Gaussian Mixture Selection and Data Selection for Unsupervised Spanish Dialect Classification

Rongqing Huang, John H.L. Hansen

Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Sciences
University of Texas at Dallas, Richardson, TX, USA
huangr@colorado.edu, john.hansen@utdallas.edu

Abstract

Automatic dialect classification has gained interests in the field of speech research because it is important to characterize speaker traits and to estimate knowledge that could improve integrated speech technology (e.g., speech recognition, speaker recognition). This study addresses novel advances in unsupervised spontaneous Latin American Spanish dialect classification. The problem considers the case where no transcripts are available for train and test data, and speakers are talking spontaneously. A technique which aims to find the dialect dependence in the untranscribed audio by selecting the most discriminative Gaussian mixtures and selecting the most discriminative frames of speech is proposed. The Gaussian Mixture Model (GMM) based classifier is retrained after the dialect dependence information is identified. Both the MS-GMM (GMM trained with Mixture Selection) and FS-GMM (GMM trained with Frame Selection) classifiers improve dialect classification performance significantly. Using 122 speakers across three dialects of Spanish with 3.3 hours of speech, the relative error reduction is 30.4% and 26.1% respectively.

Index Terms: Gaussian mixture selection, dialect classification, accent classification, data selection, GMM

1. Introduction

Dialect/accent is a pattern of pronunciation and/or vocabulary selection in a language used by the community of native/non-native speakers belonging to some geographical region. Dialect is one of the most important factors next to gender that influence automatic speech recognition (ASR) performance [3, 4]. Dialect knowledge could be used in various components of the ASR system such as pronunciation modeling [9], lexicon adaptation [12], and acoustic model training [7] and adaptation [2]. Dialect knowledge could be applied in automatic call center and directory lookup service [14].

Our efforts on dialect identification focus on classifying unconstrained audio, which means unknown gender, unknown speaker, and unknown text. If transcripts exist for the associated training audio, we have previously proposed a word-based dialect classification (WDC) algorithm which turns the text-independent dialect classification problem into a text-dependent dialect classification problem [5] and achieves very high classification accuracy. If the training data size is too small to train the word specific models, a context adaptive training (CAT) algorithm was proposed to solve this problem and also achieves high classification accuracy [6]. If there are no transcripts in the training data, the above algorithms cannot be applied, and therefore an unsupervised algorithm must be formulated. The Gaussian Mixture Model (GMM) based classifier has been applied to unsupervised dialect classification [11] and text-independent speaker recognition [10] successfully. In this study, the GMM-based algorithm is the baseline system.

Spanish dialectology is fundamentally different than English dialectology. Spanish dialects are concentrated on certain phonemes being at certain positions [1, 8]. For example, /s/ at syllable final is dropped by Cuban and is reinforced by Peruvian. The GMM classifier trained and tested on human labels which only includes this information can achieve 98% accuracy on 20-second audio files [13]. The focus in this study is to identify that part of the audio which corresponds to the dialect difference in order to produce the most discriminative model. This is the core idea of data selection or frame selection for model training in dialect classification. The Gaussian mixtures are applied to model the acoustic space of the training data. We expect that some of the mixtures correspond to the dialect difference in Spanish, while other mixtures correspond to the non-dialect-dependent acoustic events. If we can identify the dialect related Gaussian mixtures and use them to form a new model, it will be a more discriminative model. This represents the idea behind Gaussian mixture selection for model training in dialect classification.

2. Baseline Classification Algorithm

Since there are no transcripts for the train and test data, it is difficult to build a supervised generative model such as an HMM. The GMM classifier is a popular method for text-independent speaker recognition [10] and dialect classification [11]. We use the GMM classifier as our baseline system. Fig. 1 shows the block diagram of the baseline GMM training system, where N is the number of pre-defined dialects. The GMM model for dialect i is trained with data from dialect i. A GMM based gender classifier is trained similarly and is applied prior to dialect classification. Fig. 2 shows the block diagram of the unsupervised GMM based dialect classification system. We will describe the silence remover and feature extraction in the experimental section.

