Proceedings of the Institute of Acoustics

A PHONETICALLY MOTIVATED ANALYSIS OF THE PERFORMANCE OF THE ARM CONTINUOUS SPEECH RECOGNITION SYSTEM

S R Browning, R K Moore, K M Ponting & M J Russell

Speech Research Unit, RSRE, St Andrews Road, Malvern, WORCS WR14 3PS, UK.

1. INTRODUCTION

This paper presents a phonetically motivated analysis of a version of the continuous speech recognition system developed as part of the 'ARM' (Airborne Reconnaissance Mission) project at the RSRE Speech Research Unit. This version of the system is speaker-dependent and is based on phoneme-level hidden Markov models. In order to take account of the context-sensitivity of the acoustic realisation of phonemes, approximately 1500 word-internal triphone models are used. A triphone is a model of a phoneme in its left and right context. The version of the system on which this investigation was based scores an average of 86.8% word accuracy with word level syntax (perplexity 497). The 'ARM' system is more fully described in Russell, Ponting & Peeling [1]. The aim of the work described here is to investigate to what extent errors can be explained by phonetic effects; those which cannot may indicate where models could be improved. For instance if /p/ is misrecognised as /b/, this is understandable from a phonetic point of view, as the two are acoustically very similar; however if /p/ is consistently misrecognised as /z/ or /@U/ it would be difficult to explain on acoustic-phonetics grounds, and would probably indicate that there is something wrong with the model(s). The following section describes the background to the investigation, and how the phoneme performance is evaluated. The remainder of the paper contains an analysis of specific types of errors, motivated by the desire to find phonetic explanations of them. The transcriptions in this paper are in the SAM-PA notation (Fourcin et al [2].

2. BACKGROUND

2.1 The system

For the purposes of the analysis described in this paper the system was configured as a phoneme recogniser with no dictionary and no syntax. There is, however, some measure of constraint in that the right context of each triphone must match the left context of the next. This arrangement produced an overall phoneme error rate of 26.2% (The total number of phonemes was 6873). The system currently recognises the speech of three speakers, being trained separately for each, and the differences between speakers have also been studied. The system was trained using approximately fifteen minutes of speech (airborne reconnaissance mission reports) from the three speakers. Each speaker has their own dictionary to take account of dialectal variations. Speakers 1 and 2 are male; Speaker 1 is basically RP, while Speaker 2 has Midlands overtones. Speaker 3 is female and has north-eastern colours in her accent.

In addition to the triphones for each context-sensitive phoneme, a number of short words are modelled explicitly at the word level. Non-speech sounds, such as breath noise or lip smacks are also modelled explicitly with a set of single state models. Both the word models and the noise models are treated exactly in the same way as the triphones.

2.2 Evaluating the performance

The system has so far been tested on ten ARM reports from each speaker, containing a total of approximately 2290 phonemes per speaker, 6873 in all. Phoneme recognition performance is measured by aligning the output of the system with a phonemic transcription of the test material. The latter is obtained by replacing each word in the orthographic transcription of the data with its phonemic transcription from the dictionary. Errors are classified as substitutions, deletions or insertions. Substitutions occur when a phoneme is misrecognised as another phoneme, deletions when a phoneme has been missed by the system, and insertions when the system has recognised an extra phoneme. Recognition performance is measured in terms of correctness and accuracy. The first is simply a measure of how many times the system produced the same label as the dictionary transcription, while the second is a more stringent measure, which is calculated by subtracting the number of insertions from the number of correctly recognised phonemes, and as such is a more satisfactory indicator of the recognition performance.

It is in practice extremely difficult to assess performance, as in many cases the speaker will not actually produce the somewhat idealised pronunciation represented in the dictionary. For example, in the sequence *"six six"* the speaker is likely to produce only one /s/ (though it may be somewhat lengthened) for the two which phonemically occur over the word boundary. In this example if the system recognises only one /s/ it is penalised for having deleted a phoneme. There are numerous examples of this nature, and these will be discussed under the appropriate categories below. In order to ensure that our evaluation system is both consistent and automatic, we score strictly against the dictionary transcription. We are, however, currently looking at including alternative transcriptions in the dictionary, which will allow us to take account of many of these so–called errors.

