Dialect identification using Gaussian Mixture Models [•]

Pedro A. Torres-Carrasquillo, Terry P. Gleason and Douglas A. Reynolds
Lincoln Laboratory, Massachusetts Institute of Technology
ptorres@ll.mit.edu, tgleason@ll.mit.edu, dar@ll.mit.edu

ABSTRACT

Recent results in the area of language identification have shown a significant improvement over previous systems. In this paper, we evaluate the related problem of dialect identification using one of the techniques recently developed for language identification, the Gaussian mixture models with shifted-delta-cepstral features. The system shown is developed using the same methodology followed for the language identification case. Results show that the use of the GMM techniques yields an average of 30% equal error rate for the dialects in the Miami corpus and about 13% equal error rate for the dialects in the CallFriend corpus.

1. INTRODUCTION

Over the last few years, the area of language identification has seen the development of new techniques including novel approaches using Gaussian mixture models, support vector machines, and improvements on the more classical approaches based on phone recognition and language modeling, such as the PPRLM system[1-3]. The problem of dialect identification, which is closely related to the language identification problem, has not received the same level of research interest. The task of dialect identification is that of identifying the spoken dialect from within a set of utterances in a known language, (e.g., North versus South, in the case of American English). Due to the similar nature of dialects within a language, dialect identification poses a more difficult problem than language identification.

The potential applications for dialect identification are similar to those in the area of language identification, including among others pre-processing of the incoming speech for other downstream processing such as automatic speech recognition systems.

The work presented in this paper is focused on applying some of the techniques developed for language identification community to the area of dialect identification. The organization of this paper is as follows: Section 2 describes the two corpora used for the dialect ID experiments. Section 3 describes briefly the system based on GMM acoustic scores. Section 4 discusses the dialect ID results obtained for both corpora and Section 5 presents conclusions and proposals for future work.

2. CORPORA

The CallFriend corpus [4] is a collection of unscripted conversations for 12 languages, including two dialects for three of the languages, recorded over domestic telephone lines. The corpus consists of a training partition used to train the language models of the system, a development partition for parameter tuning and an evaluation partition used to test performance. The 12 languages are: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Of particular interest for the work shown in this paper are the dialects included for three (English, Mandarin, and Spanish) of the 12 languages. Each of these languages includes two dialects: North and South for English, Mandarin and Taiwanese for Chinese, and Caribbean and Non-Caribbean for Spanish.

The training partition of the CallFriend corpus includes 20 conversations for each dialect in the three languages, English, Mandarin and Spanish, resulting in 40 conversations per language. Each training conversation in the CallFriend corpus is about 30 minutes long. The development partition and the evaluation partition for the CallFriend corpus include 80 testing utterances per dialect, except for the case of English where an additional group of about 320 utterances are included in the evaluation partition. This set of 320 utterances include utterances from the King corpus, the OGI corpus and Switchboard. Each of the speech utterances available in the development and evaluation partitions and used for this work is about 30 seconds.

A second corpus, the Miami corpus, consists of two dialects within the Spanish language. This corpus, which is described in [5], includes utterances for Cuban Spanish and Peruvian Spanish. The corpus is partitioned in three disjoint sets called E, F, and G as shown in Table 1.

Set	Dialect	Number of utterances
E	Cuban	37
	Peruvian	23
F	Cuban	39
	Peruvian	20
G	Cuban	16
	Peruvian	8

TABLE 1. Partition of the Miami corpus into three sets E, F, and G.

The speech utterances in the Miami corpus are about 3 minutes long captured from an interview of native speakers. The effective size of the utterances is around 1.5 minutes as the interviewer's voice is eliminated from the utterances. Each of these utterances was recorded using a boom microphone and recorded with a Sony digital audio tape with a sampling rate of 48 kHz and 16-bit quantization.

[•] This work is sponsored by the Air Force Research Laboratory under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

3. GMM-SDC system

The Gaussian mixture model (GMM) using shifted delta cepstral features (SDC) or GMM-SDC, system employed in this work has been previously described in more detail in [2, 3] and is shown in Figure 1. The GMM-SDC system consist of a set of GMMs trained for each dialect of interest. For example, for the case of discriminating between the English dialects, two models are trained, one for each dialect. Similar to the work performed for LID, the features used for the system are based on shifted-delta cepstra, which are a set of features derived from the classical set of delta cepstra.

