Pronunciation Variation Modelling using Accent Features

Michael Tjalve[1,2] & Dr. Mark Huckvale[1]

[1]Department of Phonetics and Linguistics
University College London, U.K.
[2]Infinitive Speech Systems, Ltd.
{m.tjalve, m.huckvale}@ucl.ac.uk

10.21437/Interspeech.2005-487

Abstract

In this paper, we propose a novel method for modelling native accented speech. As an alternative to the notion of dialect, we work with the lower level phonological components of accents, which we term *accent features*. This provides us with a better understanding of how pronunciation varies and it allows us to give a much more detailed picture of a person's speech.

The accent features are included during phonological adaptation of a speaker-independent Automatic Speech Recognition system in an attempt to make it more robust when exposed to pronunciation variation thus improving recognition performance on accented speech.

We employ a dynamic set-up in which the system first identifies the phonetic characteristics of the user's speech. It then creates a model of the speaker's phonological system and adapts the pronunciation dictionary to best match his/her speech. Recognition is subsequently carried out using the adapted pronunciation dictionary.

Experiments on British English speech data show a significant relative improvement in error rate of 20% compared with the traditional non-adaptive method.

1. Introduction

Pronunciation variation, the fact that speakers pronounce the same words in different ways, is generally considered to be one of the biggest challenges in Automatic Speech Recognition (ASR) today [1]. Traditionally, there have been two general trends to dealing with pronunciation variation: 1) add alternatives to a global pronunciation dictionary, which is then applied to all speakers and 2) perform speaker adaptation on the acoustic models.

Adding pronunciation variants to the dictionary is known to introduce more substitution errors [see e.g. 1, 2, 3, and 4]. Moreover, the potentially large number of alternative pronunciations for each word is likely to have a negative impact on the computational cost due to the increased search space [1].

The limitation of speaker adaptation of the acoustic models is that it can only deal with acoustic variation due to physiological differences and does not explicitly offer the possibility of dealing with accent variation as such. If the same pronunciations are used for all speakers, the wrong phone models may be adapted during speaker adaptation, which is likely to make recognition performance worse.

An alternative approach, which reduces the risk of confusion between entries and therefore potentially improves recognition performance, is to adapt the pronunciation dictionary to the user. This paper proposes a novel method to performing speaker-dependent pronunciation dictionary adaptation. In the current work, we are thus only interested in the pronunciation variation, which is rooted in accented speech. Although our method works independently of speaker adaptation of the acoustic models, it should be considered as an extension of traditional speaker adaptation.

2. Pronunciation dictionary adaptation

Ideally, the pronunciation dictionary should exclusively contain pronunciations used by the speaker, since we are only recognising one speaker at a time. However, this conflicts with the nature of a speaker-independent ASR system where variation across speakers needs to be covered.

Adapting the recogniser to the speaker allows us to move from speaker-independent towards speaker-dependent speech recognition using the same system. Although the adaptation phase operates within the phonetic domain, we can use it to extract information about the speaker's phonological system directly from the speech signal. This is based on the assumption that there is some consistency in the way people pronounce words.

The starting point of the method proposed in this paper is a dynamic pronunciation dictionary containing multiple pronunciations. During the adaptation phase, the system identifies the phonetic characteristics of the user's speech. It then applies this information in the creation of a new speaker-dependent dictionary, an *idiodictionary*, containing only the pronunciations used by the speaker (see Section 5).

Humphries et al. [5] developed a somewhat similar approach to dealing with pronunciation variation. They automatically generated context-dependent vowel substitution rules, which were used to adapt the pronunciation dictionary to better match the speaker. Their rules have to be made context sensitive with respect to the unmarked pronunciations. This denies the possibility of the influence of orthography (e.g. /r/ before consonant) or of stress (e.g. flapping rule). Our approach benefits from being more flexible, but the accent features have to be assigned by a phonetician.

Bael and King [6] also worked with a dynamic pronunciation dictionary from which variation rules were generated to create accent-specific pronunciation dictionaries. Their accent dictionaries only allow a coarse coverage of accent variation and they obtain approximately the same result using accent-dictionaries compared with using a multiple-pronunciation dictionary.

The experiments reported here were carried out on British English speech data, but apart from the specific accent

features, the method is language-independent and should work equally well on any other language.

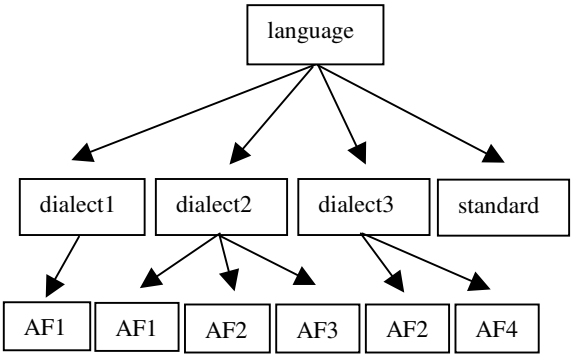
3. Accent features

Research into pronunciation variation in ASR most often focuses on differences between dialects. The speakers' accents are categorised according to their geographical affiliation, e.g. Northumberland accent versus Southern English accent [2, 6]. The term dialect describes the pronunciation of a group of people, but in ASR we are recognising one speaker at a time, not a mixture of speakers. Moreover, many speakers' accents do not belong to a particular identifiable dialect but are rather a mix of dialects.

