A Framework for Collecting Realistic Recordings of Dysarthric Speech - the *homeService* Corpus

Mauro Nicolao[1], Heidi Christensen[1], Stuart Cunningham[2], Phil Green[1], and Thomas Hain[1]
[1]Computer Science; University of Sheffield, United Kingdom
[2]Human Communication Sciences, University of Sheffield, United Kingdom
{m.nicolao, h.christensen, s.cunningham, p.green, t.hain}@sheffield.ac.uk

Abstract

This paper introduces a new British English speech database, named *the homeService corpus*, which has been gathered as part of the *homeService* project. This project aims to help users with speech and motor disabilities to operate their home appliances using voice commands. The audio recorded during such interactions consists of realistic data of speakers with severe dysarthria. The majority of the homeService corpus is recorded in real home environments where voice control is often the normal means by which users interact with their devices. The collection of the corpus is motivated by the shortage of realistic dysarthric speech corpora available to the scientific community. Along with the details on how the data is organised and how it can be accessed, a brief description of the framework used to make the recordings is provided. Finally, the performance of the homeService automatic recogniser for dysarthric speech trained with single-speaker data from the corpus is provided as an initial baseline. Access to the homeService corpus is provided through the dedicated web page at `http://mini.dcs.shef.ac.uk/resources/homeservice-corpus/`. This will also have the most updated description of the data. At the time of writing the collection process is still ongoing.

Keywords: dysarthric speech corpus, speech disorders, speech corpus collection, realistic environment recordings, automatic dysarthric speech recogniser.

1. Introduction

The world of digital devices is becoming more and more oriented towards speech-enabled interfaces. Speech is an attractive alternative to more traditional input devices, such as remote controls, keyboards, or PC mice especially when environmental or physical constraints make conventional interfaces difficult to use.

The success of speech-driven digital devices often depends on the recognition accuracy of the interface. Good results have been obtained for speakers with typical speech, however speech recognition can be challenging for people who have disordered speech associated with neuro-motor conditions. These people are also less likely to be able to use conventional interfaces due to their impaired motor function. Performance is adversely affected by the increased variability that characterises *disordered* and *dysarthric* speech. Hence it is essential to have access to representative audio signals with which the acoustic model, at the core of state-of-the-art automatic speech recognition (ASR) techniques, can be trained or adapted (Sharma and Hasegawa-Johnson, 2012; Christensen et al., 2012).

Very few databases exist, and 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 atypical that it is hard to find good matches in terms of acoustic similarity to include into a good baseline model (Christensen et al., 2014). Most of the existing dysarthric speech databases, such as the TORGO (Rudzicz et al., 2011) and UAspeech (Kim et al., 2008) corpora were collected in controlled conditions, i.e. a quiet laboratory environment with a fixed, common word set for all speakers, read speech, and with no real interaction with controlled devices.

This paper introduces a new British English speech database, named *homeService*, which has been gathered as part of the *homeService* project (see Section 2.). The corpus contains a collection of natural interactional speech data from speakers with dysarthria.

The data has been collected as speakers have used speech-commands to operate an environmental control system installed in their home, over aa extended period of time. This approach to data collection (*in-the-field*) aims to overcome the lack of suitable data when potential users are likely to have difficulties providing large amounts of enrolment data.

2. The homeService system

The homeService project is an impact showcase for the UK EPSRC Programme Grant Project, Natural Speech Technology (NST) (The Natural Speech Technology (NST) Programme Grant, 2015), a collaboration between the Universities of Edinburgh, Cambridge and Sheffield.

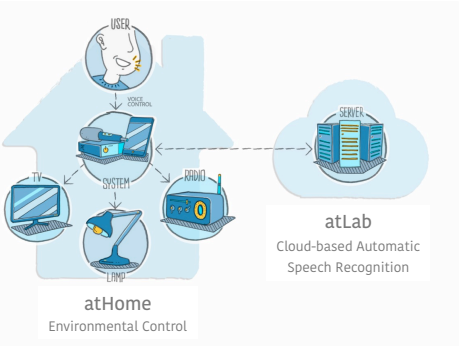

Figure 1: Diagram of the two distinct parts in the homeService system: the atHome in users' home and the atLab 'in-the-cloud' components. One user is illustrated, but the cloud-based ASR server enables simultaneous speech recognition from many participants.