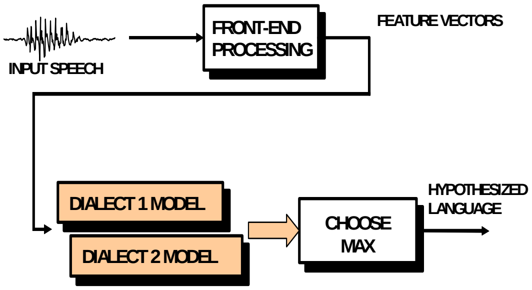

FIGURE 1. GMM-SDC system as used for the dialect ID problem.

The training and testing procedures are as follows. During training, each of the training conversations is processed by the front-end and a feature set using a 7-1-3-7 SDC parameterization is obtained for each frame. The full set of feature vectors is used to build a Universal Background Model (UBM) using all the training conversations in the CallFriend corpus. The dialect-dependent models are then trained by adapting the UBM similar to the technique described in [6].

During testing, an incoming speech utterance is processed by the front-end and scored against each dialect-dependent model using a fast scoring scheme as described in [7]. The model scoring highest is hypothesized as the model of the incoming utterance.

For the case of the CallFriend corpus, where a significant amount of data is available, we will evaluate the use of a backend classifier instead of a maximum likelihood decision. Employing this backend classifier enhances the classification performance by correcting some of the errors produced in the maximum likelihood decision criteria.

4. EXPERIMENTS

This section presents the results obtained for both the CallFriend corpus and the Miami corpus. The set of results presented for the CallFriend corpus is more extensive as the amount of data available allows more flexibility for further analysis and improvements.

4.1. Miami corpus

The first set of results shown is based on the experiments conducted for the Miami corpus. The results obtained are presented using two scenarios. Either one of two sets (E or F) is used for training the GMMs with the remaining set (E or F) along with set G used for evaluation. The GMM order used for this set of experiments is 256 as higher order models did not trained reliably. Results for the Miami corpus are summarized in Table 2, with the 95% confidence intervals for these values at approximately ±10.0 for testing on sets E and F, and ±20.0 for testing on set G.

Training set	Evaluation set	Classification Error (%)	EER (%)
E	F	28.51	35.59
E	G	45.83	41.67
F	E	23.73	32.20
F	G	41.67	41.67

TABLE 2. Summary of results for the Miami corpus.

The performance obtained for the Miami corpus is worse than that obtained in [5]. The system evaluated in [5] resulted in a better classification error by about 15%, when the E partition was used for training, and about 5% better when the F partition was used for training. Two subtleties in the Miami corpus need to be studied further. First is the amount of training data available and the availability of almost twice as much data for Cuban compared to Peruvian. Additional experiments conducted showed that the classification is highly biased toward Cuban as the GMM order is increased. This result might be an indication of a need for balancing the training data set to include similar amount for each dialect. This limitation in the GMM orders that could be reliably trained is also considered an important factor in the performance obtained Second, the choice of SDC parameterization values needs to be evaluated further to understand its effect for this data set.

4.2. CallFriend corpus

For the speech utterances in the CallFriend corpus, three experiments were conducted dealing with the classification of the dialects on each of the languages available, English, Mandarin and Spanish. The results for the experiments are shown in Table 3, with the 95% confidence intervals for these values at approximately ±2.5.

From the results in Table 3, the results for the CallFriend corpus are better than those obtained for the Miami corpus. Although such a comparison is not necessarily fair as the conditions for each corpus are different, the increased performance obtained can be due to either the higher amount of training data, which allows for a GMM order of 2048 to be trained, or the use of a parameterization for the SDC features that was chosen because of its performance for the language identification case on previous experiments for this corpus. Additionally, the collection for each corpus is different which needs to be addressed in future experiments.

Additional experiments were conducted to assess the use of a backend classifier (BE) in the dialect identification case. The first experiment conducted was performed by scoring the development set against the two dialect models available. For each case, the experiments resulted in a decrease in performance for two of the three set of dialects evaluated.

Dialects	Classification Error (%)	EER (%)
English	28.45	15.06
Mandarin	28.21	11.54
Spanish	33.33	13.73

TABLE 3. Summary of results for the CallFriend corpus.

