Speech-Enabled Environmental Control in an AAL setting for people with Speech Disorders: a Case Study

Heidi Christensen[1], *Mauro Nicolao*[1], *Stuart Cunningham*[2], *Salil Deena*[1], *Phil Green*[1], *and Thomas Hain*[1]

[1]Computer Science; University of Sheffield, United Kingdom
[2]Human Communication Sciences, University of Sheffield, United Kingdom
{heidi.christensen, m.nicolao, s.deena, s.cunningham, phil.green, t.hain}@sheffield.ac.uk

Keywords: automatic recognition of disordered speech, assistive technology, cloud-based processing

Abstract

The homeService research project is concerned with developing personalised speech-enabled interfaces, in an AAL setting, for users with severe physical impairments and associated disordered speech. By putting state-of-the-art speech recognition systems into people's homes, invaluable lessons can be learned from doing long-term trials 'in-the-wild'. The experiences gained from the first homeService user's case story is described here. Each system is initially deployed with acoustic models adapted using a relatively small amount of enrolment data. During use, data is subsequently collected as the user interacts with the system and this data is used to update the models at a later stage. This paper contrasts results from experiments carried out online with the live system, and offline with the collected data. Particular emphasis is put on the amount of adaptation data as well as the use of manual vs. automatic annotations in the context of trying to ensure that the implementation and personalisation strategy will scale with many users.

1 Introduction

Speech-enabled interfaces can provide an attractive alternative way of accessing digital devices for people who cannot use traditional methods such as a remote control, keyboard or mouse. The success of such interfaces is highly dependent on the recognition accuracy that can be achieved at the time of deployment; if the performance is too poor, the user is likely to lose interest and will not be motivated to use the system. Obtaining a high enough performance can be particularly challenging for people who have disordered speech associated with neuro-motor conditions such as cerebral palsy.

The modelling power of the acoustic model and hence the recognition performance is adversely affected by the increased variability in the acoustic signal that characterises disordered or *dysarthric* speech. To cope with this increased variability, it is essential to have access to representative data with which to train the model, and it is standard practise to adapt a speaker independent model using adaptation data recorded prior to the system being deployed [10, 4]. Such adaptation data is recorded during an *enrolment* phase.

Enrolment data is a prerequisite to achieving a sufficiently high performance at runtime, however, there is a choice of when to stop the enrolment phase and progress to the deployment phase, where the system is *online*. This represents a trade-off between reducing the amount of time and effort spent recording enrolment data and getting a good enough performance when the system first goes online. Deploying too early will run the risk of not giving the user a successful system and, conversely, spending a lot of time recording data without a tangible system runs the risk of disengaging the user. For dysarthric users, speaking is often an effort.

Another challenge that needs to be overcome is to do with how to personalise a system for a particular user and his situation whilst at the same time arriving at a deployment strategy and protocol that can scale to potentially hundreds of users. In a research project it is acceptable to hand-graft systems to some extend, but for real impact and feasibly deployable systems a strategy for how to roll out systems to many users needs to be considered at the early stages of design and implementation.

The homeService project aims to help answer some of these questions. It is concerned with how speech technology can assist people with severe speech disorders and restricted upper-limb mobility to live more independently in their homes. The project implements a cloud-based environmental control system where users can control electronic devices such as TVs, radios, lamps etc. through the use of voice-commands [5]. As well as having the users provide speech command word examples during the enrolment phase, once the system is online, all interactions with the system are also recorded. Over time, this data is used to adapt the system further to the voice and environment of the user.

This paper describes a case story view of the lessons learnt from enrolling the first homeService user, M02, and providing him with an online system to work with. This first system has been 'live' in the user's home for three months at the time of writing, and all the voice-commands the user gives when interacting with the system are saved. This has enabled us to do both *online* experiments such as looking at the immediate, 'live' effect of changing acoustic models, vocabulary etc., and behind-the-scenes, *offline* experiments such as investigating various training scenarios with the collected data.