3. ANALYSIS OF RECOGNITION RESULTS

In this section the phoneme recognition results are analysed in some detail. First the overall performance will be examined, followed by discussion of the main errors under the headings substitutions, deletions and insertions. Lack of space prevents us from presenting all the details of the performance for all three speakers, so as the results are broadly similar, most of those presented in this paper will be for all speakers combined, with speaker–specific exceptions discussed where they occur. The results appear in full detail in Browning [3].

3.1 Recognition Performance

The phoneme recognition results for each speaker and for all speakers combined are shown in Table 1.

Speaker	% correct	% substitution	% deletion	% accuracy	total no. of phonemes
1	75.5	13.2	11.3	68.8	2290
2	82.0	10.9	7.1	76.8	2290
3	80.8	11.6	7.6	76.0	2293
All speakers	79.4	11.9	8.7	73.8	6873

Table 1. Summary of phoneme recognition results.

From this it can be seen that the results for all three speakers are in the same range, although Speaker 2 and Speaker 3 have slightly better performance than Speaker 1. This general trend is evident in most of the more detailed analyses of phoneme performance; particular differences between speakers will be pointed out below.

The performance on individual phonemes for all speakers is shown in Table 2.

	Total	% Cor	% Sub	% Del	No. of Ins	% Acc		Total	% Cor	% Sub	% Del	No. of Ins	% Acc
s	408	85.8	8.6	5.6	23	80.2	i	224	86.1	11.2	2.7	9	82.1
z	171	80.1	12.9	7.0	25	65.5	I	394	72.1	18.3	9.6	32	64.0
S	93	93.5	1.1	5.4	0	93.5	E	249	81.5	10.8	7.6	7	78.7
f	234	87.6	7.7	4.7	10	83.3	{	165	88.5	10.3	1.2	4	86.1
v	90	56.7	13.3	30.0	13	42.3	A	111	97.3	2.7	0.0	2	95.5
T	107	74.8	11.2	14.0	10	65.5	Q	56	89.3	3.6	7.1	0	89.3
D	1	0.0	100.0	0.0	0	0.0	O	111	91.9	3.6	4.5	1	91.0
h	39	66.7	17.9	15.4	2	61.6	U	9	77.8	0.0	22.2	1	66.7
tS	27	81.5	18.5	0.0	0	81.5	u	165	75.2	18.8	6.0	5	72.2
dZ	33	72.7	18.2	9.1	2	66.6	3	9	100.0	0.0	0.0	2	77.2
p	123	88.6	7.3	4.1	34	61.0	@	450	62.2	16.9	20.9	40	53.3
b	45	57.9	17.8	24.2	15	24.5	V	88	72.7	19.3	8.0	8	63.6
t	693	85.9	7.6	6.5	23	82.6	eI	147	93.9	6.1	0.0	0	93.9
d	246	64.6	15.9	19.5	51	43.9	aI	153	91.5	5.9	2.6	2	89.2
k	291	91.1	3.4	5.5	12	87.0	oI	3	100.0	0.0	0.0	0	100.0
g	123	80.5	14.6	4.9	0	80.5	aU	48	89.6	6.3	4.1	1	87.5
m	147	72.2	15.6	12.2	12	64.0	@U	168	76.8	19.6	3.6	4	74.4
n	513	70.8	15.4	13.8	21	66.7	I@	51	92.2	7.8	0.0	1	90.2
N	54	51.9	27.8	20.3	1	50.0	e@	6	50.0	50.0	0.0	0	50.0
l	225	80.5	8.4	11.1	17	72.9	<at>	63	33.3	65.1	1.6	0	33.3
r	309	90.9	2.9	6.2	15	86.0	<oh>	18	44.4	55.6	0.0	0	44.4
w	132	85.6	10.6	3.8	4	82.6	<of>	33	39.4	54.5	6.1	0	39.4
j	42	85.7	9.5	4.8	2	80.9	<or>	6	33.3	66.7	0.0	0	33.3