The homeService project is concerned with how speech technology can be of use for people with speech disorders and restricted upper-limb mobility.
The project implements a cloud-based environmental control system where users can operate electronic devices such as TVs, radios, lamps, etc. through the use of voice-commands (Christensen et al., 2013). The homeService system users are being provided with speech-driven environmental control systems and eventually spoken access to other digital applications. 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.
Prior to starting the project, ethical approval was obtained through the UK NHS Research Ethics procedure. All speakers taking part in the project were informed about the study and gave their consent to take part. The on-going consent was confirmed periodically during the project.
The system consists of two distinct parts: the *atHome* and *atLab* systems as displayed in Figure 1.
The atHome system is deployed in a user's home and comprised of "off-the-shelf" hardware. A low-power computer (shuttle) was connected to a microphone array (Microcone[1]) and a Samsung Galaxy tablet computer running Android was used to display the current status of the system to the user. Infra-red control signals were sent using a transmitter connected to the PC. Relying on non-specialist tools reduces the overall cost of each installation which therefore, can be deployed to a larger number of users simultaneously. Before installation of the atHome system the number of devices that the user would like the homeService system to operate and the related command words were agreed. Therefore, each speaker in the database has a personalized list of commands suitable for their needs. Based on these devices, a hand-crafted hierarchical dialogue manager is used to help minimize the possible confusions at each stage of the interaction.
The atLab system consists of a Linux computer server on the university site, which operates the ASR system and maintains the synchronisation with all the atHome systems. The homeService ASR is set up as a distributed resource running 'in-the-cloud' and can be accessed by users via a dedicated broadband link. Whilst this is now commonplace for mainstream speech technology, it is a novel approach for speech-driven assistive technology and has numerous advantages. It enables us to collect natural speech data in a real environment, synchronise acoustic models and vocabularies with the participant's usage experience, experiment with different adaptation algorithms, and so on. All these actions can be performed without affecting the users' normal life in their home. In the first months of the data collection process, this architectural choice has allowed us to work side-by-side with the first user, monitor and troubleshoot any issues remotely. Please refer to (Christensen et al., 2013) and (Christensen et al., 2015) for more detailed description of the whole homeService system.

[1]http://www.dev-audio.com

3. Data collection process

After collecting a reasonable amount of user command audio examples (*enrolment data, ER*) under a researcher's supervision, an initial adapted model is trained in the lab. This model is then used in the ASR which drives the interaction between user and system. The size of the ER data is critical. If it is too small, there is the risk of not being able to give the user a successful system. Conversely, spending a lot of time recording data without a tangible system runs the risk of disengaging the user, since, especially for dysarthric users, speaking is often an effort.
From the moment of deployment, the user is free to experiment with the system in his/her home, using it without any researcher's supervision. These interactions (*interaction data, ID*) are therefore naturally more spontaneous than those recorded in a laboratory environment. ER datasets are manually annotated with an automatic web interface by human listeners. Both manual and ASR transcriptions are available for ID datasets. If the user requests an update in terms of system functionalities or performances, new adapted models or commands can be introduced remotely, without interfering with the recording process except for a short down-time. In addition, at regular intervals, portions of ID data are also used to update the models, so that the system continues to tune into the particular characteristics of the user's voice as well as their environment.
One of the peculiarities of this data collection system is that users are directly involved with the design and specification of the functionality of their personal system. This constant collaboration with users helps to close what is referred to as the *virtuous circle*. This example of Participatory Design (Suchman, 1993) allows for the user, through testing the system, to provide additional audio data which is used to improve the acoustic models. Moreover, this method engages the user in the research team and motivates the user to continue using the systems.
Finally, it is worthwhile to highlight that the constant model adaptation to a user's speech, in order to improve ASR quality, is fundamental to achieving a satisfying recording process. As the ASR accuracy affects the quality of dialogue-based interaction and, hence, constant wrong command recognitions would frustrate the user preventing him/her from using the system in a natural manner, the best available system must be always deployed.

4. The homeService corpus

As described in §3., the homeService corpus consists of two type of speech data:

- **enrolment data, ER**: this data is obtained by the user reading lists of the words that they have chosen as commands in their system. To match the acoustic conditions in user's home, the recording takes place with the same hardware and in the same environment in which the system is supposed to function. As the user is reading from a list, the resulting speech will be less natural but is still effective for initial training.

- **interaction data, ID**: this is the data recorded as the user operates the electronic devices in the house with

the homeService speech enabled interface. Recording starts after the user presses a switch and the microphone is open for a predefined number of seconds. In contrast with the ER data, the identity of each word produced is not inherently known. The utterance is recognised and the audio is saved, so it can be manually annotated and used off-line afterwards. Another contrast with the ER data occurs as the ID is collected from real use which may affect the speaking style.

