homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition

H. Christensen[1], *I. Casanuevo*[1], *S. Cunningham*[2], *P. Green*[1], *T. Hain*[1]

[1]Computer Science, University of Sheffield, Sheffield, United Kingdom
[2]Human Communication Sciences, University of Sheffield, Sheffield, United Kingdom
h.christensen@dcs.shef.ac.uk, i.casanueva@sheffield.ac.uk, s.cunningham@sheffield.ac.uk
p.green@dcs.shef.ac.uk , t.hain@dcs.shef.ac.uk

Abstract

We report on the development of a system which will bring personalised state-of-the-art automatic speech recognition into the homes of people who require voice-controlled assistive technology. The ASR will be sited remotely ('in-the-cloud') and run over a broadband link. This will enable us to adapt the system to the user's requirements and improv the accuracy and range of the system while it is in use. We outline a methodology for this: the 'Virtuous Circle'. A case study indicates that we can obtain acceptable performance by adapting speaker-independent recognisers with 10 examples of each word in a 30-word command-and-control vocabulary. We explain the idea of a PAL - a Personal Adaptive Listener - which we intend to develop out of this study.

Index Terms: dysarthric speech recognition, 'in-the-field' speech recognition, cloud-based speech recognition

1. Introduction

With an ageing population and the increasing acceptance of community-based care, there is a growing demand for electronic assistive technology (EAT). One of the major uses of EAT is to support independent living, particularly among the elderly and the physically impaired. Devices such as environmental control systems (ECSs) allow people to control many aspects of their home environment through a single control interface. Typically these systems will be operated using a switch-scanning interface which accommodates the limited motor control abilities of users who have physical disabilities.

A major drawback of switch-scanning interfaces is that they can be time-consuming and effortful to use. It is therefore appropriate to consider alternative input-methods for EAT that can accommodate users with limited physical abilities. The use of speech is an attractive alternative to switch-scanning interfaces. Indeed the prospect of using automatic speech recognition (ASR) as an alternative input-method for EAT has been discussed in the literature for more than thirty years [1, 2].

A significant proportion of people requiring EAT have dysarthria, a motor speech disorder associated with their physical disability [3]. As a result of the effect of dysarthria on speech production, inexperienced listeners find speech from people with dysarthria difficult to recognise [4]. Machine recognition of dysarthric speech is also considered a difficult problem.

Large vocabulary speaker adaptive recognition systems have been successfully used for people with mild and moderate dysarthria as a means of inputting text. These systems, however, have been shown to be less successful for people with severe dysarthria (e.g. [5, 6]). Specific modifications to speaker adaptive speech recognition algorithms with the aim of improving the recognition of dysarthric speech patterns have been described but they have not yet appeared in a widely available form [7, 8].

Speaker dependent speech recognition has often been thought to be more appropriate for users with severe dysarthria. This is because models can be trained directly with the speaker's utterances rather than assuming their speech is similar to the typical speech the models were originally trained with [9]. Speaker dependent recognisers have been shown to perform well for severely dysarthric users in several studies [10, 11]. In these examples however, the input vocabularies were quite small, which can limit the potential usefulness of the EAT system.

In recent years, new corpora of dysarthric speech have become available [12, 13]. These data sets have enabled researchers to conduct more systematic studies than before [14, 15], and open the possibility of comparing techniques using reference test sets. These corpora are however small compared to those used in modern, mainstream ASR. One reason for their relatively small size is the fact that prolonged speaking for people with severe dysarthria can be tiring. Therefore passive data collection from this population is likely to remain limited, unlike data collection for the typical speaking population. The only way to acquire substantial amounts of data is from a system which is being actively used.

Most voice-enabled EATs described in the literature have been systems that have been developed for relatively small scale studies and with the main focus being on the observed ASR performance. There are some real challenges to be solved when porting such systems and setups to more 'realistic' scenarios, especially because of the larger number of users involved, and the need for a large degree of automation whilst still accommodating the needs of the individual users for personalisation. This paper describes recent work on designing a real 'in-the-field' ASR-based EAT system where scalability and ease of initialisation has been at the forefront of the design from the onset. We have focused on two issues: how to most effectively setup an initial system for a given speaker (finding their optimal 'operating point') and how to use cloud-based ASR servers to allow the researcher free access to maintain and update ASR models.

