STARDUST – Speech Training And Recognition for Dysarthric Users of Assistive Technology

Mark Hawley[1], Pam Enderby[2], Phil Green[3], Simon Brownsell[1], Athanassios Hatzis[3], Mark Parker[2], James Carmichael[3], Stuart Cunningham[1], Peter O'Neill[1], Rebecca Palmer[2]

1. *Department of Medical Physics and Clinical Engineering, Barnsley District General Hospital, UK,*
2. *Institute of General Practice and Primary Care, University of Sheffield, UK,*
3. *Department of Computer Science, University of Sheffield, UK*

Keywords: electronic assistive technology, speech recognition, training

Abstract

Automatic speech recognition (ASR) can provide a rapid means of controlling EAT. Off-the-shelf ASR systems function poorly for users with severe dysarthria because of the increased variability of their articulations compared to 'normal' speech. A two-pronged approach has been applied to this problem:

1. To develop a computerised training package which will assist dysarthric speakers to improve the recognition likelihood and consistency of their vocalisations.
2. To develop speech recognition systems which have greater tolerance to variability of speech utterances.

We present results of trials to evaluate the effect of the speech training aid on the speech of dysarthric individuals. Initial results have shown good speech recognition rates for people with even the most severe dysarthria. Speech command driven environmental control systems and voice output communication aids are being developed.

Introduction

A significant proportion of people requiring electronic assistive technology (EAT) have dysarthria, associated with their physical disability. Speech control of EAT is seen as desirable but machine recognition of dysarthric speech is a difficult problem due to the variability of articulatory output. Large vocabulary consumer automatic speech recognition systems have been used for people with mild and moderate dysarthria as a means of inputting text, but there is a lack of consensus over whether these systems are appropriate for people with severe dysarthria [1]. A number of speaker independent speech recognition algorithms have been reported which have been developed with the aim of improving the recognition of dysarthric speech patterns [2,3] but they have not appeared in a widely available form.

Small vocabulary speaker dependent speech recognition for control of assistive technology has been in the literature for more than twenty years [4,5]. Speaker dependent recognition is more appropriate to severe dysarthria since it allows the user to train the

system with their own utterances rather than requiring speech that is close to 'normal'. Although modern systems have a wider tolerance to variations in utterances, they are not successful at coping with the greater degree of variation present in dysarthric speech [6,7].

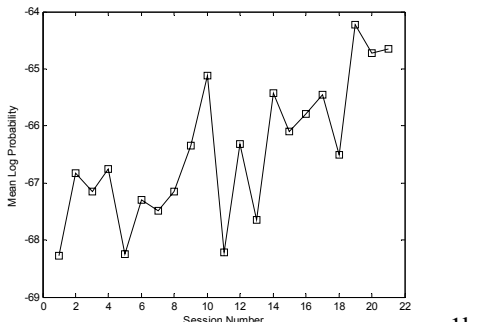
There is some evidence to suggest that speech training of the dysarthric speaker can improve their ability to accurately use speech recognition applications [8]. We have taken a two-pronged approach to addressing the problem of reliable use of speech as a control method for people with severe dysarthria:

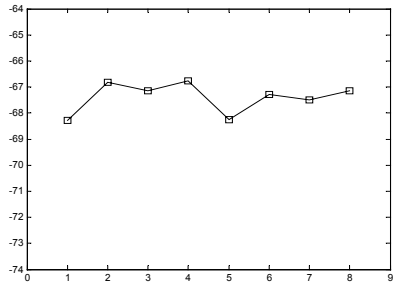
- To develop a computerised training package which will assist dysarthric speakers to improve the recognition likelihood and consistency of their vocalisations for a small vocabulary.
- To develop a speech recognition system which has greater tolerance to variability of speech utterances, by using the large corpus of data collected in the training phase.

Speech training

The speech training software is based upon the speaker dependent speech recogniser described below and runs on a personal computer. To set up the program, the recogniser is trained on a number of examples of each word from the EAT command vocabulary selected by the person and a model is built for each word. A 'best fit' utterance for each word is then determined by the software[1]. This 'best fit' utterance then becomes the target against which the person is trying to maximise their score by producing as close an articulatory approximation as possible.

1a





1c

Figure 1a the training aid display. Figure 1b and 1c - progress of 'closeness of fit' training scores over three weeks for two people with severe dysarthria.

