

# MODELLING CONFUSION MATRICES TO IMPROVE SPEECH RECOGNITION ACCURACY, WITH AN APPLICATION TO DYSARTHIC SPEECH

Omar Caballero Morales and Stephen Cox

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.

S.Caballero-morales@uea.ac.uk, sjc@cmp.uea.ac.uk

## ABSTRACT

Dysarthria is a motor speech disorder characterized by weakness, paralysis, or poor coordination of the muscles responsible for speech. Although automatic speech recognition (ASR) systems have been developed for disordered speech, factors such as low intelligibility and limited vocabulary decrease speech recognition accuracy. In this paper, we introduce a technique that can increase recognition accuracy in speakers with low intelligibility by incorporating information from an estimate of the speaker’s phoneme confusion matrix. The technique performs much better than standard speaker adaptation when the number of sentences available from a speaker for confusion matrix estimation or adaptation is low, and has similar performance for larger numbers of sentences.

## 1. INTRODUCTION

“Dysarthria is a motor speech disorder that is often associated with irregular phonation and amplitude, incoordination of articulators, and restricted movement of articulators” [4]. This condition can be caused by a stroke, cerebral palsy, traumatic brain injury, or a degenerative neurological disease such as Parkinson’s or Alzheimer’s Disease. The affected muscles by this condition may include the lungs, larynx, oropharynx and nasopharynx, soft palate and articulators (lips, tongue, teeth and jaw), and the degree to which these muscle groups are compromised determines the particular pattern of speech impairment [4]. This means that the design of an ASR system for dysarthric speakers is difficult, because as Rosen and Yampolsky [5] point out, they require different types of ASR depending on their particular type and level of disability. Rosen and Yampolsky also identify factors that give rise to ASR errors [5], the most important being *decreased intelligibility* (because of substitutions, deletions and insertions of phonemes), and *limited phonemic repertoire*, the latter leading to phoneme substitutions. In this paper, we describe a technique for incorporating a model of a speaker’s confusion matrix into the ASR process in such a way as to increase recognition accuracy. Although this technique has general application to ASR, we believe that it is particularly suitable for use in ASR of dysarthric speakers who have low intelligibility due, in some degree, to a limited phonemic repertoire, and the results presented here confirm this.

To illustrate the effect of reduced phonemic repertoire, Figure 1 shows an example phoneme confusion matrix for a dysarthric speaker from the NEMOURS database [1](see section 3). This confusion matrix is estimated by an ASR system, and so it may show confusions that would not actually be made by humans, and also spurious confusions that are actually caused by poor transcription/output alignment (see section 2.2). However, since we are concerned with machine rather than human recognition here, we can make the following observations:

1. A small set of phonemes (in this case the phonemes “ax”, “ih”, “b”, “d”, “dh”, “n” and “z”) dominates the speaker’s

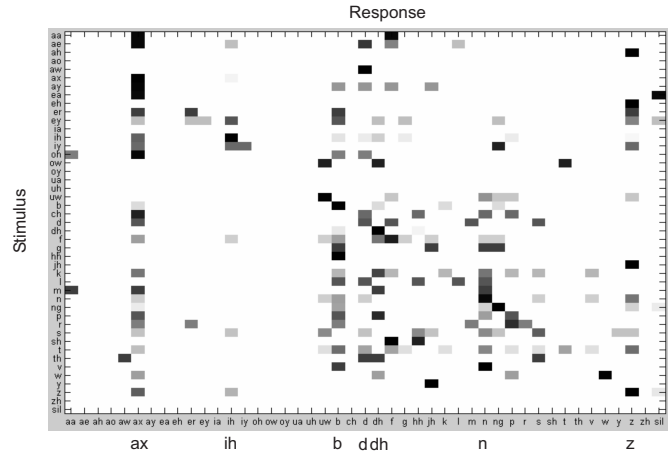


Figure 1: A phoneme confusion matrix for a dysarthric speaker

output speech.

