Automatic Speech Recognition with Sparse Training Data for Dysarthric Speakers

Phil Green, James Carmichael, Athanassios Hatzis, Pam Enderby, Mark Hawley, Mark Parker

University of Sheffield, UK
(p.green, j.carmichael, a.hatzis)@dcs.shef.ac.uk,
Mark.Hawley@bdgh-tr.trent.nhs.uk, P.M.Enderby@sheffield.ac.uk, m.p.
parker@sheffield.ac.uk

Abstract

We describe an unusual ASR application: recognition of command words from severely dysarthric speakers, who have poor control of their articulators. The goal is to allow these clients to control assistive technology by voice. While this is a small vocabulary, speaker-dependent, isolated-word application, the speech material is more variable than normal, and only a small amount of data is available for training. After training a CDHMM recogniser, it is necessary to predict its likely performance without using an independent test set, so that confusable words can be replaced by alternatives. We present a battery of measures of consistency and confusability, based on forced-alignment, which can be used to predict recogniser performance. We show how these measures perform, and how they are presented to the clinicians who are the users of the system.

1. Introduction

The work reported here is part of the STARDUST[1] project which aims to provide severely dysarthric speakers with voice access to assistive technology. Dysarthrias are a family of neurologically-based speech disorders characterized by loss of control of the articulators [Enderby and Emerson 95]. Speech produced by dysarthric speakers can be very difficult for listeners unfamiliar with the speaker to understand. Since motor-neuron disease or trauma often affects the cognitive and physical processes responsible for speech production, dysarthric symptoms often accompany neurological conditions such as cerebral palsy, head injury and multiple sclerosis. Thus many people with dysarthria are often physically incapacitated to the extent that spoken commands become an attractive alternative to normal keyboard-and-mouse input, despite the difficulty of achieving robust Automatic Speech Recognition for dysarthric material.

There have been a number of studies concerning the feasibility of ASR for dysarthric speech [e.g. Blaney & Wilson 00, Bowes 99, Deller et al. 91, Doyle et al. 97, Ferrier et al. 95, Kotler et al 97, Rosengren et al. 95, Thomas-Stonell et al. 98] which are reviewed in [Rosen et al. 00, Hawley 02]. Unsurprisingly, these studies report rather varied performance: there is a general consensus that ASR can be viable for mild to moderate dysarthria, using commercially available 'dictation' systems. However, more severe conditions defeat these systems except for a few individuals. For severely dysarthric speech, recognisers trained on a normal speech corpus cannot be expected to work well. Although some systems embody algorithms which adapt their statistical models to the speaker [e.g. Leggetter and Woodland 95], adaptation techniques are insufficient to deal with gross abnormalities. In an attempt to reduce speech production inconsistency and hence enhance the success of voice-driven assistive technology, the ASR component in STRAPTk is closely coupled with therapy. Recognition will improve if the material becomes more consistent, an objective which is greatly facilitated if the speaker is provided with some type of visual feedback informed by a matching score indicating the incoming utterance's degree of fit to what the recogniser has been trained to expect. More fine-grained visual feedback, related to phone-level articulation, can also be achieved via *phonetic maps* [Hatzis, Green & Howard 97] which are trained to relate speech acoustics to chosen positions on a two-dimensional display. The following section discusses the STARDUST application as an ASR problem. We then present the evaluation tools we have developed along with a selection of baseline results. We conclude with a discussion about the application of these tools to the assessment of speech disorders. A companion paper [Hatzis et al, 03] covers the STARDUST software and presents clinical results.

2. The STARDUST ASR Application

In STARDUST, the aim is to provide severely dysarthric people with the ability to control assistive technology by voice. Since it is not usual for any one client to require access to more than a few devices, the recognisers built for these patients normally require a *small vocabulary of isolated command words*, e.g. 'Open', 'TV', 'Channel', 'On', 'Off').

Since there is so much variation between individuals, *speaker-dependent recognisers* are trained for each client. Small-vocabulary, speaker-dependent, isolated-word recognition is a relatively easy ASR task, however the material to be recognised is significantly more variable (or *less consistent*) than normal. Furthermore, only small amounts of training data are available: many clients are rapidly fatigued by the effort required to produce multiple utterances of their command words when prompted. After identifying a list of devices suitable for the client to control, an appropriate vocabulary is selected (normally not exceeding ten words) and the collection of training data is achieved in recording 'sessions' where the client repeats each word of the vocabulary 10 times; on average, between two to four sessions are recorded for each client. STARDUST supports a longitudinal study with 8 clients currently enrolled in the pilot programme.