We present the homeService system in which we are developing state-of-the-art ASR. homeService is part of the UK EPSRC Project in Natural Speech Technology project, a collaboration between the Universities of Edinburgh, Cambridge and Sheffield. homeService users are being provided with speech-driven ECS and eventually spoken access to other digital appli-

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 29–34,
Grenoble, France, 21–22 August, 2013. ©2013 Association for Computational Linguistics

cations. We are in the process of recruiting around 10 users to a longitudinal study: each user will be engaged with homeService for at least 6 months.

From our experience in previous projects [10, 16], which included user requirement studies, we will continue to work with users in a collaborative way: the users effectively become part of the research team. As part of this process, users will inform the design and specification of the functionality of their personal system. In addition we will work with users to close what we have referred to as the 'virtuous circle'. By working with each user we will establish an initial 'operating point': a task which is sufficiently simple that we can expect good performance from the ASR and yet sufficiently useful that the user's interest is maintained. We deploy this system and provide software which enables the user to practice with it. Practice improves the user's pronunciation consistency and, crucially, provides more data for ASR training. The exercises provide the user with feedback, not based on the match to a standard pronunciation but on how well a new utterance fits the user's current model. When the performance of the system has improved sufficiently, we widen the vocabulary and range of target devices homeService controls. This process is iterated: the 'virtuous circle'. This is an example of Participatory Design [17].

As part of the ethical approval obtained for the study, the informed consent of users will enable us to collect examples of speech data from their interactions with the homeService system. These interactions will be stored and used to create a database which will become available to the research team but will not be made publicly available due to privacy issues. To further reduce any concern users might have about the system's ability to 'listen' to them, the interface will clearly indicate when the microphone is open - typically only a couple of seconds for each voice command. At any time participants will also be able to "opt-out" of the recording process, or even request recordings be deleted and not used in the database.

The ASR will run remotely 'in-the-cloud', and be connected to the homeService users' home by a dedicated broadband link. This is a novel approach for providing speech-driven EAT which will enable us to collect speech data, train new statistical models, experiment with adaptation algorithms, change vocabularies and so on without having to modify the equipment in the user's home. This will reduce the amount of researcher time spent travelling to visit users, but more importantly will enable us to modify the system rapidly. This means new models can be deployed when they are ready, and new data can be analysed as soon as it is collected. We explain the homeService setup in more detail in section 2.

The development of the 'in-the-cloud' recognition system is described in section 3. In section 5 our participatory design methodology is further developed. Some preliminary results of the speech recognition system are presented in section 4.

2. homeService setup

A schematic diagram of the homeService system is shown in figure 1. The system consists of two distinct parts: the atHome system and the atLab system. The atHome system will be deployed in a user's home and comprises a PC and a series of input and output devices to enable the system to receive spoken commands and interact with devices in the home environment, for example through the transmission of infrared signals. The atLab system resides at the university and comprises the main server which operates the ASR system and maintains the system state for each atHome system.

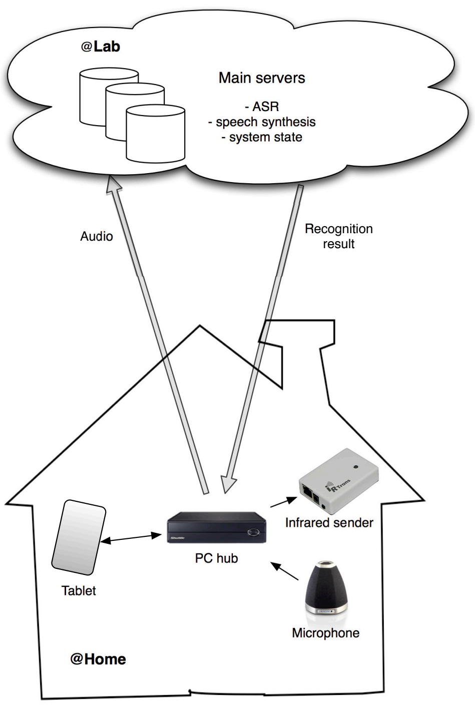

Figure 1: *Diagram of the homeService system with its two distinct parts: the atHome component in a user's home and the atLab 'in-the-cloud' part. For simplicity, only one user is drawn here but the cloud-based ASR server enables us to scale to many simultaneous users.*

The system hardware consists of 'off-the-shelf' items such as a microphone array, an Android tablet for display and an infrared transmitters, which reduces the overall cost of each installation, and means the system will not need to rely on specialist hardware. In the following sections the components of the system are described in detail.

