Proceedings of the Institute of Acoustics

DEVELOPMENT OF A MULTILINGUAL SPEECH INTERFACE FOR A CAR PHONE

W.A.Ainsworth and S.R.Pratt

Department of Communication and Neuroscience, Keele University, Keele, Staffordshire ST5 5BG, U.K.

1. INTRODUCTION

Automatic speech recognition provides a means of interacting with machinery when the hands and eyes are busy with other tasks. One such application is controlling a car phone whilst driving. During the last three years an experimental system has been developed which allows a driver to make and receive telephone calls entirely by voice. This has necessitated designing a speech recogniser which is robust in the noisy environment found inside moving vehicles and integrating this with a man-machine interface which provides various dialling modes and procedures for correcting the few remaining errors. The work has been carried out in collaboration with European partners so versions of the interface have been developed for English, French, Italian and Spanish.

2. SPEECH INPUT

Most of the effort of the project was directed to the design of a speech recogniser which operated reliably in a noisy environment [1]. The final design is shown in Figure 1. It was a speaker dependent, isolated word recogniser. It consisted of a feature extractor and a pattern matcher. The feature detector contained an end-point detector (EDP) to indicate the presence or otherwise of speech, and a non-linear spectral subtraction algorithm (NSS) to attenuate the noise [2]. The output of the feature detector was a set of mel frequency cepstrum coefficients (MFCC). These were input to a hidden Markov model (HMM) pattern matcher.

The speech input was sampled at 8 kHz with 16 bits accuracy then processed with a pre-emphasis filter

$$H(z) = 1 - 0.98z^{-1}.$$

The output of this was stored in frames of 256 samples (32 ms) overlapped by 128 samples (16 ms). Each frame was multiplied by a Hamming window

$$ham(i) = 0.54 - 0.46cos(2\pi i/256),$$

where i is the position in the frame, and then an FFT was performed by the split radix-2 algorithm [3].

The efficiency of non-linear spectral subtraction depends upon the accuracy of the end-point detector as an estimate of the noise immediately prior to the speech is required. The

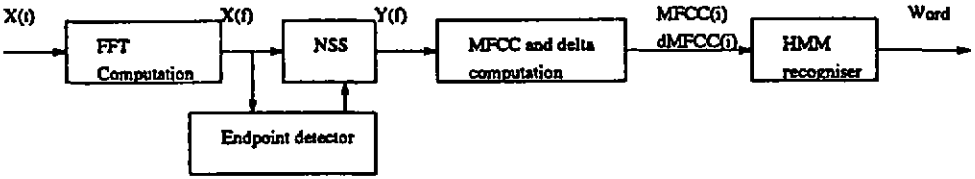

Figure 1: Block diagram of the speech recogniser

EDP was a four state automaton as shown in Figure 2. Moving from one state to another depended upon an estimate of the speech energy in each frame. This was estimated as:

$$SE_i = |X_i|^2 - |B_i|^2$$

where $= X_i^2$ is the signal energy in a frame and $= B_i^2$ is an estimate of the noise energy. This was computed as:

$$|B_{i+1}| = (1 - a)|X_i| + a|B_i|$$

where the factor a is $0 \leq a \leq 1$ and is a function of the noise level of the environment.

The version of NSS employed was defined as:

$$Y(f) = X(f) - \alpha(f)N(f) \; if \; Y(f) > 0.1N(f)$$

$$= 0.1X(f) \; otherwise.$$

The function $\alpha(f)$ due to Lockwood et al [2] is:

$$\alpha(f) = \frac{N_{max}}{N(f) + \frac{X(f) - N(f)}{N(f)}}$$

with the range of $\alpha(f)$ limited to $1 \leq \alpha \leq 4$, where X(f) is the spectrum of the signal in a frame, N(f) is the mean spectrum of 4 frames of noise immediate prior to the speech and M_{max} is the spectrum of the largest noise in that period.

From the estimate of the spectrum of the speech the MFCC were computed by grouping the FFT coefficients into 19 mel bands, taking the logarithms of the band energies then applying cosine transform [4]:

$$MFCC_j = \sum_{k=1}^{19} (E_k cos(j(k - 0.5)\pi/19))$$

where j=1, 2, ...10, the number of cepstrum coefficients and E_k for k=1 to 19 is the output of the kth filter.

The delta cepstrum parameters were computed from a window of 5 frames from frame i-2 to frame i+2.

