Proceedings of the Institute of Acoustics

TOWARDS A PHONETICALLY-MOTIVATED PRONUNCIATION DICTIONARY FOR AUTOMATIC SPEECH RECOGNITION

Sue Browning, Andrew Buckland & Mike Tomlinson

Speech Research Unit, Defence Evaluation and Research Agency, Malvern, UK.

1. ABSTRACT

It is well known that some of the inadequacies of current Automatic Speech Recognition systems stem from the mismatch between the pronunciations of words as specified in a dictionary and their actual realisation. We contend that because of this mismatch, the annotations on which the training of the models is based are not an accurate reflection of the data, and therefore the models are equally flawed. In particular, the models do not predict (explain or accommodate) phenomena such as the changes that occur due to a wide range of factors. The aim of the work described in this paper is *not* to demonstrate that we can improve speech recognition by better modelling, but rather to discuss what we mean by better modelling, and to describe some of our tentative steps towards that goal.

2. INTRODUCTION

The motivation for the work reported in this paper is to improve the 'modelling' ability of conventional HMMs through the use of expert speech knowledge. It is undoubtedly true that state-of-the-art recognition systems employing many thousands of estimated parameters perform very successfully where there are copious amounts of annotated data and the *style* of the test data is very little different from that of the training data. Such systems, however, do not extrapolate well outside the domain defined by the training data as reflected in the parameters of the model. It could be argued that such systems fail to meet the criterion of a *good* model on two counts; firstly the model is neither parsimonious nor economic; and secondly the model is not able to predict, with any reasonable degree of fidelity, phenomena not observed in the training data. The former can be characterised by the fact that the model does not explicitly, nor in a compact parameterisable form, explain phenomena such as:

- differences between accents (e.g. "grass" realised as /grAs/ or /gr{s/) [1],
- differences between speakers - their physiology and how they produce speech sounds, including speaker specific tendencies (e.g. lisp),
- changes in vocal effort - resulting in differences in spectral shape and level,
- rate of speaking (Fosler-Lussier & Morgan, 1998), and how this affects the duration, contrastiveness, reduction, elision and insertion of sounds (e.g. the word "temporary" can have 2, 3 or 4 syllables),
- speaking style (e.g. dictation, whispering, etc.),
- prosodic effects (e.g. intonation patterns),
- emotional or physical stress (e.g. formant movement),
- differences in phonetic context (e.g. co-articulation),
- differences between allophonic variants of phonemes (e.g. variants of /r/ - these are mostly predictable but need access to syllabic and prosodic information),
- systematic differences between members of phonemic classes (e.g. movement within the vowel quadrilateral),
- asynchrony of features (e.g. nasalisation, lip-rounding, etc.).

[1] These are often predictable, but not always consistent - in our data the speaker, whose accent is northern English, usually used /{/, as would be expected but on occasion used /A/

It is clear that data-driven methods do handle well effects that are due to random behaviour. However the phenomena described above, identified by speech scientists (Moore, 1995a&b), do need accommodating effectively, in terms of an appropriate representation, if inter- and intra-speaker adaptation is to be successfully accomplished (Russell, 1997).

In order to optimise word recognition rates, even without significant moves away from the conventional HMM modelling paradigm, we need to search a complex 'paradigm parameter space' (e.g. number and identity of the modelling units, the number and connectivity of states, the amount of data sharing etc.). However, it is clear that there is much interaction between the effects of movement within these paradigm dimensions, with the result that movement in only one dimension at a time may not lead to any observable improvement in recognition rate. The problem is exacerbated if we do adopt significant changes in recognition paradigm and if we require operation in a large application space. This argument has been developed by Bourlard (Bourlard, 1995), when he suggested that such systems can get stuck in local minima and that we may need to accept increases in error rate before we achieve the ultimate goal of improving automatic word recognition.

