STATISTICAL DIALECT CLASSIFICATION BASED ON MEAN PHONETIC FEATURES

David R. Miller, James Trischitta

BBN Hark Systems
Cambridge, MA 02138
dmiller@bbn.com jtrischi@bbn.com

ABSTRACT

Our paper describes work done on a text-dependent method for automatic utterance classification and dialect model selection using mean cepstral and duration features on a per phoneme basis. From transcribed dialect data, we build a linear discriminant to separate the dialects in feature space. This method is potentially much faster than our previous selection algorithm. We have been able to achieve error rates of 8% for distinguishing Northern US speakers from Southern US speakers, and average error rates of 13% on a variety of finer pairwise dialect discriminations. We also present a description of the training and test corpora collected for this work.

1. INTRODUCTION

Performance of an ASR system is strongly influenced by how well the training speech is matched to the test utterance. A mismatch in gender or dialect of the speaker will have adverse effects on performance. One effective solution to this problem is to partition training data into several groups according to the value of these characteristics and then train separate model-sets for each group. Thus, one might have one model-set for Southern females, another for Northern males, and so forth. [1] describes such a system and documents its advantages over a pooled-model system.

A multi-model-set system requires knowledge of the speaker's gender and dialect characteristics at two points, partitioning of the training data and selection of the proper model for decoding a test utterance. These characteristics can be determined by a human for the training task, since the training speech is collected from known subjects. During general usage, however, the ASR system may have no data about the speaker other than the utterances coming from that speaker. It is therefore necessary to perform a classification on the utterance in order to determine the proper model to use for decoding.

The multi-dialect system in [1] uses a clumsy and computationally intensive method to select the gender and dialect of the user. This paper describes work done on a faster method for dialect selection using linear discriminants. Section 2 describes the speech features passed to the dialect classifier. Section 3 describes the pool of speakers used in this work, as well at the specific classification tasks that have been explored. Section 4 outines the use of linear discriminants for classification. Section 5 provides our results for a variety of classification tasks. Finally, section 6 concludes with possible future directions for this work.

2. DIALECT SELECTION FEATURES

We propose to distinguish among dialects by modeling the duration and typical cepstra of each of several phonemes. As part of the front end to a speech recognizer, we computed 14-coefficient, Mel-warped cepstral vectors. We used a 20 ms analysis window, advancing every 12.5 ms. For each utterance, we generated an alignment of cepstral vectors with word-level transcriptions using an HMM based speech recognizer with a speaker-indepent, dialect-independent model. For each speaker, we then computed the average duration and average cepstrum for each phoneme that speaker said, the average being taken over multiple frames from multiple utterances by that speaker.

The per-phoneme averages are concatenated to form a single vector whose length is 15 times the size of the phoneme set. While there are roughly 50 phonemes in english, linguistic considerations indicate that many of them are roughly constant across dialects. Therefore only a limited set of phonemes (primarily vowels) are retained for the subsequent analysis. We have experimented with the composition of this set, with the results detailed below. In all cases, the inputs to the classification process are long vectors of cepstral and duration information, one vector per speaker.

3. SPEAKERS AND TASKS

In the summer and fall of 1995, a team of engineers at BBN Hark traveled to shopping malls in various parts of the country to collect speech samples. At each location, several speakers were asked to record roughly 80 utterances each. The sentences consisted of connected digit strings and com-

Group	#spkrs	total utts.
atlanta-F	51	4048
atlanta-M	32	2554
boston-F	52	3620
boston-M	48	3347
dallas-F	30	2486
dallas-M	26	2159
midwest-F	40	16604
midwest-M	36	14292
total	315	49110

Table 1: Summary of data by region and gender

mand and control phrases. These utterances were then transcribed by humans, and any corrupted utterances (corrupted by excessive background noise, laughter, microphone buzzing etc.) were discarded.

We have experimented on a subset of that data, using male and female speakers from Atlanta, Boston, Dallas, and the Midwest (a mixture of several midwest cities). Table 1 shows the break down of speakers by region. In all, the data comprises 315 speakers saying roughly 49,000 utterances. We have partitioned this data into a set of training speakers and a set of test speakers by randomly selecting 5 speakers from each of the 8 groups (listed in Table 1) to be our test set. The remaining 275 speakers went into the training set.

We have explored a number of tasks on this data:

- classify speakers as female or male;
- classify speakers as Northern (Boston + Midwest) or Southern (Atlanta + Dallas);
- for each pair of dialects, make a 2-way choice between the dialects;
- pooling male and female data, make a 4-way choice among all dialects;
- using only single gender data, make a 4-way choice among all dialects;
- make an 8-way choice among all groups.

Table 2 lists the codes we shall use to refer to these tasks.

As indicated in section 2, the observation vector for each speaker consists of 15 numbers per phoneme used in the classification. Based on linguistic considerations, we have chosen to retain only those phonemes which differ substantially across regions. We have experimented on several phoneme sets, ranging in size from 5 to 17; the specific phonemes can be seen in Table 3.

4. CLASSIFICATION METHOD

We used a rudimentary Fisher linear discriminant to classify speakers into various groups [2]. We compute the within- and between-class covariance matrices $\boldsymbol{W}$ and $\boldsymbol{B}$, and then find the eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$ having the largest eigenvalues. We use these eigenvectors as the basis for a subspace L into which we project all the data points. Test points are classified according to the Euclidean distance in L to the projected centroid of the training points for each group. In all cases, we have chosen L to have dimension one less than the number of groups involved.