As mentioned in §2., each participant's audio is recorded with a Microcone microphone array in a real environment. The complete 8-channel output (6 streams from the microphone array plus a beam-formed stereo combination of those) is stored at 48kHz sampling rate and 32bit definition. A 16kHz-sampled, single channel version of the audio is also produced and transferred via the broadband connection to the cloud-based ASR. In the current database release, only the 16kHz version of the audio is provided.
The duration of each audio segment differs from user to user depending on the length of adopted commands and on their impairment. It is normally tuned to completely contain the single user command without truncation along with some related background sounds. E.g., for M02, the duration is 4 s, but for F02, who usually introduces short pauses in her production, it had to be increased to 6 s.
Participants use the system in their real home environment. Thus, all sorts of natural house sounds are present in the recordings. This was part of the homeService design requirements as the goal is to develop a system that works in real environments.
At the time of writing, five users have been enrolled in this part of the data collection. Some information about the users is reported in Table 1. Unfortunately, two of them, M01 and F01, withdrew after the enrolment data recording, due to personal reasons. This proves, if needed, the importance of this sort of recording as it is challenging not just from the technological point of view.

User	Gender	Age	Native Language	Condition	State of recording
M01	male	47	GBEng	cerebral palsy	interrupted
M02	male	75	GBEng	motor-neuron disease	ongoing
M03	male	22	GBEng	cerebral palsy	ongoing
F01	female	55	GBEng	cerebral palsy	interrupted
F02	female	54	GBEng	–	ongoing

Table 1: List of participants involved in the homeService corpus collection.

The recording process has just started in its complete form. At the time of writing, the amount of data that has been recorded is shown in Table 2. More updated statics will be found at the homeService corpus web page.

User	Type of data	Vocabulary	Interaction #	Duration (m)	Annotated
M01	ER01train	31	230	6'34"	yes
M02	ER01train	31	130	3'16"	yes
	ID01train	76	5038	5h 38'00"	yes
	ID01test	38	1516	1h 42'00"	yes
M03	ER01train	12	114	2'47"	yes
F01	ER01train	32	97	2'19"	yes
F02	ER01train	31	314	11'58"	yes

Table 2: Amount of data collected at the time of writing.

One of the participants, M02, has been helping us develop and tune the homeService system. He has been enthusiastically recording for almost a year and, therefore, considerably more audio material is available for him.
Transcriptions are also made available for each speaker's dataset.
In the ER datasets, transcription is derived from the prompt text that was provided to the users during this stage of recording. Participant's vocabularies are distinct and overlap on very few common words only. This is due to the different needs of environmental control for each user and their capabilities of pronouncing words of different length.
In the ID datasets, on the contrary, audio have been transcribed and annotated by human listeners. An example of the word distribution is displayed in Figure 2. It reports the 44 word usage in the ID data of user M02.

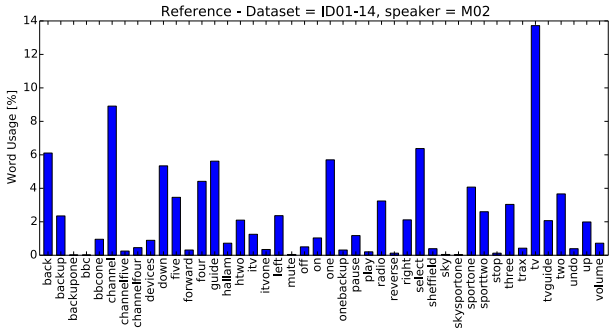

Figure 2: Word usage in the current M02 ID dataset.

All the audio data which resulted in an actual interaction with the homeService control system was annotated. When audio is not intelligible or other sound events appear in the recording, special upper-case tags are used to highlight it. Example of those are CORRUPT, DISCARD, DISTORTED, EMPTY, INCOMPLETE, NOISE, UNSURE.
In Table 2 and consequently in the corpus, the purpose of each set is also highlighted in order to establish clear training and test sets for future experiments on this corpus. Whilst ER data is supposed to be used only in the training stage (labelled *ER01train*), ∼20% of ID audio is reserved for testing. Test data (*ID01test*) was selected to have significant coverage of the words used by each user using the homeService system.
The homeService corpus is available under some usage restrictions due to the sensitive nature of such data through the web page at `http://mini.dcs.shef.ac.uk/resources/homeservice-corpus/`.

