DIALECT IDENTIFICATION: IMPACT OF DIFFERENCES BETWEEN READ VERSUS SPONTANEOUS SPEECH

Gang Liu, Yun Lei, 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, {yxl059200, John. Hansen}@utdallas.edu

ABSTRACT

Automatic Dialect Classification (ADC) has recently gained substantial interest in the field of speech processing. Dialects of a language normally are reflected in terms of their phoneme space, word pronunciation/selection, and prosodic traits. These traits are clearly visible in natural speaker-to-speaker spontaneous conversations. However, dialect cues in prompted/read speech are often neglected by the community. In this study, we consider a systematic assessment of the differences between the acoustic characteristics of spontaneous and read speech and their effects on dialect identification performance. By examining both the model space and phoneme space of read and spontaneous dialect speech, we observe that each spans different dialect spaces and with distinct characteristics that need to be addressed respectively. From this comparison, we find useful clues to design more efficient identification systems. Finally, we also propose a novel feature extraction technique, PMVDR-SDC, and obtain a +26.4% relative improvement in dialect recognition rate.

1. INTRODUCTION

Dialect identification (ID) has recently emerged to be of substantial interest in the speech processing community [1]. Dialect ID systems can be used to improve the performance of Automatic Speech Recognition (ASR) engines 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 modeling 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 *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 ASR, although speech derived from read texts, news broadcasts, and other similar prompted contexts can be recognized with high accuracy, recognition performance decreases drastically for spontaneous speech. This is due to the fact that spontaneous speech and read speech are significantly different acoustically as well as linguistically. Compared with controlled read speech, spontaneous speech can be characterized by varied speaking rate, filled pauses, corrections, hesitations, repetitions, partial words, disfluencies, "sloppy" pronunciation. Recent studies [2] show that spontaneous speech can be characterized by a shrinkage of the spectral space in comparison with that of read speech and this results in a reduction in phoneme recognition accuracy. However, little work has been done to examine the differences from the dialect identification perspective. This paper attempts to explore the acoustical differences of read and spontaneous speech and its impact on dialect ID systems.

This paper is organized as follows. In Sec. 2, we describe the database that is used for algorithm development and evaluation. The baseline system which serves as the starting point for our proposed advances is described in Sec. 3. Next, we discuss Kullback-Leibler divergence in Sec. 4, which is used to measure dialect model differences. Sec. 5 explores the differences between read and spontaneous speech and their impact on identification. Summary and conclusions are shown in Sec. 6.

2. DATABASE DESCRIPTION

The corpus employed for our study is an Arabic corpus. Utterances were digitized at a 16 kHz sample rate. Three dialects are used in our study, based on geographical origins: United Arab Emirates (AE), Egypt (EG), and Iraq (IQ). For each dialect, we have two sets of data: read (READ) and spontaneous (SPON) speech. For each dialect, READ and SPON have the same 100 speakers which are meant to suppress speaker variation. There is no overlap between any train/test utterances. The SPON speech portion was recorded in a conversation style and speakers were selected to talk about different topics from pre-defined subject pool, with

This project was funded by AFRL under a subcontract to RADC Inc. under contract FA8750-09-C-0067.

 ISSN 2076-1465

the aim of a balanced set of topics. For READ speech, topics are also balanced and selected from the same topic pool as SPON speech. Table 1 summarizes training and testing data after a silence removal process (the unit is minutes). The data size for train and test of READ are almost the same as that of SPON to provide a fair basis of comparison. In test, all audio files are partitioned into short 10 sec segmentations.

Data	Training data			Sum	Testing data			Sum
	AE	EG	IQ		AE	EG	IQ	
READ(min)	87	91	92	270	52	54	53	159
SPON(min)	96	88	86	269	53	53	53	159

Table 1:The summary of Arabic corpus (READ vs. SPON)(in min.)

3. GMM-BASED CLASSIFICATION ALGORITHM

In this study, only supervised classification is considered (All the speech files are transcription-free and labeled only with dialect category information). Here, we employ the Gaussian Mixture Models (GMM) based dialect classification algorithm as our baseline system. Figure 1 shows the flow diagram of the baseline GMM training process, where a closed set of *N* dialects are considered.