3. Gaussian Mixture Selection on GMM Re-training (MS-GMM)

The primary differences for most Spanish dialects are on certain phonemes at certain positions [1, 8]. For example, /s/ at the syllable final position is dropped by Cuban and is reinforced by Peruvian. The GMM classifier trained and tested on human labels which only includes the above information can achieve

This work was supported by US Air Force under contract No. FA8750-04-1-0058.

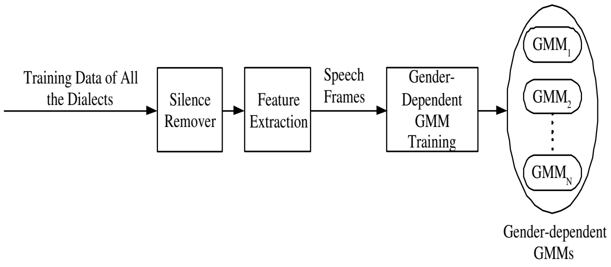

Figure 1: *Baseline GMM based unsupervised training system*

98% accuracy on 20-second audio files [13]. The Gaussian mixtures represent the acoustic space of the training data. We expect some Gaussian mixtures will represent dialect-dependent acoustic characteristics, and others to be less dialect-dependent and could in fact cause confusion for dialect classification. In this section, we will formulate a scheme to detect the most dialect-dependent Gaussian mixtures and sort the mixtures according to their discriminating abilities. The new GMM is then obtained by selecting the top discriminative mixtures.

Let Λ_i be the GMM of dialect i, and $\lambda_{ij} \sim \mathcal{N}(w_{ij}, \mu_{ij}, \Sigma_{ij})$ be the jth Gaussian mixture of Λ_i, where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, M$, and N is the number of dialects, M is the number of mixtures for each GMM.[1] The number of speech frames in the training data is T_i for dialect i, and the total number of speech frames in the training data is $T = \sum_{i=1}^{N} T_i$. The discriminating ability of the Gaussian mixture λ_{ij} is defined as,

$$\zeta_{ij} = \sum_{t=1}^{T} \delta_{ijt} Pr(\mathbf{o}_t|\lambda_{ij}), \tag{1}$$

where $Pr(\mathbf{o}_t|\lambda_{ij})$ is the weighted probability of mixture λ_{ij} generating the speech frame $\mathbf{o}_t$, which is defined as,

$$\begin{aligned} \Pr(\mathbf{o}_t|\lambda_{ij}) = w_{ij}(2\pi)^{-n/2}|\Sigma_{ij}|^{-1/2} \times \\ \exp\{-\frac{1}{2}(\mathbf{o}_t - \mu_{ij})'\Sigma_{ij}^{-1}(\mathbf{o}_t - \mu_{ij})\}. \end{aligned} \tag{2}$$

Here, δ_{ijt} in Eq. 1 is defined as,

$$\delta_{ijt} = \begin{cases} 1 & \text{if } i = \arg\max_{c} Pr(\mathbf{o}_t|\Lambda_c), \\ & \text{and } \{i,j\} = \arg\max_{\{c,d\}} Pr(\mathbf{o}_t|\lambda_{cd}), \\ & \text{and } \mathbf{o}_t \in \text{ class } i; \\ -1 & \text{if } i = \arg\max_{c} Pr(\mathbf{o}_t|\Lambda_c), \\ & \text{and } \{i,j\} = \arg\max_{\{c,d\}} Pr(\mathbf{o}_t|\lambda_{cd}), \\ & \text{and } \mathbf{o}_t \notin \text{ class } i; \\ 0 & \text{else.} \end{cases} \tag{3}$$

where $c = 1, 2, \ldots, N$, $d = 1, 2, \ldots, M$, n is the number of dimensions of the feature vector, and

$$Pr(\mathbf{o}_t|\Lambda_c) = \sum_{d=1}^{M} Pr(\mathbf{o}_t|\lambda_{cd}). \tag{4}$$

