Automatic speech recognition for people with disordered speech: results from online and offline experiments

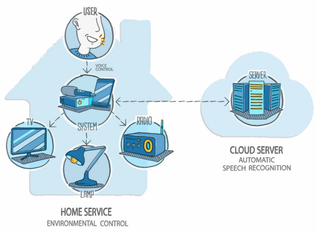
Mauro Nicolao, Heidi Christensen, Salil Deena, Stuart Cunningham, Phil Green, Thomas Hain

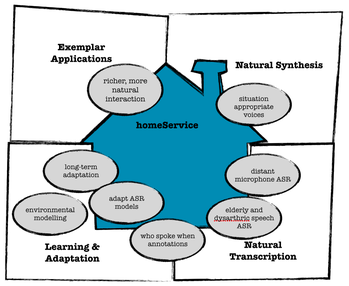
The University Of Sheffield.

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

homeService project

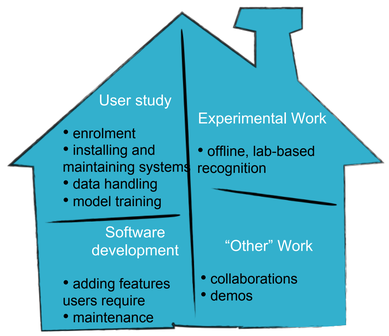
Project Goal: recognising disordered speech in people's homes using state-of-the-art techniques developed.





The many sides of homeService

- Challenging application for Natural Speech Technology:
- How can we best use methodologies and data from main-stream, typical speech ASR?
- How do we best tune an ASR system to the non-typical elements of a dysarthric speaker?
- How do we find the best operating point for a personalised, homeService ASR system?



- use of state-of-the-art training strategies for dysarthric speech
- use of typical speech knowledge for dysarthric speech
- automatic derivation of pronunciation dictionaries for dysarthric speech
- setting up initial system: choice of vocabulary, enrolment data requirement, etc.

- system currently deployed in one extremely enthusiast participant's (M02) home, recording 30-40 interactions a day
- several hours of real interaction data already recorded,
- two participants (male and female) in the enrolment process (system deployment expected by end of June)
- every participant comes with different needs and specific requests
- a well established protocol for system installation at participant's house and development to their new needs.

Future work

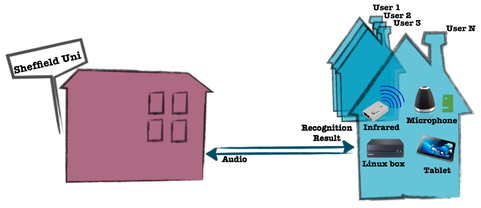
- complete the recruitment of more participants,
- change towards a keyword activation system instead of push-to-speak,
- move from command word recognition to a more natural phrase recognition,
- add Deep Neural Network (DNN) models to the atLab recogniser,
- always deliver to participants the best available system,
- involve participants in the research as much as possible, following their feedback.

Experimental data

- UAspeech: largest, English database of dysarthric speech, 16 speakers, 18hrs.
- Enrolment data (ER), speaker dependant data: offline recordings of the keywords the participant uses to control the system. Several repetitions of a list of commands in the vocabulary ($\sim$ 30 items)
- Interaction Data (ID), speaker dependant data recorded by the participant while functioning the system. Several not uniformly distributed repetitions of given list of commands

Data set	ASR system	Word set	Date	# entries (time)	Use
M02-ER01	manual	d0	20.09.14	130 (03'16")	First stage of MAP adaptation
		d1	09-12.03.15	120	
M02-ID01	mapER01	d2	13-19.03.15	174 (57'38")	Second stage of MAP adaptation
		d3	20-29.03.15	384	
M02-ID02	mapER01	d3	30.03.15-08.04.15	216 (14'24")	Offline experiments
M02-ID03	mapER01	d3	09-29.04.15	713 (47'32")	Online experiments
M02-ID04	mapER01+ID01	d3	24.04.15-11.05.15	209 (13'56")	Online experiments
M02-ID05	mapER01+ID01	d4	11-18.05.15	211 (14'04")	Online experiments

