A comparative study of adaptive, automatic recognition of disordered speech

Heidi Christensen[1], *Stuart Cunningham*[2], *Charles Fox*[1],
Phil Green[1], *Thomas 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,s.cunningham@sheffield.ac.uk,charles.fox@dcs.shef.ac.uk,
p.green@dcs.shef.ac.uk, t.hain@dcs.shef.ac.uk

Abstract

Speech-driven assistive technology can be an attractive alternative to conventional interfaces for people with physical disabilities. However, often the lack of motor-control of the speech articulators results in disordered speech, as condition known as *dysarthria*. Dysarthric speakers can generally not obtain satisfactory performances with off-the-shelf automatic speech recognition (ASR) products and disordered speech ASR is an increasingly active research area. Sparseness of suitable data is a big challenge. The experiments described here use UAspeech, one of the largest dysarthric databases available, which is still easily an order of magnitude smaller than typical speech databases. This study investigates how far fundamental training and adaptation techniques developed in the LVCSR community can take us. A variety of ASR systems using maximum likelihood and MAP adaptation strategies are established with all speakers obtaining significant improvements compared to the baseline system regardless of the severity of their condition. The best systems show on average 34% relative improvement on known published results. An analysis of the correlation between intelligibility of the speaker and the type of system which would represent an optimal operating point in terms of performance shows that for severely dysarthric speakers, the exact choice of system configuration is more critical than for speakers with less disordered speech.

Index Terms: dysarthric speech, speech recognition, speaker adaptation

1. Introduction

Dysarthria is the blanket term for a range of disorders which arise from a loss of control of the speech articulators. It is the most common speech disorder affecting 170 per 100,000 of the population [**?**] There are a number of underlying causes: congenital conditions such as cerebral palsy or acquired neurological conditions as a result of stroke or traumatic brain injury. There is a taxonomy of dysarthrias [**?**] and there are established assessment procedures for speech and language therapists [**?**]. People with severe dysarthria can be close to unintelligible to unfamiliar listeners, though they can generally communicate successfully with family and friends. Dysarthria often co-occurs with physical disability, and the inability to use conventional keyboard-and-mouse interfaces or operate assistive technology makes speech control an attractive alternative, even though the speech is degraded.

There have been a number of small-scale studies of ASR for dysarthric speech using conventional techniques (see [**?**, **?**] and [**?**] for reviews), with patchy results. Some success has been achieved for mild-to-moderate dysarthria but there is an inverse relationship between the degree of impairment and the accuracy of 'off-the-shelf' speech recognition. Perhaps the best performance with more severe conditions has been reported by [**?**], which was based on small-vocabulary, speaker-dependent (SD) whole-word recognisers built from limited amounts of training data.

Speech recognition technology has increasingly relied on large corpora, but until recently only the Nemours database [**?**] was commonly available for research into dysarthric ASR. It contains a total of 814 sentences from 11 male speakers with mild-to-moderate impairments. In the last five years, however, this has changed with the appearance of the UAspeech database [**?**], which is used here, and TORGO [**?**]. These corpora are still small by modern LVCSR standards: the baseline recogniser used in this paper was trained on just over 170 hours of normal speech whereas UAspeech has around 18 hours[1] and the TORGO recordings amount to 23 hours, not all of which is disordered speech. Nevertheless UAspeech and TORGO make it possible to apply at least some of the modern training and adaptation techniques to dysarthric speech, and thus to arrive at phone-level context-dependent models.

The problem of very limited training data reflects the situation researchers and clinical scientists face when deploying ASR systems 'in-the-wild' for dysarthric users. People who would like to use speech-driven assistive technologies, but who have physical disabilities which also affect their speech often find it difficult to contribute a sufficient amount of data; for some dysarthric speakers supplying even a couple of minutes can be very tiring and also lead to distress, especially for people with degenerative illnesses such as Parkinson's disease.

The question is then how to best make use of such limited data. Adaptation techniques, notably maximum likelihood linear regression (MLLR) [**?**] and maximum a posteriori (MAP) [**?**] are used in LVCSR to tune speaker independent (SI) recognisers to the speech of an individual, resulting in an SD system, with a relatively small amount of adaptation data. There is, however, an assumption in these procedures that the target speech is not a gross mismatch to that used to train the SI models. The viability of this assumption for dysarthric data can be expected to depend on the severity of the condition. In [**?**], MAP adaptation was used on UAspeech data with some success and in section **??**, we compare the results of this study to ours. [**?**] report encouraging results on TORGO. Their baseline was a SI monophone model recogniser trained on a mixture of normal and dysarthric data. Acoustic model adaptation by MLLR

[1]After leading and trailing silence is cut off as described later in the paper.

resulted in a 16% absolute reduction in error rate, which was somewhat improved by speaker-dependent pronunciation models.