The larger the value of ζ_{ij}, the larger the discriminating ability of the jth mixture in the ith GMM. For each GMM, the mixtures are sorted based on the discriminating ability measure. The new GMM is formulated by selecting the top discriminative mixtures in the old GMM and the weights are recalculated in order to ensure $\sum_j w_{ij} = 1$ in the new GMM. The evaluation process is exactly the same as the baseline GMM classification system. In our study, we formulated four variations of the above scheme. If we remove the probability term in the right hand side of Eq. 1, ζ_{ij} is actually the discriminative speech frame count for jth mixture in the ith GMM. We name the original ζ_{ij} as the probability score. The frame count and probability score calculated above are the raw values. If we remove the -1 term in Eq. 3, the calculated frame count and probability score are referred to as the normalized values. In this case, ζ_{ij} is always non-negative. In summary, the Gaussian mixture discriminating ability is measured in four difference scores: the raw frame count, the normalized frame count, the raw probability score, and the normalized probability score.

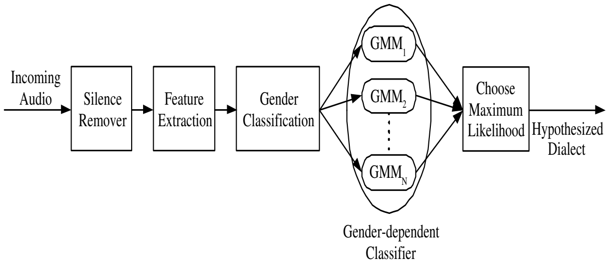

Figure 2: *The GMM based unsupervised dialect classification system*

4. Frame Selection on GMM Re-training (FS-GMM)

In addition to extracting the dialect-dependent information in the acoustic space represented by the Gaussian mixtures, we can extract the dialect-dependent discriminative information in the training data directly. Alternatively, we can remove the most confusing speech frames in the training data and train new GMMs with the remaining data (i.e., this algorithm is termed FS-GMM). The speech frame $\mathbf{o}_t$ is called a confusing frame if

$$\begin{aligned} i = \arg \max_{c=1,\ldots,N} Pr(\mathbf{o}_t|\Lambda_c), \\ \text{and } \mathbf{o}_t \notin \text{ class } i. \end{aligned} \tag{5}$$

When the number of consecutive confusing frames over time is greater than a pre-defined threshold (0.1 s in our study), we believe these frames to be "garbage" frames. After removing all the garbage frames, a new set of GMMs (i.e., discriminative GMMs) are trained by using the remaining (i.e., discriminative) speech frames. The garbage frames from all the dialects are grouped together and a garbage GMM is trained. The final GMM is obtained by combining the discriminative GMM with the garbage model. There are prior probabilities for model combining, which determines how much weight will be assigned to the discriminative GMM and the garbage GMM. Fig. 3 shows the block diagram of the GMM re-training based on frame selection.

In the combined new model, the Gaussian mixtures from the garbage model will not help discriminate the classes, but they will contribute to map the confusing acoustic events. The dialect-dependent acoustic events are mapped to the Gaussian mixtures from the discriminative model.

[1] The number of mixtures for each GMM can be different. We use the same number of mixtures for the GMMs of the dialects in our study.