Table 2. Individual phoneme recognition – all speakers

A number of phonemes (/D, oI, 3, e@ and U/) occur so rarely in the 'ARM' reports that their results are unreliable indicators of performance, so these will be ignored in this analysis. Looking at individual phonemes /A/ was recognised most reliably, closely followed by /S/ and /O/. The models with the poorest performance were those for whole words, which tended to be confused with one (or more) phonemes. Of the phoneme models the least correct were /N/ and /v/. However, in terms of accuracy /b/ and /d/ also score badly, because of the high number of insertions of those phonemes.

In trying to find general trends in phoneme recognition performance the phonemes have been classified into phonetically motivated groups, namely 'manner' and 'place of articulation'. Under 'manner' there is a broad classification into vowels and consonants, which should be self-evident, and a finer one where consonants are split into more specific clusters. The class labels and their members are shown in the keys accompanying Tables 3 and 4. (I have disregarded the word-level models in this classification.)

3.1.1 Manner of Articulation. There is no significant difference in the recognition performance between vowels and consonants, with vowel correctness 79.5% (n=2607) and consonants 80.6% (n=4146). However, consonants are more than twice as likely to be inserted as vowels; 267 insertions compared with 119, making the accuracy for the consonants slightly lower; consonants 74.2%, vowels 75.0%.

The results analysed in terms of manner of articulation are presented in Table 3.

		%	%	%	No. of	%
Class	Total	Cor	Sub	Del	Ins	Acc
Plosive	1521	82.4	9.0	8.6	126	74.1
Affricate	60	76.7	18.3	5.0	2	73.3
Str Fric	672	85.4	8.6	6.0	47	78.4
Wk Fric	471	76.9	10.6	12.5	35	69.4
Liq/Glide	708	86.3	6.5	7.2	30	82.1
Nasal	714	69.6	16.4	14.0	27	65.8
Vowel	2607	79.5	12.9	7.6	119	75.0

Class		Members
Plosive	:	p b t d k g
Affric	:	tS dZ
Str Fric	:	s z S
Wk Fric	:	f v T h
Liq/Glide	:	l r w j
Nasal	:	n m N
Vowel	:	i I E A { O Q u U V 3 @ aI eI oI aU @U I@ e@ (U@)

Table 3. Results analysed by manner of articulation

Liquids/glides and strong fricatives were recognised most correctly and accurately for all speakers. Nasals were quite clearly the worst, though the accuracy of weak fricatives was poor because of the high number of insertions. Both of these classes may be acoustically weak, and /v/ especially is easily missed, which might explain their poor performance. It is not surprising that strong fricatives should be well modelled, as they are generally acoustically prominent (compared to weak fricatives, especially). More unexpected was the good performance of liquids and glides which are often thought to be problematic for systems with limited ability to model temporal dynamics. The explanation for this may be provided by the variable frame rate analysis which is used; areas which are acoustically stable are compressed into a smaller number of frames/states, while those that vary rapidly, such as /r/ and /w/ are modelled using comparatively more states, giving the improved time resolution needed to identify these sounds.

3.1.2 Place of Articulation. Table 4 gives the analysis of the results grouped by place of articulation.

		%	%	%	No. of	%
Class	Total	Cor	Sub	Del	Ins	Acc
Labial	879	78.5	11.0	10.5	91	68.1
Alveolar	2565	80.5	10.0	9.5	159	74.3
Pal–Al	195	86.7	8.2	5.1	4	84.6
Velar	507	82.4	9.9	7.7	13	79.9
Front	1032	80.0	13.7	6.3	52	75.0
Central	547	64.5	17.1	18.4	50	55.4
Back	452	86.6	8.8	4.6	9	84.5
Fronting	303	92.8	5.9	1.3	2	92.1
Backing	216	79.6	16.7	3.7	5	77.3
Centring	57	87.7	12.3	0.0	1	86.0