3. ASSESSING MODELLING ABILITY

How can we proceed? What is required is a method that is likely to indicate potential improvements. We believe that this can achieved by assessing, at the *modelling* level, what benefits actually accrue from changes in paradigm parameter space. Further, this procedure will eventually lead to significant improvements in performance on speech exhibiting a large range of naturally occurring variation not represented in the training data – the argument for extrapolating outside the training data (variation) by using speech knowledge priors (Moore, 1995a&b). This leaves us with the problem of specifying tests for improved modelling, and the decision as to what particular aspects of the conventional HMM system should be addressed before we engage in large jumps of paradigm.

We are not proposing here a reality bypass for ASR researchers. It has been pointed out that we need to maintain a watch on recognition rates to keep a hold on reality (Moye, 1998) – so we should observe word recognition results periodically. However the authors of this paper will resort to claiming the 'Bourlard amendment' in defence of any apparent worsening of performance as reflected in increases in error rate.

The suggestion here is that we should look for other, albeit subjective, measures of 'improvement' in the ability of a 'new' model set to model some data which exhibits the above phenomena, observing, both qualitatively and quantitatively if, and where, any improvements in modelling occur. The following items are proposed as candidate metrics and factors to observe:

- spectra and expected durations, of the cepstral models,
- time alignments at state level on training data,
- state level durational distributions (noting extreme examples),
- overall phoneme class recognition scores,
- individual phoneme class recognition scores,
- phoneme class and allophone confusion matrices.

4. DEVELOPING AN EXPANDED BASEFORM SET AND PRONUNCIATION DICTIONARY

An important aspect of much of the work in developing ASR systems is the reliance on the notion of an immutable 'standard' baseform set and associated pronunciation dictionary. This is one of the causes of the mismatch between the pronunciations of words as specified in the dictionary and their actual realisation. We contend that because of this mismatch, the annotations on which the training of the models is based are not an accurate reflection of the data, and therefore the models are equally flawed.

Past work has mainly concentrated on using phonological rules or data-driven methods to automatically expand the dictionary, but these have met with limited success (Strick & Cucchiarini, 1998). Therefore we have adopted the alternative approach of using expert knowledge to simultaneously develop a (general) baseform set and a (data specific) pronunciation dictionary. We then go on later to address the question of model structure in terms of state topology and state tying.

Our approach to the expansion of the baseform set was motivated by close observation of the data with the application of phonetic knowledge and experience. This experience also includes intimate acquaintance with the failings of a typical ASR system (Browning et al., 1990), and the associated data sets. The data used in these investigations is from a constrained task with a vocabulary of around 500 words. This consists of spoken reports from a single speaker, each report containing around 60 words of continuous speech. Although this is obviously a very small data set, it was chosen because of our depth of familiarity with the data and to make detailed hand labelling possible.

There are a number of conflicting requirements of baseform units; consistency, so that different instances of the same unit have similar characteristics (there is a similar requirement when deciding the number of states per modelling unit); and economy, to encourage the modelling of the underlying structure of speech. As discussed above, reliance on modelling surface behaviour reduces the ability to generalise. It is also important, where contrasts exist, to maintain separable units.

Phonetic knowledge tells us that reduced vowels are different from unreduced ones, so to expect a single model to accommodate both makes the model 'fuzzy'. We also know that unvoiced plosives may be released and/or aspirated depending (among other things) on context, so it seems unreasonable to expect a single model to cope with up to three different structures; but this is what our models for /p, t, & k/ have been forced to do. Observation of the data led to the suggestion of a third allophone of /I/ (in addition to the reduced and non-reduced variants) exclusively for its occurrence in the syllable "-ing". Our treatment of /plosive + r/ clusters as a single unit was also motivated by the observation that it was impossible to separate the two, and that the /r/ in these was different from other instantiations of /r/. This analysis resulted in the proposed expanded baseform set, consisting of 92 elements.

