Eigen-channel Compensation and Discriminatively Trained Gaussian Mixture Models for Dialect and Accent Recognition*

Pedro A. Torres-Carrasquillo, Douglas Sturim, Douglas A. Reynolds and Alan McCree

MIT Lincoln Laboratory, Information Systems Technology Group, Lexington, MA, USA
{ptorres,sturim,dar,mccree}@ll.mit.edu

10.21437/Interspeech.2008-226

Abstract

This paper presents a series of dialect/accent identification results for three sets of dialects with discriminatively trained Gaussian mixture models and feature compensation using eigen-channel decomposition. The classification tasks evaluated in the paper include: 1) the Chinese language classes, 2) American and Indian accented English and 3) discrimination between three Arabic dialects. The first two tasks were evaluated on the 2007 NIST LRE corpus. The Arabic discrimination task was evaluated using data derived from the LDC Arabic set collected by Appen. Analysis is performed for the English accent problem studied and an approach to open set dialect scoring is introduced. The system resulted in equal error rates at or below 10% for each of the tasks studied.

1. Introduction

A dialect is “a regional variety of language distinguished by features of vocabulary, grammar, and pronunciation from other regional varieties and constituting together with them a single language” [1]. An automatic dialect identification (DID) system usually classifies the incoming speech utterances from a known language into one of the dialects, or more generally, classes of interest within that language.

In recent years, Gaussian mixture models (GMM) have been proposed as a basic component of DID systems. For example, Torres-Carrasquillo [2] used GMMs to study DID for English, Mandarin and Spanish. More recently Huang [3] has proposed discriminative training and frame selection techniques to improve the performance of DID for some dialects of Spanish and British English with some success.

Possible applications of DID include its use for surveillance and as a preprocessor for other automatic speech processing systems. For example, a DID system could be used as a preprocessor of speech utterances to be analyzed by an automatic speech recognition (ASR) system. Dialect specific speech recognition systems can then be used to process each utterance. A DID system could be employed in a similar fashion to preprocess speech as part of a speaker detection system.

In this paper, results are presented for a GMM system that employs recent advances in the areas of discriminative training using maximum mutual information (MMI) and feature compensation using eigen-channel compensation via factor analysis.

The outline of the paper is as follows: Section 2 describes the system including the feature processing and compensation. Section 3 describes the data and tasks. Section 4 presents results and discussion on each of the classification tasks. Section 5 presents an approach to open set dialect scoring and Section 6 describes conclusions and future work.

2. Dialect/Accent Identification System

The Gaussian mixture modeling (GMM) system used in this paper is based on a combination of recent advances in techniques developed for language and speaker recognition over the last few years: 1) GMM using shifted delta cesptra (SDC) features, 2) feature compensation using eigen-channel decomposition and 3) discriminative training.

2.1. Feature processing

The feature processing used follows the same general framework presented in similar systems previously described in [4, 5]. A frame rate of 10ms is used with a processing window of 20ms and processed through a filter bank with vocal tract length normalization (VTLN) warping factor α. The output is then filtered using RASTA and converted into cepstral coefficients. The set of cepstral coefficients are then concatenated using a 7-1-3-7 SDC parameterization [4] and combined with the static cepstral coefficients into a 56-dimensional feature vector. Non-speech feature vectors are then eliminated based on speech activity and the speech features are zero mean and unit variance normalized per file. The final set of feature vectors are then compensated using feature-based eigen-channel subtraction via factor analysis (described in the following section). The feature extraction chain is shown in Figure 1.

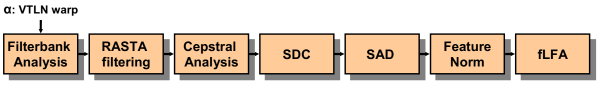

Figure 1. *Feature processing for GMM system.*

2.2. Eigen-channel compensation

In addition to VTLN, RASTA and mean/variance normalization, a feature domain implementation of eigen-channel compensation based on latent factor analysis (fLFA) is used to remove non-language dependent variations. This feature domain approach is inspired by the work described by Vair [6]. For the case of dialect identification, as in the case of speaker recognition, the session variability is modeled as an additive element to the means of the model:

$$m_i(D) = m + Ux(D) \qquad (1)$$

* This work was sponsored by the Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government

where $m_i(D)$ and m are vectors of stacked GMM means, often termed a supervector, $m_i(D)$ is the supervector for the i^{th} training session of dialect D and m is the session independent supervector from the GMM-SDC UBM. The U matrix employed for these experiments was obtained from the 40 highest energy eigen-vectors of the pooled correlation matrix from the 14 training languages and 6 dialects (20 classes) in the LRE-2007.

