"Just Speak Naturally": Designing for Naturalness in Automated Spoken Dialogues

David Williams
Vocalis Ltd.
Great Shelford
Cambridge
CB2 5LD, UK
+44 1223 847166
david.williams@vocalis.com

Christine Cheepen
Department of Sociology
University of Surrey
Guildford
GU2 5XH, UK
+44 1483 300800
christine@soc.surrey.ac.uk

ABSTRACT

This paper describes an experiment carried out in the domain of telephone banking, and investigates the notion of naturalness in human-machine spoken dialogues. The experiment showed that 'denatured' prompts which were stripped of human-like constructs were preferable to callers, and achieved transaction times similar to those resulting from a typical telebanking dialogue.

Keywords

Spoken dialogues, naturalness, usability

INTRODUCTION

This paper describes an experiment which investigates the notion of naturalness in human-machine spoken dialogues. The paper focuses on the experimental method and results. For a more detailed theoretical background see Williams and Cheepen (1997). The experimental hypothesis is motivated by the widely-held assumption in the commercial sphere that for dialogues to be perceived as 'natural' or 'friendly' by a novice user, the system output (prompts) must contain a wide variety of human-like person-directed tokens, e.g. 'please', 'thanks', 'I', 'your,' etc. This paper proposes that before embellishing a dialogue with such tokens, the type of interaction that is taking place must first be considered.

The experiment described below takes a highly goal-directed domain which is typical of current automation targets, i.e. telebanking. Here, a variety of banking services are offered to the telephone caller.

A commercially available dialogue provides the dialogue logic and speech recognition performance.

Two prompt sets are compared. The first set (which we call the original set) illustrates the typical, arbitrary use of human-like person-directed tokens in system output. The second set had these tokens stripped out or replaced by material which was not person-directed, in order to produce a 'denatured' prompt set (for an in depth discussion of this procedure, see Williams and Cheepen (1997)). There was no difference in recognition performance or dialogue logic between the two 'systems'. We proposed that there would be no objective or subjective advantage for the original system.

EXPERIMENTAL METHOD

We recruited twenty-two subjects from the general public. They were not required to have any experience of tone or speech-based automated systems. The experiment used a within subjects design; in the first session, all of the subjects used the original system, in the second session, the denatured system. The dependent variable was transaction time which was measured from the beginning of the option selection prompt to the end of the last system prompt associated with the selected option.

For each system type condition, subjects evaluated the system using a short questionnaire. At the end of both system type conditions, subjects were simply asked which of the two systems was quicker and which they preferred.

RESULTS

Unfortunately, some subjects did not call both systems as instructed. Due to the paucity of information from the recording mechanism, there was no way to identify which

subjects were culpable. Additionally, due to a system bug, a number of subjects were unable to get past the third task ('balance') to the 'statement' task: two subjects in the original system and five in the denatured system.

Objective Measure - Transaction Times

A one-way ANOVA was conducted for each task with transaction time as the dependent variable and system type as the independent variable. The results are shown in Figure 1 below.

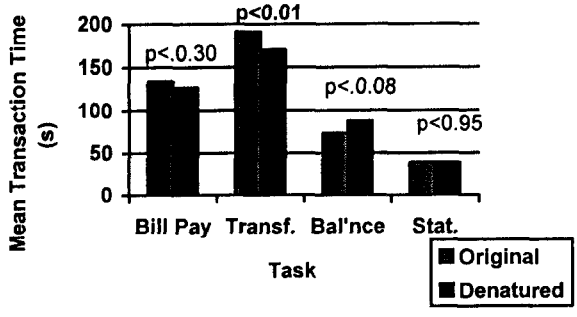

Figure 1: Transaction time comparisons for original and denatured systems

Subjective Measures - Speed and Preference

As well as obtaining these quantitative results, we also asked the subjects to complete two post-use questionnaires. Of the 22 subjects we used, 18 responded to the questionnaires. The results are shown in Table 1.

Measure	Orig.	Denat.	No Diff.	No Answer
Quicker	2	10	4	2
Preferred	1	8	7	2

Table 1: Subjective Comparisons by Subject Total

DISCUSSION

Transaction Times

In an automated system an important attribute is the length of time it takes a caller to enter the system, complete their task, and leave the system. This variable was compared between the two prompt sets. Our hypothesis, arising from our findings during a pilot experiment, was that, for a highly goal-directed domain, the denatured prompts would be more efficient overall (in terms of usability) than human-like, supposedly 'natural' prompts. We did not, however, expect that denatured prompts would necessarily display any great advantage in terms of actual transaction time.

The analysis of experimental transaction times shows no significant result in terms of the denatured prompts providing any advantage - the shorter, denatured prompts only resulted in a significantly shorter transaction time for the 'Funds Transfer' task. However, they clearly performed as well as the original prompts overall. This result validates the initial hypothesis that the different prompt sets would give very similar performances in terms of transaction times.

User Preferences

The hypothesis suggests that the denatured prompts will be preferable, or at least on a par with the original prompts. The subjective results indicate a clear preference for the denatured system. It is interesting to note that the real and perceived speed of the systems differed. To recap, the transaction time results showed no significant difference in the actual speed of the two systems The probability is that it is the prompt wordings which induce this perception, however erroneous, in the user. Given this phenomenon, it seems that the congruence of the denatured prompts with the highly goal-directed domain produces important effects on two key usability measures (Bevan and Mcleod, 1992) - perceived efficiency and satisfaction. Efficiency is provided by a feeling of speed in the dialogue, and satisfaction by the preference ratings. Examples of subject comments on the denatured system corroborate this, e.g. "Much clearer", "Seemed easier", "Straight to the point", "No fancy language and faster". However, this only makes a good argument for proposing shorter prompts in a highly goal-directed and business-oriented domain. Further work must be done in more interpersonal domains, e.g. leisure services.

ACKNOWLEDGMENTS

From an ESRC-funded research project 'Design guidelines for advanced voice dialogues', under the Cognitive Engineering Programme, project no. L127251012.

REFERENCES

1. Bevan, N., McLeod, M. (1992) *Usability Assessment and Measurement.* In Management and Assessment of Software Quality, 167-191.
2. Williams, D.M.L., Cheepen, C., (1997) *Designing for Naturalness in Automated Speech-based Dialogues: All you gotta do is act naturally.* In preparation.