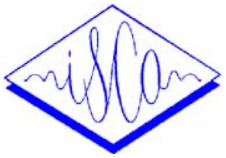

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

USER ATTITUDES TO CONCATENATED NATURAL SPEECH AND TEXT-TO-SPEECH SYNTHESIS IN AN AUTOMATED INFORMATION SERVICE

F R McInnes[1], D J Attwater[2], M D Edgington[3], M S Schmidt[4*] and M A Jack[1]*

[1]CCIR, The University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, UK
[2]BT Laboratories, Martlesham Heath, Ipswich IP5 3RE, UK
[3]SRI International, Mill Lane, Cambridge, UK
[4]Andersen Consulting

{Fergus.McInnes, Mervyn.Jack}@ed.ac.uk
http://www.ccir.ed.ac.uk
David.Attwater@bt.com
http://www.labs.bt.com
Mike.Edgington@cam.sri.com
Mark.S.Schmidt@ac.com

*Mike Edgington was at BT Laboratories and Mark Schmidt was at CCIR at the time of the experiment reported here.

ABSTRACT

Today's automated telephone services generally use recorded speech from one speaker for all output. In applications with large and varying output vocabularies, such as place names, it may be necessary to employ a second speaker to provide new vocabulary items if the original speaker is not available, or to use text-to-speech (TTS) synthesis for the whole or parts of the output. This paper reports a comparison of 10 schemes for the generation of spoken output in a travel information service, ranging from natural speech from a single speaker, through combinations of different voices and of natural and synthetic speech, to TTS synthesis throughout. The results show strong preferences for concatenated speech over TTS and for professional-quality recordings over amateur ones, and a weaker preference for a single speaker over two speakers.

Keywords: IVR, speech concatenation, speech synthesis

1. INTRODUCTION

Most current automated telephone services respond to user input by means of pre-recorded utterances. Assuming a high level of quality of the speech recording and professionalism of the speaker, this form of speech output provides optimal feedback for the user in terms of intelligibility and naturalness. Where only a small or moderate number of fixed output messages are required, these can be recorded in full and used directly. Where variable information from a fixed small vocabulary has to be spoken, simple concatenation can be used: for instance, arbitrary digit strings such as telephone numbers can be generated by concatenating recordings of single digits — though, to achieve good speech quality, care must be taken to record the words with appropriate intonations.

This approach becomes less feasible when the output vocabulary is very large (as may happen with names of places, people or companies); when the vocabulary is liable to change to include new words or phrases after the initial implementation; or when a wide range of sentence structures or intonations is required. In such cases it may be necessary to use speech synthesis instead of concatenating recorded speech; or, especially in the case of extensible vocabularies, to employ a second speaker if the original speaker is not available at the time when additional words or phrases are needed.

For user acceptance of a service, the intelligibility of its spoken output is clearly important. Other attributes such as the naturalness and pleasantness of the speech may also strongly influence users' attitudes to the service, and hence their willingness to use it. It is important for service designers to know the usability implications of possible design choices in the generation of spoken output. Therefore, as part of the Dialogues 2000 Project [1], an experiment was conducted which systematically compared a number of schemes for the generation of spoken output in the context of a simulated air travel information service. This paper reports the results, both in terms of users' expressed attitudes to the various versions of the service and in respect of the accuracy with which they recognised key information (place names and times) in the spoken messages.

2. FLIGHT INFORMATION SERVICE

A simple telephone dialogue scenario was devised in which subjects confirm the times and routes of pre-booked flights. This is a realistic scenario, as call centre flight bookings are increasing rapidly and there is often a delay of a number of days before written confirmation of bookings reaches the customer. A booking reference number, however, is always provided which would facilitate the use of automated confirmation services such as the one employed in this experiment.

Since the application requires only entry of a reference number and then a yes/no confirmation, it is feasible for the interface to use DTMF keypad input only, and the dialogue can be kept simple. A typical call would proceed as follows.

System: *Welcome to the CCIR flight time information service. Using the telephone keypad, please enter your booking reference number.*

User: Keys in booking reference number.

System: *Thank you.*

Booking reference number <123456> matches a one way flight from <Edinburgh> to <Vancouver>.

If this is correct, press 1, otherwise press 2.

User: Presses "1" to confirm the number.

