INCORPORATING LINGUISTIC KNOWLEDGE INTO AUTOMATIC DIALECT IDENTIFICATION OF SPANISH*

Lisa R. Yanguas, Gerald C. O'Leary, Marc A. Zissman
M.I.T. Lincoln Laboratory
Lexington, MA 02420-9185
lryangu | gco | maz @sst.ll.mit.edu

ABSTRACT

Automatic dialect identification, like automatic language identification, has often been approached through the use of phonetic frequencies and phonetic sequence modeling. While such statistical systems perform well on language identification problems, they are less adept at the more difficult problem of automatic dialect identification, particularly on short segments of speech. In this paper we explore issues involved in exploiting linguistic knowledge to aid in the automatic identification of dialects of conversational Spanish.

1. INTRODUCTION

Anecdotal evidence has long suggested that humans can discriminate dialects with varying degrees of proficiency. While this would imply that linguistic analyses should be helpful in automatically identifying dialects, these types of analyses have not been previously exploited in traditional language identification or dialect identification systems. In addition, purely statistical approaches, which have proven successful for the language identification task, have exhibited limited success on the more difficult dialect identification problem. If we could capitalize on the ability of humans to perform this discrimination, quantify the linguistic knowledge the human expert brings to the task, and use this to enhance our automatic systems, we should be able to improve upon the performance of those systems.

2. MOTIVATION AND OVERVIEW OF PREVIOUS WORK

We use Cuban and Peruvian Spanish dialect data from the Miami Spanish Dialect Corpus to perform our analyses. The database consists of extemporaneous speech from 180 native speakers of Spanish recorded in an interview setting, as well as read speech, and the digits 0-10. In previous work [1] a traditional language identification system that makes use of phonetic frequencies and phonetic sequence modeling was employed to discriminate Cuban speakers from Peruvians automatically. On test segments of greater than 20 seconds of speech a 16% dialect recognition error rate (84% accuracy) was obtained. This performance, however, decreased to a 34% error rate (66% accuracy) when test segments were reduced to 5 seconds.

THIS WORK WAS SUPPORTED BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

3. THEORETICAL PERFORMANCE

Much has been written on Spanish dialectology and those salient linguistic features that are notable in discriminating between and among a variety of dialects of Spanish [2,3,4]. We conducted this type of linguistic analysis on the Cuban and Peruvian data using 20-second segments from 50 speakers (25 of each dialect). We identified 49 linguistic features that occur with different rates in each of the two dialects. Examples of such features include the aspiration or dropping of [s] vs. the preservation or reinforcement of [s] in analogous environments, vowel raising and/or lowering, and the replacement of [r] by [l]. These counts are provided for the seven most prevalent of the 49 features in Table 1 below:

Feature	Explanation	Cuban	Peruvian
s_X	s is preserved	35	101
s->0	s is dropped	74	24
s->h	s is aspirated	18	29
n->ng	n is pronounced ng	23	28
d->0	d is dropped	12	10
s->ss	s is reinforced	4	12
0->h	h is inserted	11	1

Table 1: Frequency counts of occurrences of top 7 dialect features

4. THEORETICAL MODEL FOR AUTOMATIC DETECTION

We developed and present here a theoretical model for predicting the performance of an automatic dialect identification system based on the set of independent feature detectors described above that trigger on specific events in the speech signal and which are considered to be related to the dialects of interest. This model is used to calculate the expected performance of a system that could detect these features for a corpus of Cuban and Peruvian Spanish. The statistics for the model are derived from the hand marked features in a portion of the corpus. The dialect identification system consists of *K* parallel detectors, each of which is looking for a specific feature. An observation vector for a speech segment is formed

from the counts of the detections of the features from the different detectors. A decision statistic is then computed from this vector and compared to a threshold to obtain a decision on the dialect class.