In training mode, the word to be trained is displayed on the screen (e.g. 'Lamp', see figure 1a.). The user has three options, accessed by using a switch adapted to his/her needs. The user can play the 'best fit' example of the word through the computer speakers, can speak the word or can move on to the next word in the vocabulary list. If the user chooses

[1] The 'best fit' utterance is the utterance from the recogniser training set which the model would be most likely to produce. This utterance is not necessarily the most *intelligible* production of the word, but is the example that best approximates the person's most *likely* production.

to speak the displayed word the utterance is recorded and compared to the best-fit example by the recogniser. If the utterance has been recognised, the display (see figure 1a) then shows 2 bars. The height of the bar on the right represents the closeness of fit[2] score of the utterance to the model. The bar on the left represents the closeness of fit of the 'best fit' example to the model. The user is thus given an indication of how similar the utterance is to the most typical example taken from the training set. The user can then carry on practising the word, trying to raise the height of the right hand bar and use the play back facility for the best fit example so that they have an auditory target to aim for. When they have completed their attempt, the user can move on to the next word.

The aim of the training is three-fold. Firstly, in trying to make each utterance as close as possible to the target (ie maximise the closeness of fit), the user is increasing the likelihood of his utterances being recognised correctly. Secondly, in striving to imitate a stable target, the user is expected, through repetition, to reduce the overall variability in the production of these words. This is also expected to have a positive effect on the recognition accuracy. Thirdly, as each utterance is recorded during training, a large corpus of data is assembled, which can in turn be used as training data for new recognisers to improve the robustness and accuracy of recognition (see next section).

Five individuals with severe dysarthria have been provided with this training program to use at home. Figures 1b and 1c give examples showing the effects of the training program on the closeness of fit scores for two individuals using the program over a three-week period, which illustrate contrasting situations. In figure 1b, there is a general upward trend, suggesting that the user is able to gradually alter his speech to imitate a desired target in response to visual feedback. In figure 1c, the data shows no improvement in closeness of fit score. In our current experience of five users, three have shown improvement and two have shown no change. In assessing these results, it is worth noting that, often, people with stable dysarthria do not receive treatment by Speech and Language Therapists as it is thought that change can not be made, i.e. it is assumed to be a chronic condition. These results suggest that it is possible to decrease the degree of articulatory variability within the small vocabulary set used for this project in some cases of severe dysarthria. The computerised training methods utilised for this project may have implications for speech and language therapy with this client group.

Speech Recogniser

Initially, a small vocabulary recogniser has been developed to allow a limited number of control operations. The recogniser uses isolated words, as there is some evidence that people with severe dysarthria perform better with this type of recogniser than with continuous speech recognisers [9]. However, these words can be combined into command strings, which increases the number of operations that can be carried out. Since there is so much variation between individuals, speaker-dependent recognisers are trained for each individual. The HTK toolkit [10], using Continuous Density Hidden Markov Models [11], has been used for this project. The configuration of the recogniser has been optimised for the problem being addressed and additional tools have been developed to increase the accuracy of the recogniser[12].

Speech samples are initially collected using customised audio recording software and a high-quality array microphone, so that utterances are recorded as digitised data. Typically, thirty examples each of ten different command words are used as a training set to build the first recogniser. It is very difficult and time-consuming to collect speech samples

[2] The score is the log probability of the model generating the word on the Viterbi path.

as our subjects have physical problems that make the production of large amounts of speech on any given occasion very tiring. Sparsity of data tends to be problematic [12], since recognition accuracy tends to increase with the size of the training set. In the case of dysarthric speech and its greater variability, the sparse data problem is exacerbated. We are able to address this problem through the use of the training program described in the previous section. As it is used, the program records all examples of utterances used in training and these new speech samples can be used to increase the amount of training data in subsequent versions of the recogniser, thus facilitating the collection of large data sets for each individual.

Table 1 shows recognition results for three speakers, two with severe dysarthria and a control with 'normal' speech. The first column shows the level of intelligibility of these speakers for individual words and for sentences, measured using the protocol from the Frenchay dysarthria test [13]. Subject 1 has severe dysarthria and is totally unintelligible to an unfamiliar listener, whereas subject 2 is more intelligible but still severe. The results show that, as expected, the accuracy of speech recognition improves as the amount of training data increases (where recognition is not already 100%). The recognition accuracy of the STARDUST recogniser is very promising, given the severity of dysarthria for the test subjects. For comparison, the recognition accuracy of a popular commercial speech recognition environmental control system is shown[3].

