THE USE OF SHIBBOLETH WORDS FOR AUTOMATICALLY CLASSIFYING SPEAKERS BY DIALECT

A.W.F. Huggins (bhuggins@bbn.com)
BBN Hark Systems, 70 Fawcett St., Cambridge, MA 02138, USA
Yogen Patel (yogen@speech.com)
PureSpeech, 100 Cambridge Park Drive, Cambridge, MA 02140, USA

Abstract

Real-world applications using speech recognition must perform well over a range of dialects. Differences in dialect between the speakers in the training database and the target users often leads to degraded recognition performance. For the BBN Hark Hidden Markov Model (HMM) based system, we have already developed a reasonably effective technique [1] for dealing with multiple US dialects. The solution involves building separate HMM sets for each dialect from representative training speech data. This requires that training speakers be accurately classified by dialect, which is difficult to do reliably even by hand. In this paper we describe a recognition based pseudo-automatic scheme for partitioning a pool of US English training speakers into groups, such that the speakers within each group share the same pronunciation characteristics. Our scheme is speech-data driven, and involves using transcript-level word hypotheses generated by a recognizer to partition the pool of training speakers.

1 Introduction

We have implemented a multi-dialect recognition system [1] for a speaker independent, 200 word, command and control application. The "commands" are sentences which are typically connected sequences of four or five words. The underlying speech recognition engine used in our work evolved from BBN's Byblos system [2], and it uses discrete probability density functions to model triphones. The original multi-dialect system was trained on a multi-dialect speech corpus that was partitioned into dialect groups based on where the speech data was collected. Limited listening tests were carried out by a dialect expert to confirm that the location-based partitioning was reasonable.

The development of the pseudo-automatic scheme for partioning this speech corpus was driven by two major objectives. First, we felt that such a scheme could lead to improved accuracy of the multi-dialect system since we would be able to group training speakers together more accurately based on actual speaking characteristics. Second, we felt that a generic automatic data driven scheme would allow us to partition data collected in the future more easily, as we would not have to maintain nor collect detailed demographic information for each speaker in the corpus.

There has been some previous work on dialect partitioning. Cohen et al. [3] studied the two sentences that were included in the TIMIT database to capture dialectal differences (the "shibboleth" sentences). They defined eighteen segments whose alternative pronunciations they thought they could transcribe reliably, either visually from spectrograms or by ear, in the tokens of these sentences as spoken by their 630 different speakers. They found that a speaker's use of alternative pronunciations for various segments were not independent, but formed clusters, and that these clusters were also dependent on the dialect of the speaker, based on where the speaker lived from age 2 to 10. They estimated that a useful reduction would occur in the entropy of the phonological model, and suggested this could lead to improved recognition accuracy. However, they did not test this prediction, and the method was not automatic. Furthermore, only a few of the segments studied were vowels, which are usually the most salient cues to dialect for human listeners, and also, we believe, for some recognition systems.

A study by Van Compernolle et al. [4] reported an attempt to automatically cluster speakers of Dutch and Flemish by dialect and by gender. "Dutch and Flemish are identical written languages, [but] pronunciation differences are large..". First, they demonstrated substantially improved performance

from using multiple baseforms, derived from hand-classification of their 1000 speakers. Their method for automatic clustering involved (1) training two baseforms for each word (digits), based on an initial split of the training set, (2) doing recognition on each **utterance** in the training set, and assigning it to the class where it was best recognized, (3) retraining the models based on the new classification, and (4) iterating. Although this method produced improved performance on the training set, the improvement failed to carry over to the test set. The method we report on here is very similar to this; the main difference is that we used performance on a small number of words, hand-selected to discriminate between dialects, to classify **speakers** according to dialect, and then trained separate models for the two classes of **speakers**. We believe the reason their method failed was that a few useful discriminants were swamped by a larger number of cases where the main differences were due to other irrelevant aspects of the speech tokens.

2 Corpus Partitioning

We collected application-domain dependent speech data from approximately 450 adult speakers in the following three southern cities: Atlanta, Dallas, Raleigh (North Carolina), and three northern cities: Paramus (New Jersey), Detroit, and Boston. In addition, we included a few more "northern" and "southern" speakers, whose data although collected separately, consisted of essentially the same vocabulary. The speech data was recorded using a close talking microphone at a sampling frequency of 16KHz.

Our target system required developing the following four models sets; southern male, northern male, southern female, and, northern female. We manually split the corpus by gender since we wanted to focus only on developing methods to automatically partition speakers into sub-groups within each of the separate gender groups. There are already well established methods for gender recognition based on variants of speaker recognition techniques [5][6], and it was not our goal to validate these methods nor improve upon them.