Since most of these variations are not systematic, it was not possible to update the dictionary representation automatically, instead each word was listened to as it occurred in the data and transcribed accordingly. Our dictionary is therefore not only task and speaker dependent, but also completely data dependent. However, the expanded baseform set is intended to be general, in that it captures phenomena that occur in all natural speech.

Annotation was done at word level, with timing markers. That is, each word was carefully transcribed in terms of this new phone set according to how the word was actually pronounced. These new transcriptions were added to the dictionary as variants of the word. The phonetically transcribed data was then used to train new models for use in recognition experiments.

5. DEVELOPING BETTER MODEL STRUCTURES

Again, the approach taken was to combine knowledge gleaned from speech experts and the application of the data-driven HMM in an appropriate manner. The concept of a 'cookbook' was developed as a means of encapsulating expert speech knowledge, the priors, to be compiled out into a set of instructions to be followed during model creation and estimation. The initial cookbooks used in the course of the work reported here contain the expanded baseform set list, the topology specification of each model, and tags to indicate where states are identical and therefore share the same data (tying).

At this preliminary stage of the investigation the models used were context independent, as it was anticipated that a clearer picture would emerge of the benefits of baseform expansion when viewing a more limited model set. Hence here the modelling units were identical to the baseform units, except of course that the former have a defined *structure*, typically specified by a cookbook. Putative extensions to the scope of the cookbook, including extension to context dependent models, are discussed later.

6. EXPERIMENTAL PROCEDURE

6.1 Speech data

The investigation used speech data from a single male speaker dictating airborne reconnaissance mission (ARM) reports. The application has been used extensively in the Speech Research Unit for experiments using sub-word HMMs (Russell et al., 1990) and uses a 497 word vocabulary. For training, 36 reports were used, with these and 10 different evaluation reports used for recognition (comprising respectively 1991 and 541 words).

6.2 Speech processing

The speech, sampled at 20 kHz/sec and subjected to a Hamming window, was analysed using a 400 point DFT, and the log power spectrum output was quantised in 0.5 dB steps. The analysis windows were overlapped by 50% resulting in a frame rate of 100 frames per second. The representation used for the investigation consisted of the first 50 cosine (cepstral) coefficients plus a mean energy term. The spectrum was not subjected to any frequency distortion and no time-derivatives were used.

6.3 Model structure, training and recognition

All models had standard left to right topologies with self-loop, and no skip, transitions. Single Gaussian pdfs were employed with diagonal co-variances. The means and variances of conventional HMMs were initialised by uniform segmentation of the annotated utterances (words and non-speech items). The parameters were then re-estimated over 30 iterations of the Baum-Welch optimising algorithm.

Word recognition on both training and evaluation data was performed using a single pass Viterbi recognition algorithm (Bridle et al., 1982) with no language model constraints. A separate phone recognition pass, with no pronunciation dictionary, enabled unconstrained phone recognition performance to be assessed.

Five model sets were considered in this initial investigation (each additionally had 4, single-state, non-speech, models):

- baseform-0: unexpanded set of 46 baseforms[2], homogeneous 3 states per model, no tying, 142 free states,
- baseform-x: the expanded set of 92 baseforms, homogeneous 3 states per model, no tying, 280 free states,
- cookbook-v01: the expanded set with between 1 and 3 states per model and some tying, 189 free states,
- cookbook-v02: as v01 but with an extra state and tying rule for each of the 3 voiced plosive models, 192 free states,
- cookbook-v03: as v01 but with an extra state in each of the reduced vowel models, 200 free states.

6.4 Scoring and analysis

Word accuracy scores were produced using a phone mediated dynamic programming method; overall phone accuracy scores used a similar dynamic programming process. However in order to allow direct comparison between the results for two sizes of baseform set, the different allophones of each parent phoneme were treated as equivalent for scoring purposes. For selected individual phoneme classes, the percentage of correctly identified phones are also reported.

