Workpackage 3: Face to Face Communication

Objectives: To provide means for easier communication between deaf and hearing individuals in face to face transactions, such as post offices, banks, shops.

Deliverables to date:

Constrained PO system due month 7; delivered month 7.

Achievements to end of July 2000

Work to date has entailed the production of a prototype system which recognises a series of spoken phrases in a limited domain (that of post office counter transactions), translates the phrase into sign language and signs them to the customer. (Executive summary below)

Recordings of transactions performed in several post offices at different locations around the country were analysed and from this data the most frequently used phrases identified. An automatic speech recognition system was developed to recognise this restricted set of phrases, with provision to insert variable quantities within the phraes where necessary. (eg days of the week and monetary amounts.) The recogniser was implemented such that it may easily be trained to the voice of an individual user to provide high recognition accuracy despite the effects of background noise and multiple simultaneous speakers which are often found in the post office environment.

A database of motion captured data was recorded to enable the avatar developed in package 4a to sign the recognised phrases, and the avatar interfaced to the speech recognition system. The entire system was then evaluated in collaboration with the Post office and RND as detailed in WP6.

Progress towards next milestones

An investigation into the use of a less constrained recognition system, with free form speech input (still within the post office domain) being mapped to one of the pre recorded phrases. The NaturallySpeaking recognition engine from Dragon Systems is currently being evaluated as an alternative to the (now discontinued) Entropic recogniser. Methods for mapping from free form input to the signed phrases using techniques derived from telephone banking systems are also being investigated. No delays or deviations from the work plan have occurred or are expected with this workpackage.

ViSiCAST: constrained system for face-to-face communication in the Post Office

Executive Summary

A prototype system to enable a Post Office counter-clerk to communicate with a deaf or hearing-impaired customer using automatically-generated sign-language, and hence to aid completion of a transaction has been developed.

*A prio*ri, it might seem that recognising the clerk's speech and displaying it as text to the deaf customer would be an adequate aid to transactions. However, for many people who have been profoundly deaf from a young age, signing tends to be their first language and they learn to read and write more slowly. As a result, numbers of deaf people have below average reading abilities for English text. The system uses British Sign Language (BSL) rather than sign-supported English (SSE) as for the deaf community, SSE is unquestionably less popular than BSL.

Whereas systems to translate text from one spoken language to another are now readily available and work well within a restricted domain of discourse, translation from English text to a European sign-language is still a formidable research problem (this problem is being addressed within the ViSiCAST project). The approach taken to the translation problem in this system is to use pre-stored phrases and to pre-record the signs (as avatar movements) for these phrases. If only a small number of phrases is required, it is possible to record these in BSL. Phrases can be concatenated so that amounts of money can be inserted into a carrier phrase such as ``The cost is...".

Although this approach imposes considerable restrictions on the meanings that can be conveyed by the PO clerk and hence on the dialogue, it has the advantage that BSL can be used. Furthermore, the limited nature of the transactions in a Post Office means that most transactions can be completed in this way. Using pre-stored phrases also confers benefits: the speech recognition is very accurate because of the limited number of vocabulary items to be recognised and one can also be sure that the meaning of a phrase uttered is accurately translated into the target language. These gains are important ones, as the ``noise" introduced into the information channel by inaccuracies in the recognition process combined with ambiguities in the translation process can make more complex systems fail to translate correctly even simple phrases. By using pre-stored phrases, we in effect trade flexibility for accuracy. The system cannot currently recognise any signs from the customer, however this will be addressed in the third phase of the development of Tessa.

The PO clerk wears a headset microphone. The screen in front of the clerk displays a menu of topics available to him/her e.g. ``Postage", ``DVLA", ``Bill Payments ", ``Passports". Speaking any of these words invokes another screen showing a list of phrases relevant to this category which can be recognised. However, this is only an ``aide-memoire" to the clerk, and all phrases are active (i.e. can be recognised) at any time, so that switching between categories is seamless. Prior to designing the system, we obtained transcripts of recordings of PO transactions at three locations in the UK, in all 16 hours of business. Analysis of these transcriptions was essential for estimating the vocabulary which would be needed by our system to achieve a

reasonable coverage of the most popular transactions. At the end of this analysis, we prepared a set of approximately 130 phrases which we estimated were adequate to cover about 90\% of transactions.