2.3. Gaussian mixture model with shifted delta cepstra

The GMM with shifted delta cepstral (GMM-SDC) coefficients system has been previously described in [4] and represents an attempt to incorporate temporal information into the feature vectors modeled by the GMMs. The system also uses a universal background model (UBM) approach for faster scoring as described by Wong [7] for the case of language identification.

In the results presented in this paper, the GMM UBM is trained using 10 iterations of EM and 5 iterations for adapting the UBM to each language specific model. Contrary, to the case of speaker identification, the adaptation process includes means, weights and variance given the amount of data available for training. As in the work in [4], the UBM is trained using data from all the classes.

2.4. Discriminative training

The discriminative training approach implemented for these experiments is built upon the system described by Matejka [8]. The implementation used for these experiments [9]differs from Matejka's in three areas: 1) the use of a common initial training model, instead of independent models, 2) the use of higher mixture orders and 3) the use of a top-M training scheme that allows for faster turnaround without any loss in performance. The training system also parsed all utterances into segments obtained from speech activity detection and discarded those segments below 2 seconds. The MMI training process used 10 iterations for training the target language models.

Besides the speed effects on training, using a GMM-UBM also allows the final system to use fast scoring [10] at test time.

2.5. Backend classifier

The backend classification scheme employed for the system in this paper is similar to that discussed by Singer [5], where a set of diagonal covariance Gaussian classifiers is trained for each class after preprocessing the input raw scores of the language models with linear discriminant analysis. This backend classifier is trained by a held out set different than the set of data used to train the language models.

3. Corpora and Evaluation Methods

The experiments presented in this paper show DID results for three groups of dialects. Two of these dialect discrimination tasks arise from the NIST-Language Recognition Evaluation (LRE) 2007 campaign [11]. The third dialect-discrimination task comes from the Linguistic Data Consortium (LDC) Arabic corpus and collected by Appen. The two LRE-2007 sets of languages used for the experiments are the Chinese discrimination problem and the accented English identification problems. The Chinese classes included in the LRE-2007 were Cantonese, Mandarin, MinNan and Wu. The English accents considered for the 2007 evaluation were the American and Indian accents. The LDC Arabic corpus includes three dialects: Gulf, Iraqi and Levantine.

In the case of the LRE-2007 dialects, various sources of data are employed for training the language models and backend classifier. Training data for three of the Chinese-classes language models is derived from the LDC development data for Cantonese, MinNan and Wu. The Mandarin-class language model used CallFriend, Mixer, CallHome and a subset of the Fisher corpus. The American accented English language model leveraged a subset of the Fisher corpus along with CallFriend, Mixer and CallHome. Indian accented English models were trained from the Mixer, CallFriend and OGI Foreign accented English (FAE) collections. Additionally, the Indian accented English models are further augmented using Hindi and Tamil training data from both the CallFriend corpus and OGI-FAE.

For the backend classifier training for Chinese, data is derived from the LDC-2007, Fisher, OGI-22 and OHSU. In the case of English, the backend is trained with data from Fisher, CallFriend, Mixer, CallHome, OGI-22 and OHSU.

The evaluation data for the Chinese class discrimination includes 80 cuts for Cantonese, Min and Wu and 158 cuts for Mandarin. In the case of English, the evaluation data includes 80 utterances for the Indian English class and 160 utterances for the American English class.