Class		Members
Labial	:	p b m f v T D w
Alveolar	:	t d n s z l r
Pal–Al	:	S tS dZ j
Velar	:	k g N h
Front	:	i I E {
Central	:	V @ 3
Back	:	A O Q U u
Fronting	:	aI eI oI
Backing	:	aU @U
Centring	:	I@ e@

Table 4. Results analysed by place of articulation

Diphthongs which move towards a front position are most accurately recognised; while among the consonants, palatal–alveolars are the best recognised. Perhaps not surprisingly, central vowels were poorly dealt with. The /@/ vowel represents a large proportion (over 80%) of the central vowels and as this vowel is unstressed and notoriously variable, it is not surprising it is rather loosely modelled and, not only is easily confusable, but frequently inserted too. Labial consonants are only moderately well modelled, perhaps because most of the weak fricatives are in this group, and these are often acoustically indistinct. These results are strikingly consistent across speakers.

Having given a general assessment of the performance of the system, we shall now examine the errors in more detail, starting with substitutions.

3.2 Substitutions

When the system misrecognises one phoneme as another it is important to be able to explain why this has happened. If the two phonemes involved differ minimally, in one phonetic feature (/p/ and /b/, for instance) then it may be difficult to improve either model to separate them. If, however, larger differences are involved, there may be more scope for better modelling. In order to investigate what proportion of the substitution errors were phonetically predictable a confusion matrix was constructed and examined for type of error according to manner and place of articulation.

3.2.1 Manner. There is no evidence that either vowels or consonants are more subject to substitution. Consonants are recognised as consonants 93%, and vowels as vowels nearly 90% of the time.

The results of the finer manner class analysis are show in Table 5. This matrix shows how often phonemes from one class were recognised as phonemes from other classes. The matrix diagonal shows within-class recognitions; the overall class accuracy was 87.1%.

		%Recognised						
		Plo	Aff	SF	WF	L/G	Nas	Vow
	Plosive	87.4	0.3	0.8	0.7	0.6	0.3	1.0
S	Affricate	3.3	81.7	6.7	.	.	1.7	1.7
p	Str Fric	1.5	0.4	90.8	0.3	.	0.7	0.3
o	Wk Fric	5.3	0.2	0.4	80.0	.	.	1.3
k	Liq/Glide	0.7	0.1	.	0.6	87.3	0.4	3.7
e	Nasal	2.2	.	.	0.3	1.8	77.2	3.8
n	Vowel	0.6	.	0.1	0.2	0.6	0.5	89.9

Table 5. Confusion matrix for manner of articulation

Nasals were the most confused, though most of the confusions are predictable; nasals share stop-like characteristics with plosives, and a vowel-like structure with liquids and vowels. It is interesting that almost all (95%) of the nasal/plosive confusions were for Speaker 1, where /n/ was mostly misrecognised as /b/ and /d/.

Plosives were misrecognised most often as vowels. Nearly half of these unexpected confusions are with central vowels, indicating that /@/ is a major culprit in misrecognition (as well as being misrecognised itself). In general plosives are the most often substituted class.

The rest of the matrix is very much as one would expect. In general in-class recognition is good. Affricates are confused with plosives and strong fricatives with which they share many features. Weak fricatives are also confused with plosives, particular confusion being /f/ with /p/, and as these share place of articulation, being broadly speaking labial, this is not unexpected.

3.2.2 Place. The overall place class accuracy was 84.4%. The confusions are shown in Table 6.

%Recognised

		Lab	Alv	P-A	Vel	Frn	Bck	Cen	F'g	B'g	C'g
	Labial	86.1	2.6	0.1	0.5	.	0.2	1.0	0.1	0.1	.
	Alveolar	1.9	85.4	0.3	0.2	0.3	0.2	1.2	0.2	0.1	.
S	Pal–Alv	.	1.0	91.3	.	1.5	.	0.5	0.5	.	.
p	Velar	1.6	4.3	0.6	84.4	0.4	.	0.2	0.6	.	0.2
o	Front	0.1	0.8	0.4	.	85.5	1.5	2.9	1.0	0.6	0.5
k	Back	0.6	1.1	.	.	2.7	86.5	1.1	0.2	1.3	.
e	Central	2.2	2.0	0.2	0.2	5.7	1.5	67.1	0.5	1.1	0.4
n	Fronting	.	.	.	.	· 3.0	0.7	1.0	93.1	0.7	0.3
	Backing	.	0.5	.	.	2.8	0.9	7.9	2.8	79.6	0.5
	Centring	.	.	.	.	8.8	.	1.8	.	.	89.5