Having access to both online and offline setups enables us

to look in more detail at two main questions. First, how we train acoustic models with sufficient modelling power to provide the user with a reasonably performing system from day one despite having access to only very sparse data. Lack of suitable data is an inherited challenge when attempting speech recognition for disordered speech; potential users are likely to have problems providing large amounts of enrolment data, very few databases exist, and unlike with more *typical* speech one cannot assume that other speakers – typical or disordered – will provide a good enough match to the speech of a particular target user. The speech of some individuals is simply so distinct that it is hard to find good matches in terms of acoustic similarity to include into a good baseline model [3].

Second, we want to look at the issues around establishing and setting up real and successful speech-enabled systems in someone's home. These issues are not just practical to do with choice of vocabulary, hardware and software interface, but also concerns more 'soft' issues such as how you take user preferences into account and keep the user motivated and interested in continuing to use the system.

A third issue about which homeService will inform us, is the question of how to do interesting, reproducible and rigorous research in a domain where studies will only ever have a small number of available users and an even smaller number of funded researchers. In the concluding section 6, we will discuss some of the experiences we have obtained so far from homeService.

Before that however, this paper will address the first two questions by using the homeService user, M02 as a case study. We will provide as much in-depth detail as possible in the hope that this will be informative for other people attempting similar setups. The next section will describe the differences between enrolment and interaction data in the virtuous circle; section 3 will describe the homeService project and system in more details, and sections 4 and 5 will describe the experimental setup as well as the results. Finally, discussion and conclusions will be given in 6.

2 The virtuous circle

A central idea is the notion that the user, through interacting with the system, will provide additional audio data which is used to improve the acoustic models, which should in turn help motivate the user to use the system more - closing the *virtuous circle*.

The principle employed in homeService involves two different data collection strategies: the initial data collection phase, the *enrolment* phase, will collect data through basic recordings. When it is deemed that sufficient enrolment data has been collected, this data is used to adapt a set of speaker independent models to the speech of the user before the system is deployed. From then on the data collection takes place as the user interacts with the system. At regular intervals this data will also be used to update the models so the system continues to tune into the particular characteristics of the user's voice as well as their environment. There are a couple of notable differences between the two types of data:

- **Enrolment Data, ER**: this data is obtained by the user reading lists of the words that they will be using as commands in their system. To get the acoustic conditions to match as well as possible, the recordings take place in the user's home with the type and placement of the microphone matching what it will be when the system goes live. As the user is reading from a list, the identity of each word is known so this data can be used for training in a *supervised* manner.

- **Interaction Data, ID**: this is the data recorded as the user gives commands to the system. He will initiate the process by pressing his switch, which will open the microphone for a predefined number of seconds and this is the data which is saved. In contrast with the ER data, the identity of each word is not inherently known. The user will give a command to the system, which we record. The word is recognised, so we have an automatic speech recognition annotation for the word, but unless we manually transcribe the interaction words at some stage afterwards, we will not have the true annotation. Training with only automatically generated data can be thought of as *unsupervised* training. Another contrast to the ER data occurs as the ID is collected from real use which may affect the speaking style. As the user becomes familiar with the system, and the direct feedback he gets from the system about what has been recognised, he is likely to alter the speaking style to try and ensure as good a performance as possible.

3 The homeService system

The homeService project is the impact showcase for the UK EPSRC Programme Grant Project, Natural Speech Technology (NST,[1]), a collaboration between the Universities of Edinburgh, Cambridge and Sheffield. homeService users are being provided with speech-driven environmental control systems and eventually spoken access to other digital applications.

As part of this process, users are involved with 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'. This is an example of Participatory Design [8].

The system consists of two distinct parts: the atHome and atLab system as displayed in Figure 1.

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 infra-red signals. The atLab system resides at the university and comprises the main server which operates the automatic speech recognition (ASR) system and maintains the system state for each atHome system.