For the Arabic dialects, the only source of data used is the LDC corpus. In this case, the available data of about 1000 speakers for Gulf and Levantine and 300 Iraqi speakers is divided into three partitions, with 60% of the data used for training, 20% of the data used for backend classifier training and the remaining 20% used for system evaluation. The actual numbers used are 1465 training utterances (586 for Gulf Arabic, 289 for Iraqi Arabic and 590 for Levantine Arabic). The test data includes 390 Gulf Arabic cuts, 189 Iraqi Arabic cuts and 395 Levantine Arabic cuts. The test set is then divided in two to produce the backend classifier training set and the system evaluation set.

4. Results

The results for the different GMM systems are shown via DET plots and equal error rates (EER). The results presented are pooled across classes and with class test priors equalized to discount the impact of having different number of inputs for each class. The GMM mixture order for the system presented is 2048. Results for all three classification tasks are shown for 30-second test utterances. For each task there are three experiments: 1) a GMM-UBM baseline system, 2) a GMM-UBM with VTLN and fLFA, and 3) a GMM-UBM with VTLN and fLFA plus 10 MMI iterations.

The general behavior for all three classification tasks, shown in Figure 1, 2 and 3, is similar, with the fLFA-MMI system showing the best performance. However, it is interesting to notice that improvements in performance are not necessarily comparable in each case. For example, in the case of the Chinese classes most of the gain observed over the baseline GMM-UBM system is due to the feature compensation, fLFA and VTLN, and only minimal gains are shown for the discriminative training. The equal error rates for this task are 16%, 8% and 7% for the baseline, fLFA and fLFA-MMI systems respectively.

For the English accent task, the fLFA provides some gains over the baseline system with the fLFA-MMI also resulting in a small additional gain. The equal error rates for these tasks are 14.4%, 11.3% and 10.6%.

In the case of the Arabic dialect task, the behavior observed for the experiments is different. The improvements for the fLFA system over the baseline and for the fLFA-MMI combination over the fLFA system are clearly observed. The EERs for this task are 18%, 12% and 7%.

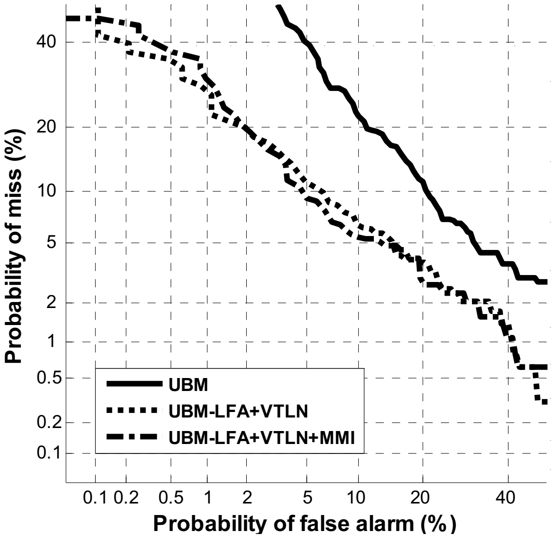

Figure 2. *Results for Chinese task on 30s test utterances.*

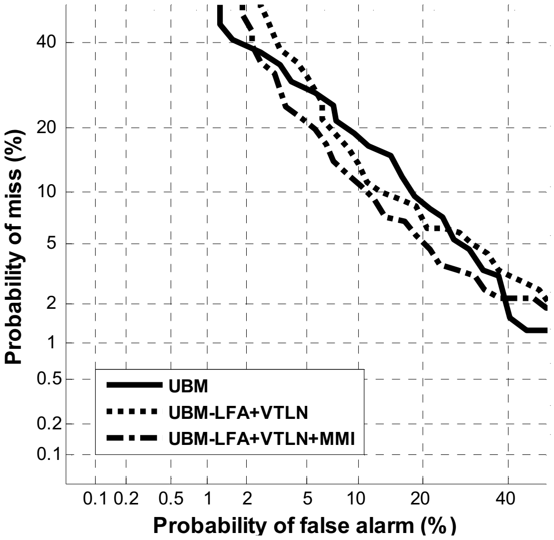

Figure 3. *Results for English task on 30s test utterances.*

The fact that the behavior of the system is different for each task is possibly due to the different number of sources used in training the language models for each task and the mismatch between the training and testing data.