2. Some vowel sounds and the consonants “sh” and “th” are never recognised. This suggests that there are some phonemes that the speaker apparently cannot enunciate at all, and for which he or she substitutes a different phoneme, often one of the dominant phonemes mentioned above.

Most speaker adaptation algorithms are based on the principle that it is possible to apply a set of transformations to the parameters of a set of acoustic models of an “average” voice to move them closer to the voice of an individual. Whilst this has been shown to be successful for normal speakers, it may be less successful in cases where the phoneme uttered is not the one that was intended but is substituted by a different phoneme. In this situation, we argue that a more effective approach is to combine a model of the substitutions likely to have been made by the speaker with a language model to infer what was said. We imagine that the speaker wished to utter a word sequence  $W_{in}$  which can be transcribed using a dictionary into the phoneme sequence  $S_{in}$ .<sup>1</sup> The sequence of phones decoded by the speech recogniser is  $S_{out}$ , and we construct a model that makes use of an extended confusion matrix estimated for the speaker plus a standard language model to estimate  $W_{in}$  from  $S_{out}$ . More details of this are given in the next section.

<sup>1</sup>For present purposes, we sidestep the issue of multiple pronunciations and hence multiple phoneme transcriptions of a word, something that occurs relatively infrequently.

## 2. INCORPORATING A MODEL OF THE CONFUSION MATRIX INTO THE RECOGNISER

### 2.1. Metamodels

Using the terminology defined in section 1, we wish to estimate the word sequence intended by the speaker,  $W_{in}$ , from the phone sequence output by the recogniser,  $S_{out}$ .  $S_{out}$  is modelled as the output of a stochastic process that operates on the “input” phone sequence  $S_{in}$ . Using Bayes Theorem, we can write

$$\Pr(W_{in}|S_{in}, S_{out}) = \frac{\Pr(S_{in}, S_{out}|W_{in})}{\Pr(S_{in}, S_{out})}. \quad (1)$$

An estimate of the joint probability of the output sequence  $\Pr(S_{out})$  and any postulated input sequence  $S_{in}$  can be made using the confusion-matrix for the speaker:

$$\begin{aligned} \Pr(S_{in}, S_{out}) &= \Pr(S_{out}|S_{in})\Pr(S_{in}) \\ &= \prod_{i=1}^M \Pr(p_{out}^i|p_{in}^i)\Pr(p_{in}^i). \end{aligned} \quad (2)$$

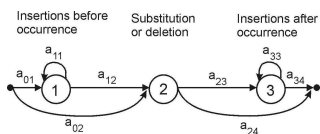
In equation 2,  $p_{out}^i$  is the  $i$ 'th phone in  $S_{out}$  (which is of length  $M$  phones) and  $\Pr(p_{out}^i|p_{in}^i)$  is estimated from the confusion matrix for the speaker. We have made the assumption that the conditional probabilities in equation 2 are independent of each other. Hence the most likely word sequence  $W_{in}^*$  is estimated as

$$W_{in}^* = \underset{W_{in}}{\operatorname{argmax}} \Pr(S_{in}, S_{out}|W_{in}). \quad (3)$$

i.e. the most likely input word sequence is found by combining a confusion-matrix model that estimates  $\Pr(S_{out}|S_{in})$  with a language model that finds the most likely word sequence.