The ASR will run remotely 'in-the-cloud', and be connected to the homeService user's home by a dedicated broadband link. Whilst this is now commonplace for mainstream speech technology, it is relatively novel to see such an ap-

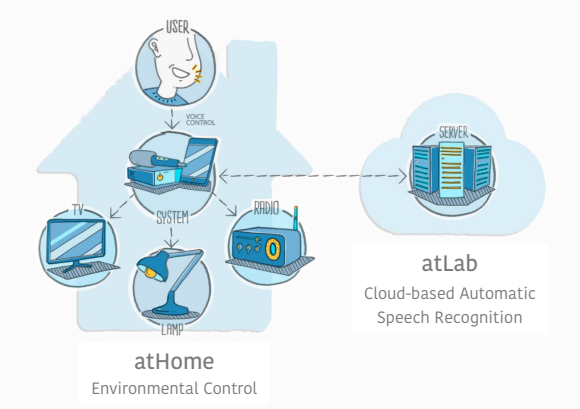

Figure 1. *Diagram of the homeService system with its two distinct parts: the atHome component in a user's home and the atLab 'in-the-cloud' part. Even though only one user is drawn, the cloud-based ASR server enables simultaneous speech recognition from many users.*

proach for providing speech-driven assistive technology. However, there are numerous advantages to choosing this seemingly more complex framework: it 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. This choice was directly motivated by the requirement to look for technological solutions that would scale with the number of users as outlined in section 1. In the first months, working with the first user, being able to monitor and troubleshoot from remote has proven invaluable.

The system hardware consists of 'off-the-shelf' items such as a microphone array, an Android tablet for display and an infra-red transmitter, which reduces the overall cost of each installation, and means the system will not need to rely on specialist hardware. Please refer to [5] for more detailed description of the whole homeService system.

3.1 System installation

After the user has consented to taking part in the study, a visit is conducted where the general setup of the system is decided with the user; the number of devices that the user would like the homeService system to be able to operate is discussed, and for each device, the command words are chosen. This is done with consideration to which words work well for a particular user's speech disorder. The idea of the virtuous circle design is explained to the user, which typically will lead to a selection of one or two devices for the initial setup of the system with a view to adding more devices later. M02, whose system is described here, initially chose to only have his TV/skybox combination operated by his homeService system, and the plan is to add his radio at a later stage. After the vocabulary has been settled upon, the user is asked to repeat each word 5-10 times. M02 went through his initial vocabulary of 33 words five times and for the first, live system this was reduced to 31 words with on average 4.2 repetitions of each word.

Apart from the choice of which devices and words, the definition of the command 'hierarchy' must also be decided with the user. Rather than allowing the user to say any command at any point in time, the user has to navigate through a command, grammar hierarchy. This helps limit the confusability of the recogniser at runtime and hence should improve system performance. At a later stage, when recognition performance has improved, abolishing the grammar hierarchy in favour of a flat, 'word-loop' grammar can be considered.

After the first visit, the initial baseline models are trained in the lab. As so little data has been collected at this stage, the first offline tests use the enrolment data both as training and test set. Once a reasonable performance has been achieved with the models adapted to the enrolment data, 1-2 installation visits are carried out and the user is then ready to work with the system on their own. With this first user, the initial phase was characterised by many smaller adjustments to the vocabulary and grammar. We imagine there will be a similar phase for all users although depending on the user's personality, how much interest they take in optimising and personalising the system will likely vary a fair amount. This is a critical phase where we found it important to communicate extensively with the user to monitor how they are getting on with the system and finding the process as a whole.

