THE USE OF ACCENT-SPECIFIC PRONUNCIATION DICTIONARIES IN ACOUSTIC MODEL TRAINING

J.J. Humphries and P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK
e-mail: {jjh11,pcw}@eng.cam.ac.uk

ABSTRACT

Speech recognition systems are increasingly being built to cover an ever wider range of speaker accents. However, electronically available pronunciation dictionaries (PDs) specific to these accents often do not exist and would be time consuming and expensive to build by hand. This paper explores the use of pronunciation modelling for the synthesis of accent-specific PDs directly from acoustic data, and their use in acoustic model training. It is shown that this is particularly effective when the amount of acoustic data from the new accent region is insufficient to build a new recogniser, and it is necessary to retrain an existing system: a further 15% reduction in word error rate can be achieved over and above the 20% reduction resulting from acoustic model retraining alone. This paper also presents an empirical evaluation of an American English PD which has been synthesised from a British English PD.

1. INTRODUCTION

Most automatic speech recognition (ASR) systems available today are tailored to suit a rather narrowly defined group of speakers, for example those of British or American English. The diversity of accents of English which can be heard around the world is great [12] and constantly evolving. As the number of applications that use ASR technology increases, so the range of speakers that use such technology rises, and this poses a serious problem to speech recognisers.

Whilst acoustic (e.g. [9]) or phonological (e.g. [4, 3]) adaptation can help a recogniser to cope with such variations, it is sometimes more desirable to build new recognisers for specific accent regions, either to be used independently, as part of a multi-model (e.g. [1]) or in conjunction with an accent-switch (e.g. [7]). In order to build the acoustic models of a new recogniser, acoustic training data is required, along with corresponding phonetic transcriptions. These transcriptions are time consuming and difficult to produce by hand and so a common approach is to make word level transcriptions available instead. These can be used in conjunction with a pronunciation dictionary (PD) to produce transcriptions at a phonetic level. For some accents of English there exist well established, electronically available, sources of PDs, such as British or American. For other accent groups, e.g. Indian, Australian or Hispanic speakers of English, this may not be the case.

In [3], we presented a system to automatically model accent variation and showed how this information could be used to produce new PDs tailored to new accents, at either a *speaker dependent* (SD) or *speaker independent* (SI) level. There the emphasis was on speaker adaptation but this paper extends that work and in particular focuses on the use of SI accent-specific PDs for building SI accent-specific recognisers.

After a brief summary of the pronunciation modelling technique, this paper looks in detail at an SI PD for American accented speakers of English which has been automatically generated from a British PD via a British recogniser. This synthesised PD is analysed empirically, producing results that extend and compliment the qualitative, linguistic analysis presented in [3].

The remainder of this paper then looks at the effect of using such a synthesised PD in training a new set of acoustic models so as to create an accent-specific recogniser. Two situations are considered: the first whereby there are large amounts of acoustic data available for the new accent region, and yet there is no accent-specific PD. Secondly the case whereby the amount of new acoustic data is small, compounded by the lack of an accent-specific PD.

2. PRONUNCIATION MODELLING

The pronunciation modelling scheme used here was described fully in [3] but is summarised by the overview shown in Figure 1. Speech from the new accent region is transcribed using a phone-loop recogniser for the reference accent region and is compared to canonical transcriptions of the new data derived from the reference accent PD. This comparison, performed using dynamic string alignment [6], results in a series of context dependent *pronunciation variation observations* (PVOs) which are clustered using decision trees, one for each base phone. These trees are grown by making splits based on phonetic feature measurements of the PVOs and each leaf contains a set of phone substitutions, deletions and insertions, each with associated probabilities which are estimated from the tree training data. The trees then provide the necessary information for the creation, from the reference PD, of a new PD that is a better representation of the pronunciations observed in speakers from the new accent region. Each word in the new PD may contain multiple pronunciations, each of which carries a probability which can be utilised by a suitable recogniser. Typically an average of 4 pronunciations per word is found to be effective.

3. EXPERIMENTAL SET-UP

Experiments presented in this paper use two HTK-based [13] large vocabulary continuous speech recognition systems:

1. A British English recogniser, trained from the WSJCAM0 [2] SI training set (92 speakers over a total of 7861 utterances). A subset of the BEEP PD was used (this was produced at Cambridge University Engineering Department and provides British accented pronunciations of English).
2. An American English recogniser, trained from the WSJ0 [11] SI-84 subset (7185 utterances). The PD was kindly pro-