Table 6. Confusion matrix for place of articulation

As might be expected from the earlier comments on recognition results, the central vowels are weakest – they are confused with a wide range of different classes, and are the most widely substituted class too. From this table it can be seen that much of the poor recognition of labials is likely to be due to them being confused with each other, with alveolars being the most likely substitute.

3.2.3 Contextual Effects. Substitution errors can sometimes be explained by context, though these effects are currently based on informed intuitions as a statistical analysis has not yet been performed. But examples such as /n/ recognised as /N/ in "*machine gun*"; /m/ as /n/ before an alveolar in "*platforms*" ; /g/ as /d/ and /d/ as /p/ in the sequence "*target grid ref*"; /s/ as /z/ in voiced environment "*zero seven*"; and the sequence /st/ as /zd/ in the voiced environment "*fuel station*" are not hard to find. A more detailed description of these errors is contained in Crowe [4], and a more quantified analysis will be found in Browning [3]. These examples nearly always involve minimal difference between target and recognised phoneme, such as place of articulation or voicing, and serve to bear out the hypothesis that a major part of the substitution errors made by the system have a phonetic explanation.

3.2.4 "Non–errors". In addition, as has already been mentioned above, some substitution errors are due to the quite legitimate variations which occur in fluent speech. The alternation of /i/ with /I/ in final unstressed syllables, such as in *facility*, and *twenty*, and /@/ with practically any unstressed vowel is well known, and caused many substitution errors.

3.3 Deletions

Deletions account for 42% of the recognition errors, so it would be useful to find out why they occur. Many of the deletion errors are not errors at all but are genuine elisions by the speaker. For example, the unstressed /@/ vowel is often elided, but in the present analysis if an /@/ appears in the dictionary transcription it will be scored as a deletion if it isn't recognised, even if in reality it wasn't there. The same is true of word–final stops, which are frequently omitted, particularly in fast speech (e.g. "*target category*" is realised as /tAgIk(t@gri/). These errors are again analysed in more detail in Crowe [4] and Browning [3].

It is probably for this reason that /@/ is the most deleted and is twice as likely to be deleted as any other vowel. Among the consonants /v/ scores poorly, as does /b/ (see Table 2). We have already discussed the possible reasons for the poor performance of /v/, and of weak fricatives and nasals in general, but it is not so clear why a sound such as /b/ should be missed, but since this is consistent across speakers, it is possible that the models are defective in some way. There also appears to be a problem with /m/ specific to Speaker 1. 28.6% of this speaker's /m/s were deleted,as compared to 4.1% for both Speaker 2 and Speaker 3. Again, there is at the moment no explanation for this.

A scored deletion is often the result of the system labelling two phonemes as one, for instance, part of the second /n/ in "*niner*" is often labelled as part of the /aI/. This may be due to the fact that "*niner*" occurs frequently in the database, so it will have a significant influence on the (aI:n_n) triphone (/aI/ with /n/ as its left and right context). The triphone may therefore end up modelling part of the /n/.

3.4 Insertions

Insertions occur when the system has put in an extra phoneme label. These often occur when a long phoneme has been recognised as two separate phonemes. Sometimes these phonemes will be identical, as when /@U/ is transcribed as /@U @U/; others are phonetically related as when /s/ following a voiced sound (and usually word initial) is transcribed as /z s/. It is also common for diphthongs to be recognised as two vowels, so "*eight*" gets recognised as /eI i t/, "*many*" as /mEniI/. Off glides from vowels are often recognised as vowel+/@/, e.g. /O/ in "*four*" as /O@/, and /@U/ in "*zero*" as /@U @/. Examples like these seem to account for a large number of the vowel insertions, though we have as yet no quantifiable results. (See Browning [3].)