'In-the-field' work with ASR can be challenging at the best of times, but assistive technology users have a very high demand for reliability. If a system exhibits poor performance it may lead to the user losing confidence in the approach and he/she starting to question the strategy behind his or her assistive technology. It can take a long time for such a user to want to engage and spend time and effort with another attempt at using ASR in their home. It is therefore crucial to have informed ways of establishing - from a small amount of initial enrolment data and possibly assessments from health professionals - a suitable best 'operating point' : a task which will be useful for the user and for which the ASR will perform well.

The work described here is a two-fold step towards this goal investigating: i) to what degree does the 'optimal' system vary with the speaker, and ii) how to leverage art ASR algorithms developed and refined on typical speech. Section **??** gives a description of the data and acoustic modelling, section **??** presents results and section **??** presents conclusions.

2. Experimental setup

2.1. Data

The UAspeech database contains synchronised audio and visual streams and the publicly available part of the recordings includes speech from 15 speakers (4 female and 11 male). The speakers were asked to repeat single words from 5 groups: 10 digits, 29 Nato alphabet letters, 19 command words ('delete', 'enter' etc.), 100 common words ('the','will' etc.), and 300 uncommon words chosen to be phonetically rich and complimentary to the remaining words ('Copenhagen','chambermaid' etc.). The speakers came into the lab to complete the recordings in three blocks, and at each block, all words were repeated once, except the uncommon words where each block contained 100 unique words. In total, each speaker has produced around 70 minutes of speech. Full details of the corpora can be found in [**?**].

The speakers all have a type of disordered, dysarthric speech, and accompanying the database are percent intelligibly scores as obtained from listening tests with unfamiliar listeners. These range from 4% to 95%. Further meta-data in the form of intelligibility classes ('very low', 'low', 'medium' and 'high') are also supplied as well as broad diagnostic classes ('spastic','athetoid','mixed' and 'not diagnosed' dysarthria).

2.2. Data pre-processing

For the acoustic modelling, the data is encoded in 12-dimensional PLP features with c_0, and with added first and second order time derivatives, giving a 39 dimensional feature vector in total. Due to the way the UAspeech data was recorded (subjects sitting in front of a laptop on which prompts were presented at regular intervals), the original audio files in the database distribution contain silence, and hence this data was aligned and resegmented allowing for a 0.2 sec silence boundary around each word. This reduced the overall amount of data from around 60 hours to around 18 hours. In a real system, a good voice activity detector could provide a similar advantage.

The UAspeech database was recorded using a 7-channel microphone array. Not all channels have been supplied for each speaker in the publicly available database; for the work described here, we have chosen to use all available channels for each speaker. Following previously published work using the UAspeech for ASR (e.g. [**?**]) the data was divided into training and test data with a 2:1 split, using blocks 1 and 3 for training and block 2 for testing. Table **??** shows the amount of data available before and after alignment and resegmentation as well as the total number of word segments in each word category.

	Duration		*Number of words*				
	Org.	Rseg.	D	L	C	CW	UW
Train	39.44	12.44	1,725	4,498	3,307	17,395	17,350
Test	19.34	5.97	878	2,262	1,646	8,765	8,724

Table 1: *Duration of data [hours] and number of words in the training and test partitions after alignment. 'Org.' is the originally distributed files, 'Rseg.' is the resegmented files post alignment. The word categories are: D: digits, L:Nato alphabet, C: command words, CW: common words, UW: uncommon words.*

2.3. 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.

2.4. Decoding

As the database consists of single words, it was decided to restrict the decoding so that only one word could be recognised per utterance. A uniform language model was used, as well as a word grammar network containing silence models at the start and end, and all possible test words in parallel in the middle.

Some initial testing was carried out with a decoding strategy which enabled more than one word to be output in the transcript. The results of not controlling the insertions and deletions was a fall in absolute, mean accuracy over all systems of between 2 and 4%. Deletion rates stayed $<$0.4% but the insertion rates were high at up towards 10%. For dysarthric speakers it is possible that the prevalence of false starts can lead to higher insertion rates. Also, it was noted that one of the speakers in UAspeech has a stammer as well as having dysarthric speech.

3. Results

Our initial work with the UAspeech database concentrated on establishing the performance of a number of baseline systems such as a speaker dependent (SD) and a speaker independent (SI) system. The speaker independent systems are trained using a 'round robin' style where in turn, the data from each speaker is held out of the training data, a SI1 model is tested on data for the speaker who was missed out. A second type of SI system was established; they differ in that this system was trained and tested on all speakers and subsequently tested with the same *known* speakers. This system is called SI2 in the following.

To compare with this, we also decoded the UAspeech test set with a typical speech model set trained on more than 177 hours of spontaneous meeting data covering an array of accented, but purely non-disordered speech; called Mtg. The Mtg models are ML trained models without CMN/CVN and are trained on the ihmtrain09 data set[**?**].

