HUMAN FACTORS IN SPEECH SYNTHESIS: FACTORS AFFECTING FRIENDLINESS AND EFFICIENCY

Christopher Miles, Dylan M Jones and Annette Simpson

School of Psychology, University of Wales, College of Cardiff, Cardiff CF1 3YG, UK

This study is concerned with the introduction of synthesized speech into the human–computer interface. A simulated hotel booking system is described which asks for information via either the screen or a DECTalk speech synthesizer. The degree of redundancy within the computer’s response set is varied. The perceived ‘friendliness’ of the system increased with the use of synthesized speech incorporating redundancy. In addition, this type of output influenced the quality of the users’ responses, such that they ‘modelled’ the computer’s speech.

INTRODUCTION

Although great technological progress has been apparent in the development of speech input/output (I/O) systems, the attitudes and responses to such systems by their potential users has largely been ignored [1]. This paper reports a study which examines the influence of synthesized speech and degree of redundancy within that speech on both attitudes and responses of users towards the interaction.

One feature of speech which may influence its perceived ‘friendliness’ is the inclusion of metacommments [2]. Metacommunication incorporates a degree of redundancy, and a particular type of redundancy important in this respect, is referred to as Linguistic Style Elements (LSEs). Previous research has shown that one particular LSE, known as Encouraging Phrases (EPs) may influence the acceptability of the system to its users. An EP may be defined as a ‘systems reaction towards the last answer of the user, aiming to increase the user’s positive attitude towards the system’ [3]. A frequency of EPs within the computer’s response set approximating that present in human–human interaction optimizes the acceptability of the interaction to the user [3].

In addition, the use of speech output per se in human–computer interactions has been shown to increase the quantity of words used by the receiver [4]. The quality of the speech used by receivers under these conditions may alter also, and this possibility is examined in the present study. It should be noted that a possible disadvantage in the introduction of speech synthesis within the human–computer interaction is that it may decrease the tolerance of users to system errors. This is possibly due to the development of ‘human–like’ expectations of the computer, within the user, in response to the speech medium [5].

The present paper reports a study which examines the combined influence of modality of computer output (speech vs screen) and level of ‘friendliness’ (via the number of EPs) on two measures. First, the attitude towards the system is assessed via a 20–item adjective checklist. Second, the nature of the users’ spoken responses is analysed. The interaction was based upon a simulated hotel booking system, with the computer
adopting the role of hotel reception. Such a simulation is typical of those where a computer may be used, it involves a database enquiry facility, the transaction is typically short, and maybe successfully conducted within a constrained vocabulary.

METHOD

SUBJECTS

24 volunteer members of the general public with an age range of 18 yrs - 57 yrs.

DESIGN

A (3x2) factorial design with frequency of EPs as a between-subjects factor (EPO; EP1; EP2), and mode of presentation as a within subjects factor: visual and spoken. EPO = no EPs; EP1 = one EP per three questions; EP2 = two EPs per three questions. The frequency of EPs was governed by analysis of tape-recorded "genuine" hotel telephone bookings. The naturally occurring frequency of EPs was calculated at one EP per 3 questions.

APPARATUS

The entire experiment was conducted using a DECTalk speech synthesizer controlled by a VAX computer. Subjects' verbal responses were recorded on a REVOX reel-to-reel tape recorder.

PROCEDURE

All subjects were provided with detailed instructions of the accommodation they were required to book prior to starting the experiment. For the visual modality, questions were presented in a predetermined sequence on the computer screen. All subjects were required to speak their responses into a microphone. Vocal responses were recorded on the tape-recorder which was out of sight of the subjects. Depression of the 'return' key initiated the next question. For the vocal modality, the same sequence of questions was spoken by the DECTalk and subjects responded in the same way as previously. The choice of DECTalk voice was determined by a pilot study measuring perceived intelligibility and naturalness [6]. In order to assess subjects' attitudes towards the interaction, they completed a 20-item semantic-differential questionnaire after completing the experiment [3,7].

DEPENDENT MEASURES

Four measures were taken from each set of subject's responses.

i) Total number of words in a response set.

ii) Percentage of words comprising an EP within a response set. An EP was defined as "any reply which encouraged the other participant to think highly of the speaker", and is a definition consistent with that reported in the Introduction.

iii) The number of EPs in each response set.

iv) The number of subjects who said 'goodbye' to the system at the conclusion of the interaction.

RESULTS

With respect to the analysis of the attitude scores, Analysis of Variance (ANOVA) showed the 'friendly' adjective to be jointly sensitive to both presentation mode and EP level. Specifically, speech output containing encouraging phrases (both EP1 and EP2) was rated as the most 'friendly' form of communication (F (2,9) = 4.7, P < .05).

ANOVA (2x3) examining computer response modality and EP level as factors was performed upon each of the four dependent measures. Total number of words within a response set did not vary across conditions (F<1). However, there
was a greater percentage of EP words spoken (F(1,21) = 14.4, P<01) in response to the synthesized speech. This measure was independently sensitive to EP level also (F(2,21) = 4.9, P<.05), increasing as the EP level increased. The total number of EPs produced was sensitive to the mode of computer response (F(1,21) = 13.5, P<.01), with a greater number associated with the synthesized speech. Finally, the number of 'goodbyes' increased with the increase in EP level (P<.05, Chi-square), regardless of modality of computer response.

DISCUSSION

The introduction of redundancy into the computer's spoken response set (via the inclusion of EPs) increased the perceived 'friendliness' of the interaction to users. The use of speech in isolation did not have such an influence. Nor did the introduction of EPs into visually presented responses facilitate perceived 'friendliness'. It is unlikely therefore, that the incorporation of speech synthesis per se into the human-computer interface will increase the 'acceptability' (in terms of 'friendliness') of the system. It is necessary for qualitative aspects of the speech, for instance its resemblance to natural language to be considered.

In contrast, the inclusion of redundancy within the synthesized speech influenced the nature of the user's responses rather differently. The effects of response modality and degree of redundancy acted independently rather than interactively. Synthesized speech influenced the 'naturalness' of the users' responses, regardless of the 'naturalness' of that synthesized speech. In addition, degree of redundancy influenced the 'naturalness' of users' responses, independently of modality. At a theoretical level the implication is that the quality of users' spoken output is differentially sensitive to both modality and content of the computer's output.

Clearly, the data suggest a relationship between users' subjective reactions to their interaction with the system, and the 'naturalness' of their own vocal response. The precise parameters of this relationship remain to be identified. Although there is some asymmetry in the relationship, it appears to be robust when one considers the constrained vocabulary set required to successfully complete the present task.

It is necessary to extend this research to other simulations, because the nature of the task may influence the role of speech in the interaction [4]. For instance, we are presently developing a financial services simulation which will engage the user in a more cognitively demanding, problem-solving type scenario. Other studies, using a home-shopping simulation task indicate that the role of speech becomes less important as users become familiar with the task demands. In addition, we have data suggesting that subjects' ratings of preferred (synthesized) voice, in terms of prosody and speech rate, are dependent upon the task context [8]. All of these factors require further research before the introduction of speech into the human-computer interface can be clearly understood.

We acknowledge Digital Equipment Corporation for the use of a DECtalk to enable this research to be undertaken.

REFERENCES