2.1 Overview of Method

A major uncertainty in making this technique work was whether an appropriate set of shibboleth words was available in our corpus. The speech data collection mentioned above took place before we had even considered developing an automatic dialect partitioning scheme. Thus, although there are some classic words and phrases that highlight dialectical differences in American English [7], we did not have the benefit of being able to include these words and phrases in the data collection. We had to pick a set of shibboleth words from the existing vocabulary.

We selected the 52 highest frequency words in our corpus, and generated several different phonetic transcriptions for each word. When generating the phonetic transcriptions, we attempted to capture the potential range of phonetic variation across all the speakers in the training corpus, but, of course, we were constrained by our existing phoneme set, whose main purpose was to capture phonemic regularity rather than phonetic variability. Each word had typically between four and twelve different transcriptions. For each speaker, the sentences that contained these words were passed through a grammar-constrained HMM **phoneme** recognizer. We were unable to use triphone models, which would have been preferable, because many of the alternative phonetic transcriptions contained triphones for which there was little or no training, and this would have resulted in strong biasing of the recognition path selected. The HMM phoneme models in the phoneme recognizer were previously generated using data drawn from a different 'uniform speaker' corpus. Relative counts were accumulated for each pronunciation of each of the shibboleth words by looking at the resulting decoded phoneme sequences for each of the different spoken words.

To convert the word counts into a distance measure, we calculated a chi-square value for each pair of speakers for each of the words, corrected it for degrees of freedom, and took the square root. We also combined the separate speaker*speaker distance matrices for each word into a euclidean pooled matrix. The resulting distance matrices were analyzed by hierarchical clustering to classify the speakers.

As an initial validation of our approach, we ran the same analysis on smaller matrices in which all the speakers were pooled according to where their data was collected, and confirmed that the clustering obtained reflected our expectations in terms of dialects. We did this three times as a result of inspection of the clustering plots. After the first pass, we removed all but the thirteen most frequent words (the digits 1-9, "oh" and "zero", plus "point" and "phone"). And after the second pass, we removed all pronunciation alternatives that received low counts. For most words, this left only two alternatives.

2.2 Computing Word Counts

As mentioned before, for each speaker we carried out grammar constrained phonetic recognition on those sentences that contained the shibboleth words. The different pronunciations of each of the shibboleth words were represented by tagged multiple phoneme paths in the grammar. For each sentence the recog-

Speaker	Relative Count	
	NINE_1	NINE_2
spkr1	0.90	0.10
spkr2	0.80	0.20
spkr3	0.10	0.90
spkr4	0.20	0.80

Table 1: Relative counts of the different pronunciations of the word 'nine', invented data

nizer generated a word level transcript, where each of the shibboleth words had a 'tag' or 'identifier' attached to it indicating the best match pronunciation.

A simple example may help to make this more clear. Consider a set of 4 male speakers, and consider the word "nine" as the shibboleth word to be used in grouping the speakers. Let us assume that the range of potential pronunciations of the word 'nine' are covered by the following two alternate phonetic spellings,

NINE: n-ay-n OR n-aa-n.

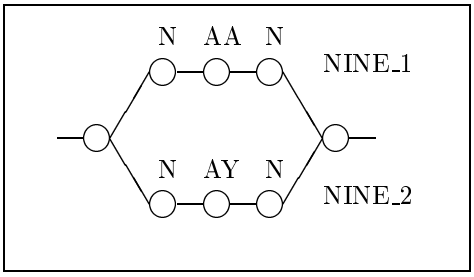

Fig 2.1

As shown in figure 2.1, the grammar will contain two paths, each corresponding to one of the phonetic spellings. All the sentences containing the word nine, are then passed through a phoneme recognizer where the recognition path is constrained by this grammar. By following the highest scoring decoded phoneme sequence for each sentence, we are able to compile a table of relative counts of the occurrences of the different forms of the word "nine" (table 1).

2.3 Clustering Speakers

As a result of the elimination rounds constituted by our initial validation tests, we had fourteen distance matrices, thirteen showing distances between each pair of speakers for each of the final words, and one a euclidean combination of all thirteen words. Inspection of the cluster plots for the individual words showed that, for some words, the clusterings differed fairly dramatically from the dialect partitioning that we were hoping for. These differences would probably repay further study. However, for the purposes at hand, we took only the two highest level clusters generated from the euclidean pooled data for all thirteen words for the male subjects, and used them to partition the male speakers into a "northern" and a "southern" group. We did the same for the female speakers.