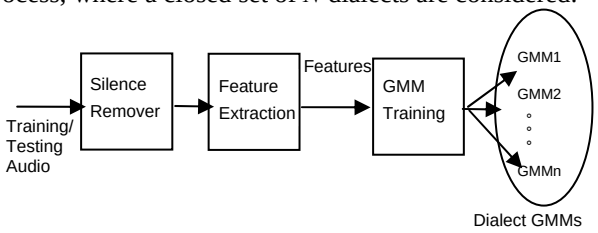

Figure 1: Baseline GMM based dialect training system.

The dialect GMM model is trained with SPON/READ training set from each dialect. The training method is generalized Maximum Likelihood Estimation (MLE). For training, silence frames are first removed from the input audio stream using an energy threshold, followed by feature extraction. In our study, we will employ two types of feature extraction front ends: Mel-Frequency Cepstral Coefficients (MFCC) and Perceptually Minimum Variance Distortionless Response (PMVDR) [3]. The testing phase follows almost the same process, except it uses the trained models from the training stage to get likelihood scores from the SPON/READ testing data.

4. MEASUREMENT OF DIFFERENCE BETWEEN MODELS

In order to compare the differences between read and spontaneous speech and determine its impact on dialect identification, we need some statistical tool for quantification. The GMMs are taken as the representation of the utterance space, so we can measure the difference between these models. The Kullback-Leibler (KL) Divergence (KLD) is suited for this task, which is based on relative entropy, and is often used as a measure of difference between two probability distributions.

Given a set of *N* dialect models in the system, denoted as $\{\Lambda_n, 1 \le n \le N\}$, let each Λ_n be viewed as a point in the dialect model space Λ. For any given two dialect models in the space and their corresponding training data, we can estimate the distance between these two models using the symmetric KLD, which is defined as the sum of relative entropy between model Λ_i and Λ_j, and also between model Λ_j and Λ_i:

$$KL(\Lambda_i, \Lambda_j) = E_{\Lambda_i(X)}[log\frac{\Lambda_i(X)}{\Lambda_j(X)}] + E_{\Lambda_j(X)}[log\frac{\Lambda_j(X)}{\Lambda_i(X)}] \quad (1)$$

where $\Lambda_i(X)$ and $\Lambda_j(X)$ are the likelihoods of occurrences of observation $\boldsymbol{X}$, given that it belongs to model Λ_i and Λ_j respectively. The GMM model $\Lambda_i(X)$ can now be described as:

$$\Lambda_i(X) = \sum_{j=1}^{M} \omega_j N(\mu_j, \sigma_j^2) \quad (2)$$

where ω *is* the weight, μ is the mean, σ^2 is the variance of the probability distribution function in the GMM model, and *M* is the total mixture number. Since there is no analytic close-form KLD expression for GMMs, we use the approximation introduced by Do [4]. We will use the KLD to quantify the differences between two dialect models.

5. EXPERIMENTS

5.1 Dialect Model Space Analysis

MFCCs are used in this analysis. The length of a frame is 20 ms, with a 10 ms overlap. The final classification performance is averaged on all test utterances. With two sets of train-test data, we obtain the following results (Table 2).

Train \ Test	READ	SPON
READ	**17.87%**	37.11%
SPON	37.33%	**20.01%**
Data pooling	25.33%	25.24%

Table 2: Error Rate (%) of dialect identification using READ & SPON speech utterances. Data pooling means to pool all the train data to form one train set.

From Table 2, it can be seen that the performance of dialect model based on the SPON speech data set is slightly worse than for the READ speech, which is counter to

traditional assumptions [5]. There may be two reasons: (1) READ speech is well pronounced and thus may better reflect clear/exact phoneme targets in the acoustic space for the model; SPON speech is shaped by many individual pronunciation habit such as hesitations, repetitions, partial words, which may negatively impact overall system performance; (2) In conversation, speakers are more likely to pick up their partner's pronunciation style than in a read style, thus for the same size of sample data, this may result in a more concentrated sample space compared with read. We also note that performance is much worse when system trained on SPON data is tested against READ data (or vice versa), which may be due to some variation that exists between prepared and spontaneous speech even for the same person and therefore leads to mismatch between SPON and READ.

