A SYSTEMATIC STRATEGY FOR ROBUST AUTOMATIC DIALECT IDENTIFICATION

*Gang Liu, John H. L. Hansen**

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas , Richardson, Texas 75083, USA
gang.liu@student.utdallas.edu, John.Hansen@utdallas.edu

ABSTRACT

Automatic dialect Classification is very important for speech based human computer interface and customer electronic products. Although many studies have been performed in ideal environment, little work has been done in noisy or small data corpus, both of which are very critical for the survival of a dialect identification system. This paper investigates a series of strategies to address the question of small and noisy dataset dialect classification task. A novel hierarchical universal background model is proposed to address the question of limited training dataset. To address the noisy question, we initiate the use of perceptual minimum variance distortionless response (PMVDR), combining with shifted delta cepstral (SDC) algorithm. Rotation forest is also explored to further improve the system performance. Finally, compared with the baseline system, the proposed best system shows relative gains of 31.8% and 28.7%, in the worse noise and clean condition on a small data set, respectively.

1. INTRODUCTION

Accent/Dialect identification (DID) has recently emerged to be of substantial interest in the speech processing community [1]. DID systems can be used to improve the performance of Automatic Speech Recognition (ASR) engines based humane computer interface (HCI) by employing dialect dependent acoustic and language models. Traditional speech recognition systems are not robust to variations due to speaker dialect/accent. Dialect classification is one solution which can characterize speaker traits and help in the development/selection of dynamic lexicons by selecting alternative pronunciations, generate pronunciation modelling via dialect adaptation, or train and adapt dialect dependent acoustic models. Dialect knowledge is also helpful for data mining and spoken document retrieval. In this study, we employ the definition for the term accent/dialect as a pattern of pronunciation and/or vocabulary of a language used by the community of native speakers belonging to some geographical region.

In dialect identification, although with huge and noise free data corpus, dialects can be identified with high accuracy, system recognition performance decreases drastically for noisy or small training data set task. Although many researches have been investigated in the ideal scenario [$1 \sim 4$], there is little work has been done on noisy and small training data corpus. Many real life applications are con-strained by limited data collection, and may need to be deployed in noisy environment after development. This paper attempts to address this challenge in a systematic method.

The rest of this paper is organized as follows. In Section 2, we describe the database that is used for system development and verification. The baseline system, which serves as the workbench for our proposed advances, is described in Section 3. Next, we introduce a series of schemes to address the small data-set and noise disturbance challenge in Section 4, with one ultimate goal: promote an identification rate. Section 5 describes the experiment set up for algorithm verification and results analysis. Summary and conclusions are shown in Section 6.

2. CORPUS

The corpus used in our study is a Latin American Spanish accent speech database with three different accents from Cuba, Peru and Puerto Rico (PR). There are no associated transcripts for the training or test data in all accents. All the data are spontaneous speech (i.e., one person from the dialect region talking spontaneously), which were recorded in an interview style. The interviewer gave sample topics such as "describe your favourite movie", and the subject would respond. The interviewer would give some hints during the collection in order to keep the subject talking smoothly. The subject used a head-mounted microphone, which also captured the speech from the interviewer at a much lower amplitude since the interviewer sat across from the subject and far away from the microphone. Table 1 summarizes the data used for system development and evaluations.

Table 1: THREE-ACCENT SPANISH CORPUS

Data	Training Set			Testing Set		
	Cuba	*Peru*	*PR*	*Cuba*	*Peru*	*PR*
speaker#	29	29	26	13	13	12
Gender (female, male)	(14,15)	(12,17)	(18,8)	(7,6)	(8,5)	(5,7)
Total Duration (min.)	52	53	36	21	23	17

3. BENCHMARK SYSTEM

The Gaussian Mixture Models (GMM) classifier is a popular method for text independent speaker recognition and has been used for language identification and DID. We use this approach as our baseline system. Figure 1 shows the block

*This project was funded by AFRL through a subcontract to RADC Inc. Under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

diagram of the baseline GMM training/testing system. The silence removal module sets aside silence in the audio files that are used for training and testing. All the acoustic files are provided with gender information and we only use it for in the universal background model (UBM) training and adapting process. Next, gender and dialect dependent GMM are trained for each dialect category. While testing, the incoming audio is classified as a particular dialect based on the maximum posterior probability measure over all the possible GMM candidates. Throughout this paper, we use 128 mixtures for all GMM to provide a fair basis for comparison of different feature extraction front-end and backend classifiers.