The speech recogniser used was the Entropic HAPI (HTK Application Interface) system, which incorporates the HTK (Hidden Markov Model (HMM) Toolkit) recogniser [1,2,3]. A network of legal phrases is supplied to the recogniser, which uses a dictionary to decompose each word within a phrase into a sequence of triphones. Decoding of the speech signal is done using a Viterbi decoder that uses the speech models and the network supplied to output the most likely sequence of words given the acoustic input. The network constrains the speech recogniser to a finite number of predefined paths through the available vocabulary. These paths define the set of allowed phrases and consist of a start node (usually denoting silence, or background noise) followed by a number of word nodes or sub-networks (that define, for instance, the legal ways of saying the integers between one and 100), finishing with an end node (again denoting silence). Sub-networks are useful ways of defining phrase segments which can vary. For instance, a sub-network called ``one2hundred'' represents the legal ways of and this can be inserted at any appropriate point into the network. There are other sub-networks called ``amounts-of-money'', ``days-of-the-week''. By constraining the grammar in this way, recognition accuracy is significantly improved over using a looser grammar. Also, because the recogniser operates on a ``best-match'' basis, a phrase which is phonetically ``close'' but not identical with a phrase in the network will be recognized as the latter, which confers some flexibility on the speech of the clerk. (For instance ``Put that on the scales, please'' would be recognised as ``Please put it on the scales''). The network was constructed using a graphical network-building tool, grapHVite. This tool enables easy construction and editing of a network of phrases.

An important point about the operation of the recognition system is that both the speech models and the network can be varied. The speech models are adapted to the voice of each user using maximum-likelihood linear regression (MLLR) adaptation [4], a process which takes about twenty minutes, and the individual's models are then stored for later use. Speaker adaptation of the models greatly increases the recognition accuracy and hence the usability of the system.

Once the speech has been recognised, the correct sign sequence is selected by means of a lookup table which maps each spoken phrase to a motion capture file. For phrases with variable quantities such as ‘days of the week’ and monetary amounts, a series of files including signs for the appropriate day or value are played by the avatar, and the avatar seamlessly blends the various files together into a single sign sequence.

The motion-data stream is displayed using a virtual human. In common with many avatars, a three-dimensional ``skeleton'' is driven directly from the motion-data. The ``skeleton'' is wrapped in, and elastically attached to, a texture mapped three-dimensional polygon mesh that is controlled by a separate thread (event loop) that tracks the ``skeleton''. Currently we employ the RIVA TNT chip, by nVidia, to render the resulting 5000 polygons at 40 frames/s using Direct-X on a Pentium class PC. As a full three-dimensional model, the pose of Tessa can be changed on-the-fly by the user as can the identity of the virtual human and other characteristics. To give the best chance of creating smooth movements on a PC, Tessa was developed using

DirectX. Tessa is capable of signing in real time with a refresh rate of 35 frames per second.

The system has been formally evaluated by several members of the deaf community, in collaboration with the RNID. Informal demonstrations have also been conducted as part of the Post Office's DDA awareness road shows.

Bibliography

[1] S. Young, J.Odell, D. Ollason, V. Valtchev and P. Woodland. *The HTK book.* Cambridge University Technical Services Ltd., 1997.

[2] J. Odell, D. Ollason, V. Valtchev and D. Whitehouse. *The HAPI book*. Entropic Cambridge Research Laboratory Ltd., 1997.

[3] S.J. Cox. Hidden Markov models for automatic speech recognition. *British Telecom technical journal,* 6(2):105-115, April 1988.

[4] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuos density hidden Markov models. *Computer Speech and Langauge.* 9(2):171-185, April 1995.