A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment

Mark S. Hawley, Stuart P. Cunningham, Phil D. Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O'Neill

***Abstract*—A new form of augmentative and alternative communication (AAC) device for people with severe speech impairment—the voice-input voice-output communication aid (VIVOCA)—is described. The VIVOCA recognizes the disordered speech of the user and builds messages, which are converted into synthetic speech. System development was carried out employing user-centered design and development methods, which identified and refined key requirements for the device. A novel methodology for building small vocabulary, speaker-dependent automatic speech recognizers with reduced amounts of training data, was applied. Experiments showed that this method is successful in generating good recognition performance (mean accuracy 96%) on highly disordered speech, even when recognition perplexity is increased. The selected message-building technique traded off various factors including speed of message construction and range of available message outputs. The VIVOCA was evaluated in a field trial by individuals with moderate to severe dysarthria and confirmed that they can make use of the device to produce intelligible speech output from disordered speech input. The trial highlighted some issues which limit the performance and usability of the device when applied in real usage situations, with mean recognition accuracy of 67% in these circumstances. These limitations will be addressed in future work.**

***Index Terms*—Augmentative and alternative communication, automatic speech recognition, dysarthria, voice output communication aid.**

I. Introduction

SPOKEN language communication is a fundamental factor in quality of life, but as many as 1.3% of the population cannot use natural speech reliably to communicate, especially with strangers [1]. For instance, the speech of people with moderate to severe dysarthria—the most common speech disorder affecting 170 per 100 000 of population [2]—is usually unintelligible to unfamiliar communication partners. For these people, their speech impairment can preclude them from interacting in a manner that allows them to exploit their potential in education, employment and recreation.

Speech impairment is often associated with severe physical disabilities as a result of progressive neurological conditions such as motor neurone disease, congenital conditions such as cerebral palsy, or acquired neurological conditions as a result of stroke or traumatic brain injury. Current technological tools for communication, voice-output communication aids (VOCAs), generally rely on a switch or keyboard for input. Consequently, they can be difficult to use and tiring for many users, and they do not readily facilitate natural communication as they are relatively slow and disrupt eye contact [3]. O'Keefe *et al.* [4] report that users need a device which is physically easy to operate in a wide range of positions and environments. Many people with VOCAs often prefer to speak rather than use the aid, even if their speech is largely unintelligible, as it is a more natural form of communication [5]. In addition, Todman *et al.* [6], found that listeners rated users of a communication aid as more socially competent if they had a more rapid rate of delivery. It is therefore desirable that a new communication aid retain, as far as possible, the speed and, ideally, the naturalness of spoken communication.

Despite its apparent attractiveness as an access method, the potential complications of recognizing impaired speech have meant the prospect of spoken access to technology remains unfulfilled. Commercially available automatic speech recognition (ASR) systems can work well for some people with mild and even moderate dysarthria [7] and [8], but these studies show that there is an inverse relationship between the degree of impairment and the accuracy of speech recognition. For people with severe speech impairment, commercial speech recognition systems are not a viable access solution. Moreover, the small-scale laboratory experiments reported in [7], [8] do not represent the range of environmental conditions that are likely to be encountered in realistic usage, which is known to degrade recognition accuracy.

Thus, while ASR has been used for many years as a method of access to technology by some people with disabilities but unimpaired speech, it received little attention as a potential input channel for VOCAs. Previous prototypes of voice-input voice-output communication aids (VIVOCAs) have been reported, but have not been tested extensively with users or reached the stage of becoming available as commercial products [9], [10]. A different approach has been proposed by Wisenburn and Higgin-

Manuscript received October 25, 2011; revised April 02, 2012; accepted May 11, 2012. Date of publication August 03, 2012; date of current version January 04, 2013. This work was sponsored by the U.K. Department of Health New and Emerging Application of Technology (NEAT) programme under Grant E090. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health.

M. S. Hawley, P. Enderby, and R. Palmer are with the School of Health and Related Research, University of Sheffield, S1 4DA Sheffield, U.K. (e-mail: mark.hawley@sheffield.ac.uk; p.m.enderby@sheffield.ac.uk; r.l.palmer@sheffield.ac.uk).

S. P. Cunningham and S. Sehgal are with the Department of Human Communication Sciences, University of Sheffield, S1 4DA Sheffield, U.K. (email: s.cunningham@sheffield.ac.uk; s.sehgal@sheffield.ac.uk).

P. D. Green is with the Department of Computer Science, University of Sheffield, S1 4DA Sheffield, U.K. (e-mail: p.green@dcs.shef.ac.uk).

P. O'Neill is with the Assistive Technology Team, Barnsley Hospital NHS Foundation Trust, Barnsley, S75 2EP Barnsley, U.K. (e-mail: peter.oneill@sheffield.ac.uk).

Digital Object Identifier 10.1109/TNSRE.2012.2209678

botham who explored the potential for using speech recognition in a VOCA to recognize the speech of a conversation partner [11], [12]. This process was then used to present suggested utterances to the speech impaired user based on what their conversation partner had said.

In recent years there has been a realization that, for people with severe speech impairment, an alternative approach to ASR must be followed. The authors' previous work has been successful in developing speech controlled interfaces to home control systems (also known as environmental control systems or ECS) for people with severe dysarthria [13]. In this work, we applied statistical ASR techniques, based on hidden Markov models (HMMs), to the speech of severely dysarthric speakers to produce speaker dependent recognition models, and developed a novel methodology for recognizer-building. This approach relied on a user-training phase in which the user practised speaking to the recognizer, whilst receiving consistent visual feedback based on the similarity between their current attempt and the distribution of their previous attempts. This enabled the user to become more efficient at producing the target utterances, by reducing variation in their vocalizations, while at the same time facilitating the collection of additional speech examples that were then used to train the final recognizer. These enhancements resulted in speech recognition being a viable means of controlling assistive technology for small input vocabularies, even for people with severe speech disorders [13]. More recently, Sharma and Hasegawa-Johnson [14] have demonstrated that maximum a priori (MAP) adaptation from speaker-independent ASR can improve recognition rates, sometimes producing better performance than the equivalent speaker-dependent ASR, though this has not yet been applied in an assistive technology context.

