USING ACCENT-SPECIFIC PRONUNCIATION MODELLING FOR IMPROVED LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

J.J. Humphries, P.C. Woodland

e-mail: {jjh11,pcw}@eng.cam.ac.uk
Cambridge University Engineering Department, Trumpington Street, Cambridge, UK

ABSTRACT

A method of modelling accent-specific pronunciation variations is presented. Speech from an unseen accent group is phonetically transcribed such that pronunciation variations may be derived. These context-dependent variations are clustered in decision trees which are used as a model of the pronunciation variation associated with this new accent group. The trees are then used to build a new pronunciation dictionary for use during the recognition process. Experiments are presented, based on Wall Street Journal and WSJCAM0 corpora, for the recognition of American speakers using a British English recogniser. Speaker independent as well as speaker dependent adaptation scenarios are presented, giving up to 20% reduction in word error rate. A linguistic analysis of the pronunciation model is presented and finally the technique is combined with maximum likelihood linear regression, a well proven acoustic adaptation technique, yielding further improvement.

1. INTRODUCTION

Most *speaker independent* (SI) speech recognition systems comprise a set of acoustic models (for example hidden Markov models, HMMs) whose parameters are estimated using speech data from a large set of speakers. Two principal differences exist between speakers: *acoustic* differences, related to the size and shape of the vocal tract, and *pronunciation* differences which are generally referred to as *accent* and are often geographically based. In practice, it is difficult to obtain training data for speech recognition systems such that all regional accents are incorporated; in England alone there are at least ten broad regional accents [11]. SI speech recognition systems do not perform as well as *speaker dependent* (SD) systems, largely because of the need to represent speaker variations within a single model.

It is increasingly common for SI speech recognition systems to adapt to the current speaker, thus improving performance to the levels of an SD system. Most successful systems to date have achieved this through adaptation of the acoustic models by re-estimation of model parameters [8]. Such techniques usually make the assumption that all speakers are pronouncing words in a predefined manner, as described in the *pronunciation dictionary* (PD). We believe that this is a poor assumption and indeed the use of pronunciation modelling for isolated word recognition was demonstrated to improve performance in [7]. This paper extends that technique to continuous large vocabulary speech and deals with phone insertion and deletion in addition to substitution. It presents experiments for the recognition of American accented speech using a recogniser trained on British speakers. The paper then goes on to describe the combination of pronunciation adaptation with a successful acoustic adaptation technique.

2. PRONUNCIATION MODELLING TECHNIQUE

The first stage in the modelling process is to obtain accurate phone level transcriptions of the *non-native* data in terms of the phone set of the *native* recogniser.

A native triphone-based phone recogniser is used to transcribe the non-native utterances. These errorful phone level transcriptions are then aligned, using a dynamic programming (DP) technique, to a phonetic transcription derived from a (given) word level transcription and a dictionary native to the phone recogniser. Forced alignment is used to select the appropriate pronunciation from the dictionary if more than one is available. In this way, a list of context dependent phone substitutions, insertions and deletions are generated, describing how the non-native speakers' pronunciations differ from those assumed by the native system (measured in terms of the native phone set).

The phone replacement observations are of the form $l - m + r \rightarrow s$ where l and r are respectively the left and right contexts of a phone m, which is replaced by s. The replacement, s, may be another phone, null (representing phone deletion) or a group of phones (representing phone insertion, perhaps combined with substitution).

A decision tree is then used to cluster the measured pronunciation variations by considering the phonetic features of the contextual information (l and r). Details of the principles of tree building are well covered in the literature, e.g. [1] and so only implementation specific details are presented here. Two algorithms have been investigated: the CART algorithm of Breiman *et al.* [1] and the algorithm of Gelfand *et al.* [5] which will be referred to as GRD throughout this paper.

The tree building algorithms partition the set of pronunciation rules by making a series of binary splits, selected from a set of around 70 questions for each of the left and right contexts, to reduce a measure of the misclassification error rate of the tree. Each node, t, of a resulting tree defines the probability $p_t(s_i)$ of each possible substituting phone class $s_i \in \mathcal{S}$ such that $\sum_{i=1}^{S} p_t(s_i) = 1$. At present, the set $\mathcal{S}$ is restricted to phone substitutions, deletions or single insertions. Multiple insertions are currently ignored.

2.1. Choice of tree building algorithm

Experiments suggest that with large amounts of training data, similar trees are grown with both algorithms, but GRD is faster. *n*-fold CART (typically with $n = 8$) was found to be more robust for tree growing from small data sets (since the amount of data held out for testing at any one point is much smaller than that held out by GRD).