Many researchers have used spectral based features such as MFCC for the purpose of DID. In our study, an analysis window of 25msec duration is used, with 10msec skipping rate. We use traditional 39-dimensional feature vector consisting of 12-dim MFCC, 12-dim ΔMFCC, 12-dim ΔΔMFCC, 1-dim Energy, 1-dim ΔEnergy, and 1-dim ΔΔEnergy. We use this future together with the GMM classifier to provide a benchmark system

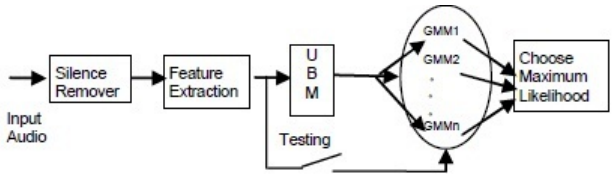

Figure 1: Baseline GMM based DID system.

4. SYSTEMIC STRATEGY FOR ROBUST SYSTEM

To build a robust system, we need address the two core issues: (a) small training set, and (b) resistance to noise, which are critical for a system to survive robustly in real life. To solve the former, we propose a modified UBM scheme to make full use of available data. To solve the latter, we propose PMVDR and SDC.

4.1 Hierarchical Universal Background Model

The universal background model (UBM) is an effective framework widely used in speaker recognition, and now people also use it for automatic speech recognition (ASR) and dialect identification. It relies on a general GMM, which represents the whole acoustic space, associated with a set of HMM state-dependent probability functions modelled as transformations of this GMM.

Traditionally, researches group all the training acoustic files to train a general GMM, i.e., UBM, which is then adapted with different category acoustic files to derive class-specified acoustic models-GMMs. But this approach fails to take the gender into consideration, which has great impact on dialect classifier performance. To make use of the advantage of UBM in addressing small dataset, and also to take gender factor into consideration, we propose the following hierarchical UBM structure (The arrows mean adaptation with acoustic files).

4.2 PMVDR

Our previous research [7] showed that perceptual Mini-mum Variance Distortionless Response (PMVDR) feature extraction is better able to model the upper spectral envelope at

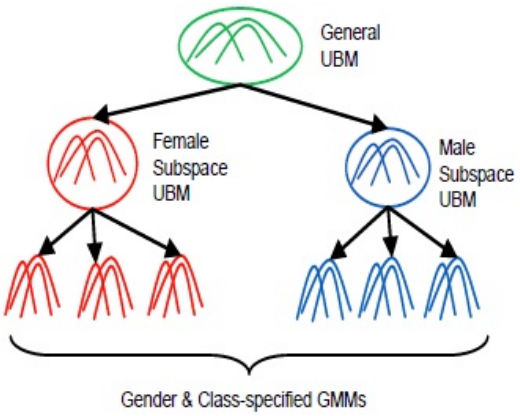

Figure 2: Hierarchical UBM structure.

the perceptually important harmonics, which may include important dialect clues. PMVDR cepstral coefficients provide improved accuracy over traditional MFCC parameters by better tracking the upper envelope of the speech spectrum. Unlike MFCC parameters, PMVDRs do not require an explicit filterbank analysis of the speech signal. We have found this new feature representation provides not only robustness against noise in speech recognition, but also higher accuracy in clean speech tasks. Here, we propose to test this feature in the context of DID. A block diagram of the PMVDR feature extraction [7] is shown in Figure 3.