2.1. Components

2.1.1. The PC

The atHome software is designed to run on a Linux-based PC. This PC will act as the main hub for the atHome system. It maintains the communication between the atLab part of the system and the peripherals in the atHome part of the system. The software controls the recording of audio from the system microphone, sends the audio back to the lab via a broadband link, provides feedback to the user, and controls the sending of infrared signals to various devices in the home. The software also sends updates to the screen of the tablet, and when appropriate, will play synthesised speech output.

Although, from an operational point of the view, the PC is at the heart of the atHome system, the design philosophy of the atHome system ensures that the PC is as unobtrusive as possible. Consequently, from the users' perspective the system microphone and the tablet PC are the key parts of the system.

The requirements for the PC are that it should to be relatively small, quiet and discrete, with a low energy consumption.

For this a Shuttle XH61v with a core i3 3220 was chosen (30.5 x 6.4 x 21.6 cm).

2.1.2. Microphone

For speech data capture, we use a high-quality USB microphone array (Dev-audio Microcone). It has a hexagonal design with 6 microphones placed in each of the six sectors, each covering approximately 60° of the surroundings. The Software Development Kit gives access to each of the 6 individual microphone channels as well as a stereo output of the beam-formed and noise-reduced signal, which will help us to reduce cross-talk from other speakers, the TV and so on. The Microcone also has a pleasing design, which is important as it will have to have a relatively prominent and very visible position in the users' homes throughout the full study.

2.1.3. Infrared transmitter

Remote control of the devices (such as TV, radio, lights etc.) is performed by an USB infrared (IR) emitter (IRtransWiFi IRDB). To make it personalised for each home, there is a configuration step where the emitter is trained with the IR commands from the original remote controls of the home devices. The researcher has to perform this step manually, using the software provided with the IR emitter. After this step is completed, the system is able to associate system actions (e.g. "turn on TV") to the specific IR commands for the devices it is controlling.

2.1.4. Android tablet

The Android tablet acts as a personalised, visual interface for the user. This has several advantages; during system operation it will

- display a representation of the system state,
- display the options available for the user (this directly corresponds to the current ASR vocabulary),
- act as a touch input if necessary.

In addition, the tablet will have an app which will enable the system to acquire additional training data from the user. Software for user practice exercises will run on the tablet.

The configuration of the display is loaded from a XML file, where the description of each device is written by the system developer. This permits the personalisation of the display.

2.1.5. atLab Server

The audio signal which is to be recognised is transferred across to the atLab part of the homeService system over the broadband link and subsequently passed on to the ASR server, also running at the university. When the recognition result is known it is 'acted' upon by the atLab software: for the environmental control system this means determining the next state of the system including possible infrared-codes which need transmitting and whether the tablet screen activity needs updating. All of the information concerning the state is then communicated back to the home of the user and acted upon. The two main communication links in the system (to the home and to the lab) are governed by individual APIs.

The atLab software runs on a dedicated server at the university. Apart from being the main interface to the individual users, it also handles the communication to and from a *bank* of ASR servers (one for each user) which will provide online speech recognition based on models and setups that are personalised to each user.

3. ASR

One of the main design aims was to base the system on 'in-the-cloud' ASR. This provides the research team with full control over the specifics of the ASR for each user; it is relatively straight-forward to change for example acoustic models, vocabularies and lexicons without disturbing the user unnecessarily. It also gives the researchers more scope for monitoring the state of the atHome systems, and crucially, for much more immediate trouble-shooting. Software components can easily be taken down and re-started. In the future, we also envisage having short remote chat-sessions with the users/carers to discuss any issues about the system.

It is important to bear in mind that this easy access design does impose constraints on the research team. For instance, given that data will be collected from the microphone for speech events while the system is in use, all users must be carefully briefed about how these recordings will be made and stored before they can provide informed consent to take part in the study. In the future it is envisaged that the system will be used in 'open mic' sessions when all the audio from the microphone will be gathered at agreed times of the day. Again, careful briefing of the users will be required as are procedures for users to retrospectively opt-out of these data collection sessions.

Each user has a dedicated ASR server which will be preloaded with personal acoustic and language models as well as grammars. To maximise performance we intend to use grammars which restrict the vocabulary according to the given state the system is in. For example, if the system is operating in the environmental control mode and the user has just turned on the guide on the TV, a state-dependent grammar would contain words needed for navigating the guide, e.g. 'up', 'down', 'left', 'right', 'ok' and 'exit' as well as certain *power* or *meta* words which would allow the user the change state, for example by saying 'home' or 'back'.