In practice, it is too restrictive to use only the confusion-matrix to model  $\Pr(S_{out}|S_{in})$  as this cannot model insertions well. Instead, a hidden Markov model (HMM) is constructed for each of the phonemes in the phoneme inventory. We term these HMMs *metamodels* [2]. The function of a metamodel is best understood by comparison with a “standard” acoustic HMM: a standard acoustic HMM estimates  $\Pr(A'|p_{in})$ , where  $A'$  is a subsequence of the complete sequence of observed acoustic vectors in the utterance,  $A$ , and  $p_{in}$  is a postulated phoneme in  $S_{in}$ . A metamodel estimates  $\Pr(S'_{out}|p_{in})$ , where  $S'_{out}$  is a subsequence of the complete sequence of observed (decoded) phones in the utterance,  $S_{out}$ . The architecture of the metamodel of a phoneme is shown in Figure 2. Each state of a metamodel has a discrete probability

Figure 2: The architecture of the metamodel of a phoneme.



distribution over the symbols for the set of phonemes, plus an additional symbol labelled DELETION. The central state (2) of a metamodel for a certain phoneme models correct decodings, substitutions and deletions of this phoneme made by the phone recogniser. States 1 and 3 model (possibly multiple) insertions before and after the phoneme. If the metamodel were used as a generator, the output phone sequence produced could consist of, for example:

1. a single phone which has the same label as the metamodel (a correct decoding) or a different label (a substitution);
2. a single phone labelled DELETION (a deletion);
3. two or more phones (one or more insertions).

As an example of the operation of a metamodel, consider a hypothetical phoneme that is always decoded correctly without substitutions, deletions or insertions. In this case, the discrete distribution associated with the central state would consist of zeros except for the probability associated with the symbol for the phoneme itself, which would be 1.0. In addition, the transition probabilities  $a_{02}$  and  $a_{24}$  would be set to 1.0 so that no insertions could be made. When used as a generator, this model can produce only one possible phone sequence: a single phone which has the same label as the metamodel.

We use the reference transcription  $S_{in}$  of a training set utterance to enable us to concatenate the appropriate sequence of phoneme metamodels for this utterance. The associated recognition output sequence  $S_{out}$  for the utterance is made by recognising the utterance with a phone recogniser, and is used to train the parameters of the metamodels in this sentence (note that the phone recogniser itself can be built using unadapted or MLLR adapted phoneme models). By using embedded re-estimation over the  $\{S_{in}, S_{out}\}$  pairs of all the utterances, we can train the complete set of metamodels. In practice, the parameters formed, especially the probability distributions, are sensitive to the initial values to which they are set, and it is essential to “seed” the probabilities of the distributions using data obtained from an accurate alignment of  $S_{in}$  and  $S_{out}$  for each training-set sentence. The process for finding this alignment is described in section 2.2. After the initial seeding is complete, the parameters of the metamodels are re-estimated using embedded re-estimation as described above. Before recognition, the language model is used to compile a “meta-recogniser” network, which is identical to the network used in a standard word recogniser except that the nodes of the network are the appropriate metamodels rather than the acoustic models used by the word recogniser. At recognition time, the output of the phone recogniser  $S_{out}$  is passed to the meta-recogniser to produce a set of word hypotheses.

### 2.2. Improving alignment for confusion matrix estimation

Use of a standard dynamic programming (DP) tool to align two symbol strings (such as the one available in the *HResults* routine in the HTK package [6]) can lead to unsatisfactory results when a precise alignment is required between  $S_{in}$  and  $S_{out}$  to estimate a confusion matrix, as is the case here. This is because these alignment tools typically use a distance measure which is “0” if a pair of symbols are the same, “1” otherwise. To illustrate this, consider the top alignment in Table 1, which was made using *HResults*. It is not a plausible alignment, because

1. the first three phones in the recognised output are unaligned and so must be regarded as insertions;
2. the fricative *sh* in the transcription has been aligned to the vocalic *y*;
3. the sequence *bea* in the transcription has been aligned to the sequence *axb*.