2.1. Recogniser Configuration and Performance

We use the HTK toolkit [Young et al. 95] to build isolated word recognisers for dysarthric speech using Continuous Density

[1] Speech Training and Recognition for Dysarthric Users of Speech Technology.

Hidden Markov Models [Rabiner 89]. The configuration is quite conventional:

- Whole-word rather than phone-level models,
- Typically 11 HMM states,
- Typically 3 mixture Gaussian distributions per state,
- ‘Straight-through’ model topology allowing only self-transitions and transitions to the next state,
- Acoustic vectors consisting of Mel Frequency Cepstral Coefficients, typically with differences but without overall energy (dysarthric speakers often have difficulty maintaining a steady volume).
- Training data labelled at the word level
- Sampling rate for audio data of 16KHz, with a 10ms frame rate.

Baseline results for normal and dysarthric speakers on 10-word vocabularies are encouraging. The table below gives word accuracy on previously unseen test data after training on 20 examples of each of 10 words by the same speaker. All material was recorded before the dysarthric speakers had received any therapy:

Speaker	Recogniser Accuracy (%)
MP (Normal)	100
AH (Normal)	100
GR (Severely Dysarthric)	87
JT (Severely Dysarthric)	100
CC (Severely Dysarthric)	96

Table 1: Accuracy Rates for Isolated-Word Speaker-Dependent Recognisers

The above performances were achieved with recognisers trained on 20 utterances per word[1]. This small quantity of training data presents unusual problems: normally a corpus is available which is sufficiently large to be split into training and test sets, with performance measured on the test set dropping only slightly compared to performance on the training set. Here, in contrast:

- It is unlikely that good results on a small training set will hold up under everyday conditions.
- To produce the best-performing recogniser one should use all, or nearly all, the available data for training rather than reserving some for evaluation.
- Predictions of recogniser performance cannot be based on an analysis of test-set confusion matrices.

In addition, the intended users of STARDUST software are clinicians[2] rather than speech technologists. These clinicians must configure and train speaker-dependent recognisers, not merely use them. STARDUST therefore provides Graphical User Interface (GUI) tools to facilitate the collection, selection, and labelling of data along with the actual building of the recognisers themselves. Once such a recogniser has been constructed, our clinicians require a report forecasting – in non-technical terms – the likely performance of the recogniser in the field. This is the topic of the next section.

3. Recogniser Evaluation Measures

3.1. Consistency Measures

In STARDUST it is possible to modify a client’s recognition vocabulary. This is important because it is usual for a dysarthric speaker to produce some words more consistently than others (‘TV’, for example might be an easier proposition than ‘television’). While clinical assessment can help in identifying such words, a quantitative measure of *word-level consistency* is needed: HMM-based recognisers do not decode speech in the same manner that human listeners do and their results are sometimes counter-intuitive. Similarly, it would be useful to measure the *overall consistency* of the speech in the training corpus, across all chosen words. Overall consistency could be used to assess the severity of the disorder and to chart the client’s progress as therapy proceeds. At a finer level, it is useful to track *utterance-level consistency*: the correlation between the probability scores returned by a client’s individual productions of a given word and the norm for that word. With this measure the clinician can identify outlier utterances for removal from the training set.

3.2. Predicting Confusions: Confusability Measures

In addition to consistency measures, robust recognition performance could be facilitated if some means of predicting recognition errors could be devised, the aim being to identify words which can be expected to be easily confused with each other. Conventionally, test-set confusion matrices are used for this purpose, but these are unlikely to be very informative over sparse data, and in any case it is advisable to use all the data (except outliers) for training purposes. An alternative is to devise a measure of word-level *confusability*. Previous work on word confusability measures has been reported in [Roe and Riley 94, Tan et al, 99], but both these studies rely on making use of the normal phonetic structure of a word, which is inappropriate for disordered speech.

3.3. Formulating Consistency and Confusability Measures

An alternative to phonetically-based metrics is to define probability-based measures based on *forced alignment* against trained models, based exclusively on the training set and the models. The following scheme uses forced-alignment of training set utterances against the models.