Here, we employ KLD to further compare the acoustic models differences. The comparison results are summarized in Table 3. In our experiment, the range of KLD value can vary from 0-50, but most pairwise values are within 15-30. Smaller divergence value means less difference and harder to classify than the bigger one.

KL divergence	**AE-EG**	**AE-IQ**	**EG-IQ**
SPON	19.36	19.74	19.79
READ	20.46	21.01	20.84
KL divergence	**AE-AE**	**EG-EG**	**IQ-IQ**
SPON-READ	20.44	20.39	20.66

Table 3: KL divergence of different dialect acoustic models.

From Table 3 we observe that the differences among the READ speech models are greater than the SPON speech ones, which is why the READ-READ test results are better than those of the SPON-SPON. In other words, KLD can work as a ruler to measure the difference between models.

In addition, the KLD between the counterparts of READ and SPON are all less than that for READ (except IQ-IQ for SPON-READ which is greater than AE-EG for READ). When we derive SPON models from READ models via Maximum a posterior or vice versa, we can obtain almost the same result for both scenarios, which implies that there is a portion of the dialect acoustic space that cannot be approached by limited adaptation data and we therefore need find other methods to address this problem.

5.2 Phoneme Analysis

For language identification, different languages have unique grammar and phonemes. Unlike language identification, different dialects may share similar phonemes and pronunciation traits. So , further investigation at the phoneme level is necessary. We partition the speech data into different phoneme islands (isolated phonemes) with an HMM phoneme recognizer. First, we compare the average length of the top 5 most frequently occurring Arabic phonemes in both READ and SPON speech (Fig. 2).

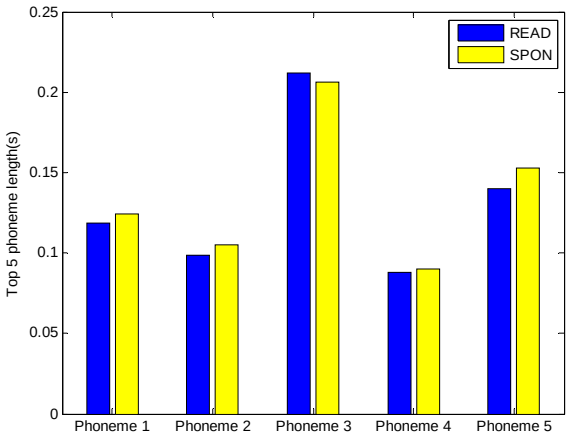

Figure 2: The duration comparison of the top 5 most frequently occurring Arabic dialect phonemes in SPON/READ speech. The five phonemes are listed in decreased frequency order from 1 to 5.

From Figure 2, we observe that the durations of the SPON speech phonemes are almost always longer than that of READ, which may mean the performance of READ speech dialect identification at the phoneme level will be inferior to that of SPON speech since for the same phoneme, SPON phoneme will have more data/information. Following a similar process as in Sec. 5.1, we obtain the phoneme-based dialect KLD measurement results (Table 4). To compare with the prior acoustic model analysis, we also summarize the results from Table 3.

KL divergence	**READ-READ**	**SPON-SPON**
GMM-phonemic	16.20	**16.37**
GMM-acoustic	**20.77**	19.63

Table 4: Average KLD of different acoustic/phonemic models.

From Table 4, we notice that SPON phonemes carry more meaningful dialect cues than READ phonemes, although the overall GMM-phonemic KLD is smaller than the GMM-acoustic values and need improvement in modeling at the phoneme level.

With this in mind, we can more efficiently design a shifted delta cepstra (SDC) [6] feature extraction front-end, in which the traditional search method for parameter optimization is time consuming hill-climbing algorithm. For READ speech, we need a smaller N for the k-d-P-N scheme [6]. In the following system schemes, we optimize the parameter of MFCC-based acoustic models as: 7-1-3-3 for READ and 7-1-3-5 for SPON. Also, our previous research [3] showed that PMVDR feature extraction is better able to model the upper spectral envelope at the perceptually

important harmonics. So, we also explore this potential feature in the current case. For the PMVDR characterized data, the k-d-P-N parameters are optimized as: 12-1-3-3 and 12-1-3-4 for READ and SPON speech, respectively.