4 Experimental setup

The goal of the homeService experiments is to increase the quality of the interaction between users and their home environment. The main source of improvement is naturally the cloud-based automatic speech recogniser. It comprises of two main parts. First, the hierarchical list of the commands which can be adjusted according to the participant's needs. Second, the acoustic model that can be adapted to capture the speaker's characteristics. The former part can be easily deployed in the functioning system and it rarely drops its performance. The development of the latter, on the other side, may need several trial-and-error cycles in order to always provide the user with a better performing system, as per ethical regulation.

The approach selected to improve the system performance can be summarised in three phases: collect data from users with their currently available online system, use some of the data to further adapt the acoustic model, and test the adapted model in a series of offline experiments on another set of recorded data. Finally, once the new model is proven to be suitable for the task, it can be deployed in the online system. In order to confirm offline performance enhancement, previous and new models were alternated every second day. Such an *interleaved* trial design reduces the variability along the time domain of the speaker and enables the contrast of two system on very similar conditions. Essentially, when doing user trials with a small number of users, each individual user is regarded as their own case story and few conclusions can be drawn from comparing the different users to each other as their conditions, systems and

situations vary so much. Instead, from a research point of view, each user becomes an N-of-1, single user trial [9]. As such they will act as their own baseline, so the initial test phase with the enrolment models acts as a baseline for subsequent tests. Unlike e.g. what the case is for drug trials, we are in a position with speech technology to seamlessly replace acoustic models overnight giving us an opportunity to explore interleaved testing with two or more different models.

Particular emphasis is spent on the distinction between online and offline experiments.

Online experiments use the system itself as both source of adaptation data and testing environment. The collected data characteristics are a function of the recogniser accuracy. E.g., if the system mis-recognises a command, this will affect which commands options that become 'live' and are presented to the user. Therefore, online experiments can not be reproduced using e.g. a different acoustic model, because different recognition errors would mean different grammars would be invoked. As a consequence, all offline experiments are carried out using a flat, word-loop style grammar.

This is another motivation why offline experiments are fundamental in the online system improvement cycle.

The following section describes the recorded data.

4.1 Data

The audio data that was used in this paper consists of audio files recorded in real home environment by a single user, *M02*. He has motor neuron disease and a moderate speech impairment. His system is set up to enable him to control his TV and skybox with speech commands.

As mentioned in Section 4, the type of recording is strongly influenced by the parameters of the system that was used for the recordings. Table 1 reports the list of data sets that were collected and their characteristics.

Data set	Date	# entries	time	Use
M02-ER01	20/09/14	130	03'16"	Training of baseline system
M02-ID01	09/03/15–29/03/15	662	46'39"	Second iteration of speaker's adaptation
M02-ID02	30/03/15–08/04/15	216	14'24"	Offline and online experiment evaluation
M02-ID03	09/04/15–29/04/15	713	47'32"	Online experiments, interleaved system with previous model
M02-ID04	24/04/15–11/05/15	209	13'56"	Online experiments, interleaved system with new models

Table 1. *Descriptions of data sets recorded by participant M02.*

M02-ER01 is the only data set which was not recorded through the online system. It represents the first enrolment data used to generate the baseline system.

Each of the data sets represent slightly different word usage characteristics. Figure 2 shows a representative histogram of which vocabulary words were used by M02 for ID02, the test set used for the offline experiments described in this paper. Clearly, the word usage is far from linear with some words being used a lot (e.g. 'TV') and others hardly at all (e.g. 'on' and 'off'). Some of these differences are due to the characteristics of his system, such as him needing to say 'TV' every time he wants to access any TV-related commands. Other high-usage words are purely a result of him investigating a particular part of the menu at that point, e.g. using the 'sport one' command a lot more than 'sport two'.