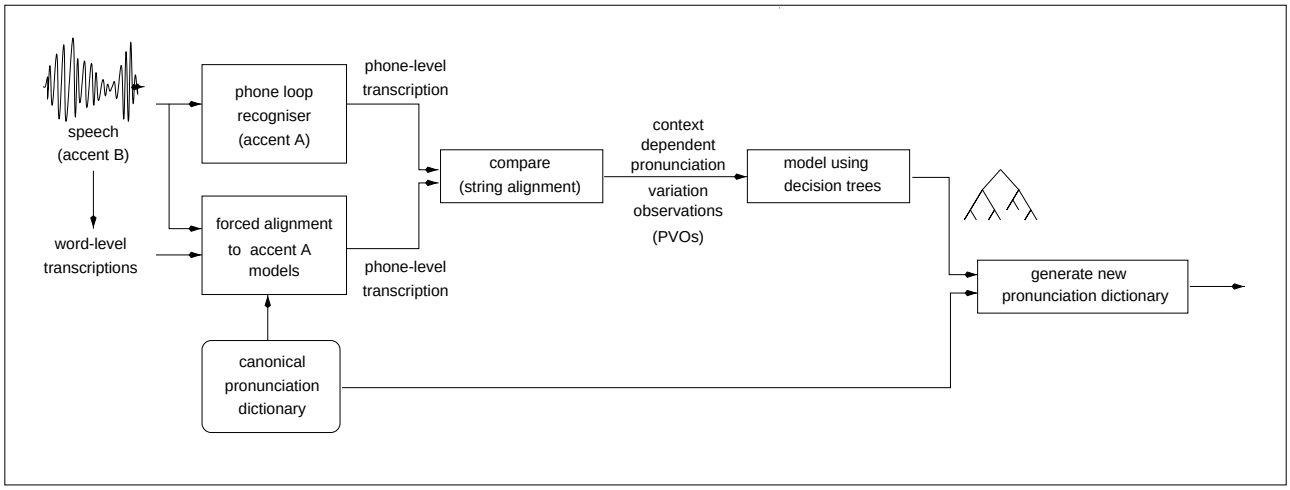

Figure 1: Overview of the pronunciation modelling process.

vided by LIMSI (their 1993 version) and contains American English pronunciations.

Both recognisers used state-clustered eight-mixture cross-word triphone HMMs [10] and a bigram language model. All speech data was parameterised as 12-dimensional MFCCs with energy, delta and acceleration coefficients.

4. EMPIRICAL EVALUATION OF A SYNTHESISED AMERICAN DICTIONARY

The pronunciation modelling technique was used to build a phonological model of the American accent. This was performed with respect to the British recogniser described above, using the American speech data of WSJ0 (see above) in conjunction with corresponding word-level transcriptions. The pronunciation model produced describes a mapping from British pronunciations to those more suited to the American accent. Using this model, an American PD was synthesised from a 5k-subset of the BEEP dictionary. This new PD contained multiple pronunciations for each word, which were weighted by probabilities taken from the model. By adjusting a probability threshold, the average number of pronunciations per word was varied.

The resulting PD was compared to the 'source' BEEP PD and an American PD (LIMSI), which may be regarded as a 'target' PD. To compare entries in PDs D_1 and D_2 the algorithm shown in Figure 2 was used.

The counts for each word were scaled by the word's unigram probability obtained from the language model. This was done to reflect the fact that a pronunciation difference in a frequently used word is likely to have a more significant impact on recogniser performance. Note that since the LIMSI and BEEP PDs are based on slightly different phone sets, it was necessary to map them both onto a common set before a direct comparison could be made.

Table 1 shows the values measured for a three-way comparison between BEEP, LIMSI and the synthesised American PD. In terms of phone substitutions, the synthesised American PD is closer to the 'target' LIMSI PD than the BEEP PD from which it was derived

for each word w_i in the 5k wordlist
{
for each pronunciation $p^a_{(w_i,D_1)}$ of w_i in D_1
{
use dynamic string alignment to find closest pronunciation $p^b_{(w_i,D_2)}$ of w_i in D_2.
record the number of phone *substitutions*, *deletions* and *insertions* between them, scaled by the pronunciation probability of $p^a_{(w_i,D_1)}$.
}
}
calculate average number of *substitutions*, *deletions* and *insertions* per pronunciation per word.

Figure 2: Algorithm to compare two PDs.

(0.033 substitutions per phone rather than 0.164). The LIMSI PD contains fewer phones than the BEEP PD, requiring 0.183 insertions per phone to map to BEEP. This is reduced slightly to 0.147 in the synthesised PD, although this measure was also found to be sensitive to the phone-set mapping used.

This result in conjunction with the linguistic analysis (presented in part in [3]) shows that the pronunciation modelling process is correctly capturing many of the accent effects.

