USING ACCENT-SPECIFIC PRONUNCIATION MODELLING FOR ROBUST SPEECH RECOGNITION

J.J. Humphries†, P.C. Woodland† & D. Pearce‡

† Cambridge University Engineering Department, Trumpington Street, Cambridge, UK
‡ The Hirst Division of GEC-Marconi Materials Technology, Borehamwood, 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 a decision tree which is used as a model of the pronunciation variation associated with this new accent group. The tree is then used to build a new pronunciation dictionary for use during the recognition process. Experiments are presented for the recognition of Lancashire & Yorkshire accented speech using a recognizer trained on London & South East England speakers. The results show that the addition of accent-specific pronunciations can reduce the error rate by almost 20% for cross accent recognition. It is also shown that worthwhile gains in performance can be obtained using only a small amount of accent-specific data.

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 by using speech data from a large set of speakers. There are two principal differences which exist between speakers: *acoustic* differences which are 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 get full coverage of all the regional accents (in England alone there are at least ten broad regional accents [5]). SI speech recognition systems do not perform as well as *speaker dependent* (SD) systems, largely because of the need to model 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 [4]. Such techniques usually make the assumption that all speakers are pronouncing words in a predefined manner, as described in the *pronunciation dictionary* (PD). The work presented here suggests that this is a poor assumption, and explores a method for modelling accent-specific pronunciations and demonstrates how this information may be used to improve SI recognition performance.

The technique used involves retranscribing at the phone level some accent specific data. The preferred transcription for each word is then compared to its PD-entry and a list of context-dependent phone replacement rules is generated. This information is clustered in a decision tree which is then used to generate accent-specific PDs. Results are presented for a set of HMMs trained on London & South East England speakers, with adapted PDs for the recognition of Lancashire & Yorkshire accented speakers.

2. PRONUNCIATION MODELLING TECHNIQUE

The first stage in the modelling process is to obtain accurate phone level transcriptions of the accent specific data in terms of the phone set of the SI recognizer (word level transcriptions are assumed to be already present.) For the purpose of these experiments the assumption is made that accent variation is most evident in vowels rather than consonants [5]. Using this assumption, a recognition network is built around the original PD-entry allowing for up to two vowel substitutions from the original definition at any one time. For example, Figure 1(a) illustrates the PD definition[1] for the word "Aberdeen" spoken with an RP[2] accent.

For each word re-transcribed, the three-best transcriptions are ranked. Comparing these transcriptions to the PD-entries enables a list of context dependent vowel substitutions to be generated. These are rules of the form $s - v + d \rightarrow u$ (s and d are respectively the left and right contexts of a vowel, v, which is replaced by another vowel, u) and form the data used for the pronunciation modelling described next.

The data needs to be processed in two ways: pronunciation effects should be differentiated from recognizer errors and the information should be generalized over all contexts. One method of achieving this is by using a decision tree [1] clustering technique. Vowel substitution rules are clustered using contextual phonetic features and the frequency of each type of substitution.

[1] The symbols used here are those of the *International Phonetic Alphabet* (IPA) [2].

[2] *Received Pronunciation* (RP) is also known as *Standard Southern English.*

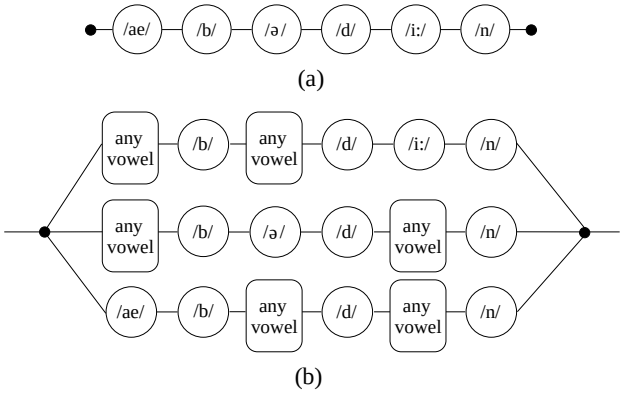

Figure 1: Pronunciation networks for the word "Aberdeen": (a) shows the standard RP definition whilst (b) shows the network used for re-transcription with up to two vowel changes at any one time.