This paper describes the development of a VIVOCA which is intended to recognize and interpret an individual's disordered speech and deliver the required message in clear synthesized speech.

II. System Description

The development made use of a user-centred design and development paradigm. An initial detailed user requirements study considered the views of both potential VIVOCA users and of speech and language therapists/pathologists who provide voice-output communication aids [15]. A wide range of user requirements were elicited and the VIVOCA was implemented to meet these requirements where feasible. The development process was iterative and the implementation was gradually refined by testing developments with a group of four potential VIVOCA users. These four users were people with moderate or severe dysarthria, with one individual having additional verbal dyspraxia. They tested the device at each stage and gave feedback to allow us to improve the VIVOCA.

Fig. 1 shows a schematic of the system and its major components. The user speaks into a microphone and the speech is processed and recognized by a speech recognizer. The recognized words are passed to a message building module. Dependent on this input, the message building module will update the screen, potentially supply audio feedback to the user, and determine the range of possible future inputs. This process continues in an iterative fashion as the user builds their message. When the message is complete it is passed to the speech synthesizer, producing intelligible spoken output via a speaker. The system components are described below.

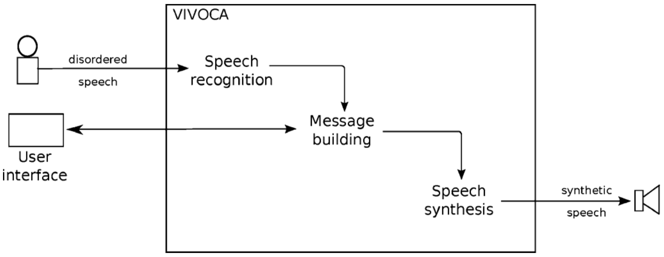

Fig. 1. Schematic diagram of the voice-input voice-output communication aid (VIVOCA).

A. Speech Recognition

In prevailing methods, automatic speech recognition (ASR) is based on statistical models (usually HMMs) of speech units. These models are trained on a large corpus (perhaps hundreds of hours) of data recorded by many speakers. For a large vocabulary system, the speech units will be at the level of individual speech sounds, phones. The resulting speaker-independent recognizer can be adapted for an individual speaker, given a small amount of enrolment speech data from that speaker. However, this ASR technique is unsuitable for speakers with severe speech disorders because the amount of material available for training is severely limited (as speaking often requires great effort), the material is highly variable, often has a limited phonetic repertoire, and is too different from the "normal" speech used in training speaker-independent models for many conventional adaptation techniques to be of assistance. Instead, we have introduced a new methodology for building small vocabulary, speaker-dependent personal recognizers with reduced amounts of training data. Using this approach, which we outline below, accurate recognition of severely dysarthric speech has been shown to be feasible for relatively small vocabularies [13].

Initial recordings were collected from the user. Depending on their preference these were collected using either a headset microphone (Sony-Erikson Akono HBH-300 headset connected via Bluetooth), or a desktop microphone (Acoustic Magic Voice Tracker array microphone), connected to a laptop computer (Dell Inspiron 1100). Signals were sampled at 8 kHz, the maximum sampling frequency on the Bluetooth audio channel.

The recordings consist of isolated productions of each of the words that are required for the recognizer's input vocabulary. These examples are used to train the initial whole word models. In this study we used HMMs with 11 states, with a straight-through arrangement. The acoustic vectors were 12 Mel-frequency cepstral coefficients (MFCCs) derived from a 26-channel filterbank with a 25 ms analysis window and 10 ms frame-rate. Energy normalization and cepstral mean normalization were also applied to the input features. This is a conventional ASR front-end. The models were trained using the HMM toolkit [16] with the Baum-Welch algorithm.

This approach is straightforward for a typical speaker, but it is more problematic for the intended users of a VIVOCA due to their speech impairment. This means there is a scarcity of training data and the consequences of this are exacerbated by the variability in the productions of speakers with dysarthria [17]. However, the approach described in detail in [13] is to use the initial recordings to estimate models that can be incorporated into a "user-training" application. This application prompts the user repeatedly to speak each of the words in the initial recognition vocabulary. Each utterance is recorded, but crucially the user is given feedback on "closeness of fit" of each attempt to their own recognition model. This is determined from the log probability of the model generating the word by the most likely path (computed by the Viterbi algorithm) [17], [18]. The user is also guided in this process by being able to listen to their "best attempt so far," so that they may attempt to replicate it. Here, "best" means that example which the current word model would be most likely to generate. Previous studies have shown that both listening to their previous best attempt, and repeated practice can have a beneficial effect in stabilizing the target utterance [19], and this has the additional benefit of providing additional training examples for retraining recognition models [20].

At the conclusion of this user-training step, the recognition models can be re-estimated using both the initial training examples, and subsequent examples collected with the user training application, producing a recognizer which is more accurate and robust to variations in the user's speech. The process can, of course, be repeated.

We have previously found that recognition accuracies above 80% for isolated words, and above 70% for commands (short strings of words) are consistently attainable for small vocabularies of severely dysarthric speech [13]. Whilst home control tasks can be carried out with a relatively small number of control inputs (and small input vocabulary of around 10–15 words), supporting speech communication requires more flexibility in its output and is therefore likely to require a larger input vocabulary. For speaker-dependent recognition it is known that word recognition accuracy falls with increasing vocabulary size [21], and this reduction is likely to be exacerbated when speech input is highly variable, as is the case with dysarthric speech. Therefore, a major challenge in this work is to be able to accommodate larger input vocabularies whilst retaining acceptable levels of word accuracy.