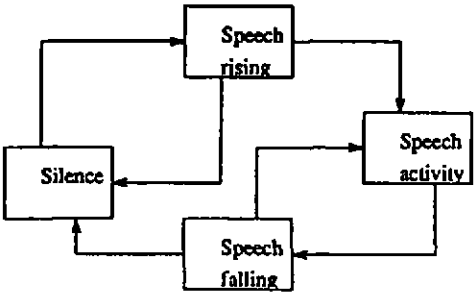

Figure 2: Block diagram of the end-point detector

The recognition algorithm used 8-state continuous density HMMs with the emission probabilities represented by a single gaussian for each state. Only self-transitions and transitions to the next state were allowed. Scoring was performed by the Viterbi method.

The method of 'incremental training' was used to estimate the model parameters. Initially the sequence of frames was distributed equally among the states and the values of the features were averaged for each state. The transition probabilities were estimated by:

$$a_{i,i} = \frac{nfr_i}{nfr_i + 1}, \; a_{i,j} = \frac{1}{nfr_i + 1}$$

where nfr_i are the frames assigned to state i.

A Viterbi recognition procedure was then performed. By backtracking through the states along the optimal path a new segmentation was obtained. The frames were then redistributed according to this segmentation and the procedure reiterated until the difference in log-likelihood between two iterations was less than a threshold. This procedure was repeated with the obvious modifications for other repetitions of each word. It was found that 3 utterances of each word gave satisfactory recognition performance. This method had the advantage that it was relatively fast and that the user could retrain a word if necessary.

The best results with speech recognisers are usually obtained when the recogniser is trained in the conditions in which it is to be used. With a car phone, however, it is considered dangerous to train the recogniser whilst driving. With the recogniser trained with the car stationary but with the engine idling it was found that the system performed with an accuracy of better than 90% when driven at 90 km/h and at about 85% at 130 km/h.

3. SPEECH OUTPUT

Audio feedback was provided by spoken messages digitised and stored in speech files. Speech was sampled at 16 bits at 8 or 16 kHz. Files were edited so that periods of silence were removed. The storage requirement was further reduced by constructing longer messages from the files for individual words wherever possible. For example, the Help message often listed the words in the active vocabulary at that stage of the dialogue. This was effected by

using the existing files for the vocabulary items and inserting 'or' before the last. The storage requirement could, of course, be further reduced by adopting one of the recognised speech coding procedures. This method of providing speech output facilitated the development of the system for new languages. Although amateur speakers were used for development and demonstrations it was recommended that professional speakers be used for all the standard utterances. It was necessary to use the speech of the user for customised words such as names. Accordingly, when the user trains the speech recogniser on a new name the first utterance is used to make a speech file for subsequent feedback.

4. DIALLING MODES

The interface allows the user to dial by one of several modes, each of which has advantages and disadvantages. Keyword (home, office, etc.) dialling is fast but requires setting up in advance and employs a fixed vocabulary. Dialling by name is also fast but the user has to train the system with the appropriate names. Short-code dialling is flexible and does not require extra training, but the user must remember the code and, for vocabularies of more than ten words, two or more words must be uttered. Dialling digit by digit is slow but it can be used for any telephone number without additional setting up or training. Number dialling is a little faster as feedback is given only at the end of the number. However, if a misrecognition occurs there is no alternative but to repeat the whole string of digits.

It is recommended that keyword or name dialling should be used whenever possible as dialling is fast and error correction is simple as feedback is provided. If it is expected that it will be necessary to dial a new number several times during the course of a journey it may be worth setting up a short-code. Digit dialling, with feedback after each digit, should only be resorted to for dialling new numbers in adverse conditions where recognition errors occur in a long string of digits.

The procedure for making a call is shown in Figure 3. When the system is activated it asks 'Answer, Telephone or Train?' to which the user replies 'TELEPHONE'. The system then requests 'Mode?' and the user says 'DIGITS, KEYWORD, MEMORY, NUMBER or NAME'. He may also say 'ESCAPE' to go back to the beginning, or 'HELP' to be told which words he may utter.

5. ERROR CORRECTION

Two methods of error correction were developed, simple repetition and repetition with elimination. When a recognition error is detected the user can say 'CANCEL' to invoke the simple repetition method. The system replies 'Cancel' and the user repeats the misrecognised word. If the user says 'CORRECTION', the system eliminates the previous attempt from the active vocabulary and says 'Repeat'. Following the next utterance by the user the system asks 'Was it ...?' to which the user replies 'YES' or 'NO' accordingly. If 'NO' is recognised the latest attempt is again eliminated. The dialogue proceeds until 'YES' is recognised, when the system repeats the correct word and exits the correction mode. The second procedure is much longer than the first but it avoids the danger of repeated errors. It is for the user to decide whether on a particular occasion the longer procedure is likely to be necessary.