The accuracy of these four baseline systems (Mtg, SD, SI1 and SI2) are presented in columns 1 to 4 in table **??**. The results

Intellibibility	Speaker	Baseline				MAP			Domain		MAPviaDomain	
		1 Mtg	2 SI1	3 SI2	4 SD	5 mapMtg	6 mapSI1	7 mapSI2	8 domSI1	9 domSI2	10 mvdSI1	11 mvdSI2
Very low	M04 (2%)	2.1	4.1	6.1	4.9	1.4	6.1	8.3	3.2	3.8	3.1	3.5
	F03 (6%)	1.4	5.5	18.2	17.5	8.6	21.4	23.0	2.9	10.1	15.8	13.7
	M12 (7%)	0.6	3.2	8.2	9.0	5.8	12.4	11.7	4.1	8.5	10.2	9.9
	M01 (17%)	2.1	12.9	23.7	18.9	13.3	29.0	29.8	7.3	16.6	20.7	19.6
Low	M07 (28%)	3.3	20.3	62.7	66.4	34.8	68.2	66.9	10.5	40.3	43.0	40.8
	F02 (29%)	3.5	7.0	30.7	29.6	24.6	36.1	36.9	6.4	23.4	30.7	30.1
	M16 (43%)	13.0	22.2	51.1	53.6	27.8	50.1	49.3	21.9	39.8	29.8	37.3
Mid	M05 (58%)	4.9	20.2	45.4	56.4	20.6	49.9	53.4	11.2	30.6	28.5	26.9
	M11 (62%)	10.2	30.3	47.4	48.2	23.0	53.9	53.0	17.8	33.1	31.5	29.5
	F04 (62%)	22.5	30.8	61.6	53.7	43.3	62.0	65.6	30.8	51.6	49.2	53.5
High	M09 (86%)	33.6	50.2	79.5	79.1	71.5	82.4	81.5	44.6	70.0	74.7	75.1
	M14 (90%)	49.6	58.1	73.6	74.9	74.9	76.6	74.9	61.6	73.1	75.6	77.8
	M10 (93%)	62.1	68.5	83.2	81.2	86.3	87.6	86.2	70.0	85.7	87.2	87.3
	M08 (95%)	57.2	64.9	81.2	85.0	76.8	83.6	81.8	68.5	80.6	78.9	78.4
	F05 (95%)	70.2	46.6	85.9	85.6	89.1	89.1	89.6	65.0	89.6	90.2	89.2
Average		22.4	29.7	50.6	50.9	40.1	53.9	54.1	28.4	43.8	44.8	44.6

Table 2: *Word accuracy rates from baseline and MAP adaptation systems by speakers. Speakers are ordered according to their intelligibility, presented in parentheses by the speaker id. All systems are tested with the UAspeech test set; the accuracy presented is the per speaker accuracy. The coloured cells indicates the system with the highest performance for that speaker. System name descriptions:'Mtg': Typical speech meeting models, 'SI1': speaker independent models; 'SI2' : speaker independent models tested with known speakers, i.e. training data from the test speaker is present in the training data; 'SD': speaker dependent models; 'mapMtg': MAP adaptation of typical speech meeting models; 'mapSI': MAP adaptation of SI models; 'domSI': domain adaptation mtg models to SI data; 'mvdSI': MAP adaptation via SI domain models.*

for each speaker are presented row-wise and ordered according to the intelligibility rating (given in parentheses); the final row of the table gives the average accuracy across all speakers. Of the four baseline systems, the Mtg models are clearly not well matched (average accuracy of 22.4 %) as no dysarthric speech has been used during training. The SI1 system shows better performance at 29.7%, but not until any speaker specific speech is used in the training, does the average accuracy rise significantly. The SI2 and SD systems have average accuracies of 50.6% and 50.9% respectively. Two observations can be made: the SI2 and SD have used vastly different amounts of data (∼12 hours vs. ∼1.2 hours), and if the accuracies are compared per speaker the best-performing model type varies.

The next step was to investigate ways of using MAP adaptation with the available data; initially through a simple MAP adaptation from the Mtg, SI1, and SI2 models. Those results are presented in table **??**; columns 5 (mapMtg), 6 (mapSI1), and 7 (mapSI2)) with average accuracies of 40.1%, 53.9% and 54.1% respectively. This represents relative increases in accuracies of around 80% for both the Mtg and SI1 - a very clear benefit of adding speaker specific data. In contrast, the SI2 to mapSI2 relative improvement is only 7%.