In the lower alignment in Table 1, these problems have been rectified and a more plausible alignment results. This alignment was made using a DP matching algorithm in which the distance  $D(p_{in}, p_{out})$  between a phone in the transcription  $p_{in}$  and a phone in the recognition output  $p_{out}$  was made proportional to  $1/\Pr(p_{out}|p_{in})$ . The effect of this is that a phoneme pair that is rarely confused has a high distance, but a phoneme pair that is often confused has a low distance, and so the DP algorithm prefers to align phoneme pairs that are more likely to be confused. Of course, the confusion-matrix entry  $\Pr(p_{out}|p_{in})$  must itself be estimated by an DP algorithm that uses a simple aligner. However, a confusion matrix pooled over 92 WSJ speakers was used to estimate these probabilities, and so gross alignment errors such as the one shown in Table 1 are swamped by the large volume of good alignments.

Table 1: *Upper pair*: alignment of transcription and recognised output using *HResults*; *Lower pair*: same, using improved aligner

TR:		dh	ax	sh	uw	ih	z	b	ea	r	ih	ng	dh	ax	b	ey	dh		
REC:	dh	ax	ng	dh	ax	y	ua	ng	dh	ax	b	l	ih	ng	dh	ax	b	uw	
TR:	dh	ax		sh	uw		ih	z	b	ea		r	ih	ng	dh	ax	b	ey	dh
REC:	dh	ax	ng	dh	ax	y	ua	ng	dh	ax	b	l	ih	ng	dh	ax	b	uw	

### 3. SPEECH DATA, RECOGNISER, AND BASELINE RESULTS

The Wall Street Journal (WSJ) database was used to build the SI speech recogniser. The training set consisted of the WSJ data from 92 speakers in set *si.tr*. This was used to construct 45 mono-phone acoustic models. The models were a standard three state left-right topology with eight mixture components per state. The front-end used 12 MFCCs plus energy plus delta and acceleration coefficients.

The dysarthric speech data was provided by the NEMOURS database [1]. This database is a collection of 814 short nonsense sentences spoken by 11 speakers (74 sentences per speaker) with varying degrees of dysarthria (we used the data from 10 speakers, as some data is missing for one speaker). Note that although each of the 740 sentences used is different, the vocabulary is shared. A subset of the first 34 sentences from each speaker were used for confusion-matrix estimation, and the remaining 40 were used for testing.

The HTK package [6] was used throughout for the experiments. For comparison with a standard technique, MLLR adaptation [6] was applied, always using the same set of adaptation sentences as were used for estimating the metamodels. The language model for these experiments was a bigram model estimated from the (pooled) 74 sentences provided by each speaker (113 different words). We take up this point in section 4.3.

## 4. RESULTS

### 4.1. Baseline

Figure 3 shows the intelligibility of each of the dysarthric speakers used in this study as measured using the Frenchay Dysarthria Assessment (FDA) test [1], and the recognition % correct when tested on the unadapted SI models (WSJ) and the MLLR adapted models (ADAPT), as described in section 3. The correlation between the FDA performance and our recogniser performance is 0.756 (unadapted models) and 0.817 (adapted). Both are significant at the 1 % level, which gives us some confidence that our recogniser displays a similar performance trend when exposed to different degrees of dysarthric speech as humans.

### 4.2. Results on dysarthric speakers

Figure 4 shows the % accuracies of the dysarthric speaker BK when using: the unadapted SI models (WSJ), the MLLR adapted SI models (ADAPT), the metamodels built using the unadapted acoustic models (META-WSJ), and the metamodels built using the adapted acoustic models (META-ADAPT). It can be seen that the baseline accuracy is very low for this speaker (10%) and even after 34 sentences have been used for speaker adaptation, the accuracy reaches only 39.9%. However, using the metamodels trained on adapted acoustic models increases the accuracy to 50% when only four sentences are used for adaptation, and the accuracy remains at about this level when more sentences are used, regardless of whether the metamodels were trained using adapted or unadapted acoustic models.