	Intelligibility Single words-sentences	STARDUST recogniser N=6	STARDUST recogniser N=20	STARDUST recogniser N=28	Commercial speech recognition ECS N=6
Subject 1	0% - 0%	64%	80%	85%	60%
Subject 2	22% - 34%	100%	100%	100%	80%
Control	100% - 100%	100%	100%	100%	98%

Table 1 - Intelligibility and recognition accuracy for two people with severe dysarthria and control with 'normal' speech. N denotes the number of examples of each word used to train the recogniser.

Assistive Technology

Two demonstration applications of the STARDUST system have been developed: an environmental control system (ECS) and a voice output communication aid. The computer is provided with two inputs: a high quality microphone, which can be head-mounted or a remote array; and a switch, adapted to the individual needs of the user. The system requires the user to press the switch to activate the ASR program. The speaker then says the command sequence, for example, '**TV volume up**' with sufficient pause between the words so that the recogniser does not treat them as one word. The recogniser recognises the individual words and parses the command, which is then converted to a code and sent via the serial port to an external infra-red sender unit.

The system is also provided with an additional interface, so that the user can choose to use the switch alone to control the interface. Thus, if the switch is held down for a period longer than a pre-set time, the computer displays a scanning interface that can be used to select the desired operation. This facility has been added because it is acknowledged that speech recognition can never be 100% accurate in a home environment. The switch-based operation is important for two reasons: firstly, because users get frustrated with speech

[3] Tests use identical data in the same controlled conditions to that used for the STARDUST recogniser at N=6.

recognition if its accuracy is perceived as insufficient. This may be temporary if, for example, the background noise level is high and a switch interface can be used to overcome this temporary problem. Secondly, some ECS equipment is safety critical, such as the need to urgently summon assistance, and a reliable back-up must be provided as an alternative to speech recognition.

A trial of the STARDUST ECS is planned. It will be provided to eight people with severe dysarthryia who currently use commercial ECS. The trial will examine the speed of use and accuracy of the STARDUST ECS compared to the triallers' current technology. A field trial will also examine the usage of the STARDUST technology (again compared to current technology) and its usability and acceptability amongst the triallers.

Acknowledgements

This research was sponsored by the UK Department of Health New and Emerging Application of Technology (NEAT) programme and received a proportion of its funding from the NHS Executive. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health or the NHS Executive.

References

1. Hawley MS (2002) Speech Recognition as an Input to Electronic Assistive Technology, *British Journal of Occupational Therapy*, 65(1), 15-20
2. Deller Jr JR. Hsu D. Ferrier LJ (1991) On the use of hidden Markov modelling for recognition of Dysarthric speech. *Computer Methods & Programs in Biomedicine,* 35(2), 125-139.
3. Jayaram G, Abdelhamied K (1995) Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research & Development*, 32(2), 162-169
4. Clark JA, Roemer RB (1977) Voice controlled wheelchair. *Archives of Physical Medicine & Rehabilitation.* 58(4), 169-75
5. Cohen A, Graupe D (1980) Speech recognition and control system for the severely disabled. *Journal of Biomedical Engineering.* 2(2), 97-107
6. Cook AM, Hussey SM (1995) *Assistive Technologies: Principles and Practice*, St Louis: Mosby
7. Mendez-Pidal X, Polikoff JB, Peters SM, Leonzio JE, Bunnell HT (1996) The Nemours database of dysarthric speech. *Applied Science and Engineering Laboratories (ASEL). AI DuPont Institute. PO Box 269. Wilmington. DE 19899. USA*, Avaiable on http://asel.udel.edu/speech/reports/icslp96a/a737.pdf
8. Kotler, A., Thomas-Stonell, N (1997) "Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment" *Journal of Augmentative and Alternative Communication*, 12: 71-80.
9. Rosen, K., Yampolsky, S (2000) "Automatic Speech Recognition and a Review of Its Functioning with Dysarthric Speech", *Journal of Augmentative and Alternative Communication*, 16: 48-60.
10. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtcho, V., Woodland, P., *The HTK Book (for Version 3.1)*, Cambridge University Engineering Department, 2002.
11. Rabiner, L., "A tutorial on hidden Markov models and selected applications in speech recognition" *Proceedings of the IEEE,* 37: 257-286, 1989
12. Green P, Carmichael J, Hatzis A, Enderby P, Hawley M, Parker M (2003) Automatic Speech Recognition with Sparse Data for Dysarthric Speakers, submitted to Eurospeech 2003
13. Enderby P.M., *Frenchay Dysarthria Assessment*, Pro-Ed, Texas, 1983.