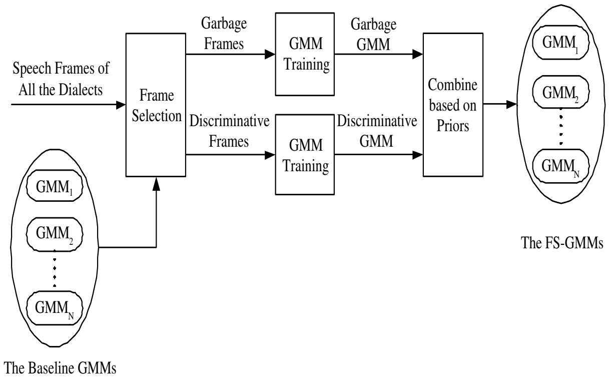

Figure 3: *Discriminative GMM training based on frame selection*

5. Experiments

The corpus used in our study consists of the Latin American Spanish dialect speech from Cuba, Peru and Puerto Rico, which is described in [14]. The spontaneous speech portion was recorded in an interview style. The interviewer gave sample topics such as "describe your family", and the subject would respond. The interviewer would give some hints to keep the subject talking. The subject used a head-mounted microphone, which also captured the speech from the interviewer at a much lower amplitude since the interviewer sat far away from the microphone. The speech from both the interviewer and the subject were recorded on the same channel. We note that there were long periods of silence in the audio. To address the above problems, we build a silence remover to eliminate the long silence and the speech from the interviewer. The silence remover is based on an overall energy measure. Table 1 summarizes information of the training and test data after the silence removal process. PR is the dialect from Puerto Rico. The speakers used for training and testing have roughly the same number of male and female speakers. Since the size of the corpus is limited, we did not set aside data for development use, and therefore all of the data is used either in training or test. We will illustrate several combinations of parameters in the experimental results. In an actual application, a development data set can be applied to select the best parameters.

Table 1: *The 3-dialect Spanish corpus used in our study*

Data	Training data			Testing data		
	Cuba	Peru	PR	Cuba	Peru	PR
Speakers	29	29	26	13	13	12
Minutes	52	53	36	21	23	17

The first experiment is to determine which discriminative measure in Sec. 3 can sort the mixtures consistently for both the training and unseen data. Therefore, the most confusing mixtures should be excluded and a new set of GMMs can be generated. Fig. 4 shows the mixture sorting using the four discriminative measures described in Sec. 3 (raw frame count, normalized frame count, raw probability score and normalized probability score). The points in the figures are the numerical labels of mixtures. Different colors mean different dialect classes (blue: Cuba; green: Peru; red: Puerto Rico). The X and Y axes represent the number of mixtures. We use 200 mixtures for the baseline GMM training in Fig. 4, Fig. 5 and Table 2. Fig. 4 is generated by first using the training data as the input to sort the mixtures of the baseline GMMs and obtain new GMMs, then using the testing data as the input to sort the mixtures of the new GMMs, the sequence of mixtures are drawn in the figure. If the top discriminative mixtures in the training data are also the top discriminative mixtures in the testing data, the points in the figure will close to the line $y = x$. If the top discriminative mixtures in the training data are the least discriminative mixtures in the testing data, the points will close to the line $y = 200 - x$. The ideal case is that all the points are on the line $y = x$. From Fig. 4, we observe that mixture sorting based on the normalized probability score can identify the top discriminative mixtures in a very consistent style for both training and testing data. It also shows that the development data is not necessary for mixture sorting, the Gaussian mixtures can be sorted using the training data on original GMMs. Table 2 shows the classification accuracy of the mixture selected GMM (MS-GMM) with the four discriminative measures. We pick the top 75% of the mixtures for the new GMMs. We observe that all four mixture selection schemes can improve the classification accuracy. The normalized probability score is the best scheme for sorting the mixtures. We will use this scheme for the following experiments.

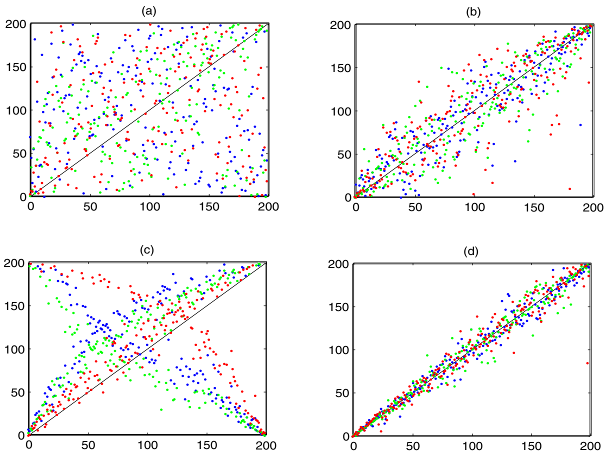

Figure 4: *Gaussian mixture sorting using the four discriminative measures. (a): based on raw frame count; (b): based on normalized frame count; (c): based on raw probability score; (d): based on normalized probability score.*