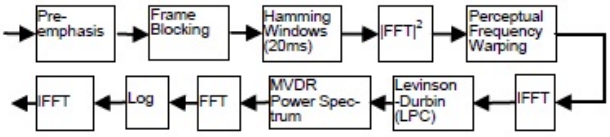

Figure 3: PMVDR feature extraction process.

It has been shown that implementing the perceptual scales through the use of a first order all-pass system is feasible. In fact, both Mel and Bark scales are determined by changing the only parameter, , of the system. The filter, H(z), and the warped frequency,$\hat{\omega}$, are approximated as $Eq.1 \sim 2$:

$$H(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{1}$$

$$\hat{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{2}$$

where ω represents the linear frequency and controls the degree of warping. For 16 kHz sampled signals, $\alpha = 0.42$ and 0.55 approximate the Mel and Bark scales, respectively.

Utilizing direct warping on the FFT power spectrum by removing the filterbank processing step leads to the preservation of almost all the information in the short-term speech spectrum. We can now summarize the remainder of the proposed PMVDR algorithm as follows:

1. Obtain the perceptually warped FFT power spectrum,
2. Compute the "perceptual autocorrelations" by utilizing the IFFT on the warped power spectrum,

3. Perform a ith order LP analysis via Levinson-Durbin recursion using perceptual autocorrelation lags,
4. Calculate the ith order MVDR spectrum according to $Eq.(1)$ in [7],
5. Obtain the final cepstrum coefficients using the straightforward FFT-based approach.

Finally, we use 39-dimensional PMVDR features and each feature vector contains 12 statics, deltas and delta-deltas along with energy, delta and delta-delta energy. We used the same windowing and frame skipping as in MFCC before further processing. Cepstral mean normalization (CMN) is also utilized on the final feature vectors. Since PMVDR removes the filterbank processing, we can avoid the demanding computation and noise sensitivity incurred by filterbank processing. This is crucial to realistic dialect identification system.

4.3 SDC

The aim of including shifted delta cepstrum (SDC) in the context of dialect recognition is to incorporate additional temporal information into the feature vector. The SDC is in fact k blocks delta cepstrum coefficients [8]. Suppose the basic set of cepstrum coefficients, $c_j(t), j = 1,2,,N-1$, is available (which can be PMVDR statics in this study) at frame t, where j is dimension index and N the number of cepstrum coefficients. The SDC feature can be expressed as following:

$$s_{iN+j}(t) = c_j(t+iP+d) - c_j(t+iP-d), i = 0,1,...,k-1 \quad (3)$$

where d is the time difference between frames for spectra computation, P is the time shift between each block, and k is the total number of blocks. The SDC coefficients can be concatenated with the basic cepstrum coefficients. Thus, we can obtain the feature vector as $\{c_j, j = 0,1,,N-1; S_{iN+j}(t), j = 0,1,,N-1, i = 0,1,,k-1\}$, which is the SDC version of features.

The popular parameter configuration of SDC $N-d-P-k$ in language identification is $7-1-3-7$. In our DID task, we fix the optimal configuration at $Dim-1-3-3$, where Dim is the dimension of basic cepstrum coefficients and is optimized at 10 for this paper.

4.4 Rotation Forest

Since GMM is a dynamic modeling technique, we also want to explore the robustness issue from the perspective of statistic modeling, such as Rotation Forest, which is one of the latest classifier ensemble methods [10]. In Rotation Forest algorithm, the training set for each base classifier (C4.5 decision tree is selected as the base classifier in this study) is formed by applying PCA to rotate the original attribute axes. Specifically, to create the training data for a base classifier, the attribute set F is randomly split into K subsets (K is a parameter of the algorithm) and PCA is applied to each subset. All principal components are retained in order to preserve the variability information in the data. Thus, K axis rotations take place to form the new attributes for a base classifier. The main idea of Rotation Forest is to simultaneously encourage diversity and individual accuracy within the ensemble: diversity is promoted through feature extraction for each base classifier and accuracy is sought by keeping all principal components and also using the whole data set to train each base classifier.

