Speech Technology for e-Inclusion of People with Physical Disabilities and Disordered Speech

Mark S. Hawley*, Phil Green†, Pam Enderby^, Stuart Cunningham§ and Roger K. Moore+

*Dept. of Medical Physics and Clinical Engineering, Barnsley Hospital, Barnsley, UK
†Dept. of Computer Science, ^Institute of General Practice and Primary Care,
§Dept. of Human Communication Science, University of Sheffield, Sheffield, UK

Abstract

Speech technology is potentially of enormous benefit to people with physical disabilities. Applications of speech technology to e-inclusion are reviewed and described in the areas of access, control, communication and rehabilitation/therapy, with particular reference to speech technology developments for people with disordered speech. To be successful, applications should effectively take into account the needs of user groups and have the ability to adapt to the needs of individuals. This is a challenging area but effective progress can be made through multi-disciplinary research and development.

1. Introduction

People with physical disabilities take advantage of a variety of methods to gain access to information technology and to electronic assistive technology for communication, mobility and daily living tasks. Many of these access methods are slow and can lead to frustration, a prime example being the use of switch-activated menu scanning (hereafter referred to as switch-scanning).

Automatic Speech Recognition is potentially of enormous benefit to people with severe physical disabilities. The tremendous richness of human speech communication gives the user many degrees of freedom for control and input. The speed of speech recognition also gives it a potential advantage over other input methods commonly employed by physically disabled people.

People with neurological conditions causing disability often have associated dysarthria, which is the most common acquired speech disorder affecting 170 per 100,000 population [1]. This may be developmental dysarthria such as that associated with cerebral palsy or acquired dysarthria associated with progressive neurological disease (e.g. Parkinson's disease, motor neurone disease, and multiple sclerosis) or sudden onset conditions such as/or head injury. In its severest form, dysarthric speech is unintelligible to others and may take the form of producing vocal utterances, rather than words recognisable to unfamiliar communication partners. The combination of speech and general physical disability can make it particularly problematic for people to interact in their environment and can severely limit independence and inclusion.

This paper examines some of the areas for e-inclusion of people with disabilities that can benefit from the use of speech technology. These areas include access, control, communication, rehabilitation and therapy. In all areas, the use of speech technology for people with dysarthric speech is particularly examined and discussed.

2. Access

Computers and the internet are becoming ubiquitous and necessary tools for modern living and people with disabilities require access to these tools on an equal basis to those who are able-bodied. Many people with severe disabilities and normal speech (e.g. spinal injuries) use speech recognition as a means of inputting text, as it can give faster input rates than the adapted keyboard or switch-scanning alternative. However, it is also the case that many such people prefer the non-speech alternative as they find speech recognition frustrating (due to less than perfect recognition) or a non-intuitive way of composing text.

There is general consensus in the research literature that speech recognition can be successfully used by people with mild to moderate dysarthria [2][3][4]. The study by Ferrier examined speech recognition using Dragon DictateTM for ten adults with cerebral palsy compared to two control subjects [4]. Intelligibility measures of dysarthric speech correlated with recognition success, showing that people with good intelligibility were more successful at using speech recognition and achieved over 90% recognition accuracy within eight dictation sessions.

Large vocabulary speech recognition systems are less successful for people with severe dysarthria. Recognition accuracy reported in the literature varies from 22% to 78% under test conditions [2][3][4][5][6]. It is not a straightforward task to compare recognition rates due to the variety of different recognition methods and recogniser training methods used and the uncertainty in comparisons of severity of dysarthria.

Success with speech recognition for people with severe dysarthria varies with the individual, though some more general factors have been noted. For example, Doyle noted that males with dysarthria generally performed worse than females in speech recognition tasks [7] whilst Ferrier noted that the more intra-word pauses and non-speech sounds that were present, the longer the individual took to achieve a given recognition rate in tests using repeated words [4].

Technology and other factors also affect implementation success. It appears that many people with disabilities which result in breathing or speech difficulties perform better with discrete word than with continuous speech recognisers. A high level of support to the user in training the recogniser is also crucial if acceptable recognition rates are to be achieved. Despite the lower recognition rates achievable by people with severe speech problems, the resulting low production rates may still be better than is achievable using alternative methods such as switch scanning [2][8].
### 3. Control

The ability to control the home is an essential aspect of independence and e-inclusion. Environmental Control Systems (ECSs) are available which address many elements of home management for disabled people, such as control of audio-visual equipment, telephones, doors and curtains as well as the ability to summon assistance. Most ECSs utilize switch-scanning or key-pad interfaces for control. Whilst key-pad interfaces give a reasonably fast control method, switch-scanning can be very slow. More recently, ECS with speech recognition have been introduced and a number of such systems are available on the market. Although recognition rates have improved since the very early systems were introduced [9], clinical experience indicates that a significant proportion of switch-scanning users still reject speech recognition as a control method, despite it being quicker, for two reasons. Firstly, speech recognition is less accurate than switch-scanning – the threshold for acceptance appears to be of the order of 80% accuracy and this is often not achieved in the home environment. Secondly, false activation by environmental noise is a frequent occurrence and a source of significant frustration.

