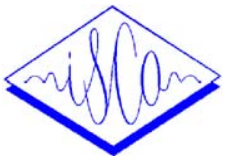

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

AN APPRECIATION STUDY OF AN ASR INQUIRY SYSTEM*

L.J.M. Rothkrantz and **W.A.Th. Manintveld** and **M.M.M. Rats**
R.J. van Vark and **J.P.M. de Vreught** and **H. Koppelaar**
Knowledge Based Systems
Technical Computer Science
Delft University of Technology
alparon@kgs.twi.tudelft.nl

Abstract Human factors play an important role in the applications of speech technology. In a Wizard of Oz experiment, 64 telephonic inquiry systems were simulated by systematic manipulation of 6 human factors. To assess the impact of those factors on the appreciation, an appreciation scale was developed based on an questionnaire.

In an experiment 414 respondents were requested to call one of the simulated systems and a system operated by human operators. The respondents rated these systems on an appreciation scale. In this paper a description of the experiment is given and the results of a statistical analysis of the appreciation scores is presented.

1 INTRODUCTION

Although in recent years, many advances have been realised in Automated Speech Processing (ASP), automated recognition of continuous speech is still far from perfect. As a result, engineers are faced with the challenge to find an optimal balance between restrictions in current speech technology and users' requirements. One of the most important, but also most ignored requirements, is the acceptance of the system [3]. Many questions still exist on how to develop optimal ASP systems for the near future.

In recent years ASP technology has been applied in telephonic inquiry systems. It was shown that optimal technical systems do not guarantee a high degree of user appreciation. For an optimal user satisfaction, we need to find out the factors which effect appreciation of ASP based systems most.

This study investigates these factors using a common application of ASP technology. The Dutch company OVR provides a telephone based inquiry system concerning public transport services in the Netherlands. Over 400 human operators handle more than 12 million calls a year, requesting information on travelling by train, bus, and ferry.

* This work is funded by OVR and Senter.

2 MODEL

Periodic survey studies showed that OVR's humanly operated inquiry system has a high degree of user appreciation. Therefore, it seems natural to base a design for an ASP system on the human dialogue model. However, current speech technology is not yet powerful enough to handle dialogues based on this complex human–human dialogue model.

As it is necessary to simplify the human model, we need to assess the impact of such simplifications on user appreciation. In a Wizard of Oz experiment [4] an ASP system for the OVR inquiry system was simulated. In a completely randomised factorial design, we manipulated six variables and assessed the degree of appreciation of the simulated system. The following six experimental modes were simulated:

- mistakes: the ASP system deliberately misinterpreted similar sounding station names (i.e. Weesp instead of Wezep).
- verification: the user was always asked for verification of filled slots.
- dialogue style: in the non–directive style the initiative was mostly left to the user (mixed–initiative), while with the directive style the system asked for each slot separately.
- break/tempo: before every operator utterance a 3 second delay was introduced, in order to simulate non–real–time operation of the system.
- voice: the sex of the simulated voice could be set to male or female.
- explicitation: in this mode the system informed the user about its current status.

3 METHOD

Subjects In total 414 respondents were recruited from the population at the Delft University of Technology, i.e. students, scientific and non–scientific

10.21437/Eurospeech.1997-487

staff. All members of three departments received a letter containing background information and an invitation to take part in the experiment. A team of 5 students selected random room numbers and invited the habitants to take part in the experiment. In 92% of the cases an appointment was made for the experimental session.

Questionnaire In order to reach a high appreciation score, an inquiry system based on speech recognition has to fulfill the following conditions [2, 4]:

- the information delivered by the system has to be correct,
- the system has to be efficient and effective, and
- the system has to be human–friendly.

A questionnaire with 21 statements on a 5–point Likert scale was designed to cover the aspects of these three topics. In the RAILTEL/MAIS project [1] a similar questionnaire was used. The items were grouped under different headings (see table 1):

- final conclusion: the user is assumed to be more satisfied if the dialogue with the ASP system results in an answer containing the correct information (i8). A highly appreciated system will be used again (i10) and there is less need to improve a highly appreciated ASP system (i17).
- communication: the user is assumed to be more satisfied if the system is more polite and friendly (i1) and if the communication (i16) and dialogue (i7) is nicely appreciated.
- user–friendlyness: the user is assumed to be more satisfied if the system uses clear instructions (i13), the system is thinking along with the user (i3), the system is not confusing (i4), not much concentration is needed (i11) and the user does not feel stressed by the system (i19).
- specific system features: the user is assumed to be more satisfied if the system is reliable (i12), makes no disturbing mistakes (i2), has a good understanding of the request (i20), asks a minimal amount of questions (i5), is flexible (i9), is fast (i6), not confusing (i15), behaves personal (i21) and an option for interruption (i14) and the system has a mode to repeat the information (i18).