We have already mentioned that consonants are more than twice as likely to get inserted as vowels (and see Tables 3 and 4). The comparatively high level of consonant insertion was common to Speaker 2 (90 consonants compared with 30 vowels) and Speaker 1 (99 consonants and 44 vowels), but not so conspicuous in Speaker 3's results which contained less insertions anyway (68 consonants and 45 vowels). Plosives and alveolars are the most inserted consonants accounting for 47% and 59% of consonant insertions, respectively. From Table 2 it can be seen that all the plosives except /g/ are frequently inserted, as is /n/. Plosives are most frequently inserted between words, and in some cases this may be due to breath noises or lip smacks. The predominance of alveolar plosive insertion may be mainly accounted for by an interesting speaker specific insertion of /d/. There are 33 instances of /d/ insertion (for no apparent reason) in Speaker 1's reports, while Speaker 2 and Speaker 3 each have 9, and this accounts for the poor accuracy of /d/ overall.

Much more encouraging are the insertions which can be accounted for by the speaker inserting a phoneme in particular context to ease the transition between sounds. Examples such as insertion of /r/ in "*4/8*" /fOrcItTs/ and between "*niner*" and "*oh*". In *2/8* /w/ is inserted /tuweitTs/, and /j/ in "*virtually unusable*" /v3tS@lI j @n.../. In these cases the system is merely recording what is there, although this currently is counted as an insertion.

4. CONCLUSIONS

There are many interesting observations to be made from this data. What has been presented here has been an attempt to pull these together and point out general trends, which might indicate what the phoneme models are doing right, as well as what they are doing wrong.

From this short discussion there have emerged two types of error: those which are genuine misrecognitions (whether phonetically explainable or not), and those which are due to the normal co-articulatory effects in fluent speech, and are thus to be expected.

Of the first the vast majority involve confusions with rather similar phonemes, or deletions of acoustically weak segments. Weak sounds such as nasals or weak fricatives predictably cause problems, as does the neutral /@/. Equally, strong and long sounds such as strong fricatives and diphthongs are well handled. The surprisingly good recognition of liquids and glides may provide an independent vindication of the use of variable frame rate analysis. A large number of the insertions and deletions could probably be prevented if our duration modelling was more sophisticated.

Although the majority of the errors appear to have a phonetic basis, there are cases where the errors are as yet inexplicable from a phonetic point of view – the unusually large number of /d/ insertions by Speaker 1, and the poor recognition of the same speaker's /m/, for example. Where there isn't a phonetic explanation of an error, it would be interesting to find out if the system's own measure of its goodness of match is consistent with our judgement of its performance.

It is important to remember that this study was based on a system which had no dictionary, though the triphones are forced to match at the edges. When lexical and syntactic constraints are available, as they are when the system is run in its usual mode, as a word recogniser, then many of the problems discussed above no longer occur. The level of performance depends on the task and vocabulary, and work is in progress to assess the extent to which the somewhat specialised vocabulary of the ARM task has influenced these results, by looking at other tasks, and bigger vocabularies, as well as at a wider range of speakers.

This study has enabled us to pinpoint a few areas where our models might be improved, but in general the errors that the 'ARM' system makes have a phonetic explanation, so it is reasonable to assume that most of the models are satisfactory.

5. REFERENCES

[1] M J RUSSELL, K M PONTING & S M PEELING, 'The Armada Speech Recognition System', Proc. Voice Systems Worldwide (1990).
[2] A J FOURCIN, G HARLAND, W BARRY & V HAZAN, 'Speech Input and Output Assessment. Multilingual Methods and Standards. Ellis Horwood Ltd., Chichester. (1990).
[3] S R BROWNING 'Analysis of the phoneme recognition performance of the ARM continuous speech recognition system' RSRE Memo (1990 forthcoming).
[4] J D M CROWE 'Contextual influences on phoneme recognition in the ARMADA speech recognition system' RSRE SP4 Research Note (1990).