System: *Thank you.*

Your flight from <Edinburgh> will leave on <Friday the second of August> at <oh eight thirty>, and will arrive at <London Heathrow> at <oh nine forty-five>. From there, it will depart for the onward journey at <eleven thirty>, arriving at <Vancouver> at <thirteen fifty>.

To hear the information again, please hold, otherwise hang up. [PAUSE]

The words in angle brackets are variable according to the flight booking, and must be concatenated with the adjoining fixed carrier phrases (except where the whole sentence is generated by synthesis).

The response scheme outlined above ensures a reasonably long exposure to the particular version of speech output. The first variable system response message (the *confirmation message*) contains information that the user will know already and provides an opportunity for familiarisation with concatenated or synthesised speech. The second variable message (the *information message*) contains a mixture of given and new information and provides a test of the intelligibility of the output.

3. CONCATENATION SCHEMES

The 10 methods of constructing the speech output (concatenation schemes) are summarised in Table 1. Each method is defined by the combination of voices used for the various components of the spoken messages. The pairwise inter-scheme comparisons identified as being of particular interest are also listed.

The message components were defined as follows:

Prompts: welcome message; prompt for booking reference number; "thank you" after reference number entry; error and reprompt messages.

Carriers: confirmation and information messages, including the digits used in reading back the reference number, but excluding the airport names, dates and times; "thank you" after positive confirmation response; "hold/hang up" message after information message.

Dates and times: the dates and times used in the information message.

Airport names: the airport names occurring in the confirmation and information messages.

(The carrier phrases, dates, times and airport names together make up the *response messages*. The prompt/response distinction separates the main interactive phase of the dialogue, in which the user provides the booking reference number, from the information readback phase.)

Three real (human) speakers and one TTS voice were employed:

R1: a professional speaker experienced in recording telephone service messages;

R2: a second professional speaker;

R3: an amateur speaker;

T1: text-to-speech synthesis based on the voice of the first professional speaker *R1*.

The text-to-speech system adopted was BT's Laureate system, which had scored well on listening effort and quality scales relative to other commercially available synthesis systems in previous evaluations [2]. No application-specific optimisation (such as use of hand-crafted pronunciations or pitch contours) was applied to the synthesis. The *R3* recordings were "amateur" not only in the choice of speaker but also in the procedure adopted, in that no coaching on intonation was provided to the speaker, in contrast to the *R1* and *R2* recordings where an effort was made to achieve appropriate intonation on each utterance for its intended context. All voices were female with British English accents (R1 and R2 from southern England, R3 from the northwest of England but resident in the south for the last 10 years).

Version	Prompts	Carriers	Dates & times	Airports	Comment	Comparisons
0	R3	R3	R3	R3	Amateur recording	1, 5
1	R1	R1	R1	R1	Professional recording	0, 2, 3, 3a, 5, 6
2	R1	R1	R1	R2	Second voice for airport names	1, 3a
3	R1	R1	T1	T1	TTS for all variables (same voice)	1, 3a, 4
3a	R1	R1	R1	T1	TTS for airport names only (same voice)	1, 2, 3, 4a
4	R2	R2	T1	T1	TTS for all variables (different voice)	3, 4a, 5, 7
4a	R2	R2	R2	T1	TTS for airport names (different voice)	3a, 4, 5, 7
5	T1	T1	T1	T1	All TTS	1, 6, 7, 4, 4a, 0
6	R1	T1	T1	T1	TTS response (same voice)	1, 5, 7
7	R2	T1	T1	T1	TTS response (different voice)	5, 6, 4, 4a

Table 1. Concatenation schemes

Scheme 1 represents the ideal situation in which a professional speaker is available to record everything.

Scheme 2 represents the case where the vocabulary has to be extended and the original speaker is not available to provide the additional recordings. It is of interest to see how this compares with the ideal case (scheme 1) and with the case where TTS based on the original speaker's voice is used to provide the additional vocabulary (scheme 3a).

Schemes 3 and 3a represent cases where the speaker for the carrier sentences is not available, or too expensive to use, to record the variable information but a TTS voice based on that speaker's voice is available. The difference between them is in whether dates and times have been recorded by the speaker or whether these have to be synthesised. In the case envisaged it seems likely that they would be available (3a), since recording the components of all possible dates and times is a small task compared with providing a full inventory of speech units for TTS; but the other possibility (3) was included for comparison purposes. The comparison with scheme 2 is of interest since the use of a second speaker will be a viable alternative to using TTS in many instances.