5. BUILDING AN ACCENT-SPECIFIC RECOGNISER

This section investigates the effect of using a synthesised accent-biased PD in the building of an accent-specific speech recogniser. For example, consider the following situation: there exists a large corpus of accented speech data (in this case American) from which a set of context dependent phone models is to be trained. This

PDs compared	average per phone mapping		
	dels.	subs.	ins.
synth. AM w.r.t. LIMSI	0.006	0.033	0.147
synth. AM w.r.t. BEEP	0.001	0.164	0.000
BEEP w.r.t. LIMSI	0.005	0.198	0.183

Table 1: An empirical comparison of a synthesised SI American English PD (1.5 pronunciations per word) to the source British English pronunciations (from the BEEP PD) and to an American PD (LIMSI). The comparison between the LIMSI and BEEP dictionaries is also shown (phone-set mappings were performed where necessary).

speech data is complete with word level transcriptions, but, there is *no* American PD available. However, there does exists a British English recogniser, complete with acoustic models and PD.

Under these conditions, the two training scenarios investigated here are:

Scenario 1 Train the American acoustic models using the American speech based on phonetic transcriptions derived from the existing British PD.

Scenario 2 Adapt the British PD using the pronunciation modelling scheme described above, with all the available American acoustic data, to give a synthesised-American PD. Use this new PD to derive phonetic transcriptions of the American speech data prior to acoustic model training. Keep the new PD for recognition tasks.

The American training data and the British recogniser were the same as those described in the previous section. The synthesised American PD was constrained to contain an average of about 7 pronunciations per word, with associated probabilities. In scenario 2 the American training data was force aligned to these multiple pronunciations (given word level transcriptions) prior to model training.

In both cases, 8-mixture, 3-emitting state Gaussian triphone models were trained, based on the BEEP phone set, using Baum-Welch re-estimation [5] of the HMM parameters. Since there were insufficient examples of some triphones for their parameters to be reliably estimated, the triphones were state-clustered using decision trees [10] according to the training data available. The two resulting speech recognition systems were then evaluated using the 425 Nov'94-S0 WSJ [11] (American accented) test utterances in conjunction with a 5k-bigram language model.

5.1. Results

The word error rate (WER) results are shown in Table 2. Whilst scenario 2 does show a 6% WER reduction, this is in fact not statistically significant. The 14.1% WER observed is still some way off the 11.6% achieved by a comparable model set trained on an American phone set with the use of the LIMSI American PD. It is known that this LIMSI PD is particularly good [8], perhaps explaining how despite adaptation of the BEEP PD, the target recogniser is difficult to beat.

5.2. Conclusions

These results show that if a PD specific to the accent of the recogniser being created is unavailable then one may be adapted from an existing PD which may have been designed for use with speech of a different accent. In doing so a small improvement in recogniser performance may be observed in comparison to using what may be regarded as an inappropriate PD (i.e. tailored to another accent).

training conditions	test PD	% WER
scenario 1	BEEP	14.94
scenario 2	synth AM	14.06
'target'	LIMSI	11.60

Table 2: Recognition results using American acoustic models trained using two PDs. Scenario 1 uses a British PD (BEEP) whereas scenario 2 uses an American PD derived automatically from the British PD ('synth AM'). The same PD used during training is used for these tests. Results using a 'target' system, i.e. one trained using an established American PD (LIMSI), are also shown for comparison.

6. BOOTSTRAPPING FROM AN EXISTING RECOGNISER

This section investigates a solution to the following problem: we have a recogniser for a particular accent region (e.g. British English) and a limited amount of sample speech from a new accent region (e.g. American English) but we know nothing about the pronunciations of this new accent. Given, therefore, that we have insufficient training data and linguistic knowledge about the new accent region to train a recogniser from scratch, is it possible to adapt the existing recogniser both acoustically and phonologically so that it performs better on the new accent?

It was shown in [3] how as few as 500 utterances are required to build a reasonable pronunciation model for a different accent. Hence, 50 utterances from each of 10 American speakers from the WSJ0 training corpus were used in the following three adaptation schemes to produce a new SI recogniser by bootstrapping from the British speech recogniser described in Section 3.

Pronunciation modelling. A pronunciation model was built using the 500 utterances with the British recogniser and PD. From this a new SI-PD was constructed. No acoustic adaptation was performed;

SI-MLLR [9] acoustic adaptation. Performed using all 500 utterances with segmentation based on BEEP pronunciations derived by forced alignment from word level transcriptions of the American speech data. No pronunciation adaptation was performed;

Combined pronunciation and acoustic adaptation. Firstly the SI-PD (from above) was used for segmentation of the 500 utterances prior to SI-MLLR adaptation. The new PD was retained during testing.