The ASR server's recognition technology is built around an in-house decoder based on weighted finite state transducers (WFSTs). This decoder was the winner in the NIST meeting recognition evaluations in 2007 and 2008. For details see [18, 19]. Every *recognition cycle* (consisting of audio being recorded, transferred across to the servers and subsequently recognised) will trigger the possibility of a change of state dependent on the current state and the newly recognised word. To further support this, the ASR server can dynamically load the next WFST from a set of pre-computed WFSTs matching all of the possible states of the system. We plan to expand this to enable online compilation of WFSTs.

4. Experimental setup

Recruitment of users is underway for the homeService study. In preparation for setting up dedicated ASR systems for each user, we have carried out a pilot-study using data from a potential user, which we recorded during previous studies. This user (F01) is a female, in her mid fifties at the time the recordings were made, who has cerebral palsy. Her speech is classified as spastic dysarthric of a severe nature. She has always been a very keen participant in our studies, and as such is a valued member of our extended research team.

We have chosen her as one of the first users in the homeService study as she has previously demonstrated that she is a highly motivated adopter of new technology; she is also a keen PC user.

She currently uses a switch mounted on the headrest of her wheelchair to access her scanning-based environmental control system and as well as to control her PC via dedicated software.

4.1. Data

F01 has provided speech recordings for two research projects in the last decade, which is of interest here. These are all isolated words initially recorded with the aim of providing training material for whole word ASR models used in an an ECS system similar to the primary homeService task. The word lists consisted of isolated words such as ”TV”, ”on”, ”off”, ”channel”, etc. In total we have 1286 individual word recordings covering a vocabulary of 33 words (approximately 38 examples of each word).

In this study we wish to train tri-phone derived word models, and the ideal training data would be sets of phonetically rich words or sentences. However, given the nature of this data set of isolated words, it is possible to quickly create a realistic test set using examples drawn from the data set.

After a process of initial alignment to remove extraneous silences, around 40 minutes of data recorded from two different projects remained; project A provided 23 minutes of 8 kHz data (for the work here, this data has been up-sampled to 16 kHz) recorded using a headset microphone (SkyTronic Tie-Clip Microphone) onto a dedicated Arm-based embedded device (Balloon 3 board with a GEWA PROG III infrared micro chip). The remaining data from project B was recorded at 16 kHz on a laptop using a microphone array (the Acoustic Magic Voice Tracker array) [10].

4.2. Acoustic modelling

All hidden Markov models (HMMs) were trained using the maximum likelihood (ML) criterion. State-clustered, triphones having Gaussian mixture models with 16 components per state were used.

4.3. F01 case study

Although the amount of data we have available from speaker F01 is relatively small compared to what one would normally need to train a high-performance, personalised ASR system, it far exceeds what we could expect to be able to record from a new homeService user in a typical enrolment session. What it does do is enable us to explore the effect of having access to different amounts of data for e.g., adaptation purposes. The experiments presented here aim to investigate the relationship between the quantity of training and recognition performance. When recruiting new users for homeService this will be a useful indicator of how much enrolment data will need to be recorded to provide a good, initial *operating point*.

4.4. Results

First though, it is useful to assess F01’s data in terms of baseline performance. Table 1 shows some baseline results for her, where we have tested all of her speech on high-performance models trained on typical speech meeting data and on good, speaker-independent models trained on the dysarthric UASpeech corpus [12]. The achieved accuracies of 8.9% and 13.5% are very low and indicate the severity of F01’s speech impairment. The UASpeech result is in a range comparable to what has been reported for some of those speakers as well [20, 15].

Table 1 also shows the results from using some of F01’s data to perform a *maximum a posteriori* (MAP) adaptation from the original, speaker-independent meeting models or UASpeech models [21]. As we have very limited data, the presented accuracy is the mean of the accuracies obtained from doing a round-robin style test using 10 folds of the complete dataset, each having a 90%/10% split into an adaptation set and a test set. The MAP-based systems performed best in precursor experiments reported in [15] and show large improvements over the baseline systems with accuracies of 74.7% and 75.5% respectively.