- We have a training set for a vocabulary of $\boldsymbol{N}$ words, $\boldsymbol{W_1 .. W_N}$
- We have trained a CDHMM $\boldsymbol{M_i}$ for each word $\boldsymbol{W_i}$.
- $\boldsymbol{w_{jk}}$ is the $\boldsymbol{k}$th repetition of the $\boldsymbol{j}$th word in the training set

By forced alignment, we can compute the per-frame log likelihood $\boldsymbol{L_{ijk}}$ of each model generating each example of each word on the Viterbi path. The *consistency* δ_i *of word* $\boldsymbol{W_i}$ is obtained by

$$\delta_i = (\Sigma_k L_{iik})/n_i, \qquad (1)$$

where $\boldsymbol{n_i}$ is the number of examples of $\boldsymbol{W_i}$ in the training set: we average the scores obtained by aligning all the examples of a word against the model for that word. The reasoning behind this is that the more variability there is in the training data for each speech unit, the larger the variances in that unit’s HMM state distributions will be. The forced-alignment likelihoods will be lower for an inconsistently spoken word than for a consistent one since its distributions will be flatter.

The *overall consistency* of the training corpus, Δ, is just the average of the δ_i:

$$\Delta = (\Sigma_i \delta_i)/N \qquad (2)$$

The *confusability* between $\boldsymbol{W_i}$ and $\boldsymbol{W_j}$ is defined by

$$C_{ij} = (\Sigma_k L_{ijk})/n_j , \qquad (3)$$

$\boldsymbol{C_{ij}}$ is the average score obtained by aligning examples of $\boldsymbol{W_j}$ against $\boldsymbol{M_i}$. The higher this is, the greater the likelihood that $\boldsymbol{W_j}$ will be misrecognised as $\boldsymbol{W_i}$.

[1] On the basis that one training vector is required for every free parameter in the model, 45 examples of each word would be required.

[2] Or even the clients themselves.

3.4. Confusability Matrices

The confusability measures C_{ij} between pairs of words can be viewed conveniently as a **confusability matrix**. To make interpretation easy, C_{ij} is used to code the greyscale shading or colour hue of each cell (a grey scale is used here for printing purposes but a colorized matrix is also included for reference). The examples below use the following scale which may represent either a standard or relative range of values.

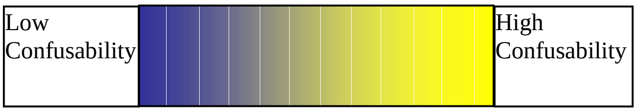


Table 2: Confusability Colur Map

Table 3 shows a confusability matrix for a recogniser trained for normal speaker MP. Table 4 is the corresponding matrix for our most severely dysarthric subject, GR. The recognisers have the same vocabulary. The diagonal of the confusability matrix corresponds to the word-level consistency measures. For a recogniser expected to perform well, cells on the diagonal of the confusability matrix (that is, the word aligned with itself) should be towards the blue end of the scale and off-diagonal (the word aligned with other models) cells should be towards the yellow end. This is the case the normal speaker MP (Table 3). For a dysarthric speaker, the blue-yellow distinction will not be so pronounced, as is apparent for GR in Table 4.

3.5. **Experiments with the Evaluation Measures**

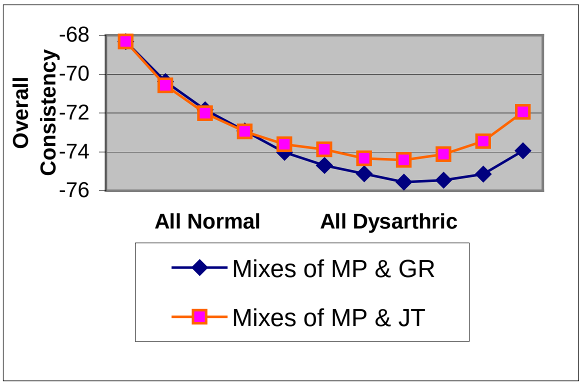


Figure 1: Consistency of Mixed Data Sets

The behaviour of the overall consistency measure *Δ* can be verified by constructing recognisers for mixtures of speech material from normal and dysarthric speakers. For the same 10-word vocabulary, corpora consisting of 20 examples of each word were constructed by mixing normal utterances (speaker MP) with utterances from each of two dysarthric speakers (GR and JT), in varying proportions. Overall consistency measures for these corpora are shown in Figure 1. The leftmost points are for all-normal speech and the rightmost points are all-dysarthric speech. In between, the proportion of dysarthric speech is increased by 10% at each step. Consistency worsens as more dysarthric speech is introduced, until the dysarthric speech begins to dominate the corpus, at which stage consistency then recovers, but not to the level of the normal speech. GR's condition is more severe than that of JT, so when his speech is mixed with normal speech in increasing proportions, consistency deteriorates more acutely and recovers less than for JT. Note that since consistency measures and confusability measures are averages over log-likelihoods, they are negative numbers representing small quantities, and a difference of ***x*** between two such measures should be thought of as ***x*** orders of magnitude.