[2] Experiments previously reported on the ARM database (Russell et al., 1990) successfully used whole word models for the 6 short function words in the ARM vocabulary. This can be viewed as a form of baseform expansion, though this was not a principled approach.

Alignments of the cepstral models with the training data were obtained by a forced recognition process, using the known sequence of model states as the syntactic constraint during recognition. Distributions of the duration of the states of the cepstral models were obtained by analysis of these alignments. Spectrograms of the trained models were generated by a single pass of the Baum-Welch algorithm using the alignments obtained from the fully trained cepstral models to provide a set of equivalent spectral models.

7. OBSERVATIONS AND DISCUSSION

Although we are not measuring our success by word recognition performance, as we said earlier in order to keep in touch with reality we still need to observe this periodically. However, it is essential to look at phone recognition performance as an indicator of how well we are doing what we set out to do. Word and unconstrained phone accuracy scores are given in Table 1, for both the training and evaluation data. However, these global results obscure much interesting detail about what is happening with individual phones, as there is much variation between these. Table 2 shows individual phone correct scores for a selection of phones on the training set; these scores will be discussed below.

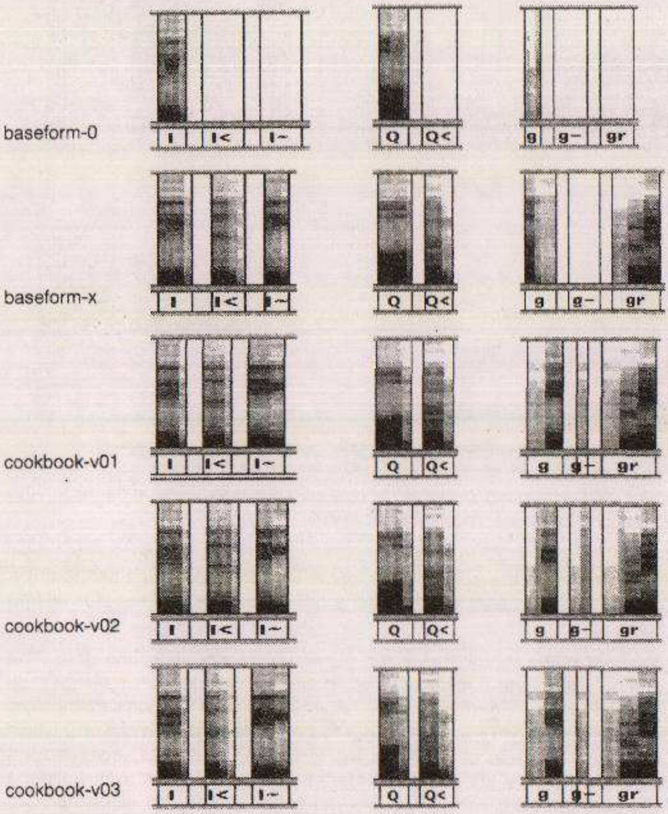

Figure 1:Spectra of models /I/, /Q/ and /g/ for all five model sets

The /I/ phoneme, is an interesting case where the score does not improve dramatically, but if we examine the models (see the first column in Figure 1), it does appear that the three allophones are modelling something different. Although it is hard to see in the figure, in the reduced form /I</ the second formant is at a slightly lower frequency than the unreduced form, and the "-ing" allophone /I~/ shows the characteristic f2/f3 merging we would expect in the velar context.

The value of tying shared states of the plosive allophones is demonstrated in the third column of Figure 1. Here the unreleased /g-/ does not occur in the training data, so is untrained in the baseform-x model set. However, because that state is forced to be tied to the closure of the released version, which does occur, then in the subsequent models it has been appropriately trained. As can be seen in Table 2, splitting the /g/ into its component allophones does indeed improve the recognition scores, though there is a drop off when the cookbooks are used. This is probably due to the fact that there are fewer free states, and therefore fewer free parameters trained for the model.