2.1. Maximizing Tree Purity

The initial stage in tree building is to pool all the training data into the root node. Binary splits of the data are made by asking questions relating to phonetic features of s, v, or d. The *purity*[3] [1], $g(l)$, of a leaf, l, of data may be calculated in terms of the probability, $p(j|l)$, of the leaf containing rules involving the substitution with each vowel, j, i.e.

$$g(l) = \sum_{\forall j \in \text{vowel set}} p^2(j|l) \tag{1}$$

where

$$p(j|l) = \frac{1}{I_l} \sum_{i=1}^{I_l} \beta(s_i - v_i + d_i \rightarrow u_i), \tag{2}$$

$$\beta(s_i - v_i + d_i \rightarrow u_i) = \begin{cases} 1 & u_i = j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and I_l is the total number of data items in leaf l. Given a tree of L leaves, the average purity, $\overline{g}$, may be calculated:

$$\overline{g} = \frac{1}{L} \sum_{l=1}^{L} g(l). \tag{4}$$

Binary splits are performed so as to maximize the purity, g, of the new leaves of the tree. This operation is applied recursively until the *stopping criterion* (discussed below) is met. The result is a binary tree whose leaves contain the original training data items, each leaf representing a cluster of similar data.

2.2. Cross-validation

Cross-validation experiments [1] are performed to determine at which point the growth of the tree should be terminated. Given a decision tree whose leaves comprise sets of items of the form $s_i - v_i + d_i \rightarrow u_i$ (where s_i and d_i are the phone context; $v_i \in V_l$, $u_i \in U_l$ for a leaf l where V_l and U_l are the sets of all possible substituted and substituting vowels) and a set of N test rules of the form $s_n - v_n + d_n \rightarrow u_n$ then the hit rate, h, is defined as:

$$h = \frac{1}{N} \sum_{n=1}^{N} \theta\left(s_n - v_n + d_n \rightarrow u_n\right). \tag{5}$$

If l_n is the leaf reached in the tree for an item $s_n - v_n + d_n$ then

$$\theta(s_n - v_n + d_n \rightarrow u_n) = \begin{cases} 1 & u_n \in U_{l_n} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

As more leaves are grown so the purity of the tree, $\overline{g}$, increases, but at the same time the hit rate, h, decreases. A suitable compromise is found by maximizing the function ϕ with respect to M (minimum leaf size):

$$\phi = \left(h_M - h_{M_{min}}\right) \overline{g}_M. \tag{7}$$

The tree splitting process therefore involves continually splitting nodes until any further splits would result in a leaf containing fewer than M items—this is the stopping criterion. $h_{M_{min}}$ is the worst hit rate achievable, namely that when node splitting is performed to the point where each leaf contains unique item types. By plotting ϕ as a function of M, the optimum minimum leaf size M_{opt} may be found, as demonstrated in Figure 2.

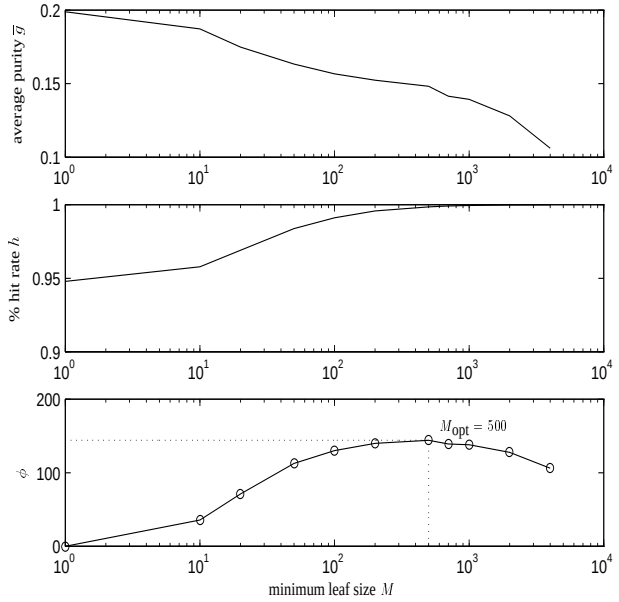

Figure 2: The top graphs show leaf purity, $\overline{g}$, and hit rate, h, whilst the bottom graph shows the function ϕ (see Equation 7).