Separate trees are grown for each base phone since it was found that for a well trained tree, the contents of almost all leaves are base phone specific, thus effectively resulting in base phone dependent sub-trees off the root node. Growing separate trees results in a question set which is one third smaller thus enabling faster tree growth.

2.2. Confidence measures

The phone error rate of the phone recogniser is known to be around 45%. Appending a confidence score to each phone output by the phone recogniser provides a way of filtering out poorly-transcribed data.

The recogniser framework used here does not allow for the output of posterior phone likelihoods, but results reported in much of the literature [3, 6, 10] suggest that a measure based on the number of competing models is useful. Such a measure was implemented by examination of the pruned search space (lattice) for each utterance. For a phone m output between time frames f_s and f_e, the following confidence measure was calculated for each recognised phone m:

$$\frac{\text{\# active search paths for phone } m \text{ between frames } f_s \text{ and } f_e}{\text{\# active search paths between frames } f_s \text{ and } f_e}.$$

This measure lies in the range 0 to 1, and examination of the distribution of confidence scores enables sensible cut-off thresholds to be chosen, below which the phone recogniser output may be considered to be unreliable. This is particularly useful for growing trees from small amounts of data where each data item has considerable influence on the outcome of the final tree.

2.3. Building an accent-specific dictionary

From the original PD, a new accent-specific PD can now be generated since the tree built using this technique enables a list of phone *replacements* (substitutions, deletions and insertions) for a specific phone within a given context to be generated. Each of these replacements carries with it a probability (the sum of all such probabilities within each leaf sum to 1), thus enabling the probability of each of the new pronunciations for a particular word to be calculated as the product of each of the individual phone replacement probabilities. A pronunciation probability threshold can be set to limit the number of pronunciations generated (and surviving probabilities then normalised such that for each word they sum to 1). The new, adapted PD may then be used in speech recognition tasks. Pronunciation adaptation has been successfully applied in a number of scenarios which are summarised in the next section.

3. EXPERIMENTS AND RESULTS

3.1. Evaluation system and data

Experimental results presented below were generated using an HTK based British English recogniser. This was trained from the WSJCAM0 [4] speaker independent training set (92 speakers over a total of 7861 utterances). The recogniser used state-clustered eight-mixture triphone HMMs [12] in conjunction with a bigram language model and a subset of the BEEP dictionary, produced at Cambridge University Engineering Department, which provides British pronunciations. The accented speech, to which we were adapting, was American speech taken from the WSJ0 [9] corpus. Three parts of this database were used:

1. s0 speaker independent training set (s0_tr) comprising 84 speakers over a total of 7185 utterances (used for SI adaptation);
2. s0 evaluation test set (s0_et) of 20 speakers over 425 utterances (used for recognition tests);
3. s0 adaptation set (s0_ad) of 40 sentences for each of the 20 speakers found in the s0 evaluation test set.

3.2. SI pronunciation adaptation

Phone level transcriptions for the 7185 utterances of the American training set (s0_tr) were produced and by DP-alignment with BEEP, derived phone level transcriptions some 500 000 pronunciation observations were generated. Initially, all observations were used to build (using GRD) an SI pronunciation tree set from which SI adapted pronunciation dictionaries were produced.

Analysis

Examination of the resulting dictionary revealed some interesting correlations with linguistic analyses of the American accent [11, 2]. A typical word entry is shown below[1] and gives three weighted pronunciations for the word *waiting*.

BEEP	top American (Am) PD entries	
pronunciation	probability	pronunciation
w eɪ t ɪ ŋ	0.49	w eɪ t ɪ ŋ
	0.38	w eɪ d ɪ ŋ
	0.13	w eɪ t ɪ n

This example shows how in American speech the /t/ is often pronounced as a /d/, an effect known as *tapping*, and gives rise to certain homophones, such as *bitter* & *bidder* and *waiting* & *wading*. Other accent features were also found to be modelled by the adaptation scheme, a few of which are listed below:

1. The longer /iː/ in the *word final* position of words such as *city* and *coffee* is often shortened in an American accent to /ɪ/, as is demonstrated by the adapted dictionary:

word	BEEP pronunciation	top Am-adapted pronunciations
CITY	s ɪ t iː	s ɪ t iː
		s ɪ t ɪ
		s ɪ d iː
		s ɪ d ɪ

2. Consider the first vowel in the word *rather*. It is reported [11, 2] that whilst some American people say /r æ ð ə/ (which rhymes with *gather*), some say /r ɑː ð ə/ (which rhymes with *father*) and others will use an open, central vowel /a/ which is different again. This is what the American model predicts:

[1] Phonetic symbols used here are those of the International Phonetic Alphabet, produced by the International Phonetic Association. Shading has been used to highlight observed effects.