For rotation forest, the feature is extracted with the *openSMILE* toolkit [5] with 6552-dimension feature set, we then perform correlation-based feature selection (CFS) to reduce the feature dimension below 60 and run classification on the machine learning platform of *WEKA* [6].

5. EXPERIMENTAL RESULTS

In order to compare the proposed hierarchical UBM (hUBM) with traditional UBM (tUBM) structure, we will replace the UBM scheme in the baseline system (Figure 1) with two options: hUBM and tUBM. To compare the ro-bustness of the different feature extraction schemes in noisy condition, we introduce additive Gaussian white noise (10dB and 0dB) to the database.

5.1 Results

Now we consider an evaluation of the effectiveness of the proposed various schemes. All GMM-based classifiers backend are based on the same experiment setup as in Section 3, and MFCC is the feature for the baseline system. The results are displayed in Table II and Figure 4.

Although speech enhancement algorithms (such as wiener filtering) may help improve the classifier performance, we stick to the feature extraction without denoise process to compare how robust individual feature extraction algorithm is in presence of noise.

Table 2: DIALECT IDENTIFICATION ACCURACY (%)

Accuracy Rate (%)	0dB	10dB	clean
MFCC + tUBM	57.9	68.4	73.6
MFCC + hUBM	57.9	68.4	76.3
PMVDR + SDC + tUBM	60.5	68.4	76.3
PMVDR + SDC + hUBM	76.3	79	79

One note we need to make is that during experiment we group all N dialects according to gender thus we finally get 2N GMM models ($N = 3$, in this paper).

To make full use of training data, we also explore a large feature set with feature selection. After derive large feature set we use the rotation forest as the backend classifier. The results are summarized in Tab. III. By linearly fusing the results of Rotation forest and PMVDR+SDC+hUBM, we can get better results, which are detailed in Tab. IV.

Table 3: ROBUST PERFORMANCE OF ROTATION FOREST

Accuracy Rate (%)	0dB	10dB	clean
Rotation Forest	53.6	63.2	89.5
Feature Dim after feature selection	52	23	16

Table 4: SCORE FUSION OF TWO SYSTEMS

Accuracy Rate (%)	0dB	10dB	clean
Score Fusion	76.3	79.0	94.7
Weight for Rotation Forest	0	0	0.4

5.2 Analysis

From Table II and Figure 4 we can see that hierarchical UBM performs better than traditional UBM in MFCC and

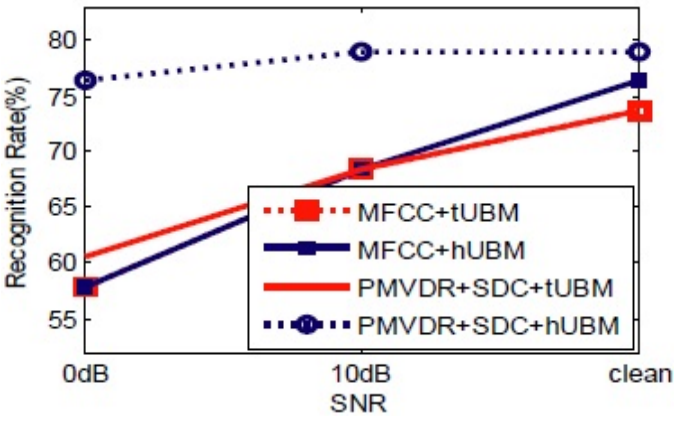

Figure 4: Robustness performance of the dynamic system schemes.

PMVDR+SDC scenarios. This proved the effectiveness of the proposed method, which can work as an alternative to the tradition UBM in small data corpus.

Also, from Table II and Figure 4, we can clearly see that PMVDR+SDC show very good robustness to noise. Compared with MFCC, PMVDR does not require an explicit filterbank analysis and thus are less sensitive to noise disturbance, similar results was also reported in other publication[7]. Here, we introduce the use of SDC to incorporate additional temporal information into the feature vector since our previous study [9] also showed it can bring more robustness to system. This benefit comes from the cepstrum substraction in Eq.(3).