System	Accuracy
Meeting (SI)	8.9 %
Meeting+MAP (SD)	74.7 %
UASpeech (SI)	13.5 %
UASpeech+MAP (SD)	75.5 %

Table 1: *Word accuracy rates for baseline systems. Please see text for further explanation.*

It is important to note that these results were obtained using more than 1100 words from speaker F01, which is far beyond what would be reasonable and realistic to obtain from a prospective user. This is not only because prolonged periods of speaking can be tiring for these users, but also it would be a considerable undertaking to make that many recordings. In our experience it would take several weeks to collect this quantity of data.

For projects like homeService, there is a notable trade-off between not asking participants to endure lengthy enrolment sessions, whilst still ensuring we can deliver a sufficiently useful level of performance in the first system we deploy. Although all users will be aware that the systems are not perfect, if it becomes frustrating to use because of too many errors we run a real risk of the users rejecting the system (and the study), thereby breaking the foundations of the ‘virtuous circle’, where good systems will lead to increased use and data collection.

We therefore wished to investigate how much adaptation would be needed to get a particular level of performance. Figure 2 shows the results of increasing the amounts of data used for adapting from the speaker-independent UASpeech and meeting models respectively.

Both curves follow the same trend, and as expected the accuracy increases with increasing amounts of data (presented as number of words out of a total of 1158 words in each of the training/adaptation folds). For the lower number of words there is a dramatic increase in performance; this can be seen to taper off approximately at around 300 words. Given F01 here has a vocabulary of just over 30 words, this corresponds to approximately 10 instances of each word.

Interestingly, both the UASpeech based and the meeting model based systems converge on approximately the same, stable level after about 400 examples, but the initial curve ascends more slowly for the meeting models, so in situations where smaller amounts of adaptation data is available the closer models from UASpeech are a better starting point.

5. Longer-term plans

As the pool of homeService users grows we will continue to monitor the design choices surrounding the cloud-based setup including ease of use for the researcher as well as whether the users’ feel comfortable with the idea of their system being monitored from outside of their home. It will also be interesting to

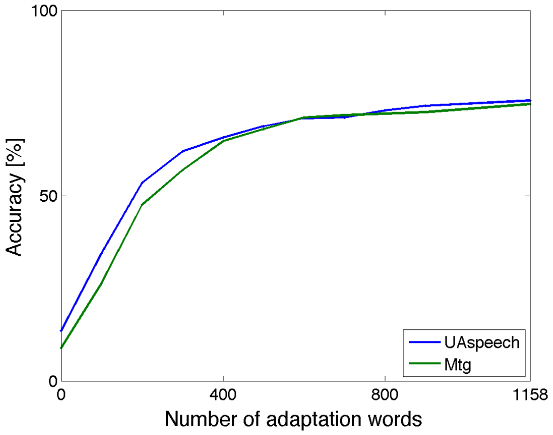

Figure 2: *Word accuracy as a function of increasing amount of data used for MAP adaptation of acoustic models; x-axis shows the number of utterances (each containing a single word) out of a possible 1158 used for adaptation.*

follow how the impact on the success of each individual user's virtuous circle.

We see the homeService systems as the first generation of PALs - Personal Adaptive Listeners. A PAL is a portable, perhaps wearable, device that belongs to an individual and adapts to the speech communication characteristics and preferences of its owner. Like human listeners, it does this whilst in use, does it quickly and extends its utility over time. A PAL is somewhat akin to a human valet: It understands its owner's needs, carries out their wishes and sometimes acts on their behalf. The technology adapts to its user, rather than the other way round. Crucially, The owner is able to teach the PAL through spoken dialogues, which develop differently for different owners. The owner-PAL relationship should be something like training a dog.

To make the step from homeService to PALs requires spoken dialogues between the owner and the device. Dialogue management techniques in commercial dialogue systems are usually hand-crafted, which makes them difficult to adapt. During the last decade it has become fashionable to approach the dialogue management problem statistically, modelling the dialogue as a Partially Observable Markov Decision Process (POMDP) and optimising the dialogue policy with Reinforcement Learning (RL) [22]. This framework provides robustness against speech understanding errors and automatic learning of dialogue policy. As the dialogue policy is learned with the data gathered from interaction with the user, it is optimised for its specific user, making it a personalised policy. RL permits online learning, so the system can also adapt its policy to changes in the user behaviour (e.g. when the user becomes more familiar with the system) and to the changes in the speech understanding system (e.g. when the ASR improves as more data is gathered). The user can also explicitly give a reward to the system after each interaction, 'teaching' the system.