Schemes 4 and 4a are similar to schemes 3 and 3a but with TTS available only in a different voice — as will often be the case in practice. The comparison with the same-voice case (3 or 3a) gives an indication of how much benefit there is in having the TTS based on the same speaker who provides the natural speech, in cases where this is feasible. The comparison with more extensive use of TTS (scheme 5 or 7) is also of practical interest, the question being whether it is better to express only the variable information in synthesised speech or to use synthesis consistently for the whole utterance in which that information is embedded.

Scheme 5 represents what might be the cheapest single-voice option assuming the availability of a TTS system, namely the use of TTS throughout the service. It might also be preferable, in terms of consistency, to the use of natural speech for parts of the service and TTS for other parts, as in *schemes 6 and 7*; comparisons with these schemes, as well as with the minimal use of synthesis (scheme 4 or 4a), are therefore of practical relevance.

Scheme 0 represents another cheap solution, the use of an amateur speaker. The comparison with scheme 5 is of practical significance; the comparison with scheme 1 shows how much is lost by using amateur recordings.

The booking reference number in the confirmation message was treated as part of the carrier for the purpose of the choice of voice: that is, it was always constructed from digits (with appropriate intonation) in the same voice which spoke the carrier sentence.

4. EXPERIMENT DESIGN

The experiment involved 100 subjects from a local panel, recruited so as to provide a cross-section of the population, who attended the research centre at the University of Edinburgh to take part. Each participant used all 10 versions of the flight time information service (differing in their concatenation schemes) in succession, with a different flight scenario (origin, destination and date) for each use to avoid habituation to the scenario details. The order of presentation and the allocation of concatenation schemes to flight scenarios were balanced across the participants. Participants were told at the outset that the experiment was concerned with how the service spoke the information.

For each use of the service, the participant was given an instruction sheet setting out the scenario and the booking reference number and asking for two previously unknown pieces of information which were to be obtained from the information message. After calling the service and writing the answers in the spaces provided, the participant completed a short questionnaire containing six statements with seven-point Likert response scales, which was printed on the back of the instruction sheet.

Finally there was an interview, during which examples of the outputs of all 10 versions of the service were played back and the participant was asked about each version in turn.

5. RESULTS

The main aim of the experiment was to discover users' opinions and preferences as to the generation of spoken messages. The responses to the questionnaire are therefore the most important component in the results.

Profiles of the mean responses to the six questionnaire items for all 10 versions of the service are shown in Figure 1. The questionnaire items were:

1. I had to concentrate hard while listening to the information.
2. It was easy to pick out the important pieces of information.
3. I feel that the service needs a lot of improvement.
4. I thought what the service said was very clear.
5. The service was pleasant to listen to.
6. Sometimes I thought the messages sounded disjointed.

For each item, the response scale ran from "strongly agree" to "strongly disagree". The responses obtained were mapped onto a scale from 1 (least favourable evaluation) to 7 (most favourable) — shown on the vertical axis in Figure 1.

The mean attitude scores, obtained by averaging over the six items in the questionnaire, are given (in descending order) in Table 2. The "Better than" column indicates where two-tailed related samples *t*-tests yielded significant results ($p<0.05$; $p<0.01$ with single asterisk; $p<0.001$ with double asterisk) for comparisons listed in Table 1.

