on the prosodic but also segmental level are surveyed in that context.

In [10] it was demonstrated how a system's verbal behaviour is able to influence the content of user utterances. We will investigate to which extent this was the case in our WOZ experiment.

2. THE WOZ EXPERIMENT

The data used for this study was gained throughout a WOZ simulation of a hypothetical reservation system for cinema tickets [5].

The participants of the experiment were asked to volunteer at the evaluation of a novel speech based reservation system. For this purpose they were placed in front of a computer terminal which was used as a dummy only. The WOZ, who was situated in another room, interacted with the users via a microphone and a loudspeaker.

All utterances of the WOZ were produced by a text-to-speech engine. In order to minimise the necessary typing and to maximise the uniformity of the system's responses a graphical user interface was used by the WOZ which included facilities for the playback of predefined token as well as templates that could be edited at run time.

Users got a written instruction on three reservation tasks (T1, T2, T3) which were identical for all users and all system variants. The tasks increased in complexity from T1 to T3 due to additional constraints on the preferred seat positions and the availability of seats in the requested positions.

I.e., while the first task was defined as follows[1]:

> **Task T1:** Film: Titanic, Time: 19:00. Reserve 2 tickets in the last row/midst.

the last task comprised additional conditions:

> **Task T3:** ... Accept seats in the middle of one of the last three rows only. Otherwise try to switch to a later show.

As the main purpose of the original study was the evaluation of dialogue strategies three different versions (S1, S2, S3) of the reservation system were simulated. These

[1]The experiment was performed in German. For the sake of convenience only English translations are quoted though.

versions varied in respect to **initiative** (S1 and S3: system initiative only vs. S2: very restricted user initiative) **confirmation strategy** (S1 and S2: direct vs. S3: indirect) and the extent to which **additional information** on free seats was actively offered to the user. The sample transcript below gives an impression of the overall 'look and feel' of the simulation.

```
System:S2 Task:T3
SYSTEM
WIZARD: Welcome to the ticket reservation system.
        How can I help you?
USER:   I would like to see the film Copland,
        at five o'clock,
        two tickets in the last row.
WIZARD: You said: the film Holland.
        You said: at five o'clock.
        Yes or no?
USER:   No.
WIZARD: Which film do you want to see?
USER:   Copland
WIZARD: Please repeat your input
USER:   Copland
WIZARD: You said: the film Copland.
        Yes or no?
...

System: S3 Task:T3
WIZARD: Welcome to the ticket reservation system.
        Which film do you want to see?
USER:   Copland
WIZARD: Please repeat your input
USER:   Copland
WIZARD: When do you want to see the film Copland?
...
```

26 subjects (19 male, 7 female) in the age from 18 to 40 with a diverse professional and educational background voluntarily participated at the experiment. This population was divided into three groups of 9/9/8 persons. Each group was assigned to one system version. Thus each subject performed the tasks T1, T2, T3 with just one system version which resulted in 77 evaluable dialogues. The dialogues were tape recorded and manually transcribed. The whole experiment comprised about three hours of interaction.

After performing the reservation task all subjects answered a questionnaire that contained rating scales as well as open questions about their impression of the system. About half of the subjects performed an additional informal interview as well.

3. PROSODY OF CORRECTIONS

Throughout our experiment a rather poor recognition rate was simulated by the WOZ. At the one hand some utterances were intentionally mis-interpreted in a consistent way. At the other hand all instances of "sloppy" or unclear pronunciation, overly conversational and fast speaking style, interjections etc. were immediately penalised

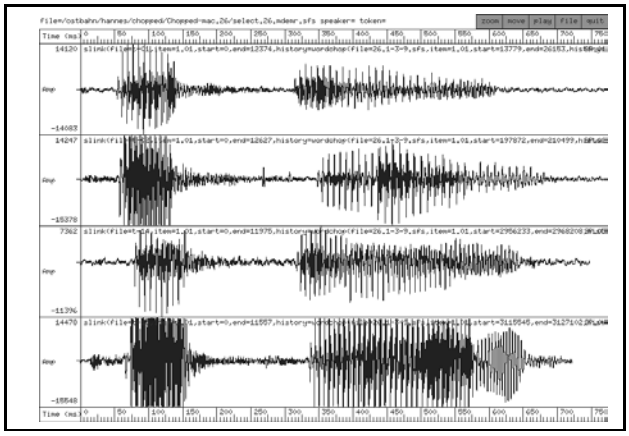

Figure 1: Increasing amplitude in corrections: Two pairs of original and repeated pronunciation of 'Copland'

by the system. In this case a "(Sorry. I did not understand.) Repeat your utterance!" request was produced. All together 137 of such explicit repeat! statements were produced which triggered *demanded repetitions* in our terminology. In 63 (46%) of these cases users responded by exactly repeating the lexical content of the original objected utterance.