Online Experiments



Online homeService

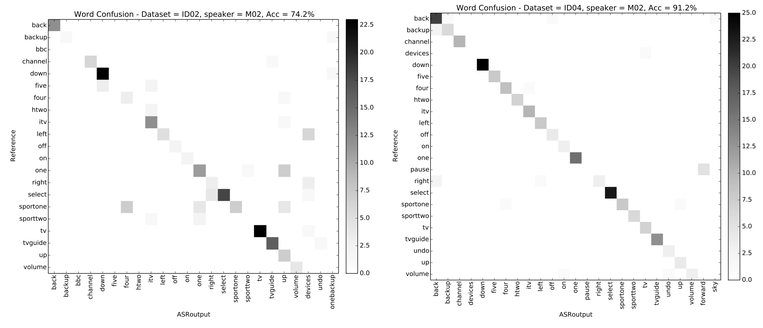
- word loop grammar restricted to one-word recognition per utterance, using a menu-based dialog manager,
- recognition dependant on dialog manager state,
- acoustic models trained with MAP adaptation on top of UAspeech.

Performance test on the real use of homeService

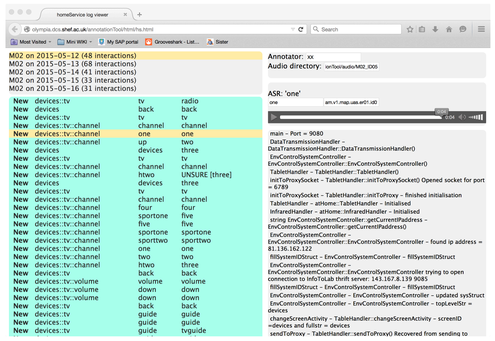
- 2 models used in the online experiments: mapER01 and mapER01+ID01,
- M02-ID03 and M02-ID04 recorded interleaving the two acoustic models,
- M02-ID05 recorded with mapER01+ID01 but different vocabulary to test model overfitting.

Data set	ASR system	Word set	# used words (tot)	Word Accuracy	OOG
		d1	18 (28)	86.87%	13.16%
M02-ID01	mapER01	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%
M02-ID05	mapER01+ID01	d4	23 (29)	82.90%	7.21%

- performance dramatically increased with adaptation to more data (mapER01+ID01),
- system is also overfitted to the d3 word set, no recognition of new d4 words (e.g. M02-ID05 vs M02-ID04)



Annotation of Interaction Data

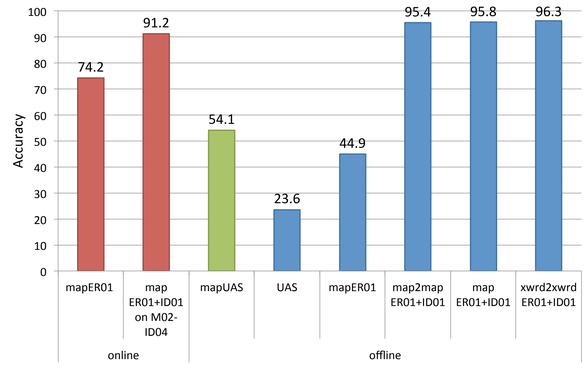


- new recorded audio from participant needs annotation to be be used as adaptation data,
- supervised annotations: humans transcribe what is said using a web interface,
- unsupervised annotations: ASR transcription from the online system,
- annotation tool developed in collaboration with Jonathan Kilgour

Offline Experiments

- Offline experiments have no dialog manager,
- more lexical confusion than online experiments, every word same probability,
- more flexibility to try different configurations to eventually deploy to users.

Acoustic model performance on M02-ID02



- UAS: UAspeech SI models
- mapER01: UAspeech SI models + MAP adaptation with M02-ER01 data, tested on M02-ID02,
- mapER01+ID01: UAspeech SI models + MAP adaptation with M02-ER01 and M02-ID01 data, tested on M02-ID02,
- MAP2MAP ER01+ID01: MAP adaptation with ID01 data on top of mapER01,
- XWRD2XWRD ER01+ID01: training from scratch using only data in M02-ER01 and M02-ID01,
- mapUAS: UAspeech SI models + MAP adaptation for each corpus speaker, tested on UASpeech,
- mapER01 and mapER01+ID01: same acoustic model as above on online test set data.

Dependency on the amount of supervised data in MAP adaptation

- Adapting with varying amounts of data from M02-ER01 and M02-ID01 and testing on M02-ID02,