The main problem with statistical dialogue management is its intractability, due to the size of the state space and to the impossibility of exact solving the POMDP, but it is possible to use approximate algorithms to build real sized dialogue systems. Another problem is the long time that takes to learn a suitable policy, but recent studies have been able to learn a policy for a non trivial tourist information system in less than 200 dialogues, which makes possible learning a policy directly from user interaction.

Adapting these techniques for PAL dialogues raises several interesting issues:

- 'teaching your PAL' should correspond to seeding the dialogue statistics.
- A PAL should not make the same mistake twice.
- The owner will know exactly what the PAL understands.

6. Acknowledgements

This research was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

7. References

[1] J. A. Clark and R. B. Roemer, "Voice controlled wheelchair," *Archives of Physical Medicine & Rehabilitation*, vol. 58, no. 4, pp. 169–75, 1977.

[2] A. Cohen and D. Graupe, "Speech recognition and control system for the severely disabled," *Journal of Biomedical Engineering*, vol. 2, no. 2, pp. 97–107, 1980.

[3] J. R. Duffy, *Motor Speech Disorders*, 3rd ed. London, UK: Mosby, 2013.

[4] S. A. Borrie, M. J. Mcauliffe, and J. M. Liss, "Perceptual learning of dysarthric speech : A review of experimental studies," *Journal of Speech, Language, and Hearing Research*, vol. 55, pp. 290–305, Feb 2012.

[5] M. S. Hawley, "Speech recognition as an input to electronic assistive technology," *British Journal of Occupational Therapy*, vol. 65, no. 1, pp. 15–20, 2002.

[6] N. Thomas-Stonell, A.-L. Kotler, H. A. Leeper, and C. Doyle, "Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy," *Journal of Augmentative and Alternative Communication*, vol. 14, pp. 51–55, 1998.

[7] J. R. D. Jr, D. Hsu, and I. J. Ferrier, "On the use of hidden markov modelling for recognition of dysarthric speech," *Computer Methods & Programs in Biomedicine, 1991, 35(2), 125-139*, vol. 35, pp. 125–139, 1991.

[8] H. V. Sharma and M. Hasegawa-Johnson, "State transition interpolation and map adaptation for hmm-based dysarthric speech recognition," in *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, 2010, pp. 72–79.

[9] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Journal of Augmentative and Alternative Communication*, vol. 16, pp. 48–6, 2000.

[10] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering & Physics*, vol. 5, no. 29, pp. 586 – 93, 2007.

[11] 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 *Proceedings of the AAATE - Challenges for Assistive Technology*, 2007, pp. 363–367.

[12] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, “Dysarthric speech database for universal access research,” in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 22–26.

[13] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” *Language Resources and Evaluation.*, pp. 1–19, 2011.

[14] H. V. Sharma, M. Hasegawa-Johnson, J. Gunderson, and A. Perlman, “Universal access: Speech recognition for talkers with spastic dysarthria,” in *Interspeech’09*, Brighton, UK, sep 2009.

[15] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, “A comparative study of adaptive, automatic recognition of disordered speech,” in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[16] H. Christensen, S. Siddharth, P. O’Neill, Z. Clarke, S. Judge, S. Cunningham, and M. Hawley, “SPECS - an embedded platform, speech-driven environmental control system evaluated in a virtuous circle framework,” in *In proc. Workshop on Innovation and Applications in Speech Technology*, 2012.

[17] S. L., *Participatory Design: Principles and Practices.* N.J.: Lawrence Erlbaum, 1993, ch. Forward, p. viiix.

[18] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, D. van Leeuwen, and V. Wan, “The 2007 AMI(DA) system for meeting transcription. in NIST rich transcription 2007,” *Lecture Notes in Computer Science*, pp. 414–428, 2008.

[19] J. G. Fiscus, J. Ajot, and J. S. Garofolo, “The rich transcription 2007 meeting recognition evaluation,” in *Multimodal Technologies for Perception of Humans*, vol. 4625/2008. Springer Berlin/Heidelberg, 2008, pp. 373–389.

[20] H. V. Sharma and M. Hasegawa-Johnson, “Acoustic model adaptation using in-domain background models for dysarthric speech recognition,” *Computer Speech and Language*, 2012.

[21] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[22] M. Gasic, M. Henderson, B. Thomson, P. Tsiakoulis, and S. Young, “Policy optimisation of pomdp-based dialogue systems without state space compression,” in *Workshop on Spoken Language Technology (SLT)*, 2012, pp. 31–36.