Table 2: *Classification accuracy of the MS-GMMs formulated by different schemes, the top 75% mixtures are selected. The labels (a)-(d) are the same as the labels in Fig. 4.*

Baseline	(a)	(b)	(c)	(d)
73.5%	75.5%	78.1%	76.4%	81.6%

Next, we show how the percentage of selected mixtures affects classification performance. Fig. 5 shows the classification accuracy of the MS-GMM formulated by picking different percentages of the top mixtures. When the top 100% of the mixtures are selected, the system becomes the baseline GMMs. From Fig. 5, we observe that (i) we have to keep at least 50% of the original mixtures; (ii) the classification performance improves if the least discriminative mixtures (i.e., the most confusing mixtures) are removed.

We also train 300, 600 and 1000 mixtures for the baseline GMMs, and the top 50% to 100% of the mixtures are selected based on the normalized probability score measure. The results are shown in Fig. 6. When the 100% of mixtures are selected, it actually is the baseline GMMs. The classification accuracy of the baseline 300, 600, and 1000 mixtures GMM is 73.7%, 74.3% and 73.7% respectively as shown in Fig. 6. We observe several interesting results: (i) when picking 50% to 95% of the sorted mixtures, the new GMMs outperform the baseline GMMs; (ii) if more mixtures are in the baseline GMMs, a smaller portion of the sorted mixtures are required to obtain the best GMMs. In Fig. 6, we observe that when 65% of the mixtures are selected for 1000-mixture baseline GMMs, the new GMMs achieve the best performance; for the 600-mixture GMMs and 300-mixture GMMs, the new GMMs achieve the best performance when 75% and 80% of the mixtures are selected respectively.

Having established the procedure for discriminative mixture selection, we can apply the same concepts for frame selection. The Frame selection based GMM (FS-GMM) retraining method will obtain the discriminative model and the garbage model. The final model is obtained by combining the discriminative model and the garbage model with pre-defined prior probabilities. We still use the 300, 600 and 1000 mixtures for the baseline GMMs in this experiment. Fig. 7 shows the classification accuracy of the combined model generated in different prior probabilities. From Fig. 7, we observe interesting results: (i) the discriminative model alone only achieves marginal improvement (i.e., the prior probability is 1 for the discriminative model in Fig. 7); (ii) the "garbage" model will help the dialect classification.

6. Conclusions

Dialect differences are reflected using a range of acoustic events. These acoustic events can be seperated into dialect discriminating content, and dialect neutral/distractive content. In this paper, we have proposed the mixture selection and frame selection algorithms (i.e., MS-GMM and FS-GMM) to identify dialect dependent structure. The retrained GMMs significantly outperform baseline GMMs. In the 600-mixture GMMs, the baseline model achieves 74.3% accuracy; the MS-GMM achieves 82.1% accuracy; and the FS-GMM achieves 81.0% accuracy. The relative error reduction is 30.4% and 26.1% respectively. This advancement has therefore taken an important step towards effective dialect classification while maintaining consistent level of memory/computing reources.

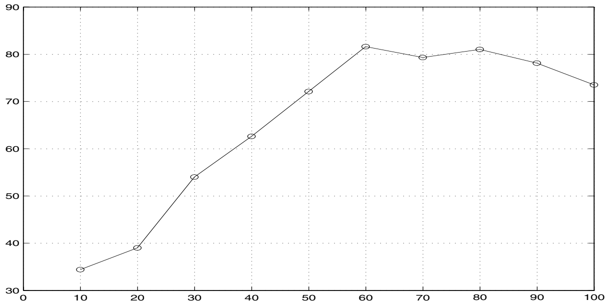

Figure 5: *Classification accuracy of the MS-GMMs by selecting different percentage of discriminative mixtures. X-axis: percentage of discriminative mixtures being selected (from 10% to 100%); Y-axis: classification accuracy (%).*