Once such a tree is built, the contents of each leaf l may be reduced to a set of S_l substitution alternatives each of the form (u_s^l, p_s^l) where u_s^l is the substituting vowel and p_s^l is its probability within that leaf.

3. BUILDING AN ACCENT SPECIFIC DICTIONARY

The tree built using the technique described above enables a list of vowel substitutions for a specific vowel within a given context to be

[3] A measure based on the *Gini* index is used here.

generated. It has previously been demonstrated that using too many pronunciation variations for each word offers little, if any, reduction in word error rate, e.g. [3]. As the trees used here stand, each leaf can contain up to around 15 valid substitutions. This increases the search space of the recognizer for a word containing more than one vowel and also introduces confusions. For this reason, a leaf probability threshold λ is set such that only vowel substitutions of the form (u_s^l, p_s^l) where $p_s^l > \lambda$ survive. Experimentation with values of λ is necessary to give an appropriate number of pronunciation alternatives for each word in the PD (typically around three to four pronunciations per word.) As an example, the PD definition of the word *Wednesday* is /w e n z d eɪ/. Figure 3(a) shows the set of possible substitutions suggested by a typical pronunciation tree for the triphone /w/-/e/+/n/ whilst Figure 3(b) shows the alternatives for the triphone /d/-/eɪ/+sil. Applying a threshold of $\lambda = 0.10$ leaves just those alternatives shown in bold. Pronunciation variations of the word *Wednesday* may then be generated as shown in Figure 3(c).

vowel v	$p(v)$
e	**0.43**
æ	**0.11**
ɪ	0.08
ə	0.07
aʊ	0.07
ʌ	0.06
eə	0.05
⋮	⋮

(a)

vowel v	$p(v)$
eɪ	**0.43**
ɪː	**0.15**
ɪ	0.09
eə	0.08
e	0.07
aɪ	0.05
ɪə	0.04
⋮	⋮

(b)

/	w	e	n	z	d	eɪ	/
/	w	e	n	z	d	ɪː	/
/	w	æ	n	z	d	eɪ	/
/	w	æ	n	z	d	ɪː	/

(c)

Figure 3: (a) The list of possible substitutions for the triphone /w/-/e/+/n/ and the associated probabilities. (b) The possible substitutions for the triphone /d/-/eɪ/+sil. (c) The variations in pronunciation of the word *Wednesday* which result from using the substitutions in (a) and (b) and a threshold of $\lambda = 0.10$.

4. EXPERIMENTS AND RESULTS

4.1. Setup

The database used (provided by GEC) is part of a telephone-based corpus of isolated words from a 2000 word vocabulary. The principal advantage of using this database is that it contains speech divided into two of England's largest accent regions as follows:

1. *London and South East* (LSE)
   - male training data: 100 speakers / total of 8540 utterances
   - male test data: 14 speakers / total of 1041 utterances
2. *Lancashire and Yorkshire* (LY)
   - male training data: 93 speakers / total of 7845 utterances
   - male test data: 17 speakers / total of 1479 utterances

Each frame of speech was represented by 12 MFCCs, normalized energy, along with first and second time derivatives. A set of SI models was trained for each accent region. Each model set comprised decision tree state-clustered triphone models with 8 mixture components per state [7].

The LSE models were used for re-transcribing the LY training data (or portions thereof) for training and cross-validation of the pronunciation variation trees. In this respect the LY training data may be considered to be the accent-specific data. LSE and LY models were used in the final recognition experiments using the LY test data both with and without accent-specific pronunciations.

The PD used was also provided by GEC and based on popular British pronunciations from the Longman Pronunciation Dictionary [6]. It contains a single pronunciation for each word.

test accent region	model accent region	
	LSE	LY
LSE	7.9	13.1
LY	17.2	9.2

Table 1: Baseline recognition results (% word error rate.)

4.2. Baseline Results

To evaluate the performance of this accent-pronunciation modelling technique, a set of baseline experiments were conducted. These involved recognition tests of both the LSE and LY speech using each of the LSE and LY model sets in turn. The single pronunciation per word dictionary used during training was used again here. As the results show (see Table 1), recognition performance degrades to give around double the error rate when an accent unobserved during training is used during testing.