Table III shows a large feature set, in the noise free environment, can make much better use of available data since it can capture more information on dialect. But the system performance degrades rapidly in the presence of noise than PMVDR+SDC+hUBM. This is not strange since noise may corrupt whole feature space and complex system may tend to produce a worse snowball effect than otherwise.

Table IV shows that higher recognition performance can be achieved by fusing the two systems in the decision process. Compared with the baseline system (MFCC+ tUBM), 36.4% and 21.5% relative gain is achieved, in the 0dB noise and clean condition, respectively. Especially, compared with the best result (82.7%) from [3], the combination in clean condition can give 14.5% relative gain.

6. CONCLUSIONS

To build a higher recognition, low noise sensitive dialect identification from limited database is challenging and also critical in real life application. We propose a series of strategy to solve this problem. We employ the hierarchical UBM to deal with small data set more effectively than tradition UBM approach. The PMVDR-SDC feature extraction, which has not been used in the context of DID, outperforms the baseline system with excellent noise robustness. To further improve the system performance, we also explore the potential benefit from the ensemble classifier-Rotation Forest, which showed superior performance in our dialect classification task. When combining the two best systems, we can get the relative gain of 31.8% and 28.7%, in the worst noise condition (0dB) and clean condition, respectively. This proves that the proposed system can well deal with the noisy and small dialect corpus classification task.

As illustrated in Table III, the worse performance of Rotation Forest in the noisy condition also proved the significance of seeking representative feature extraction algorithm. Since the noise robustness experiment in this study is performed by adding Gaussian white noise, real noise corrupted dialect corpus should be examined in the future. This is only a preliminary exploration in the front-end aiming to propose a good alternative to the popular MFCC. To compare the proposed schemes with other state-of-the-art robust techniques on a more comprehensive task such as NIST LRE will also be addressed in the future.

REFERENCES

[1] 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, May 31-June 3. 2004, pp. 297–300.

[2] W. Shen, N. Chen, D. A. Reynolds, "Dialect Recognition Using Adapted Phonetic Models" in *Interspeech 2008*, Brisbane, Australia, September 22-26, 2008,pp. 763-766.

[3] R. Huang and J.H.L Hansen, "Unsupervised Discriminative Training With Application to Dialect Classification" *Audio, Speech, and Language Processing, IEEE Transactions*, vol. 15,issue. 8, pp. 2444–2453, Nov. 2007.

[4] R. Huang, J. H. L. Hansen, and P. Angkititrakul, "Dialect/Accent Classification Using Unrestricted Audio" *Audio, Speech, and Language Processing, IEEE Transactions*, vol. 15,issue. 2, pp. 453–464, Feb. 2007.

[5] F. Eyben, M. Wollmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit" in *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009), IEEE*, Amsterdam,Netherlands, September 10-12. 2009, pp.576–581.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H.Witten, "The WEKA Data Mining Software: An Update" *SIGKDD Explorations*, vol. 11,issue. 1 pp. 10–18, 2009.

[7] U. H. Yapanela, J.H.L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition" *Speech Communication*, vol. 50,issue. 2 pp. 142–152, 2008.

[8] P.A. Torres-Carrasquillo, E. Singer, and etc. "Approaches to language identification using gaussian mixture models and shifted delta cepstral features" in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, Denver, USA, Sep. 2002, pp. 89–92.

[9] G. Liu, Y. Lei, and J.H.L. Hansen, "A Novel Feature Extraction Strategy for Multi-stream Robust Emotion Identification" in *Interspeech 2010*, Makuhari, Japan, Sep 26-30. 2010, pp. 482–485.

[10] J. J. Rodriguez, L. I. Kuncheva, C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28,issue. 10, pp. 1619–1630, Oct. 2006.