In the case of feature compensation, the use of multiple sources for training the language models could be a factor if some of the sources are underrepresented and the channel variation for an underrepresented source is not properly captured. Sources not well represented in the training of the fLFA could explain the minor gains observed for the English task, where multiple sources (not balanced across different corpora) are used for training the language models, and the training and testing sources are highly mismatched.

In the Chinese task, where the training and testing conditions are well matched and the training comes mainly from a single corpus, the feature compensation provides most of the gains. For the Arabic task, where there is a single training corpus, and training and testing conditions are well-matched improvements are observed for the feature compensated system. Additionally, it should be clear that it could be possible to tune the compensation matrix to each of the particular tasks assuming sufficient training data are available.

The results for the discriminatively trained models can also be related to the same issues discussed for the feature compensation systems. An additional problem that could explain the minor gains obtained for this stage in the Chinese task is the amount of data available for training. While the Mandarin class had multiple data sources and a large number of training examples, the other three classes only included about 20 utterances for training.

Another important issue is how the results presented in this paper compare to other systems on these tasks. Although such results are not available for the Arabic dialects task, two other system results are available for the Chinese and English task. The MIT LL system results for a SVM-GMM [12] system and for a SVM-N-gram system [13] are available and summarized in Table 1. The results show that the system presented in this paper is highly competitive to other state-of-the-art systems.

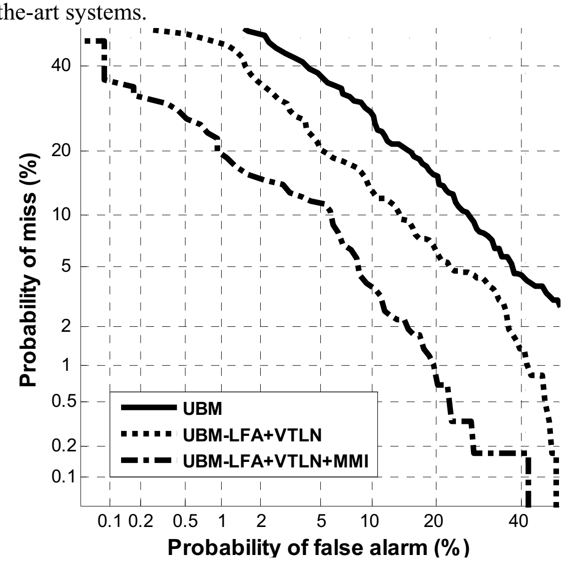

Figure 4. *Results for Arabic task on 30s utterances.*

System	Chinese EER(%)	English EER(%)
SVM N-gram	7.6	10.9
GMM-SVM	7.0	11.3
GMM-UBM w/fLFA-MMI	7.0	10.6

Table 1. *Comparison of different system on NIST-LRE 2007 Chinese and English classification tasks.*

4.1. Analysis of English accent task

Although there are three major tasks considered in this paper, the English accent task is the only one that can be analyzed in additional detail by the authors as speakers of the language. The best system obtained for the English task, the GMM-SDC with fLFA and MMI, resulted in an EER of 10.6%.

Three major sources of error are observed: Indian females producing very high scores on the American model (12 cases), African-Americans producing low American model scores (6 cases) and utterances where no perceivable Indian accent exist been misclassified as American English.

In both the Indian females case and the African American case, there is no information available to determine the richness of the training data for these sub-classes. However, for the Indian females case there is some anecdotal evidence that this sub-class is underrepresented in the CallFriend corpus. No information is available about the African-American representation on the training data.

For the case of utterances without any perceivable Indian accent, this seems like a labeling error that could be due to human error or to incorrect information gathered at the collection stage.

5. Open Set Dialect Scoring

This paper has shown results for three classification tasks using a closed-set paradigm. In this section, the “open set” dialect problem, or dialect detection in the presence of multiple languages, is addressed. For the NIST LRE07, a method of generating open set scores without training any new systems or back-ends was developed. To do this, the fact that calibrated posterior probabilities are available for both the closed set dialect and open set language tasks is exploited. Since the closed set of dialects spans the language in question, an utterance that is out of set for the dialect test must be in another language. In other words, the open set dialect posterior $P(D\mid X)$, is equal to the closed set dialect posterior $P(D\mid L,X)$, multiplied by the open set language posterior $P(L\mid X)$, or

$$P(D\mid X)=P(D\mid L,X)P(L\mid X) \qquad (2)$$

Since the NIST LRE07 priors are different for these two cases, the language posterior is adjusted to match the dialect out of set prior.