3.1. Related Work

The prosodic characterisation of spoken corrections in an English dialogue system has been investigated in [3]. There a distinction between correction of rejection errors CRE (which correspond to our demanded repetitions) and correction of misrecognition errors CME was drawn and a decision tree was used in order to distinguish between these classes and original inputs on the basis of prosodic measures only. Our own observations on the German data supported many of the findings of [3].

3.2. Amplitude

Users tended to apply different means for differentiating corrections from the original utterances which usually were varied throughout the dialogue. Thus we observed many examples of increased amplitudes such as in Fig. 1.

3.3. Duration and Pausing

As for the variation of duration unsurprisingly most of the repetitions were elongated. Usually existing pauses in the original were increased in length and additional pauses were inserted in some cases. This tendency of pause stretching even applied for word internal pauses. Fig. 2 displays 8 user turns from a single dialogue which all contain the word 'Copland'. They were extracted from the dialogue which is quoted as example for S2-T3 above and thus displays both CRE and CME. The first line displays the first turn of the user. Throughout the numerous repetitions the user monotonically increased the word internal pause up to almost 2 seconds. In the restart in line 6 pauses are also inserted before and behind the word Copland. Though this is an extreme example many users

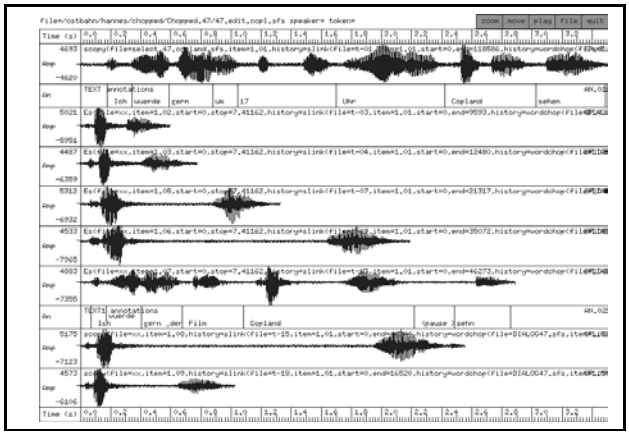

Figure 2: Examples of the pronunciation of 'Copland' throughout a dialogue. Extremely long word internal pauses are inserted.

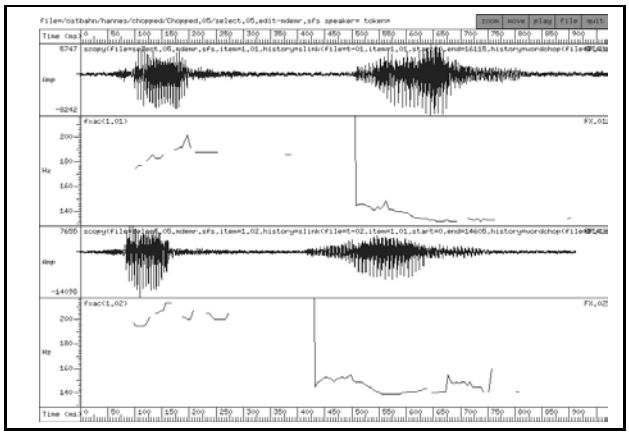

Figure 3: Increase of pitch: CRE of 'Copland'.

displayed the trend to stretch the word internal pause in 'Copland' which probably is due to both its phonological and morphological structure as this effect is not attested in a comparable degree in other words within our corpus.

3.4. Pitch

Fundamental frequency was the prosodic aspect where the most significant and interesting differences between CME and CRE were found.

Most typically CRE differ from the original utterances only by a small rise in pitch, such as in Fig. 3, where even the duration is slightly shortened in the repetition and the increased pitch bears all the load of the additional markup.

In Levow's automatic classification it was the information on the slope of rise and fall of pitch movements that had the best predictive value for distinguishing CME from CRE while the gross measure of maxima, minima, and pitch range did not prove to be a valuable measure. In addition CME displayed a tendency of increasing the slope of rises while there was a flattening effect in CRE's.

Though we did not evaluate this claim statistically this finding seems to be quite plausible and not so surprising.