B. Message Building

The message building module constructs messages, which the user wishes to communicate, from the recognized input words. Using input speech to drive output speech from a VOCA presents a new challenge which has not been addressed in any depth in previous research. The simplest, and in many ways the ideal, form of message building, given that we are recognizing word units, would be to recognize each word individually and speak out the same word in a clearer (synthesized) voice. However, since the accuracy of the recognition of severely dysarthric speech decreases rapidly as the input vocabulary size increases, this is not currently possible and we are in practice constrained to find methods which generate meaningful messages but which require relatively small input vocabularies. We need to constrain the "perplexity" of the recognition task, that is the number of words which the recognizer must choose between at a given point in the process. Doing this also has the advantage of making it easier for the user to recall what the recognizer will accept at any point.

Given this constraint, we considered a number of candidate message building methods drawn from augmentative and alternative communication (AAC) [22] as well as from other means of coding language, such as spelling or texting, for their suitability to be used in the VIVOCA.

One of the key considerations is the speed at which people are able to communicate [23]. We define communication rate as the number of words per unit time which are generated correctly according to the intentions of the user. In this definition, the correction of errors must be taken into account, and this is a major consideration when dealing with speech recognition, as error rates are high compared to most other input methods.

Modelling communication rate for different message building methods showed that, in conditions of high recognition accuracy, methods with the larger input vocabularies give greater communication rates, in line with expectations. However, due to the time cost of correcting errors, communication rate falls off rapidly with decreasing recognition rate [24]. By combining the models of communication rate with information on recognition accuracy for a range of input vocabulary sizes, we are able to estimate the communication rate of the message building methods for different levels of severity of dysarthria. For individuals with mild and moderate dysarthria, the positive relationship between input vocabulary and communication rate is retained. However, for those people with more severe dysarthria, this is not the case. As a result of the reduction of recognition rate with increasing input vocabulary, the positive relationship between input vocabulary and communication rate no longer holds. For people with severe dysarthria, message building techniques requiring a smaller input vocabulary are, counter-intuitively, more efficient.

As part of the user-centred design and development process, design meetings were held between the research team and potential users at which different message-building methods were considered, with knowledge of the modelled communication rates. Users prioritized methods which tended to have high communication rate. Some users, however, also regarded a large output vocabulary as being vital regardless of its effect on communication rate. We therefore decided to implement a hybrid translation method as a combination of phrase building and spelling, as follows.

Phrase building is used to generate frequently used phrases requiring rapid generation, such as answering the phone, conversational fillers or communicating immediate needs/problems. For example, inputting the sequence of words "want" "drink" "water" could generate the phrase "Can I have a drink of water please." Using this approach in a structured way greatly reduces the recognition perplexity. Spelling may be used for the remainder of less-frequently used words allowing unlimited output vocabulary where greater precision and conversational range are required, though at the cost of greater perplexity and much lower communication rate.

The components of the message building method, and the input and output vocabularies were individually tailored to the needs and wishes of each participant. For instance, they were able to choose a configuration that could be used to map words onto phrases or allow the spelling of words or a combination of both. It is envisaged that, once the VIVOCA becomes more widely available, this individualization will be carried out by the user's speech and language pathologist or similar clinical professional.

C. Speech Synthesis

One of the user requirements for the system output was that it should be possible to have both prerecorded and synthetic output, and that the synthetic output should be as natural sounding as possible.

To satisfy these requirements, the system software was designed to work with both prerecorded output (in the form of waveform files) and to interface with a speech synthesizer. As with the process of speech recognition, speech synthesis is a computationally demanding process. For the development of the prototype we utilized a small footprint speech synthesizer designed for mobile computer platforms called Flite [25], which is a variant of the larger and popular synthesis system known as Festival [26]. A specially compiled version of the Flite software was prepared for the Windows Mobile for Pocket PC operating system.

The acceptability of the output was evaluated with potential users. Several users preferred prerecorded output for their system, either because of the potential for delay introduced by the computationally intensive synthesis process, or the perceived poor quality of the synthetic speech. In some cases we made use of the high-quality synthesis provided by the Festival system to synthezize the required outputs and store these as waveform files that could be played out on the device when required.

D. Hardware and Software Implementation

In order to meet users' requirements, the hardware upon which VIVOCA was implemented needed to be small and light and have a suitable visual interface. When the VIVOCA development began, in 2005, the most suitable hardware was judged to be a personal digital assistant (PDA). The models used in this study were the HP iPAQ HX2700 running Windows Mobile 5.0 for Pocket PC. The PDA takes voice input from the user via a microphone, which can either be head-worn (Bluetooth or wired) or lapel-type, or the internal microphone of the PDA. The PDA's internal speaker was found to produce speech at too low a volume for practical use in any but quiet ambient conditions. Therefore, a separate amplifier and speaker were used for the spoken output. Fig. 2 shows the PDA running the VIVOCA application with input via a Bluetooth microphone.

The central processing units (CPUs) in PDAs do not have support for rapid numerical computation, and have no dedicated hardware for floating-point calculations. It is still possible to perform floating-point operations on a PDA, however, they require software emulation of the dedicated hardware found on more powerful processors. This emulation is much slower than a dedicated unit and experimentation showed that this reduction in speed introduced a significant overhead for speech recognition.