3 Experiments and Results

In order to evaluate the automatic data partitioning scheme, we carried out four sets of recognition experiments on the dialect database (described earlier). Speech data from a set of 24 speakers, evenly balanced by gender and region, was held out as a test set. The test utterances coverered a wide range of sentences from our command and control application, and many of them did not contain any of our thirteen surviving shibboleth words.

The first training run involved pooling all the training data from all the speakers and building a dialect independent and gender independent acoustic model. In the second experiment, we built a gender dependent but dialect independent acoustic model. This acoustic model thus contained two sub-models, one for female speakers, and one for male speakers.

In the third experiment, we trained a gender dependent and dialect dependent model containing the following four sub-models; male southern, male northern, female southern, female northern. In this experiment the dialect sub-models were created using the hand partitioned data. That is, we pooled together the data from Raleigh, Dallas and Atlanta to build the southern models, and that from Paramus, Boston and Detroit to build the northern models. Finally, we trained a gender and dialect dependent model, with the same number of sub-models as in the third experiment, but this time using the automatically partitioned training database to create the dialect sub-models.

Table 2 shows the sentence accuracy for each experiment. Simply using gender dependent models results in a significant improvement in performance, confirming the well-established result that gender dependent modeling alone can obtain a large reduction in error rate compared to gender-independent models. Using gender and dialect dependent models built from both the hand and automatic partitioned dialect databases reduces the sentence error even further, with the automatically partitioned data giving almost twice the error reduction of the hand partitioned data.

4 Conclusions

In this paper we have described a fairly simple, yet

Exp.	Model Set	Percent Error
1	Gender & Dialect Independent	7.8
2	Gender Dependent & Dialect Independent	5.6
3	Gender Dependent & Dialect Dependent (hand-partioned)	5.1
4	Gender Dependent & Dialect Dependent (automatic-partioned)	4.7

Table 2: Sentence error for different model sets

powerful method for partitioning a training corpus on the basis of pronunciation similarities. Our experimental results show that the models built from the automatically partioned data yield better performance than the models built from the hand partitioned data. This improvement in accuracy leads us to believe that a recognition-based scheme produces a grouping of similar speakers that is more accurate from the **recognizer's** point of view.

Pseudo-automatic partitioning schemes like the one we have presented here, are very useful for partioning speech data that does not have accompanying detailed demographic information. These sorts of methods may allow more efficient and effective management of data that is collected in unsupervised scenarios (live system data logging, etc.).

5 Further Work

We need to explore further what properties of the speech tokens controlled the clustering of those words whose clusterings appear to be irrelevant to the dialect groups. It is not at all clear that clusters that seem appropriate to the human listener are equally appropriate to the recognizer, and establishing the features underlying these clusters might lead to a better method of segregating speakers into multiple models.

It would also be interesting to see how far our approach could be pushed by deliberately including words known to distinguish between the dialects of interest. The TIMIT database provides an obvious opportunity to test this suggestion.

It will also be interesting to see how well our method works when applied to dialects in other languages. Dialect differences are relatively minor in the US, at least compared with those found elsewhere. The need for methods to classify speakers by dialect will be much more important when the attempt is made to write applications for countries where the dialectal differences are more pronounced, such as England or Germany, to name a couple at random.

References

[1] V.Beattie, S.Edmondson, D.Miller, Y.Patel, and G.Talvola. "An Integrated Multi-Dialect Speech Recognition System with Optional Speaker Adaptation". *In Proceedings of Eurospeech*, page 1123, 1995.

[2] Y.Chow, M.Dunham, O.Kimball, M.Krasner, G.F.Kubala, J.Makhoul, P.Price, S.Roucos, and R.Schwartz. "BYBLOS: The BBN Continuous Speech Recognition System". *In Proceedings of ICASSP*, page 89, 1987.

[3] M.Cohen, G.Baldwin, J.Bernstein, H.Murveit, and M.Weintraub. "Studies for an adaptive recognition lexicon". *Proceedings of the DARPA Speech Recognition Workshop, Report No. SAIC-87/1644, San Diego*, March 1987.

[4] D.Van Compernolle, J.Smolders, P.Jaspers, and T.Hellemans. "Speaker Clustering for Dialectic Robustness in Speaker Independent Recognition". *In Proceedings of Eurospeech*, page 723, 1991.

[5] Herbert Gish and Michael Schmidt. "Text Independent Speaker Identification". *IEEE Signal Processing Magazine*, 1994.

[6] Bishnu S. Atal. "Automatic Recognition of Speakers from their Voices". *Proceedings of the IEEE, Vol. 64, NO. 4*, April 1976.

[7] Charles K Thomas. *Phonetics of American English*. NY: Ronald Press, 1958.