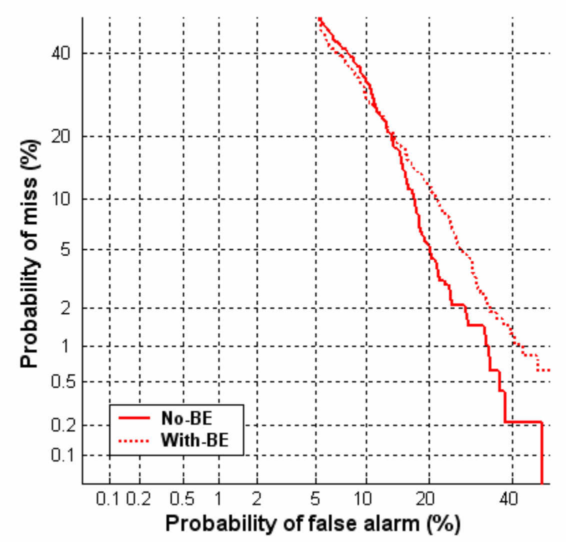

FIGURE 2. Comparison of results using a backend classifier in the case of English dialects.

The results in Figures 2, 3 and 4 show that no improvements are obtained when a Gaussian backend classifier is employed.

A second experiment was conducted to determine whether additional improvements can be obtained when auxiliary information is added to the backend classifier. In this experiment, rather than limiting the data seen by the classifier to only that of the different dialects of the same language (two scores per test utterance), the classifier is designed to utilize all the data available in the CallFriend corpus (15 scores per test utterance). All the development utterances are scored and used for training the classifier. Results for this case are shown in Fig. 5, 6 and 7.

The results shown clearly demonstrate the advantage of using the auxiliary information for the backend classifier. The impact of the backend classifier for the case of the English dialects is not quite as large as for the other two cases. This result originally was considered to be due to the presence of the 320 out of corpus utterances in the evaluation set for English, but additional experiments conducted showed that eliminating the out of corpus utterances did not improved performance.

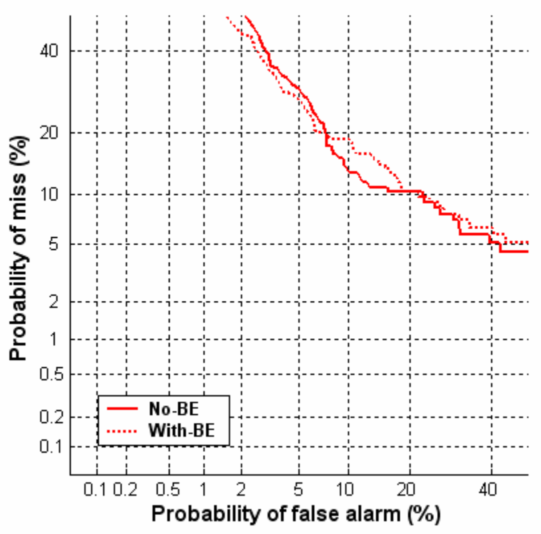

FIGURE 3. Comparison of results using a backend classifier in the case of Mandarin dialects.

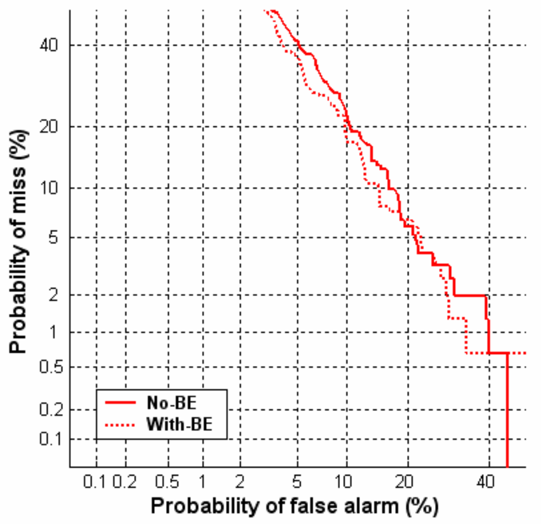

FIGURE 4. Comparison of results using a backend classifier in the case of Spanish dialects.

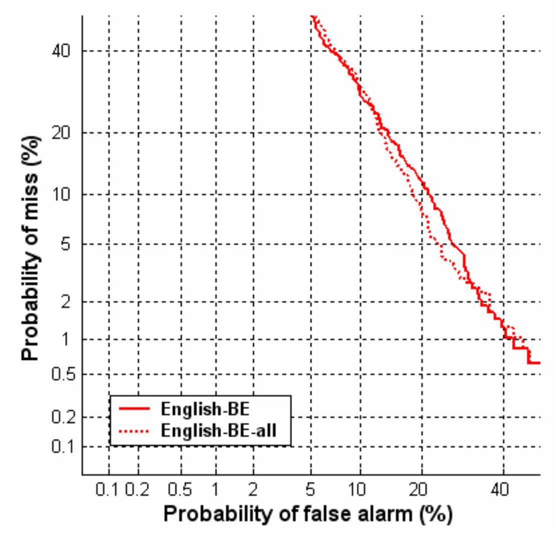

FIGURE 5. Comparison of the two backends evaluated for the English dialects.