5.3 System Combination

To achieve better dialect identification results, we propose the following 5 schemes, all of which are trained and test based on the same algorithm described in Sec. 3.:

1) Separate Training: Train READ/SPON models separately with their own training set(as in Sec.5.1);
2) Data-pooling: Pool together the training files of both READ and SPON data to form one training set, which will be used as the training set for both READ and SPON. This aims to augment the training set in 1);
3) MFCC-Combination: Score fusion of MFCC-based systems from system (2) above;
4) MFCC-SDC-Combination: Score fusion from MFCC-SDC feature-based systems;
5) PMVDR-SDC-Combination: Score fusion from PMVDR-SDC feature-based systems.

Scheme	READ	SPON
Separate Training	17.87%	20.01%
Data-pooling	25.33%	25.24%
MFCC-Combination	16.12%	14.06%
MFCC-SDC-Combination	15.26%	13.67%
PMVDR-SDC-Combination	14.86%	13.02%

Table 5: 3-way Arabic Dialect ID Error Rates (%) of different READ/SPON data/scores combination schemes. For different scheme, the test sets for READ and SPON are fixed and explained in Fig. 1

According to Table 5, we observe that although data-pooling is a natural choice, it fails to take the differences between READ and SPON dialect speech into consideration and is inferior to all other schemes. Compared with results for the MFCC-Comb scheme, the well-designed SDC versions of MFCC and PMVDR can further improve integrated system performance, especially PMVDR-SDC-Combination, achieved a +26.4% relative improvement in average recognition error rate, compared with Separate Training.

6. CONCLUSION

This paper has explored the differences between the dialect representation capability of READ and SPON speech at different levels: model space and phoneme level. The motivation for this work is that past research studies are based on SPON speech, with the belief that READ speech has limited dialect structure. One focus here has been to determine the validity of this assumption. First, our experiments showed that READ acoustic models demonstrate comparable classification performance to that of SPON acoustic models. Next, with the help of the KLD, we further quantified the difference between different models. Next, we investigated the differences at the phoneme level. At this level, we observed that dialect models constructed from SPON speech have better classification potentials than models from READ do. Interestingly, we also found the phoneme duration of the SPON speech is longer than that of READ speech, which can help search more efficiently for optimized SDC parameters. The phoneme-based KLD measurement shows that SPON phonemes carry more dialect cues than READ phonemes do. All these differences allow for performance enhancements resulting from a combination solution. From experiments, we find that the greatest improvement was achieved with the novel PMVDR-SDC based system combination, which showed a +26.4% relative improvement in average error rate.

This paper is only a preliminary attempt to distinguish the differences among READ and SPON speech. If more dialect-dependent cues/audio segments can be identified, then further efficient dialect systems can be expected from available data corpus. Also, the differences between READ and SPON speech are mainly examined from a signal processing perspective. Linguistic analysis can further shed light on ways to make better use of samples from both READ and SPON speech.

7. REFERENCES

[1] S. Gray, J.H.L. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system", IEEE ASRU-2005, pp.35-40, 2005.

[2] M. Nakamura, K. Iwano, and S. Furui. "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance". *Computer Speech and Language*, 22:pp. 171– 184. 2008.

[3] U. H. Yapanel, J.H.L. Hansen. "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition". Speech Communication 50, 142–152,2008

[4] M. Do, "Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models," IEEE Signal Process. Lett., vol. 10, no. 4, pp. 115–118, Apr. 2003.

[5] Y. Lei, J.H.L. Hansen, "Dialect Classification via Discriminative Training", INTERSPEECH-2008, pp.735-738, Brisbane, Australia, Sep, 2008.

[6] P. A. Torres-Carrasquillo, E. Singer，M. A. Kohler， R. J. Greene，D. A. Reynolds， J. R. Deller, Jr.. "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", ICSLP 2000.