4.3. Results Using Accent Modelling

As an initial experiment, all of the LY training data was used to build a pronunciation tree. This data was phonetically re-transcribed using the LSE models as described in Section 2. With three-best output, some 42000 vowel substitution rules were produced, of which 48% were self-substitutions (i.e. no change from the original PD definition). The cross-validation experiment involved using 90% of this data for tree building and the remaining 10% for testing and suggested an optimum value for M, the minimum leaf size used as the tree growth stopping criterion, of 500. A final pronunciation tree was then grown using all of the data. For various values of the substitution probability threshold, λ, multiple pronunciation dictionaries were generated. These multiple-PDs were then used in conjunction with the LSE model set for recognition of the LY test data and the results are shown in Table 2.

λ	average pronunciations per word in dictionary	WER %
0.10	5.4	14.7
0.12	3.9	**13.9**
0.14	3.3	14.5
0.18	2.0	14.6
0.22	1.5	16.3
no tree	1.0	**17.2**

Table 2: Word error rate (WER) for recognizing LY data using LSE models aided by an a accent-specific PD, shown as a function of the tree threshold λ.

The results indicate that whilst too many pronunciation variations introduce confusability which increases WER, with sufficient pronunciation variations the WER decreases by almost 20% over that obtained for the fixed PD. Note that the substitution threshold λ controls the number of pronunciations per word in the multiple-PD. It can also be seen (by comparison to Table 1) that the use of an accent-specfic multiple PD has reduced the difference between the use of accent-specific and cross-accent acoustic models by more than 40%.

4.4. Effect of Training Corpus Size

The above experiment used all of the LY training data for the purpose of pronunciation tree building. In practice this is rather a lot of data (enough to build a new model set) and so the effect of reducing the amount of tree training data was investigated, the results of which are shown in Figure 4 using 2%, 10% and 100% of the accent-specific data. This graph shows that only a small amount of data (2%) is needed to give useful improvement. The introduction of more pronunciation alternatives is traded off against an increase in confusable words, resulting in the minima which can be seen in Figure 4. Experiments using less than 2% of the training data were not investigated due to the quantity of data needed for the cross-validation process for tree optimization.

5. CONCLUSIONS

The results presented here have demonstrated the effectiveness of pronunciation adaptation in enabling a speech recognition system to be used on an unseen accent group. Further refinements to this system will involve looking at not only vowel substitutions but all phone replacements, insertions and deletions. In this way it may be possible to build trees which not only relect accent variation but also styles of speech. Extension to continuous speech is a requirement for many applications. It is also anticipated that this method could be used in conjunction with acoustic model adaptation methods.

Acknowledgements

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

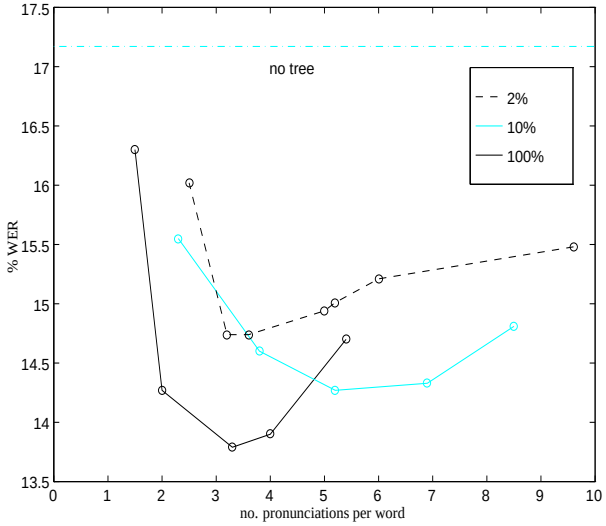

Figure 4: Graph showing word error rate as a function of the average number of pronunciations per word and the percentage of new accent corpus used to train the pronunciation tree.

6. REFERENCES

1. L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Inc., 1984.
2. P. Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich College Publishers, 1993.
3. Kai-Fu Lee. *Automatic Speech Recognition*. Kluwer Academic Publishers, 1989.
4. 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.
5. J. C. Wells. *Accents of English*. Cambridge University Press, 1982.
6. J.C. Wells. *Longman Pronunciation Dictionary*. Longman, 1990.
7. 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.