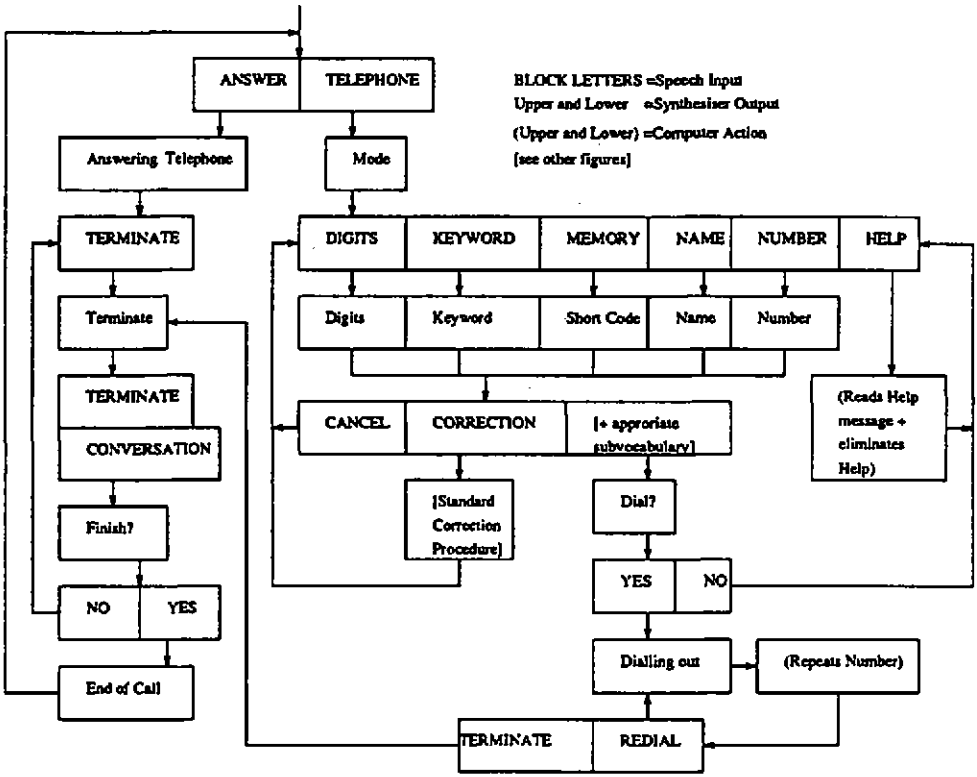

Figure 3: Flow diagram of the dialling procedure

6. DIALOGUE STRUCTURE

To answer an incoming call the user says 'ANSWER'. The system replies 'Answering telephone'. The system developed is a simulation. A real voice-activated telephone would not need a speech output message at this point; the end of the ringing tone would be a sufficient cue to the user. There are several stages in the dialogue where a speech message is used for the simulation but where a tone or other cue from the telephone system would enable the message to be deleted.

Outgoing calls can be dialled in several ways as described above. Following dialling, an action stage allows the user to correct errors, proceed with dialling, verify dialling, request help or abandon the call attempt. If he wishes to proceed with dialling he says 'DIAL' and the system asks for confirmation 'Dial?' to which the user replies 'YES' or 'NO'. If 'YES' is recognised the system confirms by speaking the message 'Dialling out'. To verify dialling the user says 'VERIFY' at the action stage and the system repeats the name, keyword or digits dialled. It is expected that a user would easily forget how far he has dialled if he is distracted by road conditions. In response to 'HELP' at this stage the system reads out the items in the active vocabulary and asks 'More?'. Following 'YES' it gives information on the simple repetition error correction procedure: 'If an error occurs, say CANCEL and try again'. The user can abort the call attempt by saying 'ESCAPE' when he will be returned to the starting point in the dialogue. 'HELP' and 'ESCAPE' also occur at other points in the dialogue with similar effect.

To avoid the embarrassment of a call being ended prematurely by words used in the telephone conversation, a specific and admittedly somewhat cumbersome sequence is required to end a call. The user says 'TERMINATE' which the system repeats, he then says 'CONVERSATION' to which the system replies 'Finish?' and the user finally says 'YES'.

7. FRENCH VERSION

Development of the man-machine interface first concentrated exclusively on the English version. Priority was given to making the dialogue as natural as possible. Occasionally this was implemented at the expense of maintaining the smallest possible active vocabularies. For example, for digit dialling, the words DOUBLE, TREBLE, HUNDRED and THOUSAND were added so that users could speak the telephone number in the normal way. When developing the versions for languages other than English we also wished to produce natural dialogues. The importance of this approach has been stressed by Leiser et al [5].