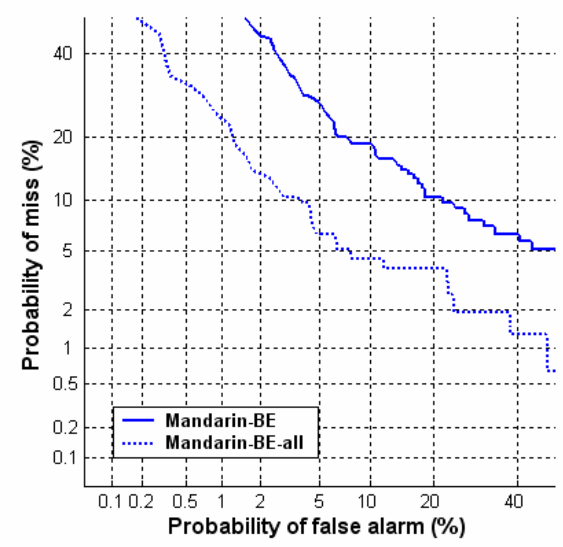

FIGURE 6. Comparison of the two backends evaluated for the Mandarin dialects.

5.CONCLUSIONS

This paper has presented results in the area of dialect identification by evaluating one of the current state of the art systems developed in the area of language identification. The results shown demonstrate the potential of the GMM technique as a candidate technique in the area of dialect identification.

The performance obtained for the GMM system evaluated in this paper is lower than that obtained in previous work with the Miami corpus[5] but the system provides very good performance for two of the dialects in the CallFriend corpus. Of particular importance is the fact that the technique has been ported without any specialization for the case of dialect identification and the results obtained are promising.

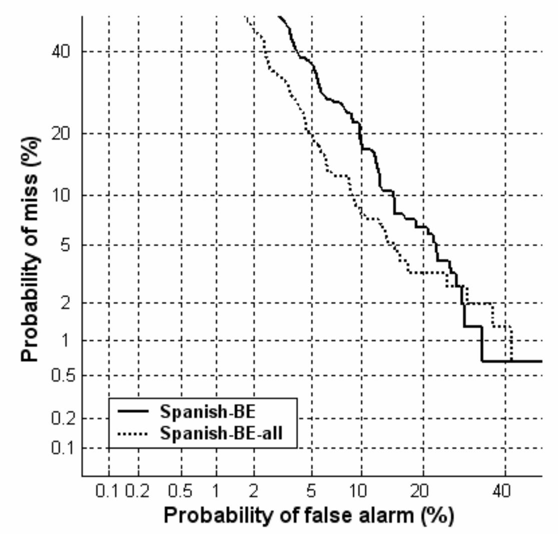

FIGURE 7. Comparison of the two backends evaluated for the Spanish dialects.

Future work in this area includes development of newer techniques and better understanding of the technique as it applies to the particular case of dialect identification, along with the evaluation of other language identification systems like PPRLM and SVM Additional work is expected to include evaluation of other algorithm developed for language identification along with better tuning of the algorithms for the problem at hand.

6.REFERENCES

[1]P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr., "Language Identification Using Gaussian Mixture Model Tokenization," at ICASSP, Orlando, Fl. , USA, 2002.

[2]P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features," at ICSLP, Denver, Co, 2002.

[3]E. Singer, P. A. Torres-Carrasquillo, T. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, Phonetic and Discriminative Approaches to Automatic Language Recognition," at EuroSpeech, 2003.

[4]"CallFriend Corpus," Linguistic Data Consortium, 1996.

[5]M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic Dialect Identification of Extemporaneous, Conversational, Latin American Spanish Speech," at ICASSP, Atlanta, Georgia, 1996.

[6]E. Wong and S. Sridharan, "Methods to Improve Gaussian Mixture Model Based Language Identification System," at ICSLP, Denver,CO, 2002.

[7]J. McLaughlin, D. A. Reynolds, and T. Gleason, "A Study of Computation Speed-UPS of the GMM-UBM Speaker Recognition System," at EuroSpeech, Seattle, Washington, 1999.