In this paper, we are exploring the potential benefit of working at a level lower than that of dialects in order to include more detail. Each dialect can be considered as consisting of a number of deviations from the standard pronunciation. We term these phonological components of dialects *accent features*. Any speaker's accent consists of a combination of these features. The accent feature idea is inspired mainly by Wells' [7] description of the pronunciation variation of the various accents of English exemplified by his *standard lexical sets*. The main benefit of using accent features is that it is possible to give a more exact picture of a person's speech.

Figure 1 gives a visual representation of how accent features provide more detailed information about pronunciation variation than the traditional notion of dialect. Note the box labelled 'standard', which refers to the standard unmarked pronunciation. By definition, this contains no accent features.

Figure 1: Pronunciation variation at different levels of detail (AF = accent feature)



In the phonological system of a given speaker, there may be some features from one dialect and other features from another dialect. If, for instance, we consider Figure 1 to be a comprehensive description of the variation in language L and speaker A's phonological system contains AF 3 and AF 4, his/her accent does not correspond to an established dialect, but is rather a mix of dialect 2 and dialect 3. From a phonological point of view, his/her idiolect – and corresponding idiodictionary – equals the standard phonological system with the alterations imposed by accent features 3 and 4.

In the experiments reported here, the following six accent features were used:

- rhoticity, e.g. <four>: /fO:/ → /fO:r/
- closing, e.g. <cup>: /kVp/ → /kUp/
- flapping, e.g. <better>: /bet@/ → /be4@/
- anteriorisation, e.g. <bath>: /bA:T/ → /b{T/
- monophthongisation, e.g. <Wales>: /weIlz/ → /welz/
- h-dropping, e.g. <have>: /h{v/ → /{v/

More features, such as yod-dropping and diphthonging, could be included, but there is a balance between the granularity of the information and the recognition accuracy (see Section 6). For the majority of speakers, a dictionary containing only unmarked pronunciations was chosen, which is to be expected.

4. The data

The experiments reported in this paper were carried out on British English speech data. Two separate data sources were chosen to avoid the training data influencing the test data and three data sets were defined

- Training set (247 speakers, 69,615 utterances, mainly commands and phonetically rich sentences, collected by Dragon Systems)
- Adaptation set (158 speakers, 25 phonetically rich sentences per speaker extracted from the ABI corpus)
- Test set (158 speakers, 100 commands per speaker extracted from the ABI corpus)

The Accents of the British Isles (ABI) corpus [8] is ideal for pronunciation variation research. With its speech data from 14 accent regions from all around the British Isles, it offers a very comprehensive coverage of British English pronunciation variation.

In order to keep the recognition task relatively simple, we built a test grammar, which distinguishes between entire phrases rather than single words. For this reason, the results in this paper are presented as sentence error rates (SER) instead of word error rates.

The pronunciation dictionary used in the experiments reported here is based on the Unisyn dictionary [6] developed at CSTR, University of Edinburgh. The Unisyn dictionary contains a very large number of words and it comes with a set of tools to create pronunciation variants reflecting various accent regions. We chose the following five major accent regions and generated an accent dictionary for each of them.

- RP
- Northern
- Scottish

- Irish
- Welsh

A large phoneme set of 68 phonemes was defined and acoustic models were built using the training set. During recognition, each speaker only makes use of a subset of this phoneme set.

5. Description of experiments

5.1. Baseline

For the baseline experiment, we chose the best accent dictionary overall. As expected, this turned out to be the RP dictionary. This dictionary was then used during recognition on all speakers. With this set-up, we obtained an overall performance of 28.23% SER.

5.2. Accent dictionary experiments

For the accent dictionary experiments, we used the predefined accent dictionaries. We then ran recognition five times on all the test data, each time with a different accent dictionary. We noted the result of the best scoring accent dictionary for each speaker individually. This represents the best match between speaker and accent dictionary and the overall score for these experiments was 25.82% SER, which translates to an improvement of about 9% compared with the baseline.

5.3. Accent feature experiments

In the previous experiments, the aim was to choose a dictionary from a number of predefined dictionaries. In the following experiments, the aim is to *create* dictionaries instead.

The accent feature experiments reported here are based on a combination of an accent feature identifier and a speaker-dependent pronunciation dictionary generator. They are composed of the following four phases:

5.3.1. Phase 1: Forced alignment

During the adaptation phase, forced alignment is carried out on 25 phonetically rich utterances per speaker using a semi-traditional global pronunciation dictionary with an exhaustive coverage of alternative pronunciations. Each pronunciation has been tagged with an accent feature code (see Figure 2). Note the first pronunciation of the word <forty>, which shows a combination of accent features. This is not uncommon.