Figure 5 shows the mean accuracies across all the NEMOURS database speakers. On average, using metamodels built using both adapted and unadapted acoustic models gives large gains in

Figure 3: Comparison of recognition performance: Human assessment (FDA), unadapted (WSJ) and adapted (ADAPT) SI models

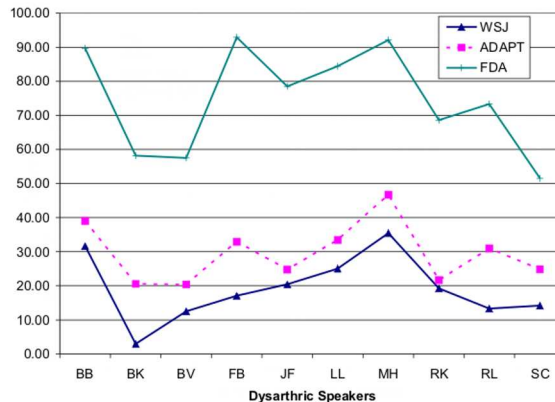
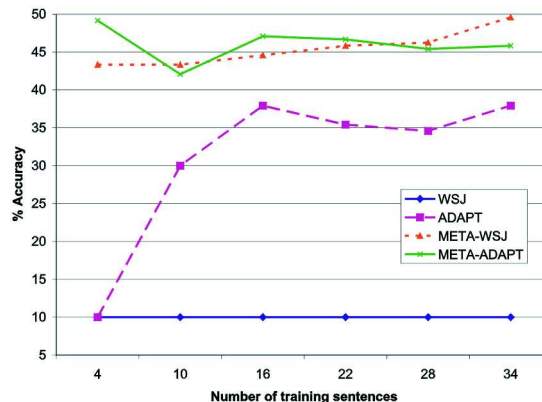


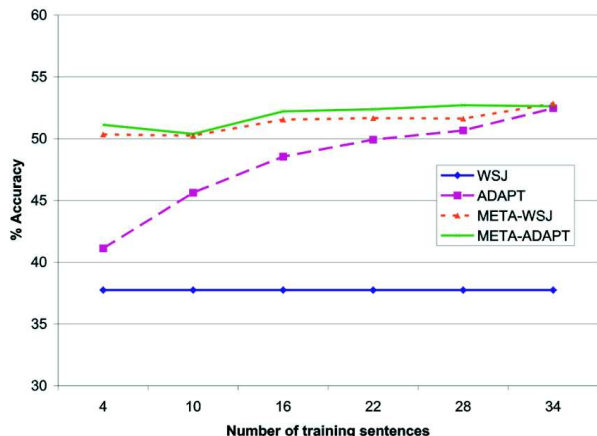
Figure 4: Dysarthric speaker BK: comparison of % accuracy for different techniques.



accuracies over MLLR adaptation when the number of sentences available for adaptation is low, but this advantage decreases as the number of sentences reaches the maximum (34), at which point performance is about the same.

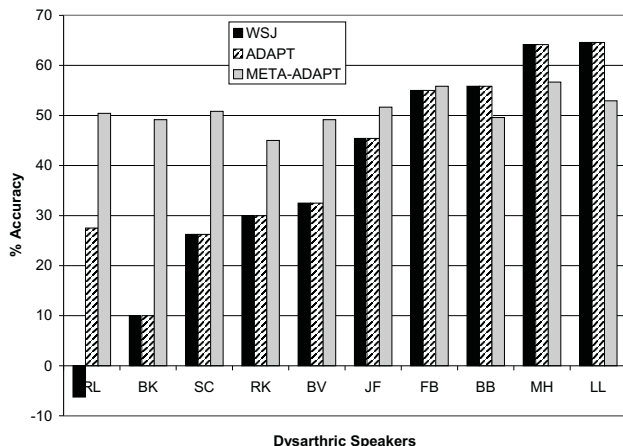
As rapid adaptation to a new speaker is highly desirable, we investigated in more detail the performance of metamodels vs. MLLR when only four sentences were available for metamodel estimation or MLLR adaptation. Figure 6 plots the % accuracy for each dysarthric speaker of the baseline system (WSJ), the MLLR adapted models system (ADAPT), and the metamodels (META-ADAPT) for this condition. The speakers have been ordered in ascending baseline performance. For speakers RL to JF inclusive, the metamodels performance is always above both the baseline and the MLLR adapted performance. For speaker FB, the metamodel performance is the same as the MLLR performance and thereafter, the metamodels give a slightly lower level of performance than the baseline, whereas the MLLR adaptation perfor-