<table>
<tr><th>word</th><th>BEEP
pronunciation</th><th>top Am-adapted
pronunciations</th></tr>
<tr><td>RATHER</td><td>r ɑː ðə</td><td>r ɪ ð[ə | ɪ]
r ə ð[ə | ɪ]</td></tr>
</table>

In this case the model is not in perfect agreement with the linguistic literature in that not all three variants are predicted. However, the /ə/ vowel is very central (and the /ɪ/ vowel very frontal) thus reflecting a shift away from the back vowel /ɑː /.

3. The distinction between the words *cot* and *caught* is not maintained by all American accents [11, 2]. Both words can be pronounced /k ɑ t/, i.e. the vowel in *caught* is brought forward. Likewise, the vowel in *cot* can be less frontal. These are both evident in the American model predictions:

<table>
<tr><th>word</th><th>BEEP
pronunciation</th><th>top Am-adapted
pronunciations</th></tr>
<tr><td>COT</td><td>k ɒ t</td><td>k ɑː [t | d]
k ɒ [t | d]</td></tr>
<tr><td>CAUGHT</td><td>k ɔː t</td><td>k ɔː [t | d]
k aɪ [t | d]
k ɒ [t | d]</td></tr>
</table>

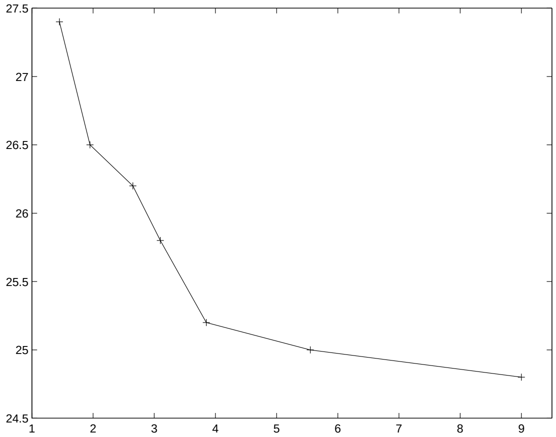
Results

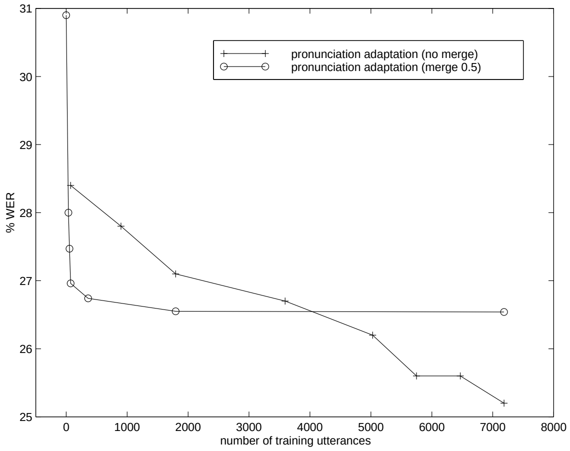
Recognition tests were performed for s0_et data using the system-native BEEP PD (baseline) and then the pronunciation adapted PDs. Table 1 shows how word error rate (WER) is reduced by 20% from the baseline WER of 30.9% when a pronunciation tree trained on all 500k phone replacement observations is used in place of the system-native BEEP dictionary. The graph of Figure 1 shows the effect on WER of varying the number of pronunciations available per word[2].

dictionary	WER (%)
baseline (BEEP)	30.9
SI adapted	24.8

Table 1: Effect of an accent specific, SI PD on recognition performance of American English speech using a British English recogniser (average of 3.9 pronunciations per word).

The amount of data required to produce a good SI dictionary was investigated. Models were generated using different amounts of training data from which PDs were built and used to rescore the baseline recognition lattices. Figure 2 shows the resulting WERs. This graph demonstrates that WER decreases roughly linearly with an increase in the training data available here. It also shows how when there is less data available, merging the new pronunciations with the original dictionary achieves a lower WER. A 50–50 merge ratio is used here, i.e. the pronunciation probabilities of each word are rescaled to sum to 0.5 and the original pronunciation re-inserted with a probability of 0.5. The graph shows that if 4000 utterances or more are available then the pronunciation tree is sufficiently reliable that no merging with the original PD is required.





_ad) were built for each speaker. The CART tree-growing algorithm was used. The 40 utterances for each speaker resulted in some 2600 pronunciation rules for training a set of SD pronunciation trees. As confirmed by Figure 2, a better PD is produced by merging the original BEEP dictionary with the new pronunciations. This resulted in an 8% reduction in WER (see Table 2) over the baseline. Performance was further increased by the use of confidence measures. It was found that discarding phone recogniser outputs which have a low confidence score, as described in Section 2.2., resulted in better trees yielding a further reduction in WER.