Whereas speaker-dependent speech recognisers can perform adequately as a control input for normal speech, their performance with dysarthric speech is significantly poorer [10][11]. Consequently, people with speech problems are effectively excluded from using this potentially faster control method. Recent research in the STARDUST (Speech Training And Recognition for Disabled Users of Assistive Technology) project has shown that by changing the recogniser building methodology, it is possible to make use of fairly conventional speech recognition techniques to greatly improve the recognition of dysarthric speech and apply this to produce efficient home control [12].

For STARDUST, a small vocabulary recogniser has been specified to allow a limited number of control operations. The recogniser uses isolated words as its recognition units, as there is some evidence that people with severe dysarthria perform better with this type of recogniser than with continuous speech recognisers [13]. Since there is so much variation between individuals, speaker-dependent recognisers are trained for each individual.

The HTK toolkit [14], using Continuous Density Hidden Markov Models [15], has been used for this project. We chose whole words as our modelling units because the phonetic abnormality of severely dysarthric speech prohibits the definition of reliable sub-word units. The models we use are quite standard and take Mel-frequency cepstral coefficients (MFCCs) as their acoustic vectors. What is different is the methodology for building recognisers, which is adapted to deal with the scarcity of training data and the increased variability of the material which is available. We have addressed this problem by closing the loop between recogniser-training and user-training. We start by recording a small amount of speech data from the client: typically 10 utterances of each word. We train a recogniser using this data and then use it to drive a user-training application, which allows the client to practice to improve consistency of articulation. The visual feedback provided to the client is derived from the goodness-of-fit to the appropriate word model, obtained by forced-alignment. As the client practices, s/he supplies more speech data, which we can then use to train a new recogniser, which can in turn be used in a further cycle of user-training, and so on. Additional tools for dealing with sparse training data were described by Green et al [10].

#### Table 1: STARDUST word and control phrase recognition accuracy for severely dysarthric users in home trials with normal domestic noise conditions.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Sex</th>
<th>Diagnosis</th>
<th>Intelligibility Rating (%)</th>
<th>Recognition Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Range</td>
<td>Mean</td>
</tr>
<tr>
<td>1</td>
<td>F</td>
<td>Cerebral Palsy (CP)</td>
<td>0-20</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>Multiple Sclerosis</td>
<td>0-50</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>CP</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>CP</td>
<td>10-40</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>CP</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We have conducted home trials with severely dysarthric ECS users to evaluate the user-training program and the speech-controlled STARDUST ECS. The training phase increased the recognition rate from 88.5% to 95.4% (p<0.001). Recognition rates were good for people with even the most severe dysarthria in everyday usage in the home (see Table 1); mean word recognition rate was 86.9% in normal domestic noise conditions. The speech-controlled ECS was less accurate (mean task completion accuracy 78.6% vs. 94.8%) but faster to use than switch-scanning systems, even taking into account the need to repeat unsuccessful operations (mean task completion time 7.7 vs. 16.9 seconds, p<0.001).

### 4. Communication

Speech technology has an important role to play in support for spoken communication for people who have unintelligible speech or no speech. Communication aids range from simple devices which play back a small number of messages stored as recorded speech, to very sophisticated devices allowing access to large vocabularies with synthesised speech output.

Much speech technology research has concentrated upon speech synthesis and particularly on methods for increasing the ‘naturalness’ of synthesised speech. As in the field of automatic speech recognition (ASR), text-to-speech synthesis (TTS) has benefited from the introduction of a data-driven approach in which corpora of annotated speech recordings are used to estimate the parameters of the underpinning models and, in the majority of contemporary systems, to provide an inventory of acoustic segments from which selected units are concatenated together to produce the output speech. Whilst current systems are impressively natural in comparison to their earlier rule-based counterparts, the reliance on recorded corpora limits the number of available voices and accents, and contributes to general lack of expressiveness. Also, model-
based TTS systems have been shown to be more intelligible and comprehensible than concatenative systems, especially in a noisy environment [16]. This means that there is still some way to go before spoken language output becomes a truly ubiquitous and widely deployed technology that is capable of talking clearly, intelligently, expressively, appropriately and realistically [17].

Given the ability for speech recognition to provide a relatively fast and reliable control method, and the availability of high quality speech synthesis, there is the potential to combine these technologies into a speech-controlled communication aid. Whilst this has been suggested previously (e.g. [18]), no successful implementation has been reported, to our knowledge. This idea is being explored in a new research project funded by the UK Department of Health, known as VIVOCA (Voice-Input Voice-Output Communication Aid) [12]. The project aims to develop a portable speech-in/speech-out communication aid for people with disordered or unintelligible speech, initially concentrating on people with moderate to severe dysarthria. The intention is to provide the user with a device that performs like a human interpreter, recognising the disordered speech and synthesising an intelligible equivalent. This will be of use to people in social situations where they interact with non-familiar communication partners, such as at work, when shopping, in the hospital, on the telephone etc. This application will require development of the work carried out under STARDUST, as follows:

- Application of techniques shown in STARDUST to be beneficial for the recognition of disordered speech to more challenging recognition situations. The systems developed in the STARDUST project were for use in the home and were trained to recognise a small number of words, typically 10. In this application, the device will be used in a wide range of environments, and will need to recognise a larger number of words.
- Development of novel user interfaces and means of translation, whether through direct translation, coding or both, to optimise interaction with the device, and to provide a more accessible form of spoken communication.
- Production of a discrete, portable, and ideally body worn device to facilitate more natural communication, including eye contact during speech production, for speakers with disordered speech.
- Development of techniques for adaptation as speech characteristics and user requirements change over time. For instance, the speech of people with progressive neurological disease (e.g. motor neurone disease) can change in ways which would confound conventional recognition techniques.