Comparing individual speaker performances for the three map systems (columns 5, 6, and 7) shows that – especially for less severely dysarthric speakers – good, >70% accuracies can be achieved using both typical speech and purely dysarthric speech as starting points. The idea that perhaps for certain speakers there would be an advantage gained from harnessing the statistics present in a large LVCSR database led us to our final set of experimental systems. For these systems, the starting point was the meeting data models ('Mtg'); these models were then adapted to the general domain of dysarthric speech using the UAspeech data (domSI1/domSI2). The final step was

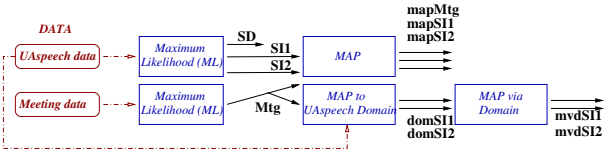

Figure 1: *Illustration of training strategies. 'ML': training using maximum likelihood criterion, 'MAP': using MAP adaptation to get SD systems, 'Domain': using MAP adaptation to get SI domain model, 'mvdSI': MAP adaptation to SD model via domain model.*

a standard MAP adaptation via this domain models resulting in an SD model (mvdSI1/mvdSI2). This MAP-via-domain training and adaptation strategy is illustrated in figure **??**.

In table **??** the results of testing each stage in the MAP-via-domain path is presented in columns 8-11. Overall, the average recognition accuracies do not compare to the mapSI systems' performance, but the average masks a very high degree of speaker variability. The colour coding in table **??** indicates the best system for each speaker; it is evident that the best choice of system varies hugely.

This is analysed in more detail in figure **??**, which shows a grey-scale visualisation of the dependance between severity of dysarthria and the accuracies from table **??**. It gives an impression of the degree to which there exists an optimal system for each speaker. Speakers are ordered by increasing intelligibility down the y-axis, and systems are ordered by increasing average, *normalised* accuracy along the x-axis. If we assume that all speakers were to exhibit a linear dependance between average system accuracy and their individual performance, the

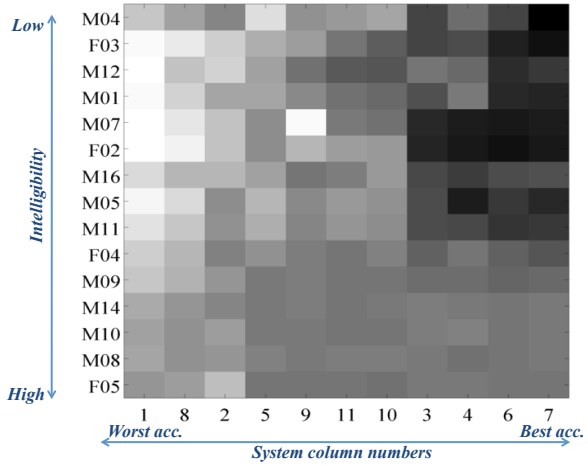

Figure 2: *Visualising the correlation between severity of dysarthria by axis and the performance of systems; image of normalised accuracies by speakerIDs and system columns from Table* **??**. *Speakers are ordered by increasing intelligibility and systems are ordered by increasing average accuracy.*

image would have increasingly darker grey colours from left to right. However, this is far from the situation. Take the case of less severe dysarthric speakers first (bottom rows). It can be seen on the limited colour variation, that their performances tend to vary little across systems. In contrast, speakers with severely disordered speech, display a strong bias towards a few systems (seen by the darker top, right-hand corner).

4. Discussion and conclusions

This paper has explored the extent to which core LVCSR training and adaptation algorithms can deal with dysarthric speech. Through a comparative study using the UAspeech database very competitive word accuracy rates have been achieved compared to results published elsewhere. [**?**] report results on a subset of 7 UAspeech speakers of which there is a 6 speaker overlap with the released data used in this study. Their MAP adaptation results are lower than ours in all cases bar for speaker F02. The results in this paper improve with an average, relative rate of 34.5% on theirs, ranging between -0.2% (F02) and 97.6% (M04). Some notable differences which may account for that difference are the cropping of extra silence from the data prior to training, and the use of highly optimised clustering, training and adaptation scripts arising from many years of LVCSR research on *typical* speech.

This leads us to conclude, in answer to our opening question, that porting LVCSR evolved on typical speech to the domain of disordered speech is a viable way of achieving good results despite the inherent differences. Although the mismatch in domains is large – exemplified by the poor performance achieved on the typical speech baseline – MAP estimation can deal with it to a large extent. For some speakers though, opting for a pure SD is a better option. In general, the study has shown how there is no 'one solution to fit all'. Particular for more severely dysarthric speakers. In future work, we plan to investigate systematic ways of arriving at the best operating point for a particular speaker in terms of system configuration. We will also investigate the effect of doing MLLR adaptation and discriminative training. This research will be supported by a longitudinal study into the use of speech-driven assistive technology for disabled and elderly users, in the homeService project, part of the UK EPSRC Programme Grant NST[**?**].

5. Acknowledgements

The research leading to these results was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).