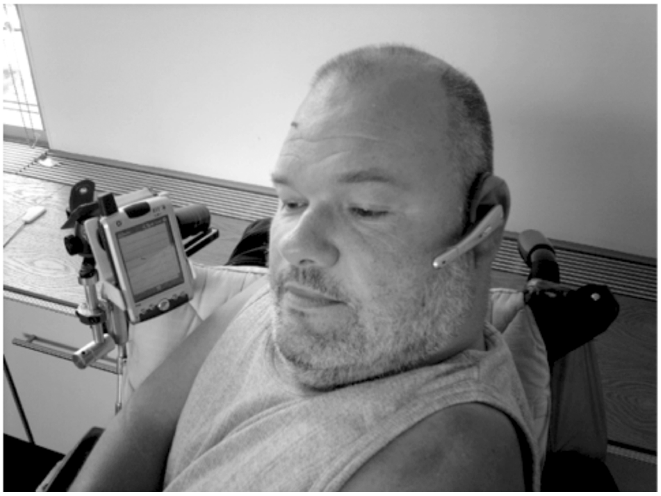

Fig. 2. A member of the project team demonstrating the prototype VIVOCA device. The user is wearing a headset microphone, and the VIVOCA software is running on the PDA mounted onto his wheelchair.

A solution is to use an alternative method to represent real numbers—namely a fixed-point representation. A fixed-point representation is a method for using binary integers to represent fractional numbers, however, for any fixed-point representation it is necessary to use a dedicated library of functions to perform basic mathematical operations. Therefore, the first step that was required, before any of the components of the system could be implemented, was the development and testing of a library of mathematical operations for our chosen fixed point representation.

We developed and tested a library for fixed-point arithmetic which has a Q format Q18.14. The library consisted of the basic operations on fixed point numbers (add, subtract, etc.), as well as more complex operations such as logarithm, exponential and the trigonometric functions, all of which are required for the signal processing and speech recognition parts of the system.

A further consequence of the limitations of the CPUs available on PDAs meant that, for the prototype system, the software to train recognition models and configure the aid was not located on a PDA, rather on a conventional PC. This decision was taken to expedite the time that would be required to train a set of models on a device with such limited processing power.

1) User Interface—User Training Application: The first interface that the user experiences is the user-training application. Fig. 3 shows a series of example screen displays from the PDA-based user training application. The interface initially prompts the user to speak a word (panel A). When the user speaks the word, a recognition score indicating the "closeness of fit" of each attempt to their own recognition model is denoted by the amount of the black circle that is filled with color. In Fig. 3 panel B, a high (89%) score is shown, and the circle is nearly entirely filled; whereas panel C shows a low (37%) score.

2) User Interface—VIVOCA Application: In order to facilitate eye contact between communication partners, we had originally envisaged that users would be able to use the

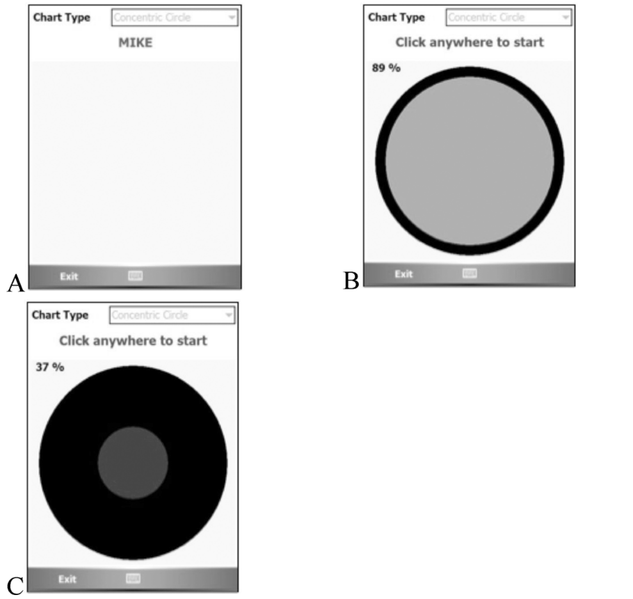

Fig. 3. The user-training application interface. In panel A the user is prompted to say the word "mike." Panel B shows the result when the word has been recognized with a high level of accuracy. Panel C shows the result when the word has been recognized with a low level of accuracy. The size of the shaded circle is proportionate to the accuracy, and is filled green when the accuracy is greater than 50%.

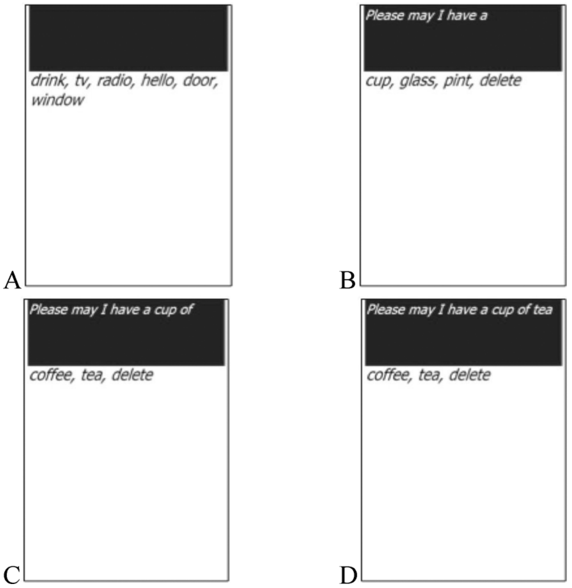
A
drink, tv, radio, hello, door, window
B
Please may I have a
cup, glass, pint, delete
C
Please may I have a cup of
coffee, tea, delete
D
Please may I have a cup of tea
coffee, tea, delete

Fig. 4. The VIVOCA message building interface. The panels show an instantiation of how a user builds up a sentence by speaking keywords to the device. Panel A illustrates the "top-level" choice of words available to this particular user. After saying the word "drink" the screen changes to that shown in panel B. Saying the word "cup" causes the screen to change to that shown in panel C. The user completes the utterance by saying "tea," and as that corresponds to the final possible choice the entire output phrase is then spoken by the speech synthesizer.