We wished to implement paired-digit dialling for the French and Spanish versions because it is so natural for speakers of these languages to think of their telephone numbers (at least in part) as a series of digit pairs, such as (the equivalent of) twenty-nine, sixteen, thirty-five, ninety-six. Unfortunately the much larger total vocabulary required for this facility could not be accommodated by the hardware, so this development could not be pursued. The facility of paired-digit dialling does present a number of problems. As well as the size of the total vocabulary mentioned above, the active vocabulary would at times be of the order of 18 words (larger than at any other point in the dialogue) and likely to prove prejudicial to the

accuracy of the speech recogniser. The ambiguity of certain sequences such as SOIXANTE followed by DIX-SEPT (60-17 or 77) is a particular problem. It is possible that the only method of providing unambiguous voice dialling is to adopt the French military practice and require the user to say 'STOP' to represent the gap between the digit pairs. The above example would then be either 'SOIXANTE STOP DIX-SEPT' or 'SOIXANTE-DIX-SEPT STOP' according to which digit sequence was intended. (It should perhaps be added that this complication normally applies only to French as spoken in France; in Belgium and Switzerland SEPTANTE and NONANTE usually replace SOIXANTE-DIX and QUATRE-VINGT-DIX.) Error-correction would be more complicated if feedback was provided after each pair of digits, because it would be necessary to inform the system which digit (or digits) was incorrect. We have implemented this dialogue for all language versions for the correction of errors in Short Code dialling. A further complication is that the telephone number (including where applicable country and area codes) cannot always be parsed as a series of pairs. An attempt has been made to make the French version idiomatic, for example 'Was four correct?' is rendered 'Le quatre est bon?'.

8. ITALIAN VERSION

The Italian version is more than an idiomatic translation of the English version. It was felt that more efficient digit dialling would be obtained if the command words were placed in the same active vocabulary as the digits. This means that to dial a number a user does not have to say the equivalent of ACTION and then select the appropriate command; he can issue a command in place of a digit. This is attractive in terms of speed of dialling a call but it runs the risk of occasional confusions. It was judged that, because the Italian digit words normally have more syllables than their English equivalents, these confusions would be much rarer.

The other changes adopted for the Italian version are mainly concerned with expanding the system output phrases to make the meaning clearer. For example, the English 'Dial?' has been rendered 'Pronto a chiamare - OKAY?' ('Ready to dial - OK?'). There is clearly some conflict between what is most appropriate for a new user who is likely to appreciate cues as to his position in the dialogue and the experienced user who knows what to expect and wants to make his call as quickly as possible. The Italian version also incorporates two tone signals where the English version uses speech messages. This is certainly an advantage and, as mentioned above, a system connected to the network could dispense with some of the speech feedback used in the simulation.

9. SPANISH VERSION

Most of the remarks on paired-digit dialling made in the section on the French version apply equally to the Spanish version of the man-machine interface. The exception is that Spanish spoken telephone numbers are not subject to ambiguous sequences such as SOIXANTE DIX-SEPT.

We have again applied the general principle of idiomatic translation for this version.

10. CONCLUSIONS

A voice activated car phone interface has been developed using a speech recogniser designed to perform well in conditions of moderately intense, stationary background noise. The dialogue has been designed to maximise hands free and eyes free operation. A priority was to make the car phone easy and natural to use for native speakers of English, French, Italian and Spanish. The method of implementation was chosen to facilitate modification to further languages. The interface has been developed in simulation; experience with a real car phone connected to the telephone network is likely to require further changes to the detail of the dialogue.

11. ACKNOWLEDGEMENTS

This work was carried out as part of Esprit II Project No. 2101 'Adverse-environment Recognition of Speech'. It is a pleasure to acknowledge the contribution of our French, Italian and Spanish Partners in the development of these interfaces.

12. REFERENCES

[1] G.BABINI, L.FISSORE, W.A.AINSWORTH, E.FRANGOULIS, R.GARCIA-GOMEZ, Y.GRENIER & P.LOCKWOOD, 'Esprit Project No. 2101 Adverse-environment recognition of speech (A.R.S.)', Esprit Conference, Brussels, 222-233 (1991)

[2] P.LOCKWOOD, J.M.GILLOT, J.BOUDY & G.FAUCON, 'Experiments with a nonlinear spectral subtractor (NSS) for speech recognition in a car', Eurospeech'91, Genoa (1991)

[3] H.V.SORENSEN, D.L.JONES, M.T.HEIDMAN & C.S.BURRUS, 'Real-valued fast Fourier transform algorithms', IEEE Trans. ASSP-35, 849-863 (1987)

[4] S.B.DAVIS & P.MERMELSTEIN, 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', IEEE Trans. ASSP, 357-366 (1980)

[5] R G LEISER, M DE ALBERDI & D J CARR, 'Generic Issues in Dialogue Design for Speech Input/Output', In J LAVER & M A JACKS (Eds.), 'Proceedings of the European Conference on Speech Technology', Edinburgh, September 1987, Vol. 2. CEP Consultants Ltd.: Edinburgh 69-72 (1987)