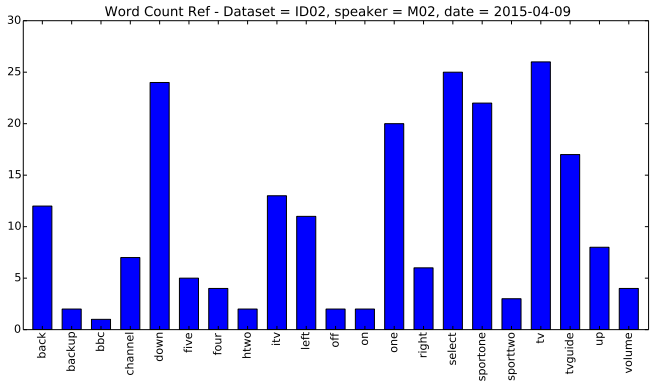

Figure 2. *Word Usage in the training set M02-ID01.*

5 Results

The datasets described in the previous section forms the basis for the online and offline results described in this section.

5.1 Online results

Data set	Acoustic model	Word set	# spoken words (Tot)	Accuracy	OOG
M02-ID01	mapER01	d1	18 (28)	86.87%	13.16%
		d2	13 (28)	55.90%	1.23%
		d3	25 (28)	76.92%	5.65%
M02-ID02	mapER01	d3	21 (28)	74.16%	3.24%
M02-ID03	mapER01	d3	26 (28)	60.97%	1.54%
M02-ID04	mapER01+ID01	d3	23 (28)	91.16%	0.46%

Table 2. *Online recognition results on data sets recorded from participant M02.*

In Table 2, online recogniser performance on different data is shown. This has been calculated by comparing the online recognition result with the manual annotation subsequently given to the word whilst taking into account the particular grammar or state the system is in. The online accuracy for all the datasets collected whilst using mapER01 ranges between 55.90% to 86.87%, however, when the new and improved models are introduced (mapER01+ID01), the accuracy increases to 91.16% percent; an average relative increase of over 28%.

The out of grammar (OOG) denotes words being said that were not included in a particular grammar. For example if the user says 'BBC one' at a time where the system grammar is expecting a device name such as 'TV', 'radio' or 'lamp'. The amount of OOG words spoken in the online experiments influence the recognition score. For this reason, these kind of errors have not been taken into account in the accuracy measurement.

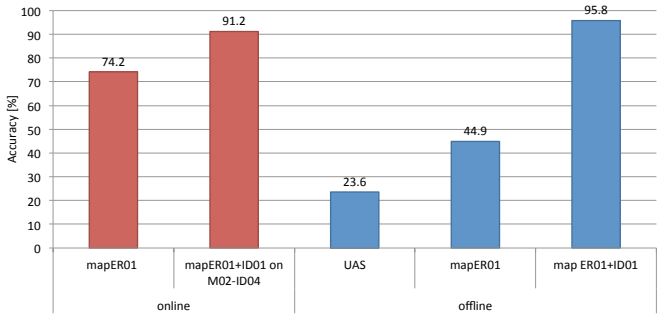

Figure 3. *Comparison among different acoustic model in online and offline experiments on M02-ID02 dataset except where stated. Please note that the online test with newly adapted model could not be tested on the same data, due to online tests characteristics.*

Before looking in more details at the offline results, Figure 3 compares a few key online and offline results. The first offline result is the result from testing a basic, speaker independent model trained on UAspeech ([7]) with the ID02 data set. This model is used as the base model from which the mapER01 is adapted, and it is seen that this MAP adaptation in itself improves the accuracy from 23.6% to 44.9%. Adding the ID01 further improves this to 95.8%. How are these improvements reflected in the online results?

The mapER01 (the baseline model at deployment) gave an accuracy of 74.2% and the mapER01+ID01 (the first updated model) gave an accuracy of 91.2%. Especially the former is considerably higher than the offline results because of the restrictions imposed by the hierarchical grammar used for the online system. However, another difference is the test set used, which does make comparisons difficult. For the online result, this is the test set that happened to be collected at the time of deployment of the various models; for the offline results it was chosen to base all the results on testing with the ID02 test set.