VIVOCA without referring to a screen, receiving essential audible prompts via an earpiece. Our user requirements research, however, indicated that users regarded a screen as important. We have, therefore, implemented both screen-based and audio interfaces. Fig. 4 shows the screen-based interface, in which the available vocabulary is listed in the white space. The user chooses the appropriate input word and the phrase, as it builds up, is shown in the top panel. For the audio interface, each available vocabulary item for each stage is presented sequentially to the user as a reminder. Clearly, the audio interface can become unwieldy for large vocabularies. In order to initiate the recognition of a series of words (the input phrase) leading to the generation of an output phrase, the user is required to press a switch to indicate that the recognizer should begin to "listen." As the users are generally people with severe physical disabilities, the switch is chosen and set up for each individual user.

III. Evaluation

The final prototype device was evaluated in a user trial designed to assess the process of configuring the device for a new user as well as the performance of the device in real communication situations. The evaluation consisted of six discrete stages, where each of the stages assessed a particular aspect of the configuration and usage of the device.

A. Participants

Ethical approval was obtained from Barnsley and North Sheffield National Health Service (NHS) ethics committees to recruit NHS patients into the trial. The principal inclusion criterion was that the participants should have moderate or severe speech impairment. Such impairment would result in low conversational intelligibility, typically less than 50%. The assessment of participants was conducted by a speech and language therapist using the Frenchay Dysarthria Assessment (FDA-2) [27].

A total of nine participants took part in the evaluation, including two of the participants from the development phase. Information pertaining to the nine participants is shown in Table I. The group of participants contained existing users of VOCAs, as well as people who did not normally use any augmentative or alternative communication method to support their communication. The group also contained participants with both progressive and stable speech impairment.

B. Procedure

The stages of the evaluation are shown in Table II. At stage 1 the researcher discussed with the participant how they might wish to use the VIVOCA device. Through these discussions a set of possible outputs was identified to cover a range of usage scenarios, and from these outputs the researcher defined a suitable vocabulary of input words that could be used to control the device to produce the outputs. The participant was then given time to reflect on the target output and the input vocabulary. On a second visit the researcher was able to make modifications to the proposed configuration in the light of the reflections of the participant.

Once the participant was satisfied with the configuration, recordings of their speech could be made (stage 2). The participant was recorded speaking around 20 repetitions of each of the words in the input vocabulary. To enable the participant to complete this stage they were supplied with a laptop and software to enable them to record their own speech in their

TABLE I
PARTICIPANTS IN THE USER TRIAL

Id	Sex	Age	Speech disorder	Aetiology	Current VOCA
E1	F	30s	Severe dysarthria	CP	None
E2	M	80s	Progressive severe dysarthria	PD	None
E3	F	30s	Severe dysarthria	TBI	Lightwriter
E4	M	50s	Moderate dysarthria	CP	None
E5	F	70s	Progressive moderate dysarthria	PD	None
E6	F	70s	Progressive severe dysarthria	MSA	Lightwriter
E7	F	40s	Progressive severe dysarthria	FA	Lightwriter
D1	M	30s	Severe dysarthria	CP	None
D2	M	40s	Severe dysarthria	CP	None

In the first column, D denotes a participant who took part in the development stages and E denotes a participant who did not. The abbreviated aetiologies refer to cerebral palsy (CP), Parkinson's disease (PD), traumatic brain injury (TBI), multisystem atrophy (MsA), and Friedreich's ataxia (FA).

TABLE II
STAGES OF THE EVALUATION

Stage	Description	Approximate duration
1	Specify input vocabulary and output speech	1-2 weeks
2	Record input vocabulary	2- 4 weeks
3	User training	2- 4 weeks
4	Device familiarisation	2 weeks
5	Trial usage	4 weeks
6	Interview and questionnaire	1 week

own time. Once they had completed sufficient repetitions of each of the words in the input vocabulary, the recordings were downloaded from the laptop and used to train the initial speech recognition system. At the completion of stage 2 the recognition accuracy of the initial models for each participant was tested using leave-one-out cross validation.

For stage 3, the participant used the user-training application for a period of 2–4 weeks, during which they were asked to practise for an hour a day where possible. Once completed it was possible to compare the recognition accuracy of the initial and final models.

The final system was configured and provided to the participant for a period of familiarization (stage 4). During this period the participant could identify changes they would like to be made to the system, and build confidence by practising using the device.

After the familiarization period, and any issues with the functionality of the participant's system had been identified and remedied, the final user trial commenced (stage 5). Although we had originally expected users to use a headset microphone as an input device, in practice wearing a headset was difficult for the users who had associated physical disability and all but one (D1) chose to rely on the internal microphone of the PDA. During the trial, participants were asked to maintain a diary to record how they used the device and their general feelings about the device. A researcher also regularly contacted the participant to ensure no difficulties had been encountered.

At the end of the user trial, at stage 6, a series of evaluations were conducted. These included testing the performance of the system. For this test, each participant completed a number of communication acts which involved them activating the device to produce a desired output using spoken commands. In order to be deemed successful, all the words in an input phrase needed to be recognized correctly. Each act was repeated three times, however for some participants it was not possible to complete every possible output of the system. Some participants' configurations (e.g., D2 and E1) had a large number of possible outputs and this had obvious risks for tiring the participant. In these cases a random subset of the possible outputs was used. Participants were also interviewed to obtain their views about the device and the trial they had completed.

IV. RESULTS

During the evaluation trial four of the participants withdrew from the project, three of whom had progressive speech disorders. These occurred at different stages in the project, and for a variety of reasons. One withdrew (E7) early in the evaluation as they found the process of recording sufficient speech examples to be too tiring. Two others withdrew (E4 and E6) as they were unable to maintain sufficient volume when speaking to activate speech recognition. Participant E2 withdrew after completing the first four stages of the evaluation as, after the familiarization stage, his speech production ability and general health deteriorated rapidly.