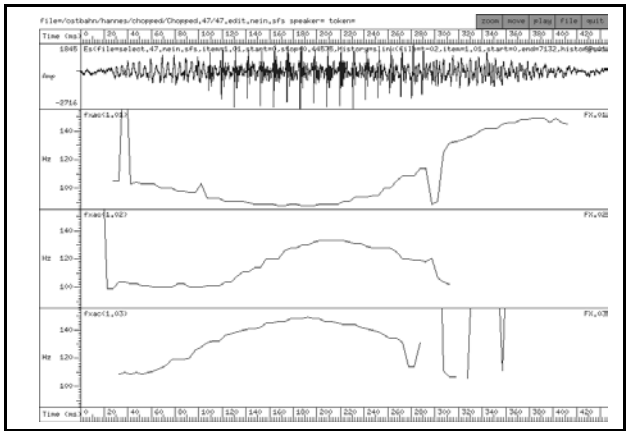

Figure 4: Alternation of pitch in error corrections: Three f0 plots for the utterance of 'no' (German 'nein' /naIn/) from one dialogue.

Comparing the three different contours in Fig. 4 it becomes clear that the form of the contour has to be somehow considered and the overall range information does not convey the necessary information.

These samples are extracted from the same dialogue as those in Fig. 2. While the first token has a rising contour frequently used by a number of speakers the other 'no' responses display the increasing irritation of the user about the system's unwillingness to accept his numerous repetitions of 'Copland'.

The changes in slope between original utterance, CME, and CRE also conform well to the differences in focus / background division in these three utterance types which are reflected in a different contribution of prominence values. If an utterance is fully rejected by the system and a CRE is requested, the user usually does not know which part of the message was not understood. If the speaker thus decides to reproduce the same content there is a tendency to produce a version which is overall better perceivable. This typically reduces the amount of deaccentuation / backgrounding. Thus the overall pitch height can be increased though the steepness and dynamics of the contour may decrease.

In the case of CME, on the other hand, often the source of the misrecognition can be identified by the user and will be foregrounded / focused by the user. The different behaviour is displayed in the following hypothetical example (capitalisation indicates prominence / stress).

```
-- CRE --
USER:    two tickets
SYSTEM:  Please repeat your input
USER:    TWO TICKETS
-- CME --
USER:    two tickets
SYSTEM:  You said three tickets?
USER:    TWO tickets
```

4. ADAPTIONS OF USER

In [10] it was demonstrated that users of natural language based systems can be strongly influenced by the system's language. They tend to adjust their utterances in respect to vocabulary and length of phrases. Thus the outcomes of the WOZ experiment inevitably will be biased to some extent by the design of the WOZ which influences the user's conception of the system's capabilities.

We performed some analysis on the basis of the transcriptions in order to check to which extent users model their behaviour throughout their ongoing interaction with the system. As most of the users did not have any prior experience with spoken dialogue systems at all we were interested to which extent users will adapt their speaking style over time.

In the simulation the WOZ used synthesised speech for his interactions. Utterances were produced with a very slow rate of speech. No pitch accents were produced, only a uniform linear declination falling from 120 Hz to 100 Hz was superimposed on the output. The slow speaking rate was even more delayed by pauses required by typing and the processing time of the synthesiser itself.

This resulted in very long delays and highly redundant system output. In a first version, e.g., the system's 'repeat!' message was even preceeded by an additional "I am Sorry. I did not understand your input.". Barge ins were not possible. Thus the production of this message required 13 seconds, comprising 5 second of pause, though the experienced user already knew the whole content when the first two words were produced.

Only one user, who clearly tried to "trick the system", started to imitate the system's unnatural intonation. It is not clear whether the decreasing speaking rate of some users was due to an adaption to the slow speed of the synthesized utterance.

4.1. Use of full sentences

In order to check the adaption of the speech mode we analyzed some of the lexical features.

With the exception of the very first turn of S2, which starts with a "How can I help you?", all the system's question are either yes-no questions or prompt for a single term such as the number of tickets.

In order to check when and how often users are using full sentences in their responses the number of user turns containing the pronoun "I" was counted.

This data shows that throughout the interaction the number of fully formulated sentences decreased, and users preferred simple phrases as their experience increased throughout the course of solving T1, T2 and T3.

Task	User Turns Total	User Turns with 'I'	%
T1	243	18	7.4
T2	274	8	2.9
T3	453	23	5.0

Counts per system of course peak at S2, nevertheless also in the other systems fully formulated answers like "I would like to reserve ..." are found.

System	User Turns Total	User Turns with 'I'	%
S1	317	7	2.2
S2	354	32	9.0
S3	302	10	3.3

4.2. Politeness Marker

Some users included a polite "please" in their utterances, which we used as another marker of a rather natural speaking style. Surprisingly the counts for these utterances, though low anyway, did not drop as sharply as expected.