The items of different aspects can be highly correlated. The questionnaire was assumed to have one underlying construct. This hypothesis was tested in [4]. The Cronbach' Alpha coefficient was 0.91. The validity and reliability were also proved to be good.

The second part of the questionnaire consisted of questions about background information and attitude towards new techniques in general and ASP as a specific case.

Dependent measures In the third part of the questionnaire, respondents have to rate the ASP system on a 10–point scale. One of the items is an overall appreciation score of the system. The reliability and validity of a one–item appreciation score cannot be high. Therefore, we defined an appreciation score based on all items of part one of the questionnaire which we assume to cover all aspects of user appreciation.

As part one of the questionnaire has one underlying construct, the sum score of all 21 items can be considered as an overall appreciation score. By using item response analysis, a weighted sum score was computed. It is known that under general conditions [5] weighted and unweighted sum scores are strongly correlated. The unweighted sum score is easy to compute and a robust measure of appreciation. So the unweighted sum score is the preferred measure of appreciation.

Wizard of Oz A Wizard of Oz experiment is an experiment in which a system is secretly simulated by a human operator. In this case the simulated system was an ASP version of OVR's travel information system. Callers are instructed to request information concerning a journey from one railway station to another. The human operator (i.e. the wizard) of the system was supplied with a standard set of prerecorded sentences and a journey planner in order to retrieve the right information to the caller. The caller was requested to call an ASP system and was not informed about the Wizard of Oz design.

Procedure In the experiment, the respondents had to inquire information concerning a certain train journey. Ten scenarios were used, which were presented non–verbally to prevent paraphrasing of the journey description.

The respondents were requested to call the ASP system and the humanly operated system using the same travelling scenario. This resulted in 414 parallel calls. After every call a questionnaire was administered by interviewers, followed by an open interview which ended the experimental sessions. Both the human–machine and the human–human dialogues were recorded and transliterated into written text.

no.	item	ASP	operator
i1	friendly impression	2.89	3.36
i2	disturbing mistakes	3.43	4.27
i3	thinking along	1.96	2.89
i4	confusing	3.73	4.30
i5	efficient	2.16	3.03
i6	slow	2.64	3.94
i7	uncomfortable	3.52	4.33
i8	info request	2.88	3.51
i9	flexibility	2.78	4.13
i10	reuse	2.20	3.05
i11	contration	3.35	3.52
i12	reliability	2.64	3.16
i13	user manual	3.84	4.24
i14	interrupt	1.92	2.55
i15	confused	3.88	4.26
i16	nice communication	1.71	3.04
i17	improvement	2.71	3.85
i18	repetition	2.69	2.70
i19	under pressure	3.95	3.92
i20	understanding	2.66	3.43
i21	impersonal	2.91	4.18

Table 1: Mean item scores

4 ASP VERSUS HUMANLY OPERATED SYSTEM

After calling both systems, part one of the questionnaire was administered. The mean item score was computed using these questionnaires (see table 1).

It can be seen that there is a significant difference in item scores. In comparison, the ASP system is too slow (i6), it needs improvement (i17) and is impersonal (i21). Also, the system makes disturbing mistakes (i2) and the dialogues with the system are quite uncomfortable (i7). The communication with the human operator is nicer (i16), the human operator supports the planning of the journey (i3), has a better understanding of the request (i20) and is much more efficient (i5). When repeatedly used, the client prefers the human operator (i10).

no.	item	ASP	operator
i1	quality	6.61	8.00
i2	flexibility	5.57	7.89
i3	speed	5.72	7.66
i4	user friendly	6.64	7.70
i5	interaction option	5.47	7.83
i6	interruption option	4.81	7.56
i7	delay/queue time	5.77	7.02
i8	dialogue style	6.19	7.69
i9	final judgement	6.10	7.78

Table 2: Mean ratings for ASP and operator

Table 2 shows that human operators outperform ASP systems by more than one point on a 10–point rating scale.