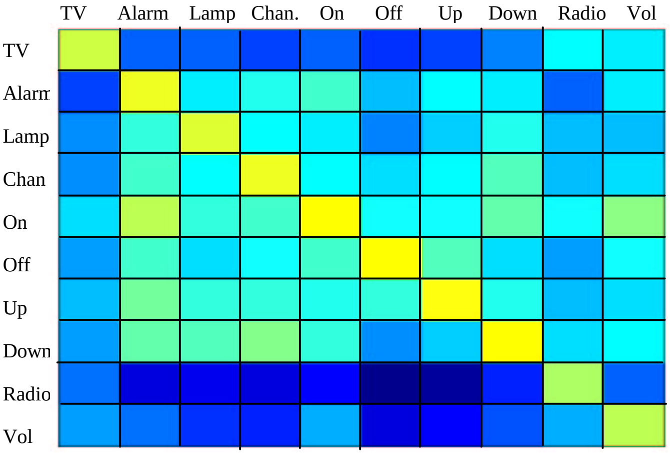


Table 3: Confusability Matrix (Standard Calibration) for Normal Speaker MP

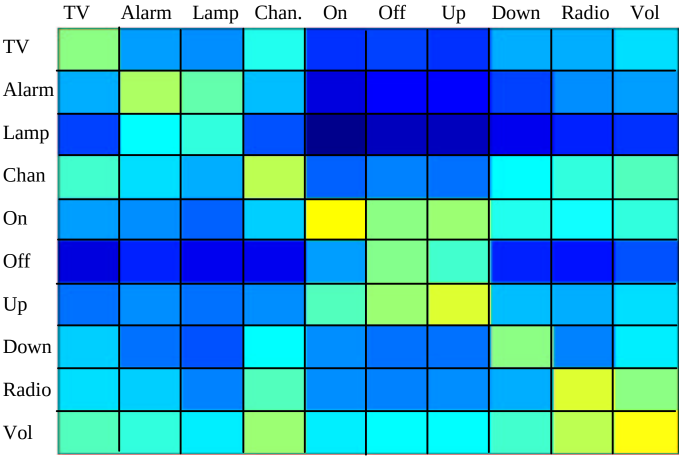
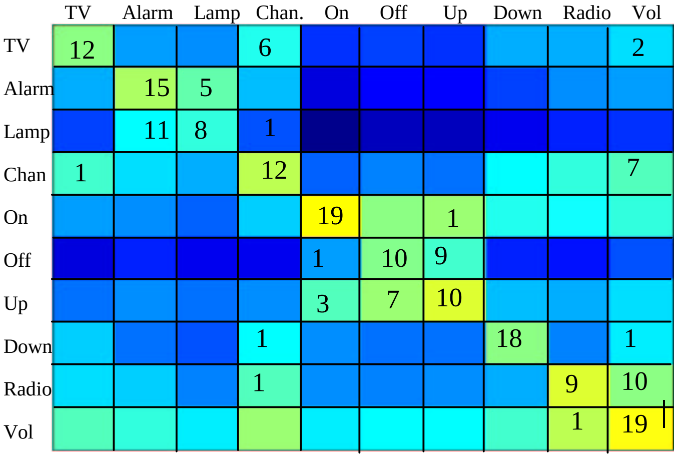


Table 4: Confusability Matrix (Standard Calibration) for Dysarthric Speaker GR

Table 5 provides informal confirmation that confusability is a good predictor of confusions. Here, the confusions on a test set (20 utterances per word) are superimposed on the GR confusability matrix of table 4.



	TV	Alarm	Lamp	Chan.	On	Off	Up	Down	Radio	Vol
TV	12			6						2
Alarm		15	5							
Lamp		11	8	1						
Chan	1			12						7
On					19		1			
Off					1	10	9			
Up					3	7	10			
Down				1				18		1
Radio				1					9	10
Vol									1	19

Table 5: Test Set Confusions Superimposed on GR Confusability Matrix

3.6. Using the Evaluation Measures

When a new recogniser has been built, the STARDUST software automatically generates an html-formatted report providing, among other statistical data, the recogniser's confusability matrix and utterance-level consistency tables. As an example of how the report is used, we notice that in Table 5, 'Alarm' and 'Lamp' show clear evidence of confusability with each other (but less so with other words) and therefore one or the other should be removed. 'Volume' returns high confusability scores for nearly all the words in the vocabulary, indicating that it should be replaced.