6.1. Results

All three systems were then evaluated over the 425 test utterances of the Nov'94-S0 WSJ evaluation in conjunction with a 5k-bigram

language model. Note that the 20 speakers in this evaluation set are different to those speaking the 500 adaptation sentences. Table 3 shows the WER for these three systems. This evaluation was also performed using a fully trained (on all WSJ0) American English recogniser for 'target' comparison.

system	acoustic models	PD	% WER
baseline	British	BEEP	30.9
SI-PD adapted	British	SI-adapted	25.8
SI-MLLR adapted	SI-adapted	BEEP	24.7
SI-MLLR, SI-PD	SI-adapted	SI-adapted	21.2
'target'	American	LIMSI	11.6

Table 3: Average WER measured for 20 American speakers on recognition systems adapted from a British recogniser using 50 utterances from each of 10 different American speakers using combinations of acoustic and phonological adaptation. Results are also shown for the baseline British English recogniser and the target American recogniser.

6.2. Conclusions

These results show that whilst SI-acoustic and SI-pronunciation adaptation schemes independently reduce the WER by 20% and 17% respectively, their combined effect produces a 31% WER reduction. It has therefore been demonstrated how acoustic and phonological models may be derived for one accent group from the existing recogniser of another accent group. Indeed, this may be achieved using as few as 500 utterances from the new accent group, which alone would be insufficient to build a continuous speech, large vocabulary recogniser from scratch. The improved WER of 21.2% is still some way off from that achieved by the 'target' American recognition system which was trained based on the LIMSI PD. Again this may be attributed to the high quality of that dictionary [8].

7. DISCUSSION AND CONCLUSIONS

The results of the previous section have shown how pronunciation modelling can be of significant benefit when an accent-specific recogniser is desired, but the amount of acoustic training data is limited, and phonological information non-existent. However, Section 5 demonstrated that when a large amount of acoustic data is available, extra phonological information is of less value. This may perhaps be because when given enough acoustic data, pronunciation variability may be absorbed to a larger degree within the acoustic models (particularly when these are multi mixture Gaussians, as was the case in these experiments). This suggests that pronunciation modelling is therefore more useful in cases where limited training data is available and it is necessary to utilise this data as effectively as possible.

Producing accent-specific PDs by hand is not only time consuming but often a subjective process. Creating such PDs automatically via pronunciation models obtained directly from acoustic data is not only fast and convenient, but has also been shown, both empirically and linguistically, to give appropriate pronunciations. Results presented here show that this process can be a valuable tool to assist the development of accent-specific speech recognition systems, particularly when used in conjunction with acoustic adaptation/re-estimation. As speech recognition technology improves, so it will be used in a wider range of more varied situations. As demands on the technology increase, so the need to cope with variations such as those caused by accent will become essential. This work brings the technology closer to this goal.

8. ACKNOWLEDGEMENTS

J.J. Humphries is funded by an EPSRC CASE award studentship in association with The Hirst Division of GEC-Marconi Research.

9. REFERENCES

[1] V. Beattie, S. Edmonson, D. Miller, Y. Patel, and G. Talvola. An integrated multi-dialect speech recognition system with optional speaker adaptation. In *Proc. Eurospeech*, pages 1123–1126, Madrid, September 1995.

[2] 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.

[3] J.J. Humphries and P.C. Woodland. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In *Proc. Eurospeech*, 1997.

[4] J.J. Humphries, P.C. Woodland, and D. Pearce. Using accent-specific pronunciation modelling for robust speech recognition. In *Proc. ICSLP*, October 1996.

[5] K. Knill and S. Young. Hidden Markov models in speech and language processing. In S. Young and G. Bloothooft, editors, *Corpus-Based Methods in Language and Speech Processing*, chapter 2, pages 27–68. Kluwer Academic Publishers, 1997.

[6] Joseph B. Kruskal. An overview of sequence comparison. In David Sankoff and Joseph B Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, chapter 1. Addison-Wesley, 1983.

[7] Karsten Kumpf and Robin W. King. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Proc. Eurospeech*, 1997.

[8] Lori Lamel and Gilles Adda. On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *Proc. ICSLP*, volume 1, 1996.

[9] 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.

[10] J.J. Odell, P.C. Woodland, and S.J.Young. Tree-based state clustering for large vocabulary speech recognition. In *Proc. International Symposium on Speech, Image and Neural Networks*, pages 690–693. IEEE, April 1994. Hong Kong.

[11] 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.

[12] J. C. Wells. *Accents of English*. Cambridge University Press, 1982.

[13] S.J. Young, Joop Jansen, Julian Odell, Dave Ollason, and Phil Woodland. *The HTK Book (for HTK Version 2.0)*. Entropic Cambridge Research Laboratory, 1995-6.