Figure 5: Mean across all dysarthric speakers: comparison of % accuracy for different techniques



mance is always higher. We used the matched pairs test described

Figure 6: Recognition accuracies of the baseline system, MLLR adapted models and metamodels, for each dysarthric speaker (four utterances used for speaker adaptation or confusion matrix estimation).



in [3] to test for significant differences between the recognition accuracy using metamodels and the accuracy using MLLR adaptation when a certain number of sentences was available for meta-model estimation or adaptation. The results with the associated  $p$ -values are presented in Table 2.

Table 2: Comparison of statistical significance of results over all dysarthric speakers

Number of sentences	Result
4	Metamodels outperform MLLR ( $p < 0.001$ )
10	Metamodels outperform MLLR ( $p < 0.001$ )
16	Metamodels outperform MLLR ( $p < 0.001$ )
22	Metamodels outperform MLLR ( $p < 0.01$ )
28	Metamodels outperform MLLR ( $p < 0.05$ )
34	No significant difference

### 4.3. Discussion

The results we have obtained indicate that the use of metamodels is a significantly better approach to ASR than speaker adaptation in cases where the intelligibility of the speaker is low and only a few adaptation utterances are available, which are two important conditions. We believe that the success of metamodels in increasing performance for low-intelligibility speakers can be attributed to the fact that these speakers often display a confusion matrix that is similar to the matrix shown in Figure 1, in which a few phonemes dominate the speaker’s repertoire. The metamodels learn the patterns of substitution more quickly than the speaker adaptation technique and hence perform better even when only a few sentences are available to estimate the confusion matrix.

The results presented must be treated with some caution in view of the very small size of the language model available for use on the NEMOURS data. In the speaker adaptation experiments, best performance was obtained using a grammar factor as high as 50, which indicates that the language model was dominating performance (N.B. all results shown here are the best results obtained after experimentation with different grammar factors). However, the language models were identical in all experiments in which MLLR adaptation was compared with metamodels, and so the significance results presented in Table 2 are justifiable.

## 5. SUMMARY, CONCLUSIONS AND FUTURE WORK

This study has shown that the use of metamodels, which incorporate a model of a speaker’s confusion matrix into the decoding process, is a promising technique to improve recognition accuracy when the speech has low intelligibility and there is limited adaptation data available for a speaker, two conditions that are often met when dealing with dysarthric speakers. Our next step is to understand why the modelling does not work as well when the baseline performance is high, and hence to develop improved models that work over a greater range of baseline performance that should be effective for normal as well as for dysarthric speakers. We will also test the effectiveness of the technique on a much larger database collected from a dysarthric speaker.

## 6. REFERENCES

- [1] Bunnell, H.T. and Polikoff, J.B. *The Nemours Database of Dysarthric Speech*. Proceedings of ICSLP, 1996.
- [2] Cox, Stephen J. and Dasmahapatra, Srinandan. High-level approaches to confidence estimation in speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(7):460 – 471, 2002.
- [3] L Gillick and S.J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 532–535, April 1989.
- [4] Kain, A., Niu X., Hosom, J-P, Miao, Q., and Santen, J. Formant re-synthesis of dysarthric speech. *Center of Spoken Language Understanding CSLU, Oregon USA.*, 2004.
- [5] Rosen, K. and Yampolsky, S. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, 16:48–60, 2000.
- [6] Young, Steve and Woodland, Phil. *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005.