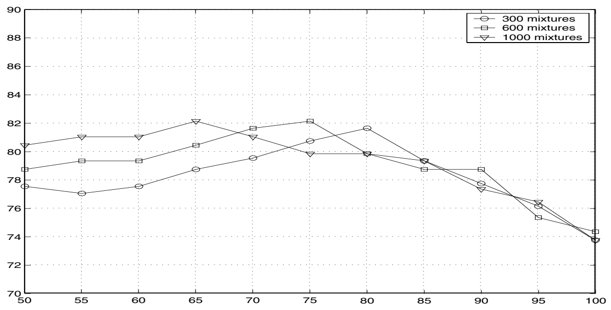

Figure 6: *Classification accuracy of the MS-GMMs by selecting different percentage of discriminative mixtures. The initial GMM has 300, 600 and 1000 mixtures respectively. X-axis: percentage of discriminative mixtures being selected (from 50% to 100%); Y-axis: classification accuracy (%).*

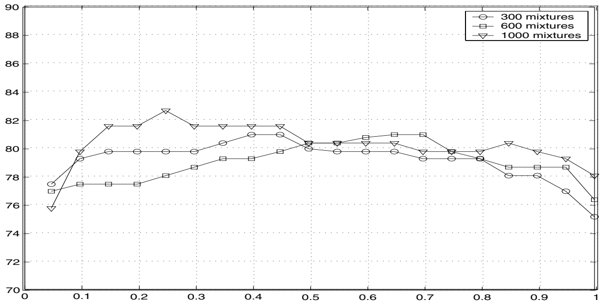

Figure 7: *Classification accuracy of the FS-GMMs by setting different prior probabilities for discriminative model and garbage model.The initial GMM has 300, 600 and 1000 mixtures respectively. X-axis: prior probability for the discriminative model in the combination (from x= 0.05 to 1), the prior probability for the garbage model is 1-x; Y-axis: classification accuracy (%).*

7. References

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

[2] V. Diakoloukas, V. Digalakis, L. Neumeyer, and J.Kaja, "Development of Dialect-Specific Speech Recognizers using Adaptation Methods", in *Proc. ICASSP* , vol.2, pp.1455-1458, Munich, Germany, April, 1997

[3] V. Gupta and P. Mermelstein, "Effect of Speaker Accent on the Performance of a Speaker-Independent, Isolated Word Recognizer", in *Journal of Acoustic Society of America*, vol.71, pp.1581-1587, 1982

[4] C. Huang, T. Chen, S. Li, E. Chang and J. L. Zhou, "Analysis of Speaker Variability", in *Proc. EuroSpeech*, vol.2, pp.1377-1380, Sep, 2001

[5] R. Huang and J.H.L. Hansen, "Dialect/Accent Classification via Boosted Word Modeling", in *Proc. ICASSP*, Philadelphia, USA, March, 2005

[6] R. Huang and J.H.L. Hansen, "Advances in Word based Dialect/Accent Classification", in Proc. *Interspeech*, Lisbon, Portugal, Sept, 2005

[7] J. J. Humphries and P. C. Woodland, "The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training", in *Proc. ICASSP*, 1998

[8] J. M. Lipski, "Latin American Spanish", *Longman*, London, UK, 1994

[9] M. K. Liu, B. Xu, T. Y. Huang, Y. G. Deng, and C. R. Li, "Mandarin Accent Adaptation based on Context-Independent/ Context-Dependent Pronunciation Modeling", in *Proc. ICASSP*, 2000

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

[11] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect Identification Using Gaussian Mixture Models", in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004

[12] W. Ward, H. Krech, X. Yu, K. Herold, G. Figgs, A. Ikeno, D. Jurafsky, and W. Byrne, "Lexicon Adaptation for LVCSR: Speaker Idiosyncracies, Non-Native Speakers, and Pronunciation Choice", in *Proc. ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*, Colorado, Sep, 2002

[13] L. R. Yanguas, G. C. O'Leary, and M. A. Zissman, "Incorporating Linguistic Knowledge into Automatic Dialect Identification of Spanish", in *Proc. ICSLP*, Syndey, Australia, November, 1998

[14] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, "Automatic Dialect Identification of Extemporaneous Conversational, Latin American Spanish Speech", in *Proc. ICASSP*, 1996