5. Automatic speech recognition experiments

The first automatic dysarthric speech recognition (d-ASR) experiments on this corpus are reported for the M02 ID data.
In all the experiments of this section as well as in the homeService system, a standard state-of-the-art HTK-based ASR engine is used. Perceptual linear prediction (PLP) acoustic features are extracted from the speech signal. A dysarthric background model is tailored to the specific speaker's

speech style with maximum a posteriori (MAP) or maximum likelihood linear regression (MLLR) adaptation. All the entries in the grammar have the same weight. The dysarthric background model is trained on the UA-Speech Database (Statistical Speech Technology Group, University of Illinois, 2013).

The homeService d-ASR has been tested on M02 ID data with both *online* experiments, which focus on the immediate, 'live' effect of changing acoustic models, vocabulary, etc., and on the behind-the-scenes *offline* experiments investigating various training scenarios.

In the online experiments, the performance of the d-ASR operating in the homeService system has been measured during the recording process. Overall, the recogniser in M02's home scored a satisfying accuracy of 77.75%. Though, the single trial numbers were extremely dependent on the amount of data available at that particular time to create the adapted models. Moreover test data is not consistent as it is dependent on the user's daily usage. Accuracy varies from 55.90% in the early trials (acoustic models adapted on M02 ER01train data only) up to 86.60% in the latest ones (models created with the whole ID01train data).

In the offline experiments, on the other hand, the test dataset has been well defined. The M02 ID01test audio samples are taken from recordings over a two-month period and represent the normal M02 usage. Results shown in Table 3 data establish the first baseline for speech recognition on this data.

Different amount of adaptation data as well as three adaptation techniques are tested. Along with the single MAP and MLLR adaptations, a cascade of MAP followed by a further MLLR adaptation has been tested. The latter was considered because it was observed to better generalise in the online experiments in case of words not available in the training set.

			Accuracy on ID01test [%]			
Training sets	Data size	# words	None	MAP	MAP2MLLR	MLLR
UAS	$\sim$ 12h03'	455	17.6	-	-	-
UAS +ER01train	3'16"	31	-	35.8	36.1	38.0
UAS +ER01train +ID01train	6h 14'	51	-	92.2	91.8	86.7

Table 3: Baseline results using M02 data for training and test.

6. Conclusions

To the best of our knowledge, this is the first database of *spontaneous* dysarthric speech recorded in a real-world environment. The task at hand (voice-enabled environmental control) means that the word diversity (voccabulary) is relatively small in the corpus, however, this is offset by the variety arising from the rich and realistic acoustic recording conditions that is a direct result of the homeService collection strategy: placing a fully functioning system people's homes and observing their interactions as they go about using the system as it was designed.

Another characteristic of the ID data in the homeService corpus is that it contains examples recorded over several months. This allows for longitudinal studies on voice variations caused by degenerative speech impairment or by the user altering his/her voice to compensate for the system's imperfections.

Even though only ER data can be released at the time of writing, for most of the users, the process to collect ID data for M03 and F02 is ongoing. However, M02 ID data alone allows us to demonstrate the quality and the amount of data that are going to be recorded with the homeService system. According to the time plan of the project, the homeService data collection is expected to be complete in Spring 2016. However, a considerable amount of data is foreseen to be available by the end of 2016. By the time of publication, we anticipate to have data collected from five dysarthric speakers using the system over a period of three months.

7. Acknowledgement

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

8. Bibliographical References

Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *Proc Interspeech 2012*, Portland, Oregon, US, Sep.

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

Christensen, H., Casanueva, I., Cunningham, S., Green, P., and Hain, T. (2014). 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.

Christensen, H., Nicolao, M., Cunningham, S., Deena, S., Green, P., and Hain, T. (2015). Speech-Enabled Environmental Control in an AAL setting for people with Speech Disorders: a Case Study. In *IET International Conference on Technologies for Active and Assisted Living*, London, UK, October.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Proceedings of Interspeech*, pages 22–26, Brisbane, Australia.

Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2011). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation.*, pages 1–19.

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

Suchman, L., (1993). *Participatory Design: Principles and Practices*, chapter Forward, pages vii–ix. N.J.: Lawrence Erlbaum.

The Natural Speech Technology (NST) Programme Grant. (2015). homepage: http://www.natural-speech-technology.org/.

9. Language Resource References

Statistical Speech Technology Group, University of Illinois. (2013). *The UA-Speech Database*. http://www.isle.illinois.edu/sst/data/UASpeech/.