4.1 Perfect Detection

Assume that we have identified a set of linguistic features $\{L_i\}$ which occur with different rates in different dialects. Assume that feature i occurs with rate λ_{iA} in dialect *A*. Another interpretation is that the feature will occur with probability $\lambda_{iA}\Delta$ in a time interval Δ. The events are described by a Poisson distribution. The probability that the event will occur exactly *n* times in an interval of length *T* is

$$P(n) = \frac{(\lambda_{iA}T)^n e^{-\lambda_{iA}T}}{n!} \quad \text{(EQ 1)}$$

For the Poisson distribution the mean and variance are given by

$$E(n) = \lambda_{iA}T \quad \text{(EQ 2)}$$

and

$$Var(n) = \lambda_{iA}T \quad \text{(EQ 3)}$$

respectively. We can compute the posteriori probability of a particular speech segment's belonging to a dialect class as

$$P(D_A|O) = \frac{P(O|D_A)P(D_A)}{P(O)} \quad \text{(EQ 4)}$$

where *O* is the observation vector of the number of detections from the different feature detectors $\{n_i\}$ and $P(D_A)$ is the prior probability of the segment belonging to dialect class *A*. If we assume that the occurrences of the different features are independent, the likelihood can be written as

$$P(O|D_A) = P(n_1, n_2 \ldots |D_A) = \prod_{i=1}^{K} P(n_i|D_A) \quad \text{(EQ 5)}$$

If we want to make a closed set decision between two classes, *A* and *B*, and if the prior probabilities are equal, then the decision statistic becomes the likelihood ratio

$$\frac{P(O|D_A)}{P(O|D_B)} = \prod_{i=1}^{K} \left(\frac{\lambda_{iA}}{\lambda_{iB}}\right)^{n_i} e^{-(\lambda_{iA}-\lambda_{iB})T} \quad \text{(EQ 6)}$$

It is more convenient to consider the log likelihood ratio for the probabilities.

$$\Lambda = \log\left(\frac{P(O|D_A)}{P(O|D_B)}\right) = \sum_{i=1}^{K} \left\{ n_i \log\left(\frac{\lambda_{iA}}{\lambda_{iB}}\right) - (\lambda_{iA}-\lambda_{iB})T \right\} \quad \text{(EQ 7)}$$

If $\Lambda > 0$ the best choice would be to guess that the segment belongs to class *A*. Otherwise, we should guess that it belongs to class B.

4.2 Interpretation

If we run the system on a large number of speech segments from the two dialect classes, we can form the distributions of the scores Λ for each of the classes. We can calculate the expected distributions of the statistic Λ from Equation 7 to predict the performance of the system. Since the distribution of each n_i is Poisson, the mean and the variance will both be $\lambda_{iA}T$ under Hypothesis *A*. If we compare the distributions under the two hypotheses, the means of the distributions for Λ are separated by

$$E(\Lambda|A) - E(\Lambda|B) = \sum_{i=1}^{K} (\lambda_{iA}-\lambda_{iB})T \log\left(\frac{\lambda_{iA}}{\lambda_{iB}}\right) \quad \text{(EQ 8)}$$

The variance of the distribution for *A* will be

$$\text{Var}(\Lambda|A) = \sum_{i=1}^{K} \lambda_{iA}T\left[\log\left(\frac{\lambda_{iA}}{\lambda_{iB}}\right)\right]^2 \quad \text{(EQ 9)}$$

The expression for *B* will be similar. The final performance will be determined by the ratio of mean separation to standard deviation.

$$\frac{E(\Lambda|A) - E(\Lambda|B)}{\sqrt{\text{Var}(\Lambda)}} = \frac{\sum_{i=1}^{K} (\lambda_{iA}-\lambda_{iB})\log\left(\frac{\lambda_{iA}}{\lambda_{iB}}\right)}{\sqrt{\sum_{i=1}^{K} \lambda_{iA}\left[\log\left(\frac{\lambda_{iA}}{\lambda_{iB}}\right)\right]^2}}\sqrt{T} \quad \text{(EQ 10)}$$

which indicates that separation improves with the difference of the rates of events in the two classes and with the length of the utterance. If we assume that Λ has an approximately Gaussian distribution, we can estimate the probability of error directly.