Task	User Turns Total	User Turns with 'please'	%
T1	243	8	3.3
T2	274	6	2.2
T3	453	13	2.9

The analysis per system with the additional count of these polite users shows that these counts are due to a small number of users only.

System	User Turns Total	User Turns with 'please'	Users of 'please'
S1	317	2	1
S2	354	15	4
S3	302	10	2

4.3. Redundant items

When asked by the system "Who many tickets do you want to reserve?" users often answered with "Two tickets". The production of the noun "ticket" was notable not only because of its redundancy but also because of its realization. This item is highly contextually bound in the dialogue. Nevertheless many user not only failed to omit or at least deaccent it but often also overarticulated the whole word.[2]

Task	User Turns concerning tickets	User Turns with bare number only	%
T1	47	11	23.4
T2	43	17	39.5
T3	51	18	35.3

[2]The second German word 'Karten' which is usually pronounced `[ka:6tn]` was pronounced as `[ka:RtEn]` where the last syllable was stressed.

System	User Turns concerning tickets	User Turns with bare number only	%
S1	42	25	59.6
S2	62	7	11.3
S3	37	14	37.8

The tables show that the specifier "ticket" is only omitted in a minority of cases. This might be due to the effect that unnaturally long pauses within the dialogue may block the linkage to prior mentioned items. At least this effect is attested for the usage of pronouns [2].

5. DISCUSSION AND CONCLUSIONS

In this paper we surveyed data obtained for the purpose of the evaluation of different versions of a dialogue system. It mainly dealt with the prosody of corrections in such an environment.

As for the practical application of these findings [3] argued, much in parallel to [8], that the automatic identification of special utterance types (CRE and CME in this case) on the basis of a prosodic analysis can be used for triggering the selection of different recognition modules.

Using a WOZ simulation for the data acquisition phase allowed for a highly controlled task setting. E.g. in all 77 dialogues user had to order "two tickets in the last row" which facilitated the analysis of the data. On the other hand it has to be kept in mind that all users of the system were basically role playing. They are no real users with real information requirements, real time constraints or even real telephone bills. Nevertheless this laboratory effect probably only has a minor impact on the aspects of prosody described in this paper.

ACKNOWLEDGMENTS

This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P13224. Financial support for ÖFAI is provided by the Austrian Federal Ministry of Science and Transport. Thanks to Erhard Rank and Friedrich Neubarth for technical support. Two anonymous reviewers supplied valuable suggestions for this paper.

6. REFERENCES

1. Dahlbäck N., Jönnson A., Ahrenberg L.: Wizard of Oz studies - why and how, Knowledge Based Systems, Vol. 6/4 pp.258-266, 1994.
2. Guindon R.: How to Interface to Advisory Systems? Users Request Help With a Very Simple Language, Proc. of ACM Conf. on Computer Human Interaction (CHI88), pp.191-196, 1988.
3. Levow G.-A.: Characterizing and Recognizing Spoken Corrections in Human- Computer Dialogue, in Proc. of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Universite de Montreal, Canada, pp.736-742, 1998.
4. Litman D.J., Pan S.: Empirically Evaluating an Adaptable Spoken Dialogue System, Proc. of 7th International Conference on User Modeling (UM'99), Banff, Canada, 1999.
5. Loderer G.: Evaluierung von Dialogstrategien eines natürlichsprachigen Dialogsystems durch Wizard-of-Oz Experimente (Evaluation of dialogue-strategies in a natural language system using Wizard-of-Oz experiments), Institut für Med.Kybernetik und AI, Universität Wien, Masters thesis, 1998.
6. Ostendorf M., Byrne B., Bacchiani M., Finke M., Gunarwardana A., Ross K., Roweis S., Shribergand E., Talkin D., Waibel A., Wheatley B., Zeppenfeld T.: Modeling systematic variations in pronounciation via language-dependent hidden speaking mode, in Proc. of ICSLP-96, 1996.
7. Pirker H., Loderer G., Trost H.: Thus Spoke the User to the Wizard, Proc. of EUROSPEECH-99, Budapest, Hungary, 1999.
8. Shriberg E., Bates R., Stolcke A.: A Prosody Only Decision-Tree Model for Disfluency Detection, in 5th European Conference on Speech Communication and Technology (EUROSPEECH 97), Rhodes, Greece, ESCA, Vol.5,pp.2383-2386, 1997.
9. Walker M.A., Fromer J.C., Narayanan S.: Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email, in Proc. of COLING/ACL 98, Universite de Montreal, Canada, pp.1345-1351, 1998.
10. Zoltan-Ford E.: How to Get People to Say and Type What Computers Can Understand, International Journal of Man-Machine Studies, 34, 527-547, 1991.