Figure 2: Excerpt of global pronunciation dictionary

```
...
eight [eIt] u
eight [et] m
forty [fO:r4i] f,r
forty [fO:rti] r
forty [fO:4i] f
forty [fO:ti] u
four [fO:] u
four [fO:r] r
...
```

5.3.2. Phase 2: Accent feature identification

An Accent Feature Identifier (AFID) has been created with the purpose of identifying the accent features of each speaker before the main recognition begins. AFID analyses the recognition results from the forced alignment showing which pronunciation has been chosen for each word. It then determines the number of occurrences of each accent feature in order to see which features are most characteristic for the speaker in question.

5.3.3. Phase 3: Generation of the idiodictionaries

In the third phase, the information about the characteristics of the speakers' speech obtained in Phase 2 is used to create a model of their phonological system. These models contain information about which accent features to activate and which to ignore and they are the key component in the creation of the idiodictionaries.

5.3.4. Phase 4: Recognition

Once the idiodictionaries are created, the system is ready for normal recognition - this time with the pronunciation dictionary adapted to the speaker.

In Phase 2, we are looking for a pattern in the speech. If an accent feature is judged to be characteristic for the speaker based on the adaptation utterances, we make the assumption that this feature will also be chosen by the speaker for words in future utterances.

The choice of features was based initially on phonetic knowledge. After studying the literature on pronunciation variation in British English, we made an exhaustive list of accent features. The ones judged to be the most significant for speech recognition were chosen. We then went through a few iterations before selecting the accent features, which we included in the experiments as well as the number of occurrences used for adaptation.

6. Analysis of the results

The results of the experiments described above are shown in Table 1. As can be seen in the table, the experiments using the predefined accent dictionaries only led to a relatively small improvement compared with the baseline experiment. The accent feature experiment, on the other hand, where the pronunciation dictionary was adapted to create idiodictionaries as described above saw a significant improvement compared with the baseline experiment. The idiodictionaries performed 12% better than the accent dictionaries overall and no speaker experienced a deterioration in performance as a results of the pronunciation dictionary adaptation. Compared with the baseline, the idiodictionaries gave an improvement of about 20%.

When all identified features for each speaker are included, accuracy deteriorates. This happens because features, which only occur a few times, cannot be considered to be particularly characteristic of the speaker in question and

do thus not provide any reliable information. We therefore defined a threshold for the minimum number of occurrences needed for an accent feature to make its way into the idiodictionary.

Table 1: Results of experiments

Baseline	SER 28.23%
Accent dictionaries	SER 25.82%
Idiodictionaries	**SER 22.66%**

Both the choice of accent features and the number of times they have to occur in the initial recognition run to be included in the idiodictionary are parameters that can be tuned towards the data in question. For the current test data, six accent features and a minimum of four occurrences gave the best results. Future experiments on other corpora will show how data-dependent these findings are.

7. Conclusions and future work

In this paper, we have presented a new method to deal with the problem of pronunciation variation in Automatic Speech Recognition. The experiments described above show that a combination of accent feature identification and pronunciation dictionary adaptation can significantly improve recognition performance.

The experiments reported here have also given new insight into how pronunciation varies. Accent features therefore seem to be a useful alternative to the notion of dialect when describing accented speech in detail.

Pronunciation dictionary adaptation alone cannot achieve the full potential of pronunciation variation modelling. We consider this method to be an extension of traditional speaker adaptation of the acoustic models and we expect that a combination of the two would improve recognition performance even further. Future experiments will investigate this claim.

In future work, we intend to use the accent feature approach during segmentation of the acoustic signal prior to training the acoustic models in order to model pronunciation variation at various levels. We also want to give the accent features different probabilities prior to adaptation rather than making a binary decision, which we think will improve performance for borderline speakers.

8. References

[1] Kessens, J.M., Strik, H. and Cucchiarini C. "Modeling Pronunciation Variation for ASR: Comparing Criteria for Rule Selection". *Proc. PMLA 2002*

[2] Lincoln, M., Cox, S. and Ringland, S., "A Comparison of Two Unsupervised Approaches to Accent Identification". *Proc. ICSLP 1998*

[3] Wolff, M., Eichner, M. and Hoffmann, R., "Measuring the Quality of Pronunciation Dictionaries". *Proc. PMLA 2002*

[4] Strik, H. and Cucchiarini, C. "Modeling Pronunciation Variation for ASR: A Survey of the Literature". *Speech Communication 29*: 225-246, 1999

[5] Humphries, J.J., Woodland, P.C. and Pearce, D., "Using Accent-Specific Pronunciation Modelling for Robust Speech Recognition". *Proc. ICSLP 1996.*

[6] Van Bael, C. and King, S., "The Keyword Lexicon – An Accent-Independent Lexicon for Automatic Speech Recognition", 2003

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

[8] http://www.aurix.com