The participants identified a range of scenarios in which they would like to use the device. Principally these were for communicating immediate needs ("I would like a drink, please"); or communicating in social or commercial situations ("Please be patient it may take me sometime to answer you.") It is noticeable that most of the participants settled on a relatively small number of goal-oriented communications. Only 2 of the participants wished to use the system in a less constrained manner by having the facility to spell words into the system (D2 and E7).

Table III shows the recognition accuracy for the initial (stage 2) and final (stage 3) recognition models for each of the evaluation users to complete both stages.

After participants had completed the user training (stage 3) the recognition accuracy improved for all participants compared, with the exception of E3, whose recognition score remained stable at 99%. This improvement is a result of both the increase in the amount of data used to train the recognition models and the reduction in variability the user has gained from repeated practise of the words.

Table IV shows VIVOCA performance for four users who completed all six stages of the evaluation.

The results show that the VIVOCA device performed better when used for phrase building than when being used for spelling. This is due to the fact that in the phrase building mode the perplexity is relatively small, as it is equivalent to the number of competing words, typically 3–10 in these trials. Conversely, when the user is spelling input using the NATO alphabet there are always at least 28 competing models (one for each letter plus space and delete). The users who tried to use the device for spelling (D1 and E1) both had lower accuracy

TABLE III
RESULTS FROM FIRST THREE STAGES OF THE EVALUATION

Id	Intelligibility			Input vocabulary size	Total number of training files		Word recognition accuracy (%)	
	W	S	C		S2	S3	S2	S3
D1	E	E	E	26	1066	1167	76	93
D2	E	E	D	14/	837/	956/	92/	98/
				28	1262	1429	76	88
E1	E	D	D	28	564	933	97	99
E2	E	C	D	35	732	1674	96	98
E3	D	D	D	14	204	540	99	99
E4	C	C	C	12	220	412	98	99
E7	E	D	D	47	1295	1661	93	95

Intelligibility (assessed at stage 1) and recognition accuracy for participants. The intelligibility ratings are from the Frenchay Dysarthria Assessment-2 [27]. The intelligibility was assessed using recordings of words (W, 10), sentences (S, 10) and conversation (C, 5 minutes) which were rated by 3 speech and language therapists. For words and sentences a rating of E means intelligibility is less than 20%, D between 20% and 50%, and C between 50% and 90%. For conversation a rating of E corresponds to totally unintelligible, D to occasionally intelligible, and C to being understood around half the time. The word recognition accuracy was determined by using 10 examples of each word which were not included in the training set. The total number of training files used at stage 2 (S2) and 3 (S3) are listed. The increase in the number of files used at these stages comes from the speech examples collected during the user-training phase. Participant D2 had a set of 14 words for phrase building, and a further set of 28words (NATO alphabet plus space and delete) that could be used to spell words.

TABLE IV
PERFORMANCE OF THE SYSTEM TESTED AT THE END OF THE TRIAL

Id	Word intelligibility	Input vocabulary size	Number of test phrases	Task completion performance (%) at Stage 6
D1	E	26	9	67
D2	E	14/	12/	75/
		28	spelling	62
E1	E	28	Spelling	52
E3	D	14	11	72

Participants repeated the command phrase to produce the desired output. The task completion performance is the percentage of outputs produced as intended. The accuracy for the users with spelling was determined as the percentage of words correctly recognised.

rates. In the case of D1 the accuracy was lower when spelling compared to phrase building (62% versus 75%).

When interviewed at stage 6, the participants remaining in the trial at this stage expressed a strong view that the device concept was good. All thought that the device offered the prospect of easier and more rapid communication, though one did not think that the device would be appropriate for her needs. Most felt that the limited number of outputs of the trial device limited its usefulness. They agreed that the device would become more useful the more outputs it could produce.

Users commented that they sometimes found it harder to use the VIVOCA to communicate than to use their usual communication method of either speaking or speaking supported by a conventional VOCA. They all related this reduced ease of communication to the accuracy of the speech recognition in the VIVOCA. Several commented that they thought that device could lead to improved communication if the accuracy of the speech recognition was higher. One stated "I wouldn't get frustrated if it got it right." Participants also stated that the recognition errors meant that the device was sometimes slower to use than their conventional VOCA.

V. DISCUSSION

A voice output communication aid controllable by automatic speech recognition has been successfully produced and tested. The development of the device followed a user-centred iterative process, whereby a group of users evaluated each stage of the development and this led to modifications and improvements to the device. The eventual aim is to develop a VIVOCA device which can be commercialized and the design features that emerged from this process should make the final device more appropriate to a wider group of end users.

Our major challenge was to achieve accurate recognition of disordered speech at the larger vocabulary sizes required to achieve a practical VIVOCA. As shown in Table III, we have been able to construct recognizers with recognition accuracy in excess of 85%, for vocabulary sizes which are larger than previously reported for such speakers under similar test conditions (e.g., [13]) and large enough to construct a basic VIVOCA. This indicates that the methodology we have adopted to recognize dysarthric speech is viable for the task of producing a voice-input voice-output communication aid, and this is a significant finding of our research.