4.3 Imperfect Feature Detection

In practice, of course, the feature detectors will make errors. Two kinds of errors are possible. The feature detector can miss an event which actually happens, or it can generate a spurious detection. If we assume that the probability of detection for an event of class *i* is P_{di}, then the rate of event detections for dialect *A* will be $P_{di}\lambda_{iA}$. We also assume that this feature detector will generate false alarms at a rate[1] λ_{fi}. The effective rate of detections for class A then becomes $P_{di}\lambda_{iA} + \lambda_{fi}$. This value can be inserted into Equation 10 to get the new result. The principal effect is on the separation of means. The log term becomes:

$$\log\left(\frac{P_{di}\lambda_{iA} + \lambda_{fi}}{P_{di}\lambda_{iB} + \lambda_{fi}}\right) \quad \text{(EQ 11)}$$

Note that as P_{di} becomes small or as λ_{fi} becomes large, the ratio in Equation 11 will tend to unity and the value of Equation 11 tends to zero. Thus the numerator of Equation 10 will become smaller, as expected.

4.4 Theoretical vs. Actual Performance

Assuming perfect feature recognition, to account for the non-independence of features, their relative importance, and inter-

1. We assume here that the rates of detection and of false alarms for a particular feature are independent of the dialect. This assumption is plausible and simplifies the notation a bit, but it is not essential to the argument.

speaker variability, we employ a Gaussian classifier. We found that a subset of 16 of the 49 originally-identified features obtains nearly perfect performance in discriminating between the two dialects, and that a close approximation of optimality may be achieved by relying on only the two most prevalent features. We compared these results with those from both the theoretical model described above and the phonetically based automatic language recognition system (PRLM-P) mentioned earlier on both 20- and 5-second segments of speech. A comparison of these results is shown in Table 2 below:

	20 seconds	5 seconds
	% error	
Automatic Dialect ID (PRLM-P)	**25%**	**34%**
Theoretical Model on Hand-Labelled Data	**1%**	**10%**
Gaussian Classifier on Hand-Labelled Data	**2%**	**14%**

Table 2: Comparison of performances of three detection systems

5. HUMAN LISTENING EXPERIMENTS

Upon observing that an expert linguist can discriminate dialects often prior to even the occurrence of the first identified feature, we conducted a series of listening experiments. Perfect automatic recognition, while difficult to begin with, will be further hindered by a number of issues, including context; i.e. cues may be evidenced only when a certain phonological or morphological rule applies. For example in Cuban Spanish:

$$r \rightarrow rr \; / \; \_ \; \# \; + \qquad \text{(EQ 12)}$$

by which a morpheme-final [r] is reinforced when followed by an agglutinated or clitic morpheme as in “cantar” (to sing)/”cantarles” (to sing to them). Thus, we considered whether an orthographic transcription is helpful in obtaining high-performance dialect identification [5, 6]. Human subjects (native English speaking, non-experts in Spanish) were first asked to identify the dialects of five speakers from four dialect groups speaking the same English shibboleth sentences. Eight out of the nine test takers made fewer than 40% errors on this task, while all nine produced less than 60% errors. The test takers, however, were unable to identify dialects of Spanish without benefit of specific and directed training. They were then asked to attempt a forced choice (i.e. Cuban vs. Peruvian Spanish) dialect identification task on read sentences from the Miami Corpus. The subjects were given focused training on each dialect along with an orthographic transcription of the texts. Each speaker in both the training and the test sets read the same two texts. Following training, the test takers were asked to identify the dialect of test utterances. All but one of the subjects performed at an error rate of less than 40% with this type of directed training and all produced less than 60% errors. From this we gleaned that non-experts can, in fact, be systematically trained for dialect identification and that they are able to outperform an automatic system.