FM	female vs male
NS	north vs south
AB	atlanta vs boston
AD	atlanta vs dallas
AM	atlanta vs midwest
BD	boston vs dallas
BM	boston vs midwest
DM	dallas vs midwest
pooled4	atlanta vs boston vs dallas vs midwest
male4	atlanta-M vs boston-M vs dallas-M vs midwest-M
female4	atlanta-F vs boston-F vs dallas-F vs midwest-F
full	atlanta-M vs boston-M vs dallas-M vs midwest-M vs atlanta-F vs boston-F vs dallas-F vs midwest-F

Table 2: Codes for classification tasks

Set 1	AY, R, AO, EH, IY
Set 2	AA, AW, OH, OW, UH
Set 3	AY, R, AO, EH, IY, OW, OH, TH
Set 4	AA, AE, AH, AO, AW, AY, EH, EI, EY, IH, IY, OH, OW, OY, UH, UW, R

Table 3: Phoneme sets for classification

There are two technical points to observe in our implementation of this discrimination rule. First, since the features described in section 2 have a wide variety of dynamic ranges and are typically not centered at zero, we have normalized each component feature to have zero mean and unit variance. Second, not every speaker has uttered every phoneme. As a result, there are some NA values in our data matrix where the speaker average for a missing phoneme should appear. In these (very few) cases we have substituted the pooled mean value of this feature, namely zero.

5. EXPERIMENTAL RESULTS

We have performed each of the tasks listed in Table 2 using each of the phoneme sets in Table 3. For each combination, we built a discrimination rule using LDA on the training speakers, and then evaluated that rule on the test speakers.

Table 4 shows the total classification error rate for each task/phoneme set combination.

Error rates varied substantially across tasks, but were generally encouraging. The technique is certainly able to discover the oft-observed differences between average male and female cepstra. More meaningful is the performance on the NS task, where we achieve an 8% error rate. In general, our average pairwise dialect classification error was 13%.

	Set 1	Set 2	Set 3	Set 4
FM	0	8	3	20
NS	8	15	10	15
AB	20	10	30	50
AD	25	30	50	50
AM	0	15	5	50
BD	5	15	10	50
BM	20	25	10	50
DM	10	5	50	50
Avg 2-way	13	17	26	50
pooled4	25	33	28	33
male4	10	50	85	75
female4	20	30	35	75
full	28	33	25	53

Table 4: % error rate for 2-way, 4-way, and 8-way classification tasks.

Given the high success of the FM classifier, we can see that by combining it with the gender specific 4-way classifiers would yield a gender independent 4-way classifier with a 15% error rate, far better than the 4-way model built from pooling the data.

We observed that the Midwest dialect group seemed most easily distinguished from the others. Examining the individual confusions made, we saw that Midwest speakers were almost always correctly identified, and that non-Midwest speakers were almost never classified as being from the Midwest population. The most obvious difference between the Midwest data and the other data is the number of training utterances per speaker; for Midwest subjects the average was 400 utterances per speaker, while for all other subjects the average was only 75 utterances per speaker. The greater number of utterances likely generated more accurate means for the Midwest speakers, and so reduced the covariance of the Midwest population. It was then easier to find hyperplanes separating this population from the others.

The variation of phoneme sets was disappointing. Phoneme set 1, thought *a priori* to consist of phonemes with discriminative information, was by far the most successful. Set 2, which has no overlap with but is the same size as set 1, was arbitrarily selected and gave worse performance. Set 3, which is an extension of set 2, also gave inferior performance. Finally, set 4, containing all 17 vowel phonemes, was virtually worthless (with 255 dimensional space and only 275 speakers, the classifier was undoubtedly over tuned to the data). The need to invoke external knowledge to select a good set of phonemes is an undesirable aspect of the current system, and one that we will seek to remove in future work.

6. FUTURE WORK

The dialect corpus collected is a rich body of data which we hope to study further in many ways. We have shown in this paper an effective means of classifying speakers into dialect groups. Using a linear discriminant model, we can do a reasonable job of reproducing the human determined classification of speakers. While the current work uses a large number of known utterances from each speaker, we plan to explore schemes using far less test data and potentially errorful recognizer-generated transcripts. In addition, we plan to explore the use of Gaussian or Gaussian mixture classifiers.

Speaker clustering is another topic which future work may address. It need not be the case that the present partitioning is the ideal one for improved ASR. We have taken an arbitrary stab at the proper way to divide a national pool of speakers into distinct subclasses. A more principled approach is to discard the notion of geographical based dialects entirely and replace it with abstract groups of similar speakers. Under this viewpoint, the problem to be solved is not one of classifying speakers into predetermined groups, but to cluster the data into consistent groups. Evaluation of a given clustering scheme would be done by training a multi-model-set speech system from the given clusters and measuring recognition accuracy for test speakers who also fell in that cluster. Once a good clustering is determined, the classification techniques described in this paper can be used to determine the cluster to which a new speaker belongs. We hope to address this topic in future work.

ACKNOWLEDGEMENTS

We would like to thank A.W.F. Huggins for his help with dialects and phoneme selection.

7. REFERENCES

1. V. Beattie, S. Edmondson, D. Miller, Y. Patel, G. Talvola. "An integrated multi-dialect speech recognition system with optional speaker adaptation". *Proc. Eurospeech*, pg 1123, 1995.
2. R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, 1992.