Figure 4 analyses the online mapER01 result in more depth by looking at the confusions that occur. Noting that these confusions are restricted by the current, online grammar it is clear that some grammars provide far more 'opportunities' for confusions. Some patterns can also be explained, such as the many examples of the word 'sport one' that we observed in Figure 2. This word has got especially many confusions and M02 has most likely been wanting to try, repeatedly, to get the recogniser to recognise the word correctly.

5.2 Effect of varying amounts of adaptation data

Offline experiments are informative both in their own right by giving insight into the data, and as a predictor for the performance in the online system. To make the offline experiments as useful for predicting the online performance as possible, they are carried out with conditions as closely matched as possible to those of the online system. This means using as a baseline model the initial model for when the system first went 'live': the model adapted using the ER01 data.

Improved performance of the models are obtained through

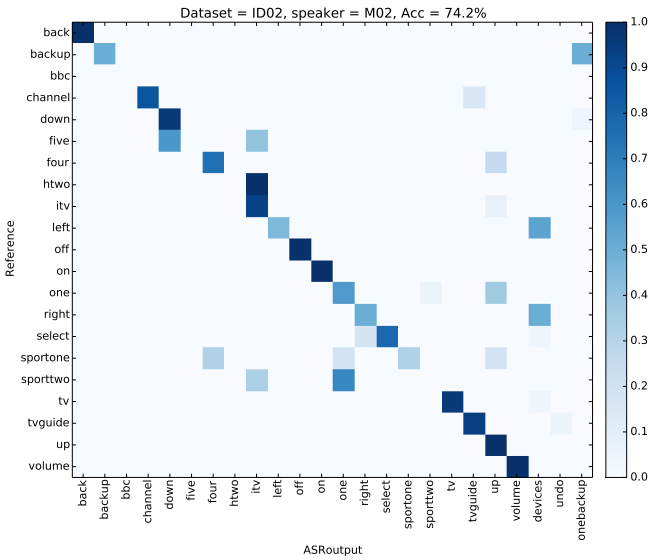

Figure 4. *Normalised confusion matrix - confusion with mapER01 models on M02-ID02.*

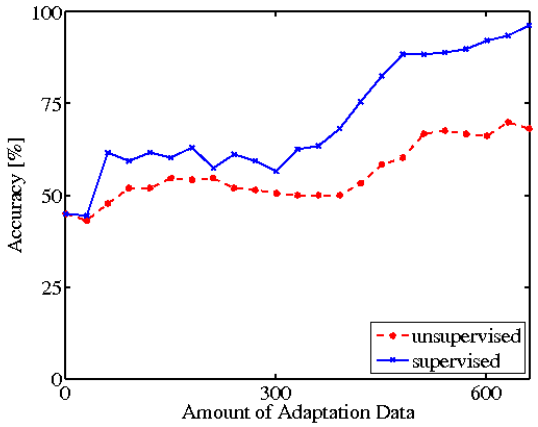

Figure 5. *Effect on accuracy of using different amounts of adaptation data; supervised vs. unsupervised annotations.*

adaptation to increasing amounts of interaction data as it is collected. Offline, it is therefore interesting to investigate the relationship between the amount of adaptation data used and the resulting accuracy of the system. Figure 5 shows how the accuracy is affected by the amount of adaptation data. The adaptation data is an increasingly larger subset of the ID01 data. The models are the mapER01 baseline models adapted to a given amount of adaptation data and tested with the full test set, ID02. The interaction data are ordered chronologically and each increase in adaptation data amount are done in stepsizes of 30 interactions. For this particularly speaker, that corresponds roughly to a day's worth of data.

On Figure 5, the solid line corresponds to the case where all the adaptation data has been manually annotated. As expected, overall, the accuracy increases when using increasing amounts of adaptation data. However, the increase is far from

linear; instead it progresses in jolts and with 'plateaus' in between. Adding just the first 30 data points, about a day's worth of interaction data, does not change the accuracy significantly, whereas using 60 interaction words increases the performance from an initial baseline of 44.9% accuracy to 61.6%. This level then defines the first plateau with performances between 57.4% and 63.4% until the 400 word mark, where a sharp increase takes the level up to around 88% for a final, gradual increase.