This study has, however, once more highlighted a major difference in recognition accuracy in controlled conditions, compared to the accuracy attained under realistic usage conditions (comparing results in Tables III and IV). Some of the reduction in performance is due to the fact that Table IV reports results for the recognition of a string of input words, whereas Table III reports results for single words. The misrecognition of any word in the string sequence results in an incorrectly completed task. However, due to the hierarchical nature of the message-building process it would be impossible to quantify this effect using a linear model of error accumulation. There are other factors influencing performance which may not simply be characterized as an example of the classic "training-testing mismatch" that has beset speech recognition technology for many years. In this study, the initial training recordings and final evaluations were conducted in the participant's home, therefore there were not major differences between the acoustic environments. There were, however, two notable differences in usage conditions, which affect the recognition accuracy: namely, the different microphones required and the need for the user to press a switch to activate the recognizer. As, for reasons of comfort and convenience, most of our users felt unable to use a close-talking microphone for everyday usage, instead choosing to use the PDA's internal microphone, the mismatch between the microphones used for training and testing were a significant cause of degraded performance. In addition, in using the VIVOCA, the user must press a switch in order to indicate to the recognizer that it should expect to receive and recognize a word or word string (an action which is not required in the recording or user training applications). This push-to-speak feature was introduced in order to reduce the possibility of activation of the VIVOCA by environmental sounds. However, due to the nature of the physical disabilities of our users, pressing a switch can require considerable effort, often produces associated body and head movement, and can even affect the quality of speech produced. This additional head movement and degraded speech exacerbates the problem of not being able to use close-talking microphones.

Previous work with speech-input home control systems had similar practical issues [13] and demonstrated similar performance degradation. Whereas for home control applications some users found the level of performance acceptable, for control of a VOCA, feedback from users has confirmed that such performance is not acceptable and that higher accuracy at larger vocabulary sizes is essential.

While users held a positive view of the technology concept, they felt that in its current form it would not substantially add to their communication ability or independence. However, the majority of participants felt that, if the prototype could be improved in terms of accuracy and range of output, it would be a viable and useful aid to their communication. We conclude that some fundamental practical issues of using speech recognition with disabled users must be addressed before the VIVOCA can become a viable tool.

Our future work will address two practical limitations of the current system which have a significant influence on recognition performance. The first of these will be to develop a VIVOCA based on a platform which supports a better quality internal microphone, which is most users' preferred option.

In a second development we aim to remove the necessity for push-to-speak operation. We will introduce a "word-spotting" ASR mode so that users are relieved of the burden of pressing a switch each time they wish to speak. The expectation is that this will also improve recognition accuracy.

VI. Conclusion

This paper has described the development of portable, voice output communication aid controllable by automatic speech recognition. The device can be configured to enable the user to create either simple or complex messages using a combination of a relatively small set of input "words." Evaluation with a group of potential users showed that they can make use of the device to produce intelligible speech output. The evaluation also, however, highlighted several issues which limit the performance and usability of the device, confirming that further work is required before it becomes an acceptable tool for people with moderate to severe dysarthria. Overcoming these limitations will be the focus of our future research.

Acknowledgment

The authors wish to thank all the participants who gave up their time to take part in this study.

References

[1] D. Beukelman and P. Mirenda, *Augmentative and Alternative Communication*, 3rd ed. Baltimore, MD: Paul H. Brookes, 2005.

[2] P. Enderby and L. Emerson, *Does Speech and Language Therapy Work?*. London, U.K.: Singular, 1995.

[3] C. L. Kleinke, "Gaze and eye contact: A research review," *Psychol. Bull.*, vol. 100, no. 1, pp. 78–100, 1986.

[4] B. O'Keefe, N. Kozak, and R. Schuller, "Research priorities in augmentative and alternative communication as identified by people who use AAC and their facilitators," *Augmentative Alternative Commun.*, vol. 23, no. 1, pp. 89–96, 2007.

[5] J. Murphy, "I prefer contact this close: Perceptions of AAC by people with motor neurone disease and their communication partners," *Augmentative Alternative Commun.*, vol. 20, pp. 259–271, 2004.

[6] J. Todman, N. Alm, J. Higginbotham, and P. File, "Whole utterance approaches in AAC," *Augmentative Alternative Commun.*, vol. 24, no. 3, pp. 235–254, 2008.

[7] L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit, "Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition," *Augmentative Alternative Commun.*, vol. 11, no. 3, pp. 165–175, 1995.

[8] N. Thomas-Stonell, A. L. Kotler, H. A. Leeper, and P. C. Doyle, "Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy," *Augmentative Alternative Commun.*, vol. 14, no. 1, pp. 51–56, 1998.

[9] R. N. Bloor, K. Barrett, and C. Geldard, "The clinical application of microcomputers in the treatment of patients with severe speech dysfunction," *IEE Colloquium High-Tech Help Handicapped*, pp. 9/1–9/2, 1990.

[10] U. Sandler and Y. Sonnenblick, "A system for recognition and translation of the speech of handicapped individuals," in *9th Mediterranean Electrotech. Conf. (MELECON'98)*, 1998, pp. 16–19.

[11] B. Wisenburn and D. J. Higginbotham, "An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: Objective results," *Augmentative Alternative Commun.*, vol. 24, no. 2, pp. 100–109, 2008.

[12] B. Wisenburn and D. J. Higginbotham, "Participant evaluations of rate and communication efficacy of an AAC application using natural language processing," *Augmentative Alternative Commun.*, vol. 25, no. 2, pp. 78–89, 2009.

[13] M. S. Hawley *et al.*, "A speech-controlled environmental control system for people with severe dysarthria," *Med. Eng. Phys.*, vol. 29, no. 5, pp. 586–593, 2007.

[14] H. V. Sharma and M. Hasegawa-Johnson, "State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition," in *NAACL HLT Workshop Speech Language Process. Assistive Technol.*, 2010, pp. 72–79.

[15] R. Palmer, P. Enderby, and M. S. Hawley, "A voice input voice output communication aid: What do users and therapists require?," *J. Assistive Technol.*, vol. 4, no. 2, pp. 4–14, 2010.

[16] S. Young *et al.*, Cambridge University Engineering Department The HTK book, 2006.

[17] P. D. Green, J. Carmichael, A. Hatzis, P. Enderby, M. S. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers," in *Eur. Conf. Speech Commun. Technol. (Eurospeech)*, 2003, pp. 1189–1192.