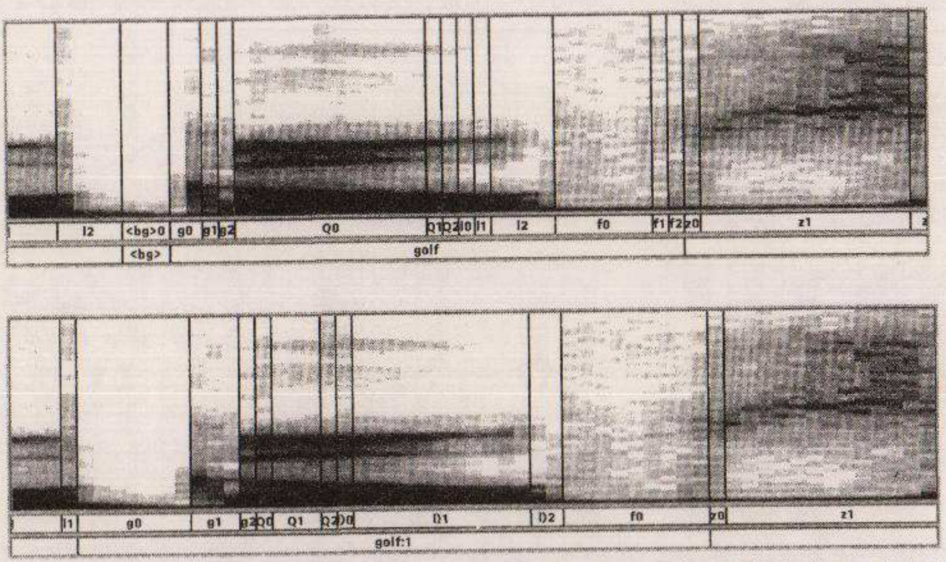

Figure 2. Spectrograms of a section of the training data, with state and word level alignments for the baseform-0 model set (upper) and the cookbook-v02 model set (lower)

There also appears to be some justification for the suggested /plosive + r/ unit. Although this is not a traditional phonetic unit, clustering in this way appears to have had a beneficial effect on the plosive and the /r/ models - see /g, t and r/ in Table 2.

Figure 2 shows the interactive effects of better modelling on the alignment of models to the data. The subjectively 'anomalous' behaviour in the baseform-0 alignment was detected by an analysis of individual state durations. The top of this figure shows the alignments of the original /g, Q, and l/ models in the word "golf", and as can be seen, the first state of the /g/ (g0) appears not to be aligned with the closure, which is evidence of poor modelling. Also, the first state of the /Q/ is too long; for three state vowel models intuition tells us that the middle state of the model should model the 'steady' part of the vowel, while the first and last states ideally would model the transitions. This has led to the misalignment of the /Q l/

sequence, which has been exacerbated by the inappropriate /l/ model. In the lower part of the figure, the result of cookbook v02 (where shared states of the plosive allophones are tied), the closure and release of the /g/ are better modelled. The steady part of the /Q/ is modelled by the middle state of the model, which is what we hoped to see, and this combined with having a model for the dark /l)/ is contributing to making that alignment better too. Looking at Table 2, we see that the scores for all three models have also gratifyingly risen.

As can be seen from Table 2, there are some losers in this, both /v/ and /V/ end up worse as a result of our efforts - for these we plead the Bourlard amendment. As expected we may have to get some other things right before sorting out the complete set of models.

Overall, the case has been made for revisiting the commonly used baseform set, and hence continuing the dialogue between speech scientists conversant with ASR problems and speech technologists. However the jury is still out on our approach to improving recognition performance outside the training data through the use of priors, and a more subjective method of assessment than that employed hitherto.