6. PRELIMINARY AUTOMATIC DIALECT IDENTIFICATION EXPERIMENTS

From the experiments we have described here, it seems clear that the solution of particularly difficult problems for automatic identification, such as in this case, dialect identification, can be improved upon when linguistic information is able to be exploited. We explained how in the theoretical realm, at least, this knowledge can significantly improve upon the performance of automatic systems. In this section we present experiments leading toward future work on using speech recognition approaches to identify dialects in a limited domain.

6.1 Dialect Identification in a Limited Domain

Encouraged by these theoretical results, we pursued a modified continuous speech recognition approach to dialect identification [7], limiting ourselves specifically to the recognition of the Spanish spoken digits from 0 to 10. We first examined spectrograms and considered the differences across the two dialects at the phone and sub-phone levels. These differences are both notable and seemingly quantifiable, and led us to surmise that duration and energy, when quantified, should provide discriminatory power between the Cuban and Peruvian dialects at the phone level, at least with respect to digits. We used a time aligner trained on the TIMIT database of American English dialects and then subsequently incorporated into the HTK matrix a mapping scheme to the true Spanish phones reflected in the data. Even so, our automatic aligner proved faulty in time-aligning phones so in addition, we hand-labelled a minimum of data for purposes of accuracy of our experiments and retrained the TIMIT front end on these data.

6.2 Phone Duration Discrimination

So as to discount the influence of speaking rate between and among speakers and across dialects, we normalized for phone duration within a given speaker. Phone duration comparisons across the two dialects showed some discriminatory power, with particular segments carrying much more information than others. We also found that we needed to consider [s] and the vowels [o] and [e] solely in word-final contexts, since they are virtually invariant when they occur word-initially (the only two possibilities here). Figure 1 below shows a comparison on hand-marked data of durations of word final [s], normalized for speaking rate.

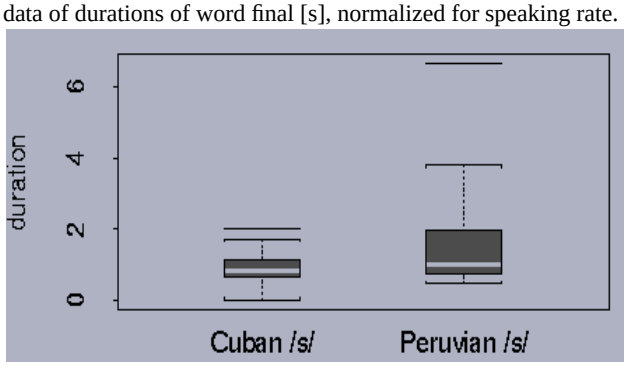

Figure 1: Gaussian distributions showing duration comparisons for word-final [s]: Cuban vs. Peruvian[2].

6.3 Phone Energy Discrimination

Initial experiments computing and comparing the overall energy of phones across the two dialects proved promising for discrimination, as well. Once again, we chose to focus on a first pass on word-final [s]. In addition to other properties we have looked at with respect to [s], it carries significant information in terms of energy and tends to behave markedly differently in each of the two dialects. We examined energy contours across four bandwidths for word-final occurrences of [s] in the read digits on hand-labelled data and observed a consistent distinction between the two dialects, as shown in Figure 2 below:

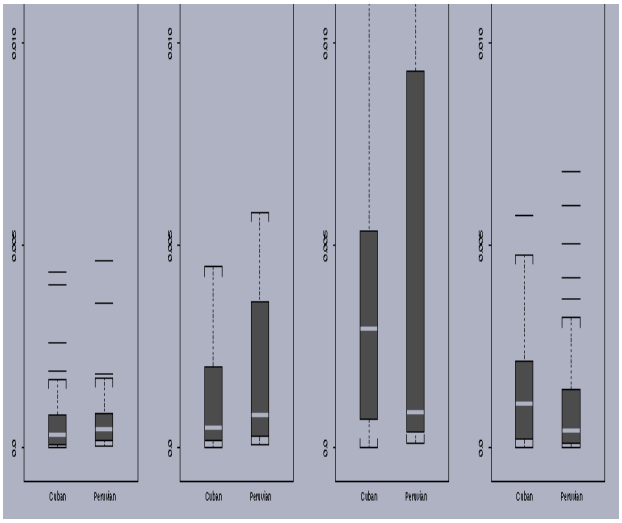

Figure 2: Gaussian distributions showing energy comparisons for four Cuban vs. four Peruvian speakers (hand-labelled data) for word-final [s] across four 2000 Hz bandwidths with flat filters at a sampling rate of 16000 Hz^2.

We then used a Gaussian classifier and principal components analysis to measure our performance on the dialect identification task in this limited domain. With the incorporation of this type of energy information based on linguistic knowledge, we find that for a given vector of [s] alone, we can correctly recognize dialect with 72% accuracy on read digits.

7. CONCLUSIONS AND FUTURE WORK

As described, we have completed some solid baseline experimentation in both phone duration and energy computation with promising results. In follow-on work to this we plan to investigate the use of additional or different types of filters, for example auditory filters, which we expect will provide us more finely detailed comparisons. In addition, further work in this area will include expanding our analyses to additional dialects of Spanish. Furthermore, our results to date encourage us to now move to the relatively harder problem of attempting to recognize digits as they occur within extemporaneous speech. Perhaps most notably, we are also working to further automate our system by not relying on hand-labelled data, but rather, by using automatically aligned data. Ultimately, we plan to employ this approach, but to broaden our domain to the more generic context of dialect recognition in extemporaneous speech.

2. A boxplot is a graphical representation showing the center and spread of a distribution. The horizontal white line in the interior of the box is located at the median of the data. The height of the box is equal to the interquartile distance, which is the difference between the third and first quartiles. The IQD indicates the spread or width of the distribution for the data. The whiskers or dotted lines extend to the extreme values of the data or 1.5 IQD from the center, whichever is less. For data with a Gaussian distribution, approximately 99.3% of the data falls within the whiskers. Outliers are indicated by horizontal lines outside the whiskers [8].

ACKNOWLEDGMENTS

The authors would like to thank Kay Berkling, Terry Gleason, Jack McLaughlin and Tom Quatieri for their invaluable help in the conduct of this research.

REFERENCES

[1] M. A. Zissman and T. P. Gleason and D. M. Rekart and B. L. Losiewicz. Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech. In Proceedings ICASSP, volume 2, pages 777-780, May 1996.

[2] D.L. Canfield. *Spanish Pronunciation in the Americas*. University of Chicago Press, Chicago 1981.

[3] E.G. Cotton and J.M. Sharp. *Spanish in the Americas*. Georgetown University Press, Washington, D.C. 1988.

[4] J.M. Lipski. *Latin American Spanish*. Longman, London, 1994.

[5] C. Teixeira and I. Trancoso and A. Serralheiro. Accent Identification. In Proceedings ICSLP, volume 3, pages 1784-1787, October 1996.

[6] V. Beattie, S. Edmondson, D. Miller, Y. Patel and G. Talvola. An integrated multi-dialect speech recognition system with optional speaker adaptation. In Proceedings Eurospeech, volume 2, pages 1123-1126, September 1995.

[7] J. Brousseau and S.A. Fox. Dialect-dependent speech recognizers for Canadian and European French. In Proceedings ICSLP, volume 2, pages 1003-1006, October 1992.

[8] MathSoft, Inc. *Statistical Sciences, S-PLUS User's Manual*. Version 3.2, Seattle: StatSci 1993. pages 3-50.