[18] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, Jun. 1986.

[19] R. Palmer, P. Enderby, and S. P. Cunningham, "The effect of three practice conditions on the consistency of chronic dysarthric speech," *J. Med. Speech-Language Pathol.*, vol. 12, no. 4, pp. 183–188, 2004.

[20] M. Parker, S. P. Cunningham, P. Enderby, M. S. Hawley, and P. D. Green, "Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project," *Clin. Linguistics Phonetics*, vol. 20, no. 2–3, pp. 149–156, 2006.

[21] J. N. Holmes and W. J. Holmes, *Speech Recognition and Synthesis*, 2nd ed. London, U.K.: Taylor Francis, 2001.

[22] S. L. Glennon and D. C. DeCoste, *Handbook of Alternative and Augmentative Communication*. San Diego, CA: Singular, 1997.

[23] J. Todman, "Rate and quality of conversations using a text-storage AAC system: Single-case training study," *Augmentative Alternative Commun.*, vol. 16, no. 3, pp. 164–179, 2000.

[24] M. S. Hawley, S. P. Cunningham, F. Cardinaux, A. Coy, S. Seghal, and P. Enderby, "Challenges in developing a voice input voice output communication aid for people with severe dysarthria," in *Proc. AAATE—Challenges Assistive Technol.*, 2007, pp. 363–367.

[25] A. Black and K. Lenzo, "Flite: A small fast run-time synthesis engine," in *4th ISCA Speech Synthesis Workshop*, 2001, pp. 157–162.

[26] A. W. Black, P. Taylor, and R. Caley, The Festival Speech Synthesis System. Edinburgh, Univ. Edinburgh, 1999.

[27] P. Enderby and R. Palmer, *Frenchay Dysarthria Assessment*, 2nd ed. Austin, TX: Pro-ed, 2008.

Mark S. Hawley received the Ph.D. degree from the University of Sheffield, Sheffield, U.K.

He is Professor in the School of Health and Related Research, Head of the Rehabilitation and Assistive Technology Research Group, and Director of the Centre for Assistive Technology and Digital Healthcare, at the University of Sheffield. He is also Honorary Consultant Clinical Scientist at Barnsley Hospital.

Prof. Hawley is a Chartered Scientist and Member of the Institute for Physics and Engineering in Medicine. He was awarded the Honorary Fellowship of The Royal College of Speech and Language Therapists in 2007 for his service to speech therapy research.

Stuart P. Cunningham received the B.Eng. degree in software engineering and the Ph.D. degree in computer science from the University of Sheffield, Sheffield, U.K., in 1997 and 2003, respectively.

He is a lecturer in Human Communication Sciences at the University of Sheffield, Sheffield, U.K. His primary research interests are in the automatic recognition of disordered speech, and the development of speech-based interfaces for assistive technology.

Phil D. Green received the B.Sc. degree in cybernetics from the University of Reading, Reading, U.K., in 1967, and the Ph.D. degree in automatic speech recognition from the University of Keele, Keele, U.K., in 1971.

He holds a Chair in Computer Science in the Department of Computer Science at the University of Sheffield, U.K., where he previously held the posts of Lecturer, Senior Lecturer, Reader, and Head of Department. He heads the Speech and Hearing Research Group, has authored around 100 publications in speech science and technology and acted as Principal Investigator for around 15 research grants.

Pam Enderby is a qualified speech and language therapist and received the Ph.D. degree on the subject of dysarthria from Bristol University, Bristol, U.K. She was awarded an honorary doctorate from the Department of Computing and Mathematics, the University of West of England, Bristol, U.K., 2000; and an MBE for services to speech and language therapy, in 1993.

Prof. Enderby is Professor of Community Rehabilitation at the University of Sheffield and has recently stepped down from being the clinical director of the South Yorkshire Comprehensive Local Research Network. She has conducted clinical research through most of her career combining posts in the NHS with that of University of Sheffield.

Rebecca Palmer received a first class bachelor's degree in linguistics and language pathology from the University of Reading, Reading, U.K., in 1999, followed by the Ph.D. degree in the treatment of dysarthria from the University of Sheffield, Sheffield, U.K., in 2005.

She is currently a Senior Clinical Lecturer at the University of Sheffield, Sheffield, U.K. Previous posts held include highly specialist speech and language therapist and rehabilitation trials manager for the NIHR Trent Stroke Research Network. Her primary research interest is the use of speech and language technology to assist improved communication of stroke survivors.

Dr. Palmer is a member of the Royal College of Speech and Language Therapists and the Health Professions Council.

Siddharth Sehgal received the B.Sc. degree in mathematics and a M.Sc. degree in applied operational research from the University of Delhi, Delhi, India, in 1996 and 1999, respectively. He received the M.Sc. degree in advanced computer science from the University of Sheffield, Sheffield, U.K., in 2004. He also has a Diploma in network centered computing from National Institute of Information Technology, Delhi, Delhi, India, in 2001. He is currently pursuing the Ph.D. degree at the University of Sheffield, Sheffield, U.K.

He worked from 2000 to 2003 as a Software Engineer with Futuresoft India Private Limited and Birlasoft, Delhi, India. Since 2005, he has been a Research Associate in the Department of Computer Science and Human Communication Sciences at the University of Sheffield, Sheffield, U.K.

Peter O'Neill received the B.Sc. (Hons.) degree in software engineering from Sheffield Hallam University, Sheffield, U.K., in 1996 and has worked in the assistive technology domain for the last 16 years. He attained his Doctorate with the title "Enhancing the Prescription of Electronic Assistive Technology," from Sheffield Hallam University, in 2006.

He has held the position of Research Associate at Barnsley Hospital NHS Foundation Trust since 1996, and has also held the position of honorary research fellow at the University of Sheffield, Sheffield, U.K.