Table 2 shows the 13% reduction in WER achieved when the bottom 30% of the data (with respect to the confidence measure) was discarded.

[2] These results were obtained by rescoring lattices resulting from the baseline experiment. Hence the WER shown is slightly greater than that which would be produced by full decoding.

dictionary	WER (%)
baseline (BEEP)	30.9
SD adapted	28.5
SD adapted + confidence threshold	26.8

Table 2: Effect of SD adapted PDs on recognition performance of American English speech using a British English recogniser.

3.4. MLLR acoustic adaptation

The original premise of this work was that differences between individual speakers may be measured in terms of acoustic as well as phonological differences. Thus, a combination of the pronunciation adaptation method described in this paper with a popular method of acoustic adaption known as MLLR (maximum likelihood linear regression), details of which may be found in [8], was investigated.

As a baseline, MLLR adaptation alone was performed in supervised batch mode, using the 40 adaptation sentences per speaker from the s0 adaptation set. Results are shown in the first column of Table 3, where it can be seen that MLLR reduced WER by 31%.

	WER (%)		
	BEEP	Am adapted dict	
	dict	SI	SD + confidence
no MLLR	30.9	24.8	26.8
with MLLR	21.3	18.6	19.8 (scheme I) 19.1 (scheme II)

Table 3: Effect of acoustic and pronunciation adaptation on recognition performance of American English speech using a British English recogniser

MLLR was then combined with the PD adaptation algorithm described in this paper. The first method for their combination, denoted *scheme I*, was used with both SI and SD adapted PDs and involved performing pronunciation adaptation prior to MLLR adaptation. A second method for combining MLLR with pronunciation modelling was also investigated (*scheme II*). This differs from scheme I in that the MLLR adaptation was performed first.

Table 3 shows that combining the two adaptation techniques results in a WER that is lower than that obtained by one technique alone. The SD results demonstrate that performing MLLR adaptation prior to pronunciation adaptation (scheme II) results in a lower WER than that achieved by scheme I. Scheme II is not applicable to the SI case where dictionary generation is an off-line process and not dependent on the speaker whose speech is currently being recognised. Note that confidence level thresholding was used for only the SD task.

4. CONCLUSIONS

This paper has extended previous work on pronunciation adaptation and has shown its successful application to the task of large vocabulary continuous speech recognition. Its use for accent specific PD generation as well as speaker dependent adaptation has been highlighted, showing WER reductions of up to 20%. A framework for combined pronunciation and acoustic adaptation has been presented, demonstrating how a 10–13% WER reduction may still be achieved on top of the 30% reduction afforded by MLLR adaptation.

Further work will investigate the inclusion of word and syllable boundary information in the pronunciation modelling process. It is also anticipated that pronunciation modelling may be used to produce PDs tailored to different speech styles, for example spontaneous speech.

ACKNOWLEDGEMENTS

J.J. Humphries is funded by an EPSRC CASE award studentship in association with The Hirst Division of GEC-Marconi Research. Thanks are due to Kamran Kordi of The Hirst Division.

5. REFERENCES

1. L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Inc., 1984.
2. Arthur J. Bronstein. *The Pronunciation of American English*. Appleton-Century-Crofts, 1960.
3. Stephen Cox and Richard Rose. Confidence measures for the Switchboard database. In *Proc ICASSP*, pages 511–514. IEEE, 1996.
4. J. Fransen, D. Pye, A.J. Robinson, P.C. Woodland, and S.J. Young. WSJCAM0 corpus and recording description. Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department, Trumpington Street, Cambridge, England, October 1994.
5. Saul B. Gelfand, C.S. Ravishankar, and Edward J. Delp. An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):163–174, February 1991.
6. Larry Gillick, Yoshiko Ito, and Jonathan Young. A probabilistic approach to confidence estimation and evaluation. In *Proc ICASSP*, pages 879–882, 1997.
7. J.J. Humphries, P.C. Woodland, and D. Pearce. Using accent-specific pronunciation modelling for robust speech recognition. In *Proc ICSLP*, October 1996.
8. C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, pages 171–185, April 1995.
9. Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc ICSLP*, volume 2, pages 899–902, 1992.
10. Ze'ev Rivlin, Michael Cohen, Victor Abrash, and Thomas Chung. A phone-dependent confidence measure for utterance rejection. In *Proc ICASSP*, pages 515–517. IEEE, 1996.
11. J. C. Wells. *Accents of English*. Cambridge University Press, 1982.
12. S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *ARPA Workshop on Human Language Technology*, March 1994. Merrill Lynch Conference Centre.