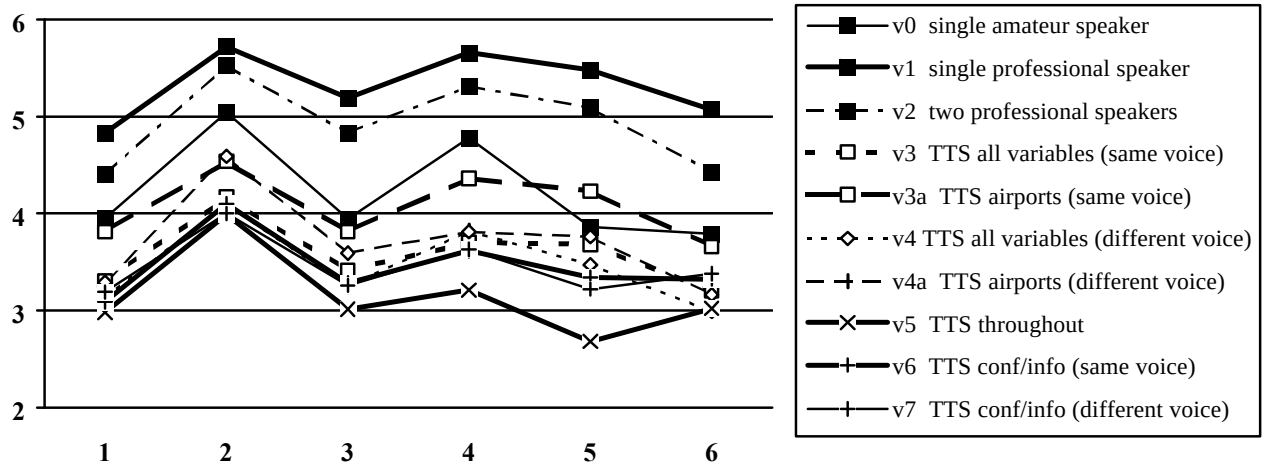

Figure 1. Profiles of questionnaire responses for all versions of the service. Broad lines are used for versions using speaker 1 (R1 and/or T1) throughout, and narrow lines for other voices or combinations; solid lines for versions which do not concatenate different voices or speech types in the same message, and other line styles for those which do.

Version	Score	Better than
1	5.325	0** 2* 3** 3a** 5** 6**
2	4.932	3a**
0	4.227	5**
3a	4.068	1** 3** 4a
4a	3.700	5**
3	3.577	
4	3.460	
7	3.447	5
6	3.415	5
5	3.148	

Table 2. Mean attitude scores

Notable features of these results are as follows:—

Natural speech was preferred to *synthetic speech* — not only in that the versions of the service using natural speech throughout (0, 1 and 2) were ranked above all those using any synthesis, but also in that those using less synthetic speech were consistently ranked above those using more. Thus versions 3 and 4, with synthesis for dates and times as well as for airport names, were rated worse than versions 3a and 4a (though not all the comparisons are significant); versions 6 and 7, with the whole response messages synthesised, were marginally lower; and version 5, with synthetic speech throughout the dialogue, had the lowest attitude score of all.

Professional recordings (versions 1 and 2) were preferred to *amateur* ones (version 0). The drop in attitude due to use of an amateur speaker and procedure (version 0) was greater than that due to mixing voices while retaining professional standards (version 2). (The *t*-test result for the comparison of versions 0 and 2 was very highly significant: p=0.00001.)

There was some preference for a *single* speaker over a *mixture* of speakers, other things being equal. This was strongest in the case with natural speech throughout (1 v. 2), and became weaker and eventually insignificant as the amount of synthetic speech was increased (3a v. 4a, 3 v. 4, and 6 v. 7).

The accuracy with which participants identified the two requested pieces of information was also measured. For the *connection airport*, significant error rate differences (p<0.05, or p<0.01 where asterisked) were found when comparing version 1 against versions 3, 3a*, 5* and 6; version 0 against version 5*; and version 2 against version 3a. In each case the accuracy was higher with natural speech throughout. On the *departure or arrival time* there were no significant differences. This suggests that synthetic speech is less intelligible than natural speech for a vocabulary containing possibly unfamiliar or unexpected items, but perhaps not for a familiar closed vocabulary such as time-of-day expressions.

6. ACKNOWLEDGEMENTS

This work was carried out as part of the Dialogues 2000 Project, made possible by support from Engineering and Physical Sciences Research Council and from BT as part of the SALT LINK Programme. The authors acknowledge the assistance of their colleagues at the University of Edinburgh and at BT Laboratories.

7. REFERENCES

[1] Schmidt, M.S., M.A. Jack and F.W.M. Stentiford (1995), Dialogues 2000 — Towards the introduction of best practices in the design of automated telephone services, *Proc. Voice Europe 95*.

[2] Page, J.H., and A.P. Breen (1996), The Laureate text-to-speech system — architecture and applications. *BT Technology Journal*, vol.14, no 1, pp. 57-67.