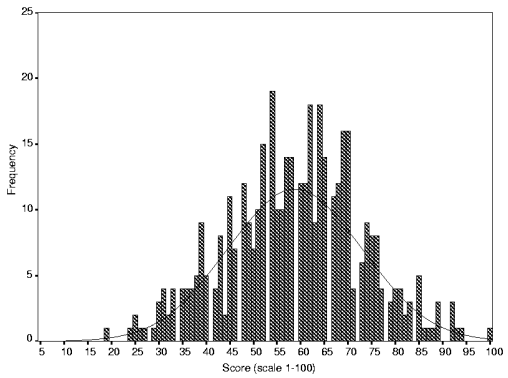

Figure 1: Distribution of ASP appreciation score

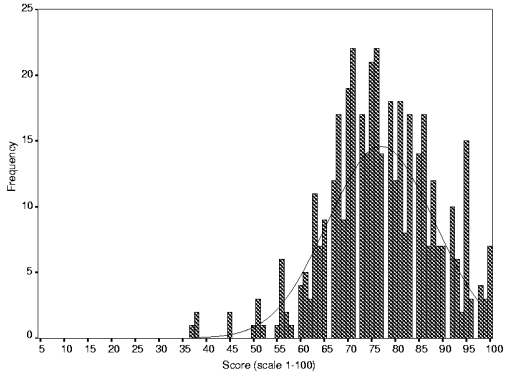

Figure 2: Distribution of OVR appreciation score

As stated before the individual item scores can be combined to an overall appreciation sum score. In figure 1 and 2 the distribution of these sum scores are displayed. Although the shape of both distributions is similar, the distribution for the human operator is shifted to the right, which indicates a significantly higher appreciation.

5 ANALYSIS OF VARIANCE

The experimental design was a completely randomised factorial design. Six dichotomised variables were manipulated, resulting in 64 cells. Unfortunately the design is not completely balanced which results in a different number of respondents per cell. The main cause of imbalance is the fact that not all variables could be totally controlled. For example, in the experimental mode mistakes, misinterpretations cannot be made, if the caller does not mention the appropriate name from the scenario. In the WOz design some parameters are set at the beginning of a recording session. However, respondents are free to choose the time to call, there is some imbalance in the distribution of the calls and as a consequence of

the corresponding mode. In the analysis of variance a special procedure has to be taken to correct this imbalance.

The H_0–hypothesis is that all the mean appreciation scores of all experimental modes are equal. First, we compute the mean appreciation score for each group to analyse the impact of the different modes on the appreciation scores (see table 3).

	score	no.	score	no.
c1	mistakes		no mistakes	
	−11.80	265	−7.64	143
c2	no verification		verification	
	−10.08	177	−10.54	231
c3	directive		non–directive	
	−8.30	233	−13.06	175
c4	break		no break	
	−11.75	196	−9.04	212
c5	male voice		female voice	
	−11.22	253	−8.91	155
c6	high tempo		low tempo	
	−10.00	204	−10.68	204
Total population			−10.34	408

Table 3: Group means for different experimental modes

To investigate if there are any systematic differences in the appreciation scores for the six manipulated variables and corresponding experimental modes an ANOVA analysis is applied (see table 4).

Source of variations	Sum of Squares	F	Significance
c1 mistakes	938.258	4.429	.009
c2 verification	1.149	.008	.927
c3 dialogue style	1370.658	10.082	.002
c4 break	32.235	.237	.627
c5 voice	49.793	.146	.703
c6 tempo	131.932	.970	.325
Main effects	3475.63	4.429	.000

Table 4: ANOVA analysis

The H_0–hypothesis has to be rejected, so there are significant differences in the mean appreciation scores for the different modes. It can be concluded that the parameters mistakes and dialogue style have significant effects on appreciation scores.

6 Conclusions

In this study a scale was developed to assess the appreciation of (non–)automated inquiry systems. This scale has a high degree of reliability and validity.

Dialogue style and quality of speech understanding have a significant impact on ASP system appreciation. As both the male voice as the female voice are equally appreciated, the appreciation is probably effected by voice quality and not by gender. Tempo of simulated systems and additional explicitations of the operators showed ambiguous results. Some people like an easy going system with additional comments, others like to speed up the system and like to reduce the number of prompts to a minimum.

The developed appreciation scale will be used to assess appreciation of new releases of OVR's ASP system. Analysis of the appreciation scores will reveal new insight in the complex relation between human factors and ASP systems.

References

[1] M. Blasband and A. Puccioni, Definition of the Evaluation Methodology for the Field Trials, in *Selected Publications, RAILTEL/MAIS Project (ESPRIT)*, 1993–1995.

[2] H. Dybkjaer, L. Dybkjaer, and N.O. Bernsen, Design, formalization and evaluation of spoken language dialogue, in *Corpus–based Approaches to Dialogue Modelling*, 9th Twente Workshop on Language Technology, pp. 67–82, University of Twente, 1995.

[3] D. Krause, Social Research in Context of Speech Systems. The Case of VERBMOBIL, in *Workshop Dialogue Processing in Spoken Language Systems*, European Conference on Artificial Intelligence, pp. 49–53, 1996.

[4] W.A.Th. Manintveld and L.J.M. Rothkrantz, The OVR–WOz experiment: Setup and Analysis, Technical Report 97–04, Alparon, Delft University of Technology, 1997.

[5] H. Wainer, Estimating Coefficients in Linear Models, *Psychological Bulletin*, 83.2:213–217, 1976.