This unusual profile reflects the fact that the user had an initial customisation phase where he was finding his way around the system. During that initial phase, a smaller set of voice commands are used as the emphasis of the user is more on learning how the system is used, and less on exploring all the different menus and commands available.

5.3 Effect of annotation type - supervised vs. unsupervised

On figure 5 in contrast to the solid lines corresponding to the 'supervised' case, the dashed line is the accuracy achieved with using varying amounts of adaptation data that has been annotated automatically (unsupervised adaptation). That is using annotations corresponding to what would have been recognised if tested with initial baseline models, mapER01. As expected, the accuracy is much lower than for the supervised case, although an improvement in accuracy can be seen when $>$400 interaction words are used. This appears to be at the same point where the supervised curve also shows increased improvement.

The two curves on Figure 5 shows the two extreme cases of either being in a position where all interactions can be manually annotated (supervised) or where all of the interaction data is automatically annotated (unsupervised). In a real system, it can be expected that the situation will be somewhere in-between: there may be resources to annotate some initial data manually up to a point in time after which there is a swap to using automatic annotations for all subsequent data.

6 Discussion and Conclusions

This paper has described details about the first homeService user's progression through the various phases involved with setting up a system tailored to his voice characteristics as well as vocabulary preferences. We have detailed our experiences with the enrolment and adjustment phase as well as the online experimental phase of attempting to interleave acoustic models. We have also described how we have used offline experiments to support the online choices. Further offline experiments have been carried out giving more insight into the use of adaptation data and annotation efforts in terms of manual vs. automatic annotations. In summary, we have found:

- Offline experiments can be used to support the choices of models for a given user's online system although, it is clear that actual system usage such as word distribution and runtime grammar and vocabulary will greatly influence the actual online accuracy achieved.
- The benefits of cloud-based setup are numerous and significant both to the user (swift troubleshooting, quick update of models, less inconvenience around system maintenance) and for the research team (easy monitoring of performance and "harvesting" of data from each user to add to the pool of data).
- The practicalities of collecting and building a database from real interaction data causes such data to differ from properly structured and carefully planned datasets. This means that care has to be taken when inferring from offline results on databases to decisions affecting a live user trial. In future work we will look at how offline experiments on the UAspeech database of dysarthric speech ([7]) ports to real, 'in-the-wild' results, e.g. [6, 2, 3].

As the future homeService users are enrolled and have systems installed we will continue to monitor and pool data and experiences.

7 Acknowledgements

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

References

[1] The natural speech technology (NST) homepage.

[2] H. Christensen, M. B. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In *Interspeech'13*, 2013.

[3] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain. Automatic selection of speakers for improved acoustic modelling : Recognition of disordered speech with sparse data. In *Spoken Language Technology Workshop, SLT'14*, Lake Tahoe, Dec 2014.

[4] 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.

[5] H. Christensen, S. Cunningham, P. Green, and T. Hain. homeservice: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition. In *4th Workshop on Speech and Language Processing (SLPAT)*, 2013.

[6] H. Christensen, P. Green, and T. Hain. Learning speaker-specific pronunciations of disordered speech. In *Interspeech'13*, 2013.

[7] 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*, pages 22–26, Brisbane, Australia, 2008.

[8] Suchman L. *Participatory Design: Principles and Practices.*, chapter Forward, page viiix. N.J.: Lawrence Erlbaum, 1993.

[9] P. A. Scuffham, J. Nikles, G.K. Mitchell, M.J. Yelland, N. Vine, C.J. Poulos, P.I. Pillans, G. Bashford, C. del Mar, P.J. Schluter, and P. Glasziou. Using n-of-1 trials to improve patient management and save costs.

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