	Word		Phone	
	train	eval	train	eval
baseform-0	85.4	85.6	63.6	61.6
baseform-x	90.3	89.1	69.3	65.9
cookbook-v01	87.5	85.4	61.6	59.1
cookbook-v02	88.2	84.8	63.1	61.8
cookbook-v03	87.3	85.4	61.2	58.8

Table 1: Word and phone recognition scores on the training and evaluation data sets (% accuracy)

	I	Q	@	V	g	t	r	l	v
baseform-0	51.2	50.0	37.8	72.1	56.7	49.2	77.8	47.8	59.7
baseform-x	57.0	77.7	57.8	53.0	80.6	71.7	88.2	70.8	57.1
cookbook-v01	46.8	67.5	60.0	49.3	65.5	61.8	85.1	67.5	42.8
cookbook-v02	49.7	70.9	56.4	50.6	66.3	62.3	86.2	67.8	42.2
cookbook-v03	49.5	66.6	58.1	51.8	62.1	61.1	85.1	63.1	39.1

Table 2: Individual phoneme class recognition scores on the training data set (% correct)

8. FURTHER WORK

One of the things we would still like to sort out is how to model the unstressed vowel formerly known as schwa. Although performance on this does improve quite a lot (Table 2), we think this is probably accounted for by the fact that 28% of sounds which were labelled schwa in the original transcriptions have been re-labelled, mostly as reduced forms of other vowels, but some as syllabic consonants[3].

[3] In the original dictionary syllabic consonants were inconsistently labelled, resulting in both the schwa model and the consonant models possibly being muddied.

It is envisaged that the content of the cookbook will be extended to encompass rules for grouping of models where the connection is more complex than the identity, both acoustically and durationally, and should also indicate contrastive cues. A further consideration for inclusion is the set of phonetically motivated questions forming the expansion rules of a context sensitive/dependent decision tree.

Further, we intend to explore the use of these ideas in conjunction with larger paradigm jumps such as those concerned with asynchrony of features (Tomlinson et al., 1997) and trajectory modelling (Holmes & Russell, 1997).

9. REFERENCES

H. Bourlard, (1995). 'Towards increasing speech recognition error rates', Keynote Paper, Proceedings of EUROSPEECH'95, Madrid, pp. 883-894.

J. S. Bridle, M. D. Brown, & R. M. Chamberlain, (1982). 'A one-pass algorithm for connected word recognition', Proceedings of IEEE ICASSP'82, Paris.

S. R. Browning, R. K. Moore, K. M. Ponting, & M. J. Russell, (1990). 'A phonetically motivated analysis of the performance of the ARM continuous speech recognition system', Proceedings of the IOA Speech and Hearing Conference, Windermere, pp. 133-40.

W. J. Holmes & M. J. Russell, (1997). 'Linear dynamic segmental HMMs: variability representation and training procedure', Proceedings of IEEE ICASSP'97, Munich, pp. 1399-1402.

E. Fosler-Lussier & N. Morgan, (1998). 'Effects of speaking rate and word frequency on conversational pronunciations', Proceedings of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, The Netherlands, pp.35-40.

R. K. Moore, (1995a). 'Computational Phonetics', Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm.

R. K. Moore, (1995b). 'Automatic speech recognition: theory and practice', In the ERASMUS Handbook on European Studies in Phonetics and Speech Communication.

L. Moye, (1998). Personal Communication – SRU Tea room discussion.

M. J. Russell, (1997). 'Progress towards speech models that model speech', Proceedings of the IEEE Workshop on Speech Recognition and Understanding, Santa Barbara, CA.

M. J. Russell, K. M. Ponting, S.M. Peeling, S. R. Browning, J. S. Bridle, & R. K. Moore, (1990). 'The ARM continuous speech recognition system', Proceedings of IEEE ICASSP'90, Albuquerque.

H. Strik & C. Cucchiarini, (1998). 'Modeling pronunciation variation for ASR: overview and comparison of methods', Proceedings of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, The Netherlands, pp.137-44.

M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland, & M. A. Fawley, (1997). 'Modelling asynchrony in speech using elementary single-signal decomposition', Proceedings of IEEE ICASSP'97, Munich.