GENERIC ISSUES IN DIALOGUE DESIGN FOR SPEECH INPUT/OUTPUT

R.G. Leiser*, M. de Alberdi*, D.J. Carr*.

ABSTRACT

In the design and evaluation of speech input/output dialogues for control of PBX features, various issues were encountered which could not be resolved by reference to the available literature.

This paper outlines solutions to some generic issues in speech dialogue design in the form of Generic Dialogue Modules. These are descriptions of procedures which may be implemented in a wide range of speech input/output applications. Where criteria are application-dependent, the factors affecting optimal adjustment are listed.

Issues discussed are template training methods, strategies for input confirmation and correction, encouraging template maintenance, provision of Help, and development of alternative language variants of speech input/output dialogues.

THE GSU PROJECT

The aim of the GSU (Generic Speech Unit) project was to provide PBX users with a simple and direct method of accessing the various valuable features by speech input to a standard telephone set. A major constraint on dialogue design was the non-availability of visual feedback. Issues which arose in the design and evaluation of dialogues were both specific to the application and generic to all applications of speech input/output where no visual feedback is available.

GENERIC DIALOGUE MODULES

A problem encountered at an early stage of this project was the paucity of generic principles for speech dialogue design. Rather than produce a list of guidelines which are prone to mis-interpretation and are often conflicting, solutions are presented below as Generic Dialogue Modules. These describe simple routines which can be implemented in a wide range of applications. Criteria which may need to be tuned for different applications are indicated, with suggestions as to how this tuning should occur.

TEMPLATE TRAINING METHODS

Speaker-dependent systems require users to provide templates

* Alcatel ESC, Great Eastern House, Edinburgh Way, Harlow, Essex
of their speech for future recognition purposes. The primary requirement for training methods is that the templates recorded represent articulation of the words in real system use.

The classic approach has been to prompt users to repeat each word. Problems may arise here as a result of the users mimicking the intonation of the prompt during training, then reverting to their own natural intonation during system use, resulting in a decrement in recognition performance.

This innate mimicry can be reduced by adoption of a training dialogue where the word to be repeated is followed by a spoken prompt before the user responds. The spoken prompt has the effect of overwriting the precategorical acoustic store (ref 1), hence depriving the user of a model to mimic. The prompt should be varied to prevent the user from identifying, attending to and rehearsing the intonation of the word to be trained as a result of its being the only variant part of a repeated message.

INPUT CONFIRMATION

As in human-human dialogue, mis-recognition of inputs is inevitable. Both user and system must be protected against disastrous effects of this.

Confirmation of input may increase reliability, but soon becomes tedious to the user. A solution is to request confirmation only when the probability of correct recognition falls below a specified level. Figure 1 shows, for a typical system, the relationship between the closeness of a recognition match and the probability that the match is correct. Figure 2 shows the relationship between the separation of the best and second best matches, and the possibility of a correct match. The system can use these relationships to estimate the probability that a matched input will be correct. A request can be made for Yes/No confirmation or for repetition, depending on just how bad the input is.

![Figure 1](image1)

Probability of correct 1st guess vs score for 1st guess

![Figure 2](image2)

Probability of correct 1st guess vs distance between 1st and 2nd guesses
The threshold criteria for either of these cases is application-dependent. In high-risk applications, the system should be sensitive to poor matches, sacrificing brevity for accuracy. For a low-risk application, it is desirable to have fewer requests for confirmation.

TEMPLATE MAINTENANCE

In systems using speaker-dependent recognition, template re-training will be necessary, particularly if templates are trained on the first session of use only. The need for re-training becomes apparent to the user through repeated mis-recognition. A basic approach to this problem is to rely on the user to recognise the problem and remember to re-train the problem word. Such deferred maintenance tasks, which bring no immediately apparent benefit, tend to be neglected.

A complex alternative is to apply continuous template updating. Where software limitations do not allow this, a simple but effective method is available, which encourages template maintenance without interfering with the task or relying on the user's memory.

A mis-recognition counter for each word is incremented every time a mis-recognition of that word is detected by the system. (The detection may be based on the Input Confirmation/Correction strategy described above.) When the count level reaches a certain criterion, the user is given the option to speak the mis-recognised command to provide an update template for the problem word. The user is then returned to the original dialogue, and the counter for the word is zeroed.

The setting of the criterion count determines the relative proportions of re-train requests and recognition problems. High criteria are desirable in noisy environments to minimise unnecessary re-training following mis-recognition due to background noise, and to prevent counter-productive training in such an environment. If infrequent use in short sessions is expected, resulting in high voice variability, the count should be low so that training can be dealt with promptly without having to wait for several mis-recognitions.

PROVISION OF HELP

Unlike unwanted Help information on a VDU screen, Help messages in a speech input/output application cannot be ignored. For this reason it is desirable that Help messages are carefully worded and provided only when necessary.

User-initiated, context-sensitive Help should be available at all points in the dialogue. System-initiated Help should be provided at two levels. If the user fails to respond appropriately to a prompt, feedback on the problem detected
should be provided followed by the original prompt. This should include some indication of the current command mode. A further inappropriate or absent response from the user initiates a reminder that Help is available followed by a listing of available commands.

This module is based on the principle of detection and repair of communication failure (ref 2).

DEVELOPING LANGUAGE VARIANTS OF SPEECH INPUT/OUTPUT DIALOGUES

Experience with the PBX project revealed that it is not sufficient to translate system messages to provide foreign language variants of speech dialogues. As well as subtle alterations to sense arising from translation, cultural differences must be considered. In particular, the requirements for politeness and brevity of system output will vary between cultures. Dialogues must be optimised by user testing with the target population.

A recommended method of dialogue evaluation is detailed elsewhere (ref 3).

SUMMARY

Solutions to generic issues in speech dialogue design are presented as Generic Dialogue Modules. This method of representation allows easy transfer of Human Factors expertise to a wide range of applications.

REFERENCES