5. Rehabilitation/Therapy

Improving motor performance requires repeated practice, with accurate and consistent feedback. Much of the skill of a speech and language therapist is devoted to this: where is the client going wrong? How does her/his articulation compare to a normal production? Is the next attempt better or worse than the last? Providing meaningful feedback is extremely labour-intensive and particularly difficult to provide through speech and language therapy for persons with dysarthria. The auditory perception system naturally adapts to a speaker's communicative efforts and thus the therapist may not be accurate in her/his feedback to the patient's attempts, which may be perceived by the therapist as improving when in fact, the patient is repeating a similar error. There is obvious attraction in the possibility of developing software to play this role objectively, allowing the client to practice without a therapist being present.

The problem is that feedback based on the goodness-of-fit of an articulation to the corresponding recognition model may not correspond to the production-and-perception based feedback which a therapist would supply: for instance the match against a statistical word model is likely to be dominated by longer-duration vowels rather than the abrupt consonants one is trying to teach. So the feedback provided in STARDUST (§3) is aimed not at therapy but at improving consistency of production. Nevertheless, experience with the STARDUST training protocol shows that it can provide a method of individualised therapy targeting a personally selected vocabulary, which can be practised without a therapist in attendance and providing consistent feedback.

Speech is a highly integrated system, and therefore it is not surprising that one area of dysfunction can have an effect upon other areas, which although not impaired might not be able to function normally as a result of secondary effects. Many patients with severe dysarthria demonstrate inconsistent attempts at articulating the same sound combinations. The range of inconsistency is associated with reduced intelligibility. Familiar listeners are able to understand persons with dysarthria, as they become attuned to the pattern of inconsistency and factors that affect this. Reducing inconsistency not only assists improve the functional consequences for access to e-technology but also can impact on vocal effectiveness. Thus technology mentioned in this paper is not only seen as replacing or improving a person's control over their environment by using technology, but has also been seen to coincidentally improve the actual impairment forming the barrier.

Rather than basing feedback on recognition decisions, the OPTACIA method [20], developed in the EC-funded project OLP, provides real-time visual feedback related to the instantaneous acoustics. As the client speaks a sprite moves between target areas defined on a window acting as a kinematic map. The conversion from acoustics to map position is effected by training a neural net, and OPTACIA method [20], developed in the EC-funded project OPTACIA can be specialized to the needs of the client. OPTACIA has been used, for instance, by Oster et al [21] in teaching pronunciation of Swedish fricatives to deaf children.

6. Discussion

As the above analysis indicates, there are many ways of applying speech technology to enhance e-inclusion. It is also important to note, however, that advancement in technology can, and often does, lead to the further e-exclusion of large numbers of disabled individuals. For example, when the GUI or graphical user interface was introduced as the ubiquitous user interface for computers in the 1980s, many users who were previously able adequately to navigate text-based operating systems found that they were unable to navigate the GUI as they were physically unable to use a mouse. An analogy can be drawn with progress towards connected speech recognition: although this is advantageous for much of the mainstream population, some individuals with speech problems can only use discrete word recognition – a
technology that is now regarded as almost redundant for mainstream users.

Key to producing successful technologies for e-inclusion is effectively to take into account the characteristics of the user population. The concept of ‘design for all’ aims to produce technologies usable by as wide a section of the population as possible. There will still be some groups whose needs cannot be incorporated in designs aimed at a wide spectrum of users, however, and assistive technology developments must be carried out specifically for the needs of the user groups in question. In order to take into account the characteristics of the user population, technology development projects should include a detailed study of the needs of the user population in order to inform technology design [19]. It is also crucial that iterative evaluation with a range of potential users is carried out in order to refine the technology.

Another key requirement is to build in the capability to tailor technology to the needs of the individual user. Whilst commercial speech-to-text recognisers adapt to individual speech characteristics to a certain extent, they are only able to cope with speech that is close enough to their ‘normal’ models. Experience with people with severely disordered speech is that recognisers based on ‘normal’ models are unsuitable and that high recognition rates can only be obtained by employing methodologies that maximize the ‘individualisability’ of the technology, including the use of speaker dependent recognisers and closing the loop between recogniser-training and user-training.

The development of speech technology for e-inclusion of people with physical disability and disordered speech is a challenging area. Making effective progress requires knowledge of the disabled user, of assistive technology and of speech technology, and thus necessitates a multidisciplinary team approach.

7. References