4. Relationship to the Assessment of Speech Disorders

The assessment of speech disorders can contribute substantial knowledge to assist in the diagnosis of the underlying neurological problems. Assessment is also conducted to assist in monitoring the effectiveness of speech and language therapy. One important component of such an assessment is an analysis of intelligibility. Intelligibility assessments are normally based on listening tests and are notoriously complex and time consuming to conduct and psychometrically weak, having poor reliability and validity. The confusability and consistency measures defined above provide complimentary (and to some extent an alternative) metrics based only on statistics of the speech acoustics. These objective measures can be obtained rapidly and have the psychometric properties of being reliable and repeatable. They can be used within clinical sessions and the results can be analysed in more or less detail, as is required. Speech consistency is not the same as speech intelligibility but may be expected to be related to it, a topic offering much scope for future studies. The relationship between intelligibility and consistency has not been reviewed elsewhere and remains a piece of work that this team will pursue.

Acknowledgements

STARDUST is funded by the UK NHS NEAT programme. Related work is supported by the EC framework 5 project OLP.

References

[1] Blaney, B. and Wilson, J. "Acoustic Variability in Dysarthria and Computer Speech Recognition", *Clinical Linguistics and Phonetics*, 14(4): 307-327, 2000.

[2] Bowes, D., "Getting it right and making it work! Selecting the right speech input and writing software for users with special needs", *Proceedings of Technology and Persons with Disabilities,* California State University Northridge, 1999.

[3] Deller, J., Hsu, D., Ferrier L., "On the use of hidden Markov modelling for recognition of Dysarthric speech", *Computer Methods and Programmes in Biomedicine*, 35:125-139, 1991

[4] Doyle, P., Leeper, H., Kotler, A-L, Thomas-Stonell, N., O'Neill, C., Dylke, M-C., Rolls, K., "Dysarthric speech: a comparison of computerised speech recognition and listener intelligibility" *Journal of Rehabilitation Research and Development,* 34(3): 309-316, 1997.

[5] Enderby, P., Emerson, J., *Does Speech and Language Therapy Work? A Review of the Literature*. Whurr Publishers, London 1995

[6] Ferrier, L., Shane, H., Ballard, H., Carpenter, T., Benoit, A., "Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition", *Journal of Augmentative and Alternative Communication*, 11: 165-74, 1995.

[7] Hatzis, A., Green, P., Howard, S., "Optical Logo-Therapy (OLT), A computer based real time visual feedback application for speech training", *European Conference on Speech Communication and Technology,* September, 1997.

[8] Hatzis, A., Green, P., Carmichael, J., Cunningham, S., Palmer, R., Parker, M., O'Neil, P., "An Integrated Toolkit Deploying Speech Technology for Computer Based Speech Training with Application to Dysarthric Speakers", submitted to Eurospeech 2003

[9] Hawley, M. "Speech Recognition as an Input to Electronic Assistive Technology", *British Journal of Occupational Therapy*, 65(1):15-20, 2002.

[10] Kotler, A., Thomas-Stonell, N., "Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment" *Journal of Augmentative and Alternative Communication,* 12: 71-80, 1997.

[11] Leggetter, C., Woodland, P., Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Computer Speech and Language,* 9: 1490-1499, 1995.

[12] Rabiner, L., "A tutorial on hidden Markov models and selected applications in speech recognition" *Proc IEEE*, 37: 257-286, 1989

[13] Roe, D., Riley, M., "Prediction of Word Confusabilities for SpeechRecognition", *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, 1994.

[14] Rosen, K., Yampolsky, S., "Automatic Speech Recognition and a Review of Its Functioning with Dysarthric Speech", *Journal of Augmentative and Alternative Communication*, 16: 48-60, 2000.

[15] Rosengren, E., Raghavendra, P., Hunnicut, S., "How does automatic speech recognition handle severely dysarthric speech?" In: P. Porrero, R. Puig de la Bellacasa, (eds.), *The European context for Assistive Technology*, IOS Press, Netherlands, 336-339, 1995.

[16] Tan, B., Gu, Y., Thomas, T., "Word Confusability Measures for Vocabulary Selection in Speech Recognition", *IEEE Automatic Speech Recognition and Understanding Workshop*, 1999.

[17] Thomas-Stonell, N., Kotler, A-L., Leeper, H., Doyle, P. "Computerized Speech Recognition: Influence of Intelligibility and Perceptual Consistency on Recognition Accuracy", *J. Augmentative and Alternative Communication*, 14: 51-56, 1998.

[18] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtcho, V., Woodland, P., *The HTK Book (for Version 2.1),* Microsoft Corporation, 1995.