The results in Table 2, for the fused system in MIT-LL submission to the NIST 2007 LRE for four of the dialect tasks, show that this approach provides good performance in terms of the hard decision error rate Cavg (which are similar to the systems’ EER). In fact, the open set dialect task is easier than the closed set task, presumably because rejecting the non-language utterances is easier than discriminating between the dialects. With this approach, the open set dialect problem can be solved with no new information other than closed dialect and open set language scores.

	Cavg x 100	
Dialect	**Closed set**	**Open set**
Chinese	4.9	4.3
English	8.8	7.9
Spanish	35.6	28.9
Hindustani	35.5	29.5

Table 2. *Comparison between closed- and open-set scores for the MIT-LL submission to the LRE-2007.*

6. Conclusions and Future Work

In this paper, results for three sets of dialects have been presented with two of these sets coming from the NIST-LRE 2007 along with introducing an approach for open set dialect scoring. Out of the three classification tasks, the GMM system studied results in single digit equal error rates for two of the three tasks and competitive performance with other state of the art dialect identification systems.

Additionally, some analysis of the American versus Indian accented English problem was presented and evidence shows the need for both richness and balance in order to study this problem and be able to produce accurate results. One issue that is still not well understood is the false alarm rates produced by female speakers with Indian accent in the English task.

Although, our system showed excellent performance for these three classification tasks there are still a number of issues that need to be studied further. For example, the results observed show that the improvements provided by feature compensation are different across the tasks and this could be due to multiple reasons including the number of sources, the lack of tuning to the specific channels used in the tasks and the role of VTLN independent from fLFA. Also, it may be possible to improve the discriminative training process by judiciously selecting the training utterances presented at each training iteration. The amount of data needed for discriminative training for the dialect tasks also needs to be examined.

7. Acknowledgements

The authors wish to thank their colleagues Fred Richardson and Bill Campbell for providing results for the NIST LRE 2007 SVM and SVM-GMM systems.

8. References

[1] Merriam-WebsterOnlineDictionary. 2008; Available from: http://www.merriam-webster.com/dictionary/dialect.

[2] Torres-Carrasquillo, P.A., T.P. Gleason, and D.A. Reynolds, "Dialect Identification Using Gaussian Mixture Models", in *Odyssey 2004.* Toledo, Spain.

[3] Huang, R. and J.H.L. Hansen, *Unsupervised Discriminative Training with Application to Dialect Classification.* IEEE Transactions on Audio, Speech, and Language Processing, 2007. **15**(8): p. 2444-2453.

[4] Torres-Carrasquillo, P.A., et al. "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features". In *ICSLP*. 2002. Denver, CO.

[5] Singer, E., et al. "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition". In *EuroSpeech*. 2003. Geneva, Switzerland.

[6] Vair, C., et al. "Channel factors compensation in model and feature domain for speaker recognition". In *IEEE Odyssey:* 2006. San Juan, PR.

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

[8] Matejka, P., et al. "BRNO University of Technology System for NIST 2005 Language Recognition Evaluation". In *IEEE Odyssey*. 2006. San Juan, PR.

[9] Torres-Carrasquillo, P.A., et al., "The MITLL NIST LRE 2007 Language Recognition System", in *InterSpeech*. 2008 Australia.

[10] Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker Verification using Adapted Gaussian Mixture Models.* Digital Signal Processing, 2000. **10**(1-3): p. 19-41.

[11] *NIST LRE-2007 Evaluation Plan*. Available from: http://www .nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf.

[12] Campbell, W.M., et al. "SVM based Speaker Verification using GMM Supervector Kernel and NAP Variability Compensation". In *ICASSP*. 2006. Tolouse, France.

[13] Richardson, F.S. and W.M. Campbell. "Language Recognition With Discriminative Keyword